The IT Leader's Guide to Managing Unstructured Content

everteam

everteam

The IT Leader's Guide to Managing Unstructured Content

Every organization creates and stores information. Traditionally, IT leaders focus on managing the structured data created by applications and stored in databases. But it's no secret that the amount of information created and retained is growing exponentially every year and the vast majority of it is unstructured data, in the form of files and documents.

Unstructured data is growing faster (velocity), in greater amounts (volume) and more formats (variety). So IT leaders are confronted with an urgent need to expand their focus to include deploying strategies, processes, and systems to manage unstructured data.

This paper is meant to provide a starting point for IT leaders who are looking for a place to start and a way to manage their initiatives.

Information Governance and Unstructured Data

Gartner's definition is one of the best, if not slightly complex definitions of information governance (IG):

".. the specification of decision rights and an accountability framework to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information. It includes the processes, roles and policies, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals."

Note that nothing in that definition limits IG to structured data. What it does say is that IG is a combination of people, process, and technology, and the role of technology is to support the strategy and enable the processes.

IT leaders have been adopting technologies to manage and leverage structured data for years, from data marts and warehouses to master data management and data lineage solutions. But the reality is that there is no one platform or product that can solve all your IG challenges, and unstructured content requires new strategies and new tools.

With advancements in artificial intelligence such as machine learning and natural language processing, the technology to support an IG strategy for unstructured data is available today. Understanding the tools available requires a framework for thinking about the available solutions and how they can support your unstructured content governance strategy.

" Increased data growth over the past decade has created an unstructured data nightmare," says Alan Dayley, research director at Gartner. " It's not just the cost to store it. Huge volumes of dark data make it harder to find what is useful and may mean we miss business opportunities. "

https://www.gartner.com/smarterwithgartner/how-to-tackle-dark-data/

An IG Framework for Unstructured Data

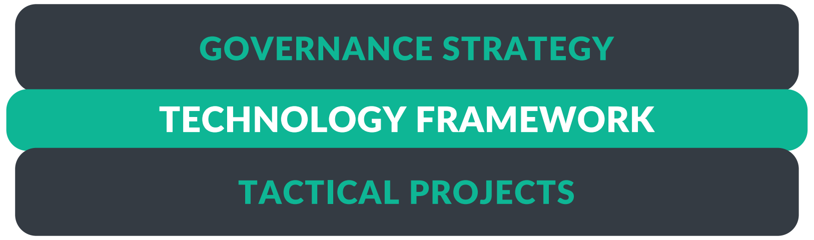
Implementing an effective IG strategy for unstructured data is an evolutionary process that will unfold over several years. While that's happening, business is, of course, ongoing; your company is still creating more information and storing it in various locations across the enterprise. The reality is that you can't afford to wait for an IG strategy to be fully implemented and all the IG processes to be in place before you start taking action to address the risks and requirements for managing unstructured content.

An ideal approach is to take on one of the basic IG use cases, which allows you to build an initial process and implement a set of software tools, and then feed the results back into the overall governance initiative. For example, a first initiative could be to decommission a legacy application or to identify and remediate the "dark content" in a set of shared folders or SharePoint Libraries. Or, a first project may be driven by the need to comply with a key regulation such as GDPR, CCPA, or NYDFS.

These use cases are all examples of information governance projects that you can work on in the near-term, in parallel with the development of an overall IG strategy. They are tactical projects that deliver immediate results, executed under the auspices of the strategic initiative.

Supporting these tactical projects is where an information governance technology framework plays a critical role, by ensuring that each project leverages and builds on a common toolset.

Think of an oreo cookie. The overall governance strategy is the top part of the cookie, and the tactical projects are the bottom part of the cookie. The technology framework is the middle layer, providing a consistent, repeatable approach to executing IG projects while leveraging a standard set of tools, technologies, and processes.



A simple way to think of the IG technology framework is as a series of seven steps that you can use to analyze the requirements for a wide variety of IG projects. Not all projects will use every step of the framework, and some projects may move back and forth between steps. The key is that each step offers a standard approach and a set of tools that you reuse across multiple projects, allowing you to leverage your investment and build reusable best practices.

An Information Governance Technology Framework for Unstructured Data

Let’s examine each component of the technology framework in detail.

CONNECT

IDENTIFY

CLASSIFY

GOVERN

ACT

ARCHIVE

ENFORCE

CONNECT

Regardless of the IG use case or project, a first step is to connect to one or more content repositories. The right way to do that is to have a standard approach to how you connect to various repositories throughout the organization.

With information located in a variety of locations and formats, previous approaches to IG would suggest that you ingest content from across all these systems to one central location. But today, a proven approach that is much more effective is to manage the content "in place;" that is to leave it where it is rather than collect and duplicate it. Your framework needs a way of connecting to all types of content stores, wherever it is.

The right technology will provide a standard set of connectors. In some cases, these are connectors to known systems such as SharePoint and Exchange. In other cases, a standards-based connector built on CMIS or JDBC will enable you to connect to enterprise content management and business application repositories in a standard way. The connector approach should be expandable to allow rapid development of custom connectors to unique or in-house systems.

With an effective approach to connecting to every system that creates data, the governance system can search across repositories, retrieve file properties and metadata and analyze the full text of every item the repository contains.

In a nutshell, an effective connector architecture will allow you to:

- Connect to sources of content across the enterprise
- Retrieve and index file properties, stored metadata, and full text
- Connect using standard connectors, standards-based connectors, and easily configured custom connectors.

everteam

IDENTIFY

The Identify step in the framework is about identifying and analyzing the content you have stored in the connected repositories. With the right software, you can take advantage of artificial intelligence technologies like machine learning (ML) and natural language processing (NLP) to identify items based on properties, locations, metadata and the content of the text to determine what content is redundant and no longer needed, trivial (not required) and obsolete (no longer used). You'll also be able to see what personally identifiable information (PII and PCI) is stored in unsecured repositories.

In a nutshell, the Identify component of the framework is all about:

- Exploring content to identify patterns for potential classification
- Locating personally-identifiable information (PII, PCI)
- Identifying duplicate items
- Identifying obsolete files

CLASSIFY

The Classify step is where the bulk of the work happens. During this step, you use the results of your analysis to enhance metadata and classify every item across all the connected repositories.

With the right tool, you don't have to do this manually. Tools that leverage machine learning and natural language processing can automate much of the classification process. You can classify the content (or reclassify it) using a defined taxonomy, including classifying based on an existing taxonomy of record types (records managers call this the record series).

For example, using named entity extraction (a type of NLP), you can identify all references to types of items, such as company names, individual names or locations. For any file where you find an item of that type (a "named entity") it is extracted and added as a label, which in turn can be used to help classify that item.

During the classification step, you can also label and set aside a set of files that need some action taken on them, such as deleting them because they are redundant or trivial and not required, or routing them to be reviewed and approved for archival because they are no longer needed. This is where you tie in a workflow to take action, forwarding the group of files to a workflow process where an auditable trail of actions is taken.

As you can see, the Classify step comprises two fundamental processes:

1. Apply what we've learned to tag and classify content
2. Submit content to a workflow to take action on it in a structured, auditable and defensible manner

There are many tools on the market that provide a subset of these discover capabilities. These file analytics tools help you understand what content you have across the organization. But this is only the first step. If you are involved in a tool selection processes, make sure it provides the ability to take the next step - cleaning, organizing, classifying and taking action on the information once you have identified it.

In the Classify step of an IT framework for governance, the system should:

- Analyze file content to identify document types and categories
- Enhance the metadata associated with each item
- Determine a classification for every item based on a taxonomy of information types
- Take action on selected content through workflows

GOVERN

The Governance component of the framework plays the critical role of holding the instructions on how every item, across all the connected systems, should be managed. This includes a comprehensive set of retention rules, information policies, and data life cycles which, together, define where an item should be stored and who has access to it at every point in its lifecycle, as well as when it should be disposed of.

A complete solution for this part of the framework may be called an information catalog or information registry. It acts as the place that every other part of the framework can go to answer the question, "how should an item of this type be handled?" That way, once items are classified, the right policies can be applied to them by other parts of the system.

A complete information catalog will also keep track of the associations for each information policy. For example, citations that answer the question of what regulations are associated with a retention rule.

The Govern part of the framework should have the ability to;

- Provide a look-up capability for other systems and users to ascertain the rules for handling every type of item stored anywhere
- Store the complete taxonomy of record types used by the organization
- Define the rules for managing each type of item across its complete lifecycle

ACT

A complete governance framework will let you take a set of actions on items or a set of items. This action may include moving the content to another repository as part of a decommissioning initiative, or to a system that is already in place for long term storage and acts as your system of record. Moving or migrating items -- or groups of items -- should be possible immediately for authorized users or based on a workflow process with review and approval steps.

In a nutshell, the Act component of the framework should support:

- Migration of content to target systems on-premise or in the cloud
- Moving the enhanced metadata to the target system as well as the source file or document
- Retaining the classification of the item to inform the target system of the record type of the items

ARCHIVE

Archive is a close relative of the Act component but includes specific capabilities to provide long-term preservation when it is desirable to move content out from the application environment -- for example when vendors charge licensing fees by volume.

An effective archive should provide the capability to store content on different media at different stages of its lifecycle to provide the most cost-effective solution; often called tiered storage.

Tiered storage not only ensures all your archived content is accessible, with the most recently archived content more readily accessible in higher storage tiers but it also helps reduce the cost of the archive without affecting compliance.

Archive solutions should also offer different levels of security to ensure employees have the proper level of access.

An effective Archive component in the IG framework ensures:

- Preservation of information and its associated metadata for the life of the item until final deletion
- Storage of items in different storage tiers based on the need to access them
- Access to dashboards that enable Records Managers and IT Administrators to easily see the volumes and state of content in the store

ENFORCE

The final component of the framework addresses the need to enforce the formal disposition process based on the defined lifecycle for each item type. These rules define who can access the file, how to organize the file, and most notably when and how to destroy it.

Managing the disposition process needs to include a workflow capability to review and approve steps in the process, and the ability to create tasks that result in the actual deletion. It is also critical to create audit records that document the entire process for every item deleted to provide auditable evidence of "defensible deletion."

Another capability that is important in this component of the framework is to manage items that need to be placed on "legal hold." A legal hold function allows an item to be preserved to meet the needs of a legal situation, where no changes can be made to an item or set of items.

The core elements of the Manage component of the framework:

- Define workflow processes that reflect the disposition process
- Automatically route records for disposition approval when they reach expiration based on defined retention rules
- Retain audit evidence for defensible destruction

everteam

Artificial Intelligence and Process Automation: Two Essential Technologies for Managing Unstructured Content

Artificial Intelligence

AI technologies such as natural language processing (NLP) and machine learning are key elements of an IG framework. These capabilities enable you to find information across repositories and classify it, enabling you to manage it using rules that govern different types of information.

As an example, NLP techniques can enable you to enhance the metadata of the objects stored in a repository based on the text contained in the file. That means that the system can identify files and records of specific types and label them appropriately based on your file plan or taxonomy. It also means that specific types of objects can be identified as records and classified by record type. Each record type can be associated with a retention rule that governs how long to retain it and when to route it for destruction.

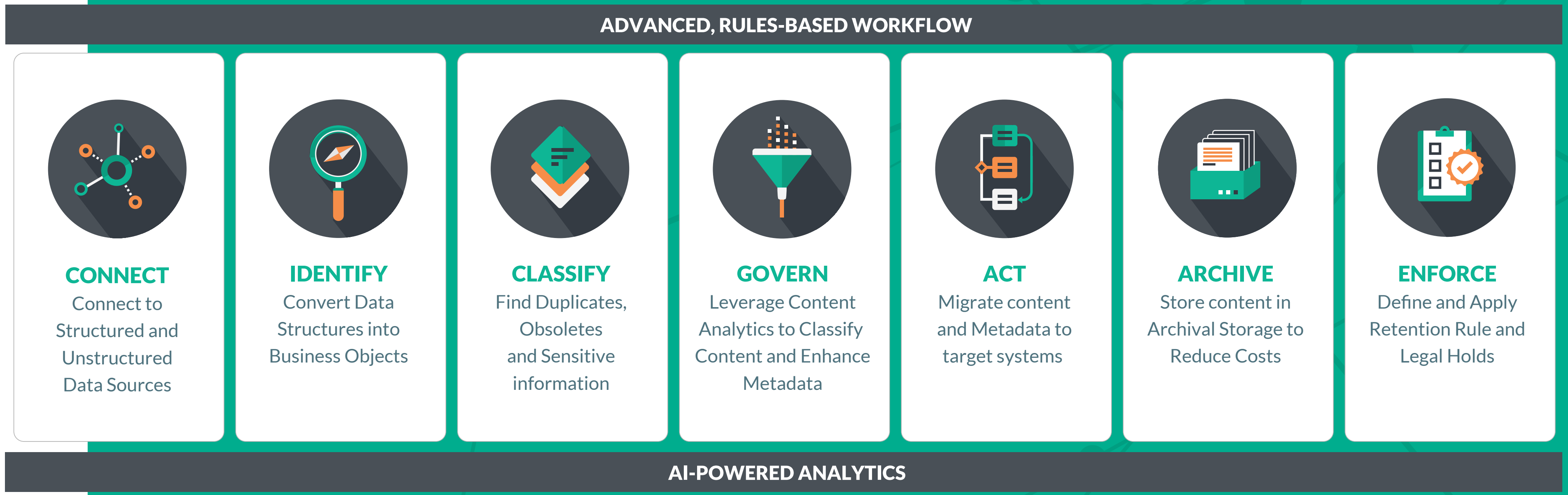
Machine learning extends those abilities by allowing the system to be “trained” to recognize different types of documents. With machine learning, a sample set of files of a particular type is provided to “train” the system, which then learns the characteristics of that type of file and looks for additional examples in the target repository. Again, the result is the ability to classify information and take actions based on that classification automatically.

Process Automation

Setting up an information governance program includes defining a set of processes and procedures that employees must follow. This is also true when working on initiatives that use the information governance technology framework. When you are working with multiple tools, processes and technologies in the framework, you need a way to orchestrate activities between them seamlessly, and that's the role of a process automation engine.

Process automation is an essential component of any successful IG technology framework. It fills two important gaps. First, it provides a workflow engine to ensure the transition between steps occurs with enough supervision and checks. Second, it provides the ability to execute data transformations. As data flows between components, it can be to be restructured and reformatted to fit the requirements of each step.

With process orchestration, you can build reusable processes that your governance applications can use with little to no change. For example, once you organize and tag a set of content as "ROT," you can kick off a process that includes setting up a workflow to route the files to a group of people for review. Once approved as ROT and ready for deletion, those files are destroyed, and audit logs created to show who approved the deletion and when they were deleted.



ADVANCED, RULES-BASED WORKFLOW

CONNECT
Connect to Structured and Unstructured Data Sources

IDENTIFY
Convert Data Structures into Business Objects

CLASSIFY
Find Duplicates, Obsoletes and Sensitive information

GOVERN
Leverage Content Analytics to Classify Content and Enhance Metadata

ACT
Migrate content and Metadata to target systems

ARCHIVE
Store content in Archival Storage to Reduce Costs

ENFORCE
Define and Apply Retention Rule and Legal Holds

AI-POWERED ANALYTICS

In our perspective, an information governance technology framework must have artificial intelligence and process orchestration capabilities available to every component. These two critical technologies allow you to automate the process of finding and classifying information, take action on it and route items from step to step based on defined policies.

everteam

Unstructured Data IG Use Cases

To help you understand how you can apply the information governance technology framework, we'll look at the six primary IG use cases you can find in your organization.

DARK CONTENT

COMPLIANCE

MIGRATION

POLICY & RECORDS MANAGEMENT

COGNITIVE SEARCH

APPLICATION DECOMISSIONING

everteam

THE DARK CONTENT USE CASE

“ I need to evaluate, clean up, remediate and categorize unknown content in shared folders or SharePoint to reduce risks and costs. ”

There are many reasons to kick off a project to find and remediate “dark content”:

- You want to reduce the volume of content being stored by finding and eliminating redundant, obsolete and trivial files (ROT).
- You need to find all personal or sensitive information in stored files across the company.
- You want to complete an audit that documents what types of content you store across the organization and its location.

Some of these projects lead to ongoing processes that you will want to run on a regular basis (like the employee leaving the company), others are one-time initiatives to look for specific things. Regardless of the type of remediation project, they all share a common set of tasks.

Let’s look at finding dark content. You may think you know where all your information is located, but you would be surprised how much content is “dark” - hidden in repositories no longer used or places your employees store them to work on them and then forget about them.

There are three big costs to dark content:

1. Corporate Knowledge: A loss of value and insights about your information that could lead to poor decision-making, a lack of innovative ideas and competitive differentiation.

2. Growing IT Costs: It costs money to store information and the more you have, the more you will pay. Also, initiatives such as application decommissioning and migrating to the cloud will take more time and could cost more money because you may be moving content that is no longer required.

3. Increases Legal and Risk Costs: Keeping information you don’t need can lead to increased legal and non-compliance risks and costs. It also results in challenges for GDPR compliance as well as barriers and increased risks for mergers and acquisitions.

A file remediation project requires you to connect to all the repositories in your organization where content could be stored and using file and content analytics, find all the information you have, organize it and then make some decisions on what to do with it. Information categorized as ROT should be defensibly destroyed and information considered as “records” must be added to a records management system, regardless of whether you manage them in place or not. Information that must be kept, but you no longer use, should be archived accordingly.

everteam

THE COMPLIANCE USE CASE

“ I need to address a compliance requirement such as GDPR, CCPA or NYDFS. ”

Most industries have to follow a number of regulations to ensure they properly manage the information they capture. There are regulations by industry such as life sciences, financial, healthcare, and then there are regulations, like the EU GDPR that apply to all companies regardless of industry.

A compliance project enables you to find all the information you need to manage to conform to regulation and apply the associated rules.

GDPR is a good example. The General Data Protection Regulation (GDPR) applies to all companies that provide products and services to people located in the EU. It focuses on the privacy and protection of consumer information. One of the articles in the GDPR states that a consumer can request a company provide all the information they have on that person, and at the consumer’s request, delete it. How can you provide this information, or delete it, if you don’t know everything you are storing about that consumer?

A GDPR compliance project involves connecting to all repositories where you store customer and prospect information, create a data inventory and organize it. If you are storing information you don’t use, you should delete it regardless of whether the person requests you delete it or not. Part of the GDPR states that you should only be storing personal information you need and have asked the customer permission to capture.

everteam

THE COGNITIVE SEARCH USE CASE

> **“ We need to get a complete 360-degree view of our information to be more efficient and drive better decision-making.”**

Cognitive search focuses on connecting all your information silos, and leveraging technologies such as machine learning, natural language processing, and entity extraction, among others, to index and analyze the information, and help you surface actionable insights. The right solution also supports the ability to tag and classify content as well as identify content as records.

For example, a knowledge worker wants to find all information related to a Customer ID across all systems. Cognitive search enables them to enter the ID and receive search results that span all records, augmenting those records with customer information from email, file shares, and other business systems, and providing a 360-degree view of the customer. Where necessary, the knowledge worker can identify and mark certain information as records through tagging and classification.

Employees need a complete view of the customer and its relationship with your organization to provide the best support. If an employee can search and find every interaction a customer has had with your company, including invoices, emails, products purchased, warranties, previous support calls, and more, they can quickly customize the support experience to help the customer resolve their issues.

With cognitive search you can connect, discover, organize and query information from multiple data sources. Not only do you get a 360-degree view of your information, but you can take advantage of AI technologies that can improve search results over time, as the system learns more about the information it is analyzing, automatically classifies and labels it, to help you understand what is the most important.

everteam

THE DECOMMISSIONING USE CASE

“ I have a legacy system from an acquisition that I want to decommission, but I need to preserve some of the content.””

Digital transformation initiatives are driving organizations to adopt new technologies and retire legacy systems. But it’s not as simple as shutting down the old one and working with the new one. You have to think about the information stored in the legacy system and make decisions about what needs to move to the new application, what you can delete, and what you archive.

An application decommissioning project will help you:

1. Reduce Costs: Often, an organization will continue to run a legacy system to keep the information somewhere, but it would reduce costs if you moved the information and shut down the system no longer needed.

2. Reduce Risks: Legacy systems that are only kept to store information are seen as a risk because they are no longer properly maintained, leaving them vulnerable to hacking or accidental exposure of information.

3. Reduce Complexity: IT doesn’t need a ton of systems to manage. It needs to focus on those systems and applications most important to supporting digital transformation efforts. When you decommission applications no longer required, you are reducing complexity for IT.

When you conduct an application decommissioning project, you are connecting to the application and analyzing and classifying its information. Redundant and obsolete information is deleted, retention rules are applied to the content you do need to keep using a records management system, and the remaining content is migrated to either an archive or the new application.

THE RECORDS MANAGEMENT USE CASE

“ I need to implement retention policies and manage our content based on retention rules, and enable legal hold. ”

When required, you can use records management capabilities to connect content with defined retention policies and rules, to enforce access rules and manage the disposition process.

As part of complying with certain regulations, you may have a set of content that you need to maintain for a period and that content should be secured and accessible to a set group of people. In this example, a records management project might entail moving certain information from an application to a records management system and applying retention rules, as well as security and audit policies.

You will need to connect to the application, search and find all the information you need to place under records management, move it to a records management system and apply the proper retention rules and other policies. You may decide to organize all the information in the system and mark which information you will manage as official records, but only move certain information as part of the project. You could then come back later and move additional records.

Another project example relates to a legal hold. If your company is part of a legal case, you are required to provide information and not change or destroy that information while the case is in process. A records management system allows you to mark information for legal hold, ensuring you cannot alter it until the hold is released. For example, if you have records marked for destruction, a legal hold will halt that process temporarily. In this example, you are only working in the Manage step of the framework.

THE MIGRATION USE CASE

“ I am moving to a cloud-based enterprise application, and I do not want to migrate unnecessary content. ”

Many organizations are talking about migrating to cloud-based applications, such as file sharing applications like Office 365, DropBox and Box. These solutions provide quick, secure access to content from anywhere at any time. But you don’t necessarily want to move all your files to a new application.

The project here focuses on the migration of your content to the cloud application. But before you do that you want to clean out the ROT, archive some content and possibly move other content to records management.

You’ll need to connect to your content repositories, index and organize your existing content, make some decisions about what to do with each content type, then migrate the necessary content to the cloud application. This is typically a one-time project for migrating to a particular cloud-based application unless you are migrating employees in phases.

everteam

Everteam Governance Solutions

The Everteam Governance Suite includes three separate but integrated products that deliver a complete solution for managing unstructured content.

everteam.discover

Everteam.discover is a solution for analysis and enrichment of your content helping you simplify regulatory compliance, preserve sensitive data, migrate documents to archives, or simply clean and sanitize your information. It is comprised of a combination of file and content analytics and leverages AI technologies such as machine learning, natural language processing and named-entity extraction to organize and find the information you need.

everteam.policy

everteam.policy is a solution for collecting, managing and publishing information policies including retention rules. It provides a central location for records managers and data curators to share a complete taxonomy of record types and the full set of access rules, defined lifecycle and retention rules that apply to each. Everteam.policy is available as an on-premise or SaaS solution.

everteam.archive

Everteam.archive includes capabilities for connecting to the systems that produce records, ingesting content from those systems when appropriate and providing the ability to search, retrieve and manage those records based on defined retention rules.

everteam

Everteam is a leading provider of information governance and enterprise content management solutions. We deliver solutions based on advanced software technologies that address business needs. Including:

- Content archiving
- Records Management
- Application Decommissioning and
- Shared Folder Remediation

Everteam is a global software vendor with headquarters in Lyon (EU) and Boston (US), and regional offices in Beirut, Dubai, and Paris.

CONTACT US:

745 Atlantic Avenue,
Boston, MA, 20111
1-617-500-1982
email: infoNA@everteam.com