Classify Documents and Records in File Shares, O365, Microsoft Teams, SharePoint, OneDrive, etc.

Transform the way you store, find, govern, and use enterprise content

netwrix

July 2020

share²

Presenters: Auto-Classify / Identify Documents and Records

- **Robert Piddocke**
  VP, Content Services

- **Jim Duncan**
  Chief Architect, ShareSquared
  Microsoft Certified Master

- **David Kruglov**
  Managing Director

Intention: Intention for the Webinar

- Teach not present
- Distinguish the meaning and need for document classification
- Present ideas and solutions for automating classification
- Foster a community of people who endeavor to improve collaboration and records management solutions
- Answer any questions you may have
- Enroll you into being in action!

Be of Service to You

Agenda: Auto-Classify / Identify Documents and Records

Overview

- What do we mean by “classification”?
- Why do we care about it?
- Use Cases

Taxonomy

- Definition
- *Demo:* Information Architecture Workbook

Auto-Classification

- Netwrix Overview: Approach, Technology, and Differentiators
- *Demo:*
  - Data collection, building weighted clues, taxonomies, etc.
  - Classification on document upload

Closing and ***Extended*** QA

Classification: What is it in this context?

- Essentially telling the system what purpose the document serves and how to “treat” it
- Identifying a document against a list of known document types
  - ➢ Supports Search by Document Type (e.g. only Invoices)
- Identifying what Record Series/Category a document belongs to
  - ➢ Supports Records Retention and Disposition
- Assigning the Content Type and capturing metadata
  - ➢ Metadata can be used to refine search, filter List Views, workflow, etc.
- Netwrix Data Classification (white-box) vs. AI/ML (black-box)
  - ➢ NDC enables you to tune it for your specific content and requirements

Classification: Why do we care?

Search and ROT

- Knowledge workers spend ~2.5 hours/day or 30% of the workday searching for info (IDC)
- On average, every misfiled document costs $125; a lost document costs $700 (IDC)
- Up to 80% of electronically stored information is ROT (AIIM)

Security

- In the US, a data breach costs a company on average $8.19 million (Ponemon Institute)
- Companies with security automation technologies experienced ~1/2 the cost of a breach

Information Governance

- Less than 50% of content is correctly indexed, tagged or efficiently searchable (IDC)
- 67% of data loss in records management is due to end user error (AIIM)
- Without effective governance, most metadata projects will fail (Forrester)

Classification: Use Cases

File Shares

- Analyze documents, assign the Content Type, and add metadata
- Metadata is added as custom properties of the document
- Metadata can be used to determine ROT and to drive an “intelligent migration”
- *Cannot expect users to identify/tag historical content due to the volume*

SharePoint/O365 and Other Repositories

- Analyze documents, assign the Content Type, and add metadata
- Metadata is stored in existing Content Types
- *End users do not do this reliably*

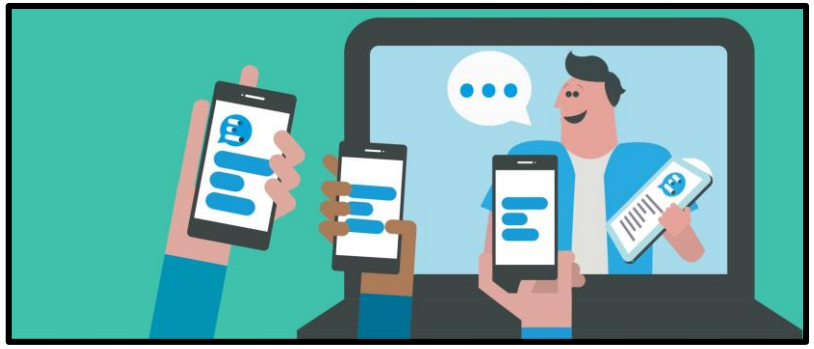
Document Uploads

- Analyze documents as they are uploaded by End Users
- Automatically determines the Content Type and adds metadata
- Enhances user adoption
- Increases standardization and reliability
- *End users resist the “extra work” which causes solution adoption issues*

Poll: What are your top initiatives?

Select all that Apply:

- Collaboration / Intranet
- End User Adoption
- Compliance / Records Management
- Content Migration / Identify ROT
- Process Automation

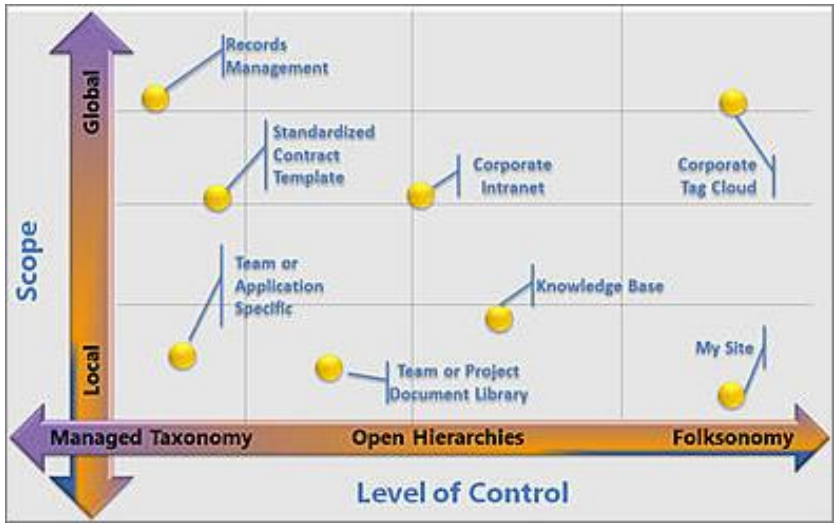


Next Up:

➢ *Taxonomy*

Taxonomy: Definition

- “A Taxonomy groups the words, labels, and terms that describe something, and then arranges the groups into a hierarchy”
- Supported via Content Types and Term Sets in SharePoint/O365
- Content Types are assigned to documents and have fields to store additional information (i.e. metadata), for example:
  - The Document is a Word document containing unstructured text
  - The Content Type assigned to it is an Invoice
  - The Amount field (metadata) indicates that the invoice is for $5,000
- Term Sets
  - A list of related terms
  - Limit the values that can be put into the fields of a Content Type
  - Provide consistency in the way things are named or tagged
- Content Types are used to drive:
  - Search
  - Workflow
  - List Views and Filters
  - Document and Records Management

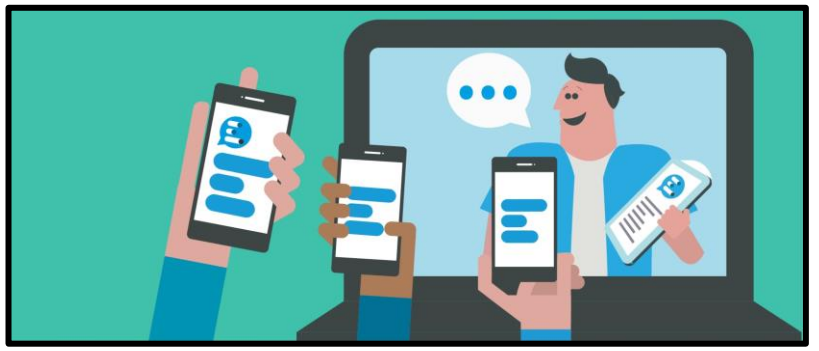


Demos: Search and Taxonomy

- Progressive Search
- IA Workbook

Poll: How mature is your Taxonomy?

- Not Addressed
- In Design
- Tried / Failed to Launch
- Could Use Enhancement
- Robust / Satisfies Requirements



Next Up:

➢ *Netwrix Overview & Demo*

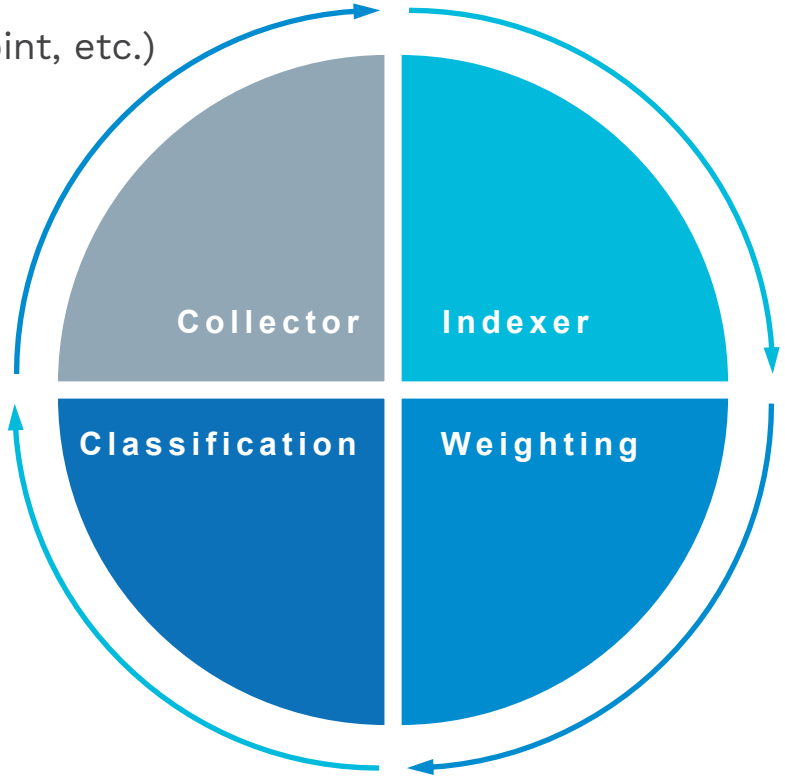
Netwrix: Approach to Auto-Classification

Collector

- Crawls repositories (File Shares, SharePoint, etc.)
  - Discovers structure
  - Prepares Content
- Performs document preparation stages:
  - Content type detection
  - Text extraction (OCR, Attachments)
  - Metadata extraction

Classification

- Performs document classification stage
  - Classifies documents based on business defined rules
  - “Tags” documents at their source
  - Executes business defined workflows
- Agnostic of document type
- Enriches search with document categories



Indexer

- Language detection
- Natural language processing (NLP)
- Duplicate detection *(Background)*
- Agnostic of document type
- Builds full text search index

Weighting

- Document parsing
- Term extraction
- Compound Term Processing
- Frequency
- Location in text
- Proximity
- Format of text

Netwrix: Data Classification Differentiators

Compound Term Processing

- Compound term processing identifies and weights multi-word concepts based on purely statistical analysis independent of vocabulary, language, or grammatical style
- Most scalable and adaptive metadata tagging and classification platform available

Technology Agnostic

- Flexible technology can access content regardless of where it is stored
- Native integration with all versions of SharePoint, SharePoint Online, Exchange, File Shares, SQL, OneDrive for Business, Box, and other CMS's

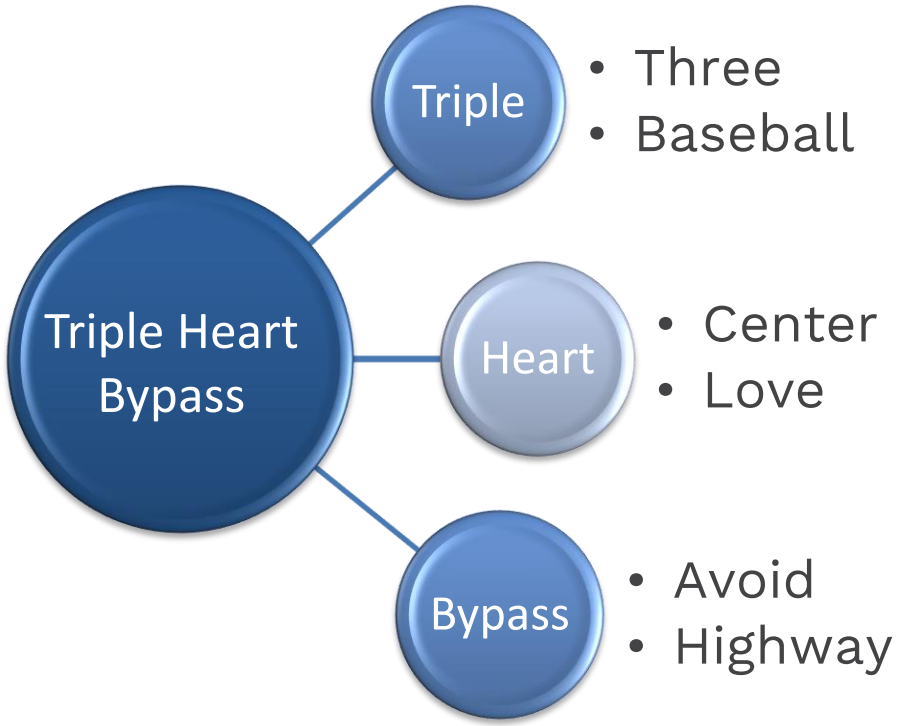
Automatic Multi-term Metadata Generation

- The conceptClassifier platform delivers metadata generation, auto-classification, and taxonomy tools
- Eliminates manual tagging

Transforms Context into Content

- Extracts meanings from Structured and Unstructured data

Netwrix: Why Are We Different?

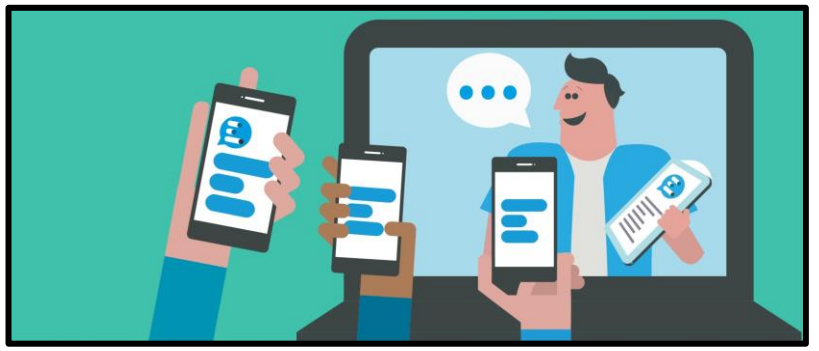


Differentiators

- Unique IP *Compound Term Processing*
- Identifies multi-word terms that form a complex entity
- Ambiguity inherent in single words is eliminated
- Works in any language, regardless of grammar or linguistic style
- Persistent index reduces collection time and load
- Powerful weighted clues improve accuracy

Poll: Would you like us to follow up with you?

- By Email
- By Phone
- Not at this time



Next Up:

➢ *Closing and Extended QA*

Resources: Webinars and Links

Upcoming Webinars

- **August 19th**: *Maximizing O365 User Adoption*
- **September 23rd**: *Electronic Forms, Workflow, and Robotic Process Automation*
- **October 21st:** *Advanced Content Services for Office 365 including Document Capture, Indexing, and OCR*

On the Web

- Webinar Presentations and Recordings
- Netrix for Records Management
- Introduction to Managed Metadata

ShareSquared: Who We Are

- Solution provider specializing in SharePoint/O365 for 15 years
- Collab/Intranet, EDRMS, Workflow, Migration, etc. Solutions
- Reseller and Implementation Partner for products such as:

Gimmal
netwrix
KnowledgeLake
The SharePoint ECM Company
RecordPoint
Plumsail
Aquaforest Searchlight 365
harmon.ie
DocuSign
wand
precision classification for search
bindtuning
VisualSP
SKYSYNC

Closing: Call to Action and *Extended* QA

Contact us to:

- Brainstorm about your project/initiative
- Schedule an in-depth demo
- Get an estimate for a project
- Get pricing for a product we resell (e.g. Netwrix, Gimmal, Harmonie)
- Discuss a Proof of Concept
- Get a sample RFP solicitation
- Get information about successful O365 projects

Contact Info:

ShareSquared, Inc.
David Kruglov
David@ShareSquared.com


share²