ASP INTELLIGENT FILE SCANNER | AI WHITE PAPER

Discover Sensitive Data Before Your AI Agents Do

Content-level discovery for the enterprise files used to train, ground and inform AI agents.

KNOW THE CONTENT	CONNECT TO POLICY	CONTROL THE DECISION
Find sensitive information inside the files selected for AI use, including nested content.	Bring classifications and linked policy relationships into enterprise governance.	Give owners evidence to approve, restrict or remediate content before expanding AI access.

Executive Premise

An AI initiative can inherit a data-governance problem before it produces its first answer. Documents, spreadsheets, email stores and archives may contain personal information, financial records, health information or confidential business content. A repository name or file label does not establish what is inside every document. Before organizations reuse that content for AI, they need an evidence-based view of its sensitivity.

> **Sensitive information can reach an AI system through training data, retrieved context or an agent's tools. Governing only the training dataset leaves other routes unexamined.**

NIST identifies data privacy among generative AI risks, while OWASP identifies sensitive information disclosure and excessive agency as distinct application risks. [1-3] ASP's conclusion is practical: inspect the source content, connect findings to governance and use them to inform controls before expanding AI access.

ASP Intelligent File Scanner Capabilities Aligned to These Priorities

- Content-level discovery across eight repository types: AWS S3, ADLS, GCS (buckets), Google Drive, OneDrive, SharePoint, SFTP and File Share.
- Inspection across 57 file types and 71 supported extensions, including PDFs, Office files, images, email stores, archives and structured/semi-structured content.
- Recursive archive and attachment inspection, including Outlook PST/OST stores and supported nested archives up to 10 levels, subject to configured safety limits.
- Editable detection rules, precision dictionaries and dependency-aware logic; file-level composite classifications when defined combinations are detected.
- Matched-value evidence supports classification review. Native Google Docs, Sheets and Slides extend discovery across collaboration content.
- Results published to Informatica Cloud Data Governance & Catalog (CDGC), where linked corporate policies and Intelligent Dashboards reveal affected sources, files and classifications.
- Incremental scans, checksums, checkpoints and retry controls support repeatable discovery as repositories change.

The scanner extends Informatica CDGC with content-level discovery. It supports data-selection and governance decisions; the AI platform and surrounding security systems remain responsible for enforcing access, filtering retrieval and controlling agent actions. [6]

Three Ways Enterprise Files Reach AI

Training, grounding and agent access are different mechanisms. Each needs its own approval boundary, even when all three use the same repository.

Route	What happens	Where discovery contributes
Training or fine-tuning	Selected examples are used to update model parameters. Sensitive content may enter the training corpus.	Inspect candidate source files before dataset assembly. Owners decide what to remove, redact or approve.
Retrieval and grounding	Retrieval-augmented generation (RAG) supplies selected content as context at inference time; retrieval itself does not retrain model weights.	Classify source material before indexing and review sensitivity when refreshing the knowledge base.
Agent tool access	An agent searches or reads connected content and may perform downstream actions within its permissions.	Identify sensitive files in the proposed access scope; use findings to inform least-privilege and handling rules.

Why the File Layer Matters

A business workflow may have well-governed database fields while related exports, signed forms and supporting attachments remain distributed across collaboration sites, cloud buckets and file shares. Copying those files into a knowledge base does not resolve their sensitivity. It changes how easily the information can be found and reused.

A useful discovery boundary therefore includes the container and its contents: a ZIP package may hold spreadsheets and PDFs; an email archive may contain messages and attachments; a shared document may combine public guidance with customer-specific details. The review should follow the content into supported nested files and capture any items that could not be inspected.

Documented Risk and Adoption Signals

- NIST's Generative AI Profile treats privacy as a lifecycle concern, supporting review of how sensitive data enters and is handled by an AI system. [1]
- OWASP describes sensitive information disclosure risks in both the model and application context. Prompt instructions alone do not establish reliable protection. [2]
- OWASP links excessive agency to excessive functionality, permissions or autonomy, reinforcing the need to constrain tools and access. [3]
- Morgan Stanley's research assistant retrieves insights from a large proprietary report collection, while Citi's agentic workspace supports more complex employee tasks. These examples illustrate growing content reuse, not a claim of control gaps. [4-5]

> **A classified repository is a better-informed starting point. It is not, by itself, an approved training corpus or a secure agent deployment.**

A Practical Discovery-to-Governance Workflow

1. Define the intended AI use

Identify the use case, repository owners, proposed corpus and allowed users or agents. Record whether files will support training, retrieval or tool access; specify approved purposes and excluded content.

2. Inspect source content

Connect to supported repositories and scan file contents, archives and attachments using the selected rules. Track successful scans separately from protected, corrupt, oversized or otherwise unprocessed items. An unscanned file is not a clean result.

3. Validate sensitivity

Review matched values and classification logic with data stewards. Tune dictionaries, dependencies and custom rules for the business context. Apply composite classifications where combinations of findings need a file-level designation.

4. Publish and decide

Publish findings into CDGC and examine linked policy relationships in dashboards. Content owners decide which files are approved, restricted, excluded or sent for remediation. Record that decision separately from the detection result.

5. Enforce and verify

Implement the decision through source permissions, ingestion controls, redaction or masking processes and AI-platform safeguards. Test that retrieval and agent identities respect the approved scope. These controls are separate from scanner classification.

6. Reassess as content changes

Use incremental discovery and checkpoints to refresh the inventory. Reconcile changed files and scan exceptions with the approved corpus, and retain evidence of review before refreshed content is made available.

Illustrative Scenario

A service team proposes a knowledge assistant over a shared support repository. Alongside general procedures, the folder contains archived customer correspondence and account-specific spreadsheets. Scanning identifies sensitive patterns in those files and their attachments; CDGC makes the classifications and associated policies visible. The owner approves general procedures, routes mixed-content documents for redaction and excludes customer archives from the initial corpus. Retrieval permissions and ingestion controls implement those choices.

This is an illustrative workflow, not a customer case study. Discovery helps identify what needs a decision; it does not independently establish consent, legal basis, document accuracy or the security of a model.

Measure Readiness Through Evidence

Measure	What it tells the organization
Inspection coverage	Successfully inspected in-scope files divided by all in-scope files; report unprocessed exceptions separately.
Sensitivity review	Files with detected sensitive content and the share reviewed by accountable owners.
Decision completion	Proposed AI-source files with a recorded approve, restrict, remediate or exclude decision.
Remediation verification	Required handling changes completed and independently checked before ingestion or access.
Inventory freshness	Age of the latest scan and elapsed time between source changes and governance review.

Business Value

For AI teams, earlier visibility can reduce the need to rebuild a corpus after sensitive content is discovered. For security and privacy teams, classifications provide a focused basis for remediation and access review. For governance teams, CDGC offers a shared view of discovered content and policy relationships. These outcomes depend on configured coverage, validated rules and follow-through by the teams that own controls.

Conclusion

AI readiness includes knowing what enterprise files contain before those files become training examples, retrieved context or inputs to agent actions. The ASP Intelligent File Scanner provides the content inspection and classification evidence needed to make those decisions with greater visibility. Combined with accountable owners and effective access and AI controls, it helps organizations expand useful AI access while governing sensitive information.

Sources

1. NIST - AI 600-1 Generative AI Profile, July 2024
2. OWASP - LLM02:2025 Sensitive Information Disclosure
3. OWASP - LLM06:2025 Excessive Agency
4. Morgan Stanley - AskResearchGPT, Oct. 23, 2024
5. Citi - Citi Stylus Workspaces with Agentic AI, Sept. 22, 2025
6. ASP - Intelligent File Scanner capabilities; September 2026 baseline

External sources establish risks and adoption examples. The workflow and business-value analysis are ASP recommendations, not claims of certification or guaranteed risk elimination. Product scope reflects September 2026; detection depends on file accessibility, supported extraction and configured rules.

Explore the ASP Intelligent File Scanner | Request a demo