It’s 5 o’clock. Do you know where your info risks are?

Kate Pugh, Columbia, AlignConsulting
Sandra Serkes, Columbia, Valora Technologies

KM World November 5, 2019
We love love our furry friends...but we really love metadata

Sandra Serkes, CEO, Valora Technologies
Guest Lecturer, Columbia University Information and Knowledge Strategy Master’s Program

Katrina Pugh, CEO, AlignConsulting
Faculty and former Academic Director, Columbia University Information and Knowledge Strategy Master’s Program
What is content bloat costing us?

1 hour, 41 minutes/week spent searching (average worker)

13.2 billion hours across all US workers

$628 billion opportunity cost for the US

What could info bloat cost us?

GDPR introduced hefty penalties per instance:
up to €20M or 4% of global turnover – whichever is higher

- Breach of customer data: Proposed fine: £183M
- Failing to protect customer data: Proposed fine: £99M
Why else do we “take control of” unstructured data?

Compliance:
Retention, Legal Hold, GDPR, Privacy, Security

Identification:
ROT, Dupes, PII, DocType, Relevance, Privilege, Custodian, Date, IP potential and more

Disposition:
Quarantine, Defensible Deletion, Redaction, Migration, Remediation
Do any of these look like you?

Large Oil & Gas Manufacturer
- Frequent acquisitions of other companies
- Millions of documents
- Not enough SMEs
- Risky content..Legal hold…Data Privacy…ROIT

Multi-store, multi-country fast-food
- Independent, franchise & retail sites
- Managers autonomously hire, fire
- GDPR concerns

Info hound (bankruptcy data vault)
- Out run by content volume
- Regulatory reporting obligations
- Sensitive data
- Missing value-added revenue…?
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata tells us:</td>
<td>Unitize, translate, transform, analyze,</td>
<td>Structures, policies,</td>
</tr>
<tr>
<td>▪ What it is (e.g.,</td>
<td>assess, triage groups of documents, based</td>
<td>procedures, processes and</td>
</tr>
<tr>
<td>contract, whitepaper)</td>
<td>on metadata and context</td>
<td>controls implemented</td>
</tr>
<tr>
<td>▪ What it is about (e.g.</td>
<td></td>
<td>to help manage</td>
</tr>
<tr>
<td>topics, insights)</td>
<td></td>
<td>information at an</td>
</tr>
<tr>
<td>▪ Inherited labels (e.g.,</td>
<td></td>
<td>enterprise level over</td>
</tr>
<tr>
<td>file type, date,</td>
<td></td>
<td>time</td>
</tr>
<tr>
<td>author, last modified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>date)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>▪ Impact, risk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Designing, contextualizing*  *Processing, tagging*  *Planning for future*
1. Metadata: Framing
Rich Metadata for Enterprise Content

What attributes & tags make sense for your organization?

+ “hidden” metadata
Metadata is “portable”

But what about similarly rich metadata (tags) for our enterprise content?
Isn’t rich metadata there…or can’t we just add it?

- Sort of..
  - Rich metadata tends to be common with structured data (databases), and uncommon with unstructured data (email, slide decks, reports, etc.)
  - Rich metadata is typically input by hand (via forms & wizards) and suffers from negligence, inconsistency, laziness, error, etc.
What does rich metadata get you?

Ability to know what you have, where it is, and how to manage it

- Lifecycle management tags
- Protected status tags (legal hold, privacy, access, etc.)
- Context for searches & decision making

Ability to see across data repositories

- Normalized tags
- Data visualization, presentation and reporting

Ability to operate by facet

- Manage by content, not by time or other “simplistic” approaches
“Wonder if anyone would mind...?”

Context matters as we define metadata!
The slow way: Hand classifying Cuban Missile Crisis transcript

Ingestion (OCR) - Automated

Cuba Tape

This is a transcription of tapes recorded in the Saturday, October 27, 1962.

The tape begins with an inconclusive discussion of questions, including plans to stop a slip (Grozny) and daylight surveillance missions, morning and afternoon, interrupted a few minutes into the meeting as the President:

JFK: (Reading) Premier Khrushchev told President Kennedy yesterday he would withdraw offensive missiles from Cuba if the United States with-drew its rockets from Turkey.

(voices unclear.)

Voice: He didn't really say that, did he?

JFK: That may not be -- he may be putting it in disguise.

(Mixed voices. Calls for Pierre [Salinger]).

JFK: That wasn't in the letter we received, was it?

Voice: No. (Voices unclear)

JFK: Is he supposed to be putting out a letter?

Salinger: Putting out a letter he wrote to you.

JFK: Let's just -- uh -- keep on going (words unclear).

Voice: It's in a different statement.

Rusk: Well, I think we better get -- uh -- (words unclear).

Voice: (Mixed voices)

JFK: What's the advantage of the second mission?

McNamara: It creates a pattern of increasing intensity.

JFK: What's the advantage of the second mission?

McNamara: It creates a pattern of increasing intensity.

Participant Reply Coding

Preliminary analytics (= 1 hr. Messy, not repeatable)

Coding – 1 Hour/page (= 84 hours)

https://catalog.archives.gov/id/193723
AutoClassifying an attachment (patent application)

DocType = Patent Application
Date = 10/18/2007
Date Format = US
Author = Patent Authors, Author City, Author Country
Assignee = RIM
Tone = Neutral to slightly positive

Embedded Graphic with Title

Other Data Capturable Data Elements:
- Patent Number
- Filing Date
- Key Phrases & Terms
- Managing PTO
- Implied/Attached Docs
- Bar Code Present
- And many more . . .

Implied status: Responsive, Nonprivileged
2. Auto-Classification: Analyze and Design
What is the secret sauce of auto-classification design?

It’s actually **RULES** that are the real key to unlocking the power of AutoClassification.
What are Rules?

- Software algorithms that determine the disposition of content
- Typically follow an IF-THEN format, often nested
  - Ex: IF the DocumentType = Contract AND the Date < 1/1/2020, THEN Mark File as ACTIVE (Retain)

So, if you could do any of these rules, which ones should you do first?
[Hint: could you do multiple simultaneously?]
Rules Engine illustration

- Folders for organization
- Ruler represents single rule
- String test (can have multiple per rule)
- Operation tab allows initial definition
- Field tab denotes which PowerHouse fields rule is executed against
- Values tab shows which values the rule should identify on
- Support for values files
- Weights & confidences create hierarchical decisioning options
- Ability to test on family ranges and attachments (include & exclude)
- Match type options: Exact, Word Stemmed, OCR, RegEx, Numeric, Date
- Rules Editor illustration
### Clear PII: SSN

- Interest Only
- Subject Property Address (street, city, state, ZIP): 5421 Talahassee Place, Perry FL 32347
- Legal Description of Subject Property (attach description, if necessary): Lot 64

### Likely PII: ("warning sign")

- Purpose of Loan: Purchase
- Property Type: Construction
- Property Type: Construction-permanent

### Clear PII: Home Phone Number

- Home Phone Number: 562-55-5555
- (904) 986-9985
- 04/11/1969
- 2
- 548-32-9995
- 02/16/1978

- Brady G. Schlesinger
- Wilma B. Schlesinger

### Not PII: Interest Rate

- Interest Rate: 6.5%

### Implied classification: Active PII, needs protection & redaction

- Address:
  - 5421 Talahassee Place, Perry FL 32347
- Legal Description of Subject Property:
  - Lot 64
- Purpose of Loan:
  - Purchase
  - Construction
  - Construction-permanent
- Home Phone Number:
  - 562-55-5555
  - (904) 986-9985
- Interest Rate:
  - 6.5%
- Implying the need for protection and redaction of PII data.
3. Information Governance: Control
The information governance lifecycle

New content

Processing
- Ingests data or processes files in place
  - Creating OCR for scanned images
  - Transcribing voice and video
  - Extracting text for native files & email
  - Speech to text for audio/video files
  - Translating content to English
  - Re-ordering or re-aligning pages
  - Applying redactions
  - Unitization

Tagging
- Extracts key Info& attributes about each document (aka Coding, Indexing)
  - Document Type, Important Dates
  - Key Names & Phrases
  - Topics, Keywords & Themes
  - File, Content and DocType attributes
  - Relation to other documents (duplicate, related, attached, contradictory, etc.)

Disposition
- (Rules) create a destination, action or status for each document
  - Retention status & duration
  - Privacy & security protections
  - Folder (taxonomy) location
  - Labelling & keywords display
  - Activity notification
  - Legal Hold

Business Assess’t
- AI and Machine learning help assess risks to reputation, ID new IP
  - Relevancy/ prioritization
  - Contract inconsistencies
  - Conflicts of interest
  - New IP
  - New Product or feature

Old content
Data Visualization of Auto Classification (Data dashboard)
Data Visualization of Auto Classification (Graph Theory) (cont.)
Data Visualization of Auto Classification
Automated Retention Handling
Automated Retention Handling (cont.)

<table>
<thead>
<tr>
<th>DocID</th>
<th>Dot Type</th>
<th>Date</th>
<th>Expiration Date</th>
<th>Expiration Handling</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>IG-0029263</td>
<td>Purchase Document</td>
<td>03/09/2016</td>
<td>03/09/2020</td>
<td>Policy Review</td>
<td>CLOSING STATEMENT</td>
</tr>
<tr>
<td>IG-0029264</td>
<td>Press Incident Cover</td>
<td>03/24/2016</td>
<td>03/24/2020</td>
<td>Defensible Deletion - 5 Years</td>
<td>Counterfeit auto parts are threat to auto safety</td>
</tr>
<tr>
<td>IG-0029267</td>
<td>Press Incident Cover</td>
<td>07/23/2016</td>
<td>07/23/2021</td>
<td>Defensible Deletion - 5 Years</td>
<td>Counterfeit LED TVs from China confiscated at Miami Seaport</td>
</tr>
<tr>
<td>IG-0029269</td>
<td>Press Incident Cover</td>
<td>06/23/2016</td>
<td>06/23/2021</td>
<td>Defensible Deletion - 5 Years</td>
<td>Counterfeit pain pills likely came to Prince Illegally</td>
</tr>
<tr>
<td>IG-0029293</td>
<td>Press Incident Cover</td>
<td>06/25/2015</td>
<td>06/25/2020</td>
<td>Defensible Deletion - 5 Years</td>
<td>Counterfeit products drive fashion industry to improve, study suggests</td>
</tr>
<tr>
<td>IG-0029294</td>
<td>Report</td>
<td>03/2016</td>
<td>04/20/2016</td>
<td>Policy Review</td>
<td>UNREGULATED E-WASTE EXPORTS FUEL COUNTERFEIT ELECTRICITY</td>
</tr>
<tr>
<td>IG-0029296</td>
<td>Purchase Document</td>
<td>04/20/2016</td>
<td>04/20/2016</td>
<td>Quarantine - 3 Years</td>
<td>Your Credit Card Account Statement</td>
</tr>
<tr>
<td>IG-0029299</td>
<td>Press Incident Cover</td>
<td>05/11/2015</td>
<td>05/11/2020</td>
<td>Defensible Deletion - 5 Years</td>
<td>Does Etsy Have A Problem With Fake Goods?</td>
</tr>
<tr>
<td>IG-0029304</td>
<td>Civil Court Document</td>
<td>05/13/2014</td>
<td></td>
<td>Never Expires</td>
<td>UNRESOLVED ISSUE</td>
</tr>
<tr>
<td>IG-0029314</td>
<td>Press Incident Cover</td>
<td>07/30/2015</td>
<td>07/30/2020</td>
<td>Defensible Deletion - 5 Years</td>
<td>Etsy investors sue for fraud over Risk of Counterfeit Goods Etsy Inc.</td>
</tr>
<tr>
<td>IG-0029308</td>
<td>Report</td>
<td>03/21/2014</td>
<td></td>
<td>Policy Review</td>
<td>EXECUTIVE SUMMARY</td>
</tr>
</tbody>
</table>
“State of the State” of Artificial Intelligence in auto-classification

Reactive machines
Does not use past experiences to make decisions

Limited memory
Learns by monitoring actions & behaviours over time

Theory of mind
Has thoughts and emotions that affect their own behavior

Self-awareness
Has consciousness, empathy and a sense of being
Scientists have found a way to decode brain signals into speech. It’s a step towards a system that would let people send texts straight from their brains.

by Antonio Regalado

MIT Technology Review, April 24, 2019
Tagging: AI recognizes tone, slang

- Google Perspective API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give real-time feedback to commenters or help moderators.

https://www.perspectiveapi.com/
Tagging: Intent recognition example: Sarcasm

Humans don’t always recognize it! It requires context (speaker, situation, world)

Approaches:
1. Machine learning (volume, then phrase matching / information retrieval)
2. Rules (e.g., unexpected juxtaposition, exaggeration)
3. Deep learning

Check out Indian Institute of Technology research Sarcasm Suite

Source: Quartz (2016) and NVDia
“How do you really carry context throughout a dialogue? This is the biggest challenge. A lot of how you understand what I’m saying depends on what I said maybe five sentences ago, or fifteen sentences ago. You’re building up a state of the conversation.”

Satinder Singh, Director of the Artificial Intelligence Lab at the University of Michigan.

(Hold and resume concept (Kore.ai))
4. Case Studies
Recap: Do any of these look like you?

Large Oil & Gas Manufacturer
- Frequent acquisitions of other companies
- Millions of documents
- Not enough SMEs
- Risky content..Legal hold…Data Privacy…ROI

Multi-store, multi-country fast-food
- Independent, franchise & retail sites
- Managers autonomously hire, fire
- GDPR concerns

Info hound (bankruptcy data vault)
- Out run by content volume
- Regulatory reporting obligations
- Sensitive data
- Missing value-added revenue…?
Case Study #1: Oil & Gas

Acquiring a company without most of the people. 200 GB of data and none of the tacit knowledge - who owns this? Should we keep it?

“Don’t buy it before you know the info risk!”
Case Study #2: Multi-store/int’l fast food retailer

<table>
<thead>
<tr>
<th>Burning Platform</th>
<th>Scope</th>
<th>Critical Metadata</th>
<th>Analytics/Intelligence</th>
<th>InfoGov Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data privacy concerns on the verge of moving to the cloud</td>
<td>~10K employees &gt;100 stores ~20 repositories 80 terabytes of records Content analysis &amp; migration</td>
<td>Custodian Document Type PII/PHI Record Type Expiry Handling Dates</td>
<td>Retention ROT Data Privacy Classification Rec File Loc. Duplication</td>
<td>Data privacy Fileshare cleanup DSARs IG maturity</td>
</tr>
</tbody>
</table>

"It’s not about migration. It’s about gov and access."
Too much data to assess, evaluate, prioritize in tight timelines

Federal, state, local bankruptcy court filings
1705 Jurisdictions
300 Formats
100K-200K cases/yr

Debtor Name & specifics
Custom fields per DocType
Jurisdictional Dates
Financial

Bankruptcy terms violations?
Defrauding regulators?
Criminal checks?
Redundancies?
Expirations?

Compliance reporting
Risk analysis

Scope
Critical Metadata
Analytics/Intelligence
InfoGov Applications

Case Study #3: Info Hound (bankruptcy data vault)

“The 2008 meltdown forced us all to rethink & retool.”
Three Cases’ lessons learned

1. You **lay bare** facts and patterns (“It’s like walking with all of the labels on your clothes on the outside”)

2. Building **the rules engine** – how you label it, and what you do with it -- **takes longer** than you’d think. (“You learn when to dig deeper and when to back off.”)

3. It’s not just “Document Intelligence,” or “eDiscovery.” It’s “**info governance**”
Discussion
Thanks for attending!

www.ValoraTech.com
info@ValoraTech.com
+1 781.229.2265

AlignConsulting

www.alignconsulting.com
katepugh@alum.mit.edu
+1 617-967-3910
Appendix
Privacy Regulations require:

- Do you know *where* it is?
- Do you know *what* it is?
- Is it *sensitive*?
- Is it *protected and secure*?
- Is it *accessible* when needed?

AutoClassification allows you to:

- **Locate it:** email, shared drive, ECM
- **Identify it:** SSN, DOB, drivers license
- **Secure it:** set security access controls
- **Redact it:** grant permission levels
- **Action it:** DSAR, Right to be Forgotten
Rules for eDiscovery

Note: If you’ve got a good IG practice through AutoClassification, eDiscovery is much simpler, faster & lower cost

eDiscovery requires us to:

- Look for content
- Filter through tons of data
- Relevant & irrelevant, Privileged & Non-priv
- Document review
- Document production

AutoClassification you’ve already:

- Located it: across all data stores
- Identified it: for Early Data Assessment
- Cleaned it: only dealing with relevant content
- Secured it: who has access to what
- Actioned it: Retention, archived, etc.
Legal Hold asks us:

- Which data? Whose data?
- For how long?
- What’s relevant? What’s not?
- Access for inside counsel? Outside counsel?
- What happens when Hold lifts?

AutoClassification you’ve already:

- Located it: we haven’t missed anything/where
- Cleaned it: already excluded irrelevant content
- Identified type & custodian: we know content, context & owner
- Identified retention dates: schedule already exists
- Set it & forget it: ongoing perpetual analysis

Note: You can still dynamically govern content under Legal Hold