VALORA technologies

eBook: AutoClassification 101

Table of Contents

What is AutoClassification?

AutoClassification is a suite of Machine Learning software that automates the analysis and classification of digital content or files. Software algorithms perform the work, not people, – thus "AutoClassification."

> The only way to understand *what* your data is, is to *classify* what it is.

Classification of documents assigns contextual attributes (rich metadata tags) based on the actual content of the file, not just the file type. File metadata defines the operating system attributes of the files, such as Creation Date or Registered User. Rich metadata defines the contextual attributes based on the content of the document, such as Document Type (Quarterly Sales Forecast, W-2 Tax Filing, Radiology Report, etc.), Document Category (health record, mortgage application, etc.) and Risk Level (based on content sensitivity, such as types of personal data or PII).

> **Classification** answers the question: "What is this content and what should I do with it?"

By tagging or classifying enterprise data and understanding the context of the content, organizations can better make decisions on what to do with information assets, where they belong, who should have access, and where and for how long to store it.

> **AutoClassification** automates the analysis and decisioning of the proper answers at all times during the files' content lifecycle.

File Metadata vs. Rich Metadata

Metadata provides detailed information about files and lets us understand what they are and what they contain, and therefore how they need to be managed. Without metadata, any two (or more) files can look and feel the same.

The ability to not only identify, but then also apply rules and actions to files based on content and context, facilitates streamlined, enterprise-wide, information governance.

> A good Information Governance strategy starts with the complete understanding of content through the application of metadata - to identify **what** the content is, **where** it lives across all data repositories, and **how** each piece of content will be handled - now and in the future.

File Metadata

Considered “tombstone metadata,” file metadata defines *what* each document is at the file level only, and is system-generated when a file is created. It defines the physical properties of a file including file application type (Word, Excel, Email) and often other standard attributes such as User Account, Created Date, and Last Modified Date. While helpful, file metadata is limited and does not give insight as to the contents of the file, such as its purpose; its intended users; important people, topics or business areas mentioned; or how sensitive its contents may be.

Rich Metadata

Rich metadata applies attributes about the file *contents*. Applying rich metadata allows us to apply further attributes related to the contents of a file to "see" and "understand" what the file is, so as to be able to define and action it appropriately. Is it a contract? A blueprint? An employee file? A record? Does it contain personal data and need to be protected? Should it be under Legal Hold? Is it past its retention period? Is it duplicative or an older version of another file?

Why AutoClassification?

AutoClassification is used where there are large amounts of disparate content across many data stores and locations. AutoClassification of data facilitates unprecedented:

Speed.

AutoClassification software processes content at a speed that no human, or team of humans, possibly can. AutoClassification software performs 10,000,000 computations per second, saving months or years of manual file categorization and attribution.

Accuracy.

Sophisticated algorithms remove the possibility of human error, oversight or malicious actions through the automated crawling and classification of every file – nothing is missed, overlooked or misused.

Consistency.

Automating the classification of data allows for consistent and perpetual reconciliation across multiple data repositories, regardless of time of day, time of year, data set size or location.

Detail.

AutoClassification typically provides dozens of fielded metadata tags, far more than would be practical to perform by hand, providing rich, faceted data for sophisticated, content-based, search, retrieval, visualization and reporting.

How AutoClassification Works

AutoClassification software uses both pattern-matching algorithms and Machine Learning to detect file contents and attributes, and then assigns contextual attributes (rich metadata) and disposition (rules) for each document or file. AutoClassification answers: What is this content? and How should it be managed throughout its lifecycle?

AutoClassification software applies a 5-step methodology for locating, identifying, analyzing, actioning and monitoring content across multiple data stores.

1) LOCATE

Inventory all content and data across the enterprise.

Just knowing what you have and where it's stored is a challenge for most organizations. AutoClassification software crawls, scans and locates content from disparate data sources across the organization, from emails and embedded attachments, to shared drives and cloud storage environments. AutoClassification leaves no file unaccounted for across disparate data sources including:

- Single and multiple shared drives
- Email repositories and servers
- Document Management Systems (DMS) & Enterprise Content Management Systems (ECM)
- Collaborative sites (SharePoint, Box, Dropbox, Drive)
- Personal drives and laptops
- eDiscovery repositories
- HR, ERP & billing systems
- Scanned paper records and contracts
- On-prem and cloud-based document repositories

2) IDENTIFY

Determine the kind, type and context of all files and documents.

Not all content is created equal. AutoClassification allows organizations to identify content that holds actual enterprise, legal or regulatory value (real content important to your business) versus content that is considered Redundant, Obsolete or Trivial (ROT), such as duplicate files, temp files, junk mail and out-of-office emails. It allows for the identification and flagging of documents that contain personal data or other sensitive or confidential information. It can identify different document *types* by the content of the files, such as contracts, mortgage applications, health information and blueprints.

3) ANALYZE

Understand your data and make informed business decisions.

Evaluate and analyze your data to better understand retention, privacy, data residency, data control and processing, cross-border requirements, and trends and activities across your business. Use classification tags to make business, legal and financial decisions around corporate and sensitive information, and ensure compliance with corporate policies, as well as state, federal and international regulations. View reports, preview documents, search PDFs and images files through automatic OCR, transcribe audio and video files, and translate foreign language files into English to understand content and context.

4) ACTION

Establish processes and workflows to classify and manage content.

This where the magic happens. AutoClassification software uses its rules-based and Machine Learning algorithms to automate the disposition and handling of all content across one or multiple data sources, and then it tells it what it is and what to do with it. These rules typically follow an IF-THEN format. For example, IF the DocumentType = Contract AND the Date < 1/1/2025, THEN Mark the File as ACTIVE (Retain). Through these rules-based workflows organizations can:

- Apply or amend rich metadata to files
- Apply retention schedules and legal hold
- Apply security access controls
- Migrate content on demand
- Delete and sequester specific content
- AutoRedact sensitive information
- Initiate custom workflows

5) MONITOR

Automate the application of processes and procedures.

AutoClassification software implements a tooled and automated approach to Content Management. It perpetually runs in the background across all data silos without interfering with other systems or business processes and with no performance draw on systems or repositories. It constantly monitors and audits data environments for new and edited content and automatically applies appropriate rules to identify, classify, protect, move or delete specific data. Machine Learning rules automatically update with changes in strategy, systems, personnel and regulations for “always on” or “evergreen” approach to content management.

Practical Applications of AutoClassification

AutoClassification software is flexible, scalable and customizable to fit the needs of each business type, content challenge and data management need. Organizations use AutoClassification software for:

Records & Information Management

Analyze, manage and automate large-scale and customizable end-to-end records management solutions, tailored to your Retention Schedules and policies.

Retention & Content Lifecycle Management

Easily implement retention schedules and workflows and ensure files are only kept for the amount of time required per policy, per business process, per repository.

Content Migration

Connect with different on-prem and cloud-based content management systems to analyze data, remove unnecessary content, classify and apply rich metadata, including Recommended File Location in the new repository or storage systems, and migrate or consolidate content from one place to another.

Data Privacy & Security

Locate and identify files that contain personal data (PII, PHI or PCI) and identify documents that may be sensitive in nature (contracts, employment agreements, etc). Helps satisfy GDPR, CCPA and other emerging privacy regulations.

Legal & eDiscovery

Gather and analyze relevant documents for Early Case Assessment (ECA), identify appropriate content to be placed under Legal Hold, and migrate final data sets to third-party review platforms.

Additional Resources

Valora Technologies website

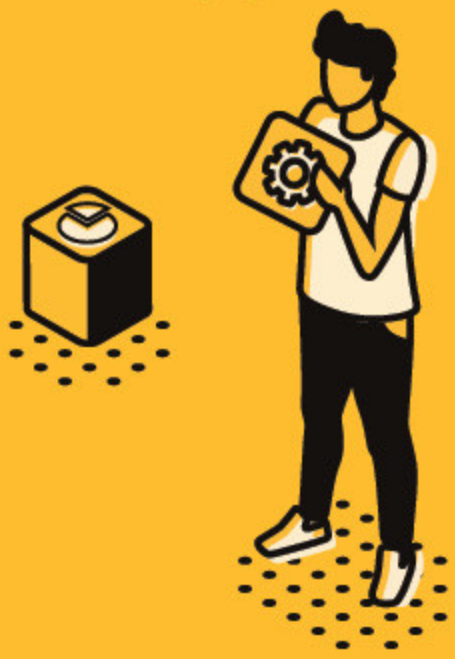
eBook: 8 Pillars of Information Governance

Webinar: What is AutoClassification and Why Should I Care?

Webinar: Universal AutoClassification Master Webinar Series

Video: Meta...What? Metadata! ~ National Archives of Australia

Schedule a Demo



VALORA technologies

Valora Technologies, Inc. is a leading provider of AutoClassification technology that helps solve corporate Enterprise Content Management and Information Governance challenges.

Through a combination of proven methodology, industry best practices and its powerful technology platform, Valora enables organizations to leverage flexible and scalable enterprise-wide management of AutoClassification, data mining, rich metadata application and strategies, content analytics, document intake and visualization, and document lifecycle management in some of the most complex data environments in the world.

For more information, visit www.valoratech.com