



# PROTECTING PRODUCTION DOCUMENTS FOR USE IN DEVELOPMENT, TESTING, AND AGENTIC AI

## Make documents and images accessible through secure data provisioning

Development teams, testing organizations, and agentic AI systems need documents that reflect the work a business actually performs. Claims packages, loan applications, identity images, and patient records supply layouts, narratives, annotations, and exceptions that determine whether an application can interpret information and complete a workflow correctly. Yet those same production materials can contain personal, financial, health, and confidential business information.

## The Problem Is How Documents Become Available for Reuse

When an organization lacks a reliable preparation process, teams can face two unsatisfactory outcomes: production documents are reused before sensitive content has been adequately protected, or access is restricted and useful material remains unavailable.

The first increases exposure; the second can delay engineering and AI work or leave important scenarios underrepresented. A third alternative is needed – Protected Provisioning.



### UNPROTECTED REUSE

Sensitive content exposed



### ACCESS RESTRICTED

Useful content unavailable



### PROTECTED PROVISIONING

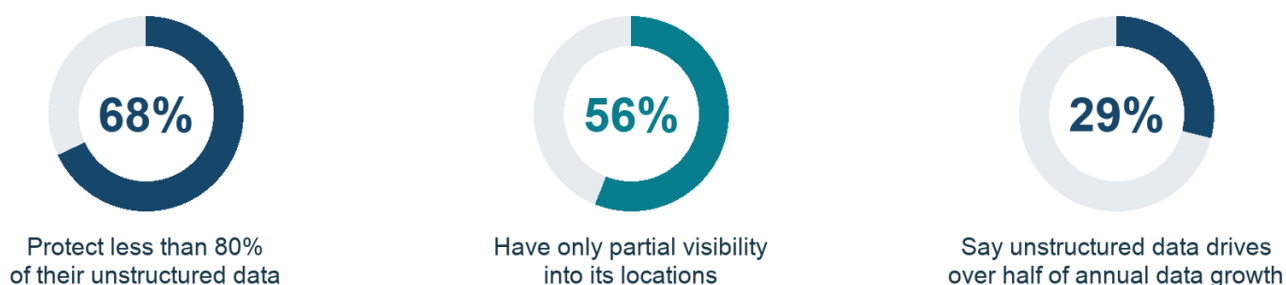
Inspect, protect, and validate

This preferred alternative is to inspect, protect, validate, and provision documents for an approved purpose. Teams receive material suited to their task, while owners retain control over what is shared and how it is used. Protection and accessibility become requirements of the same workflow.

Before production documents and images are made available for development, testing, or AI, they should be scanned to identify sensitive information such as names, account numbers, and health details. That content can then be redacted, and the prepared copies checked to confirm that the information has been removed while the document features needed for the task remain useful. Approved access and retention rules govern who can use those copies and how long they may be kept.

## The Research Shows Why Document Preparation Matters

The [Cloud Security Alliance's 2026 study](#), commissioned by Thales, shows that the ability to inspect unstructured content remains limited: only 9% of respondents reported real-time scanning capabilities, while 23% reported no scanning capability at all. These findings point to a gap in identifying sensitive information within documents and other content. Without that visibility, organizations have a weaker basis for deciding what requires protection before reuse, even when access restrictions or other security controls are already in place.



CSA 2026 • Separate questions; arcs do not represent the share of data protected  
Colored arc = respondents reporting the stated finding; gray = remaining respondents.

CSA surveyed 210 IT and security professionals in November 2025. Questionnaire co-developed with Thales; analysis conducted by CSA.

The discovery problem also appears in [BARC's 2026 unstructured-data research](#). Organizations cannot reliably select, protect, or authorize document collections they cannot locate. The two studies measure different populations and questions but point to the same operational requirement: discover content before expanding its use.



Fully know where AI-relevant  
unstructured data resides



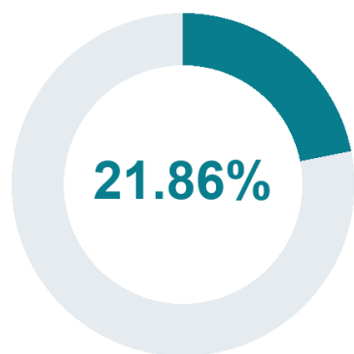
Say less than half is discoverable  
and usable for analytics or AI

BARC 2026 • Respondent percentages; separate questions about visibility and usability  
Colored arc = respondents reporting the stated finding; gray = remaining respondents.

BARC: 225 cleansed responses, February–March 2026; supported by Ohalo and Datahub. Percentages describe respondents, not the proportion of all enterprise documents.

## AI File Uploads Can Carry Sensitive Business Information

[Harmonic Security's Q2 2025 analysis](#) examined one million prompts and 20,000 uploaded files across more than 300 AI applications. About one in five uploaded files contained sensitive information. PDFs made up more than half of all uploads.



**About 1 in 5 uploaded files**  
contained sensitive information

PDFs represented more than half of all uploads.  
Other formats included Word, Excel and CSV files.

Harmonic Q2 2025 • Sensitive-file percentage covers multiple file types, not PDFs alone.

The analysis also found that files were the main source of several types of sensitive information detected in AI activity. For example, 75.3% of customer-profile findings were associated with uploaded files rather than typed prompts. The bars below show the same comparison for other information categories.

### Share of sensitive-data findings associated with files



Harmonic Q2 2025 • Each bar measures files' share within that information category.  
These percentages do not measure confirmed breaches.

Source: Harmonic-protected organizations in the US and UK, April-June 2025; browser activity only. Harmonic says sensitive content did not leave customer tenants.

A document can carry sensitive information into an AI workflow along with the content the task requires. Inspecting and protecting the material before reuse creates a controlled alternative to distributing an unprotected copy. Validate redacted or synthetic documents for their intended use, then govern access and retention.

## Agentic AI Makes Enterprise Content an Operating Resource

[Box's 2026 State of AI in the Enterprise research](#) shows why protecting documents must also preserve their availability. Agents need access to relevant enterprise knowledge, yet connecting them to trusted content remains difficult. Security and access concerns reinforce the need for approved, purpose-appropriate content rather than unrestricted production access.



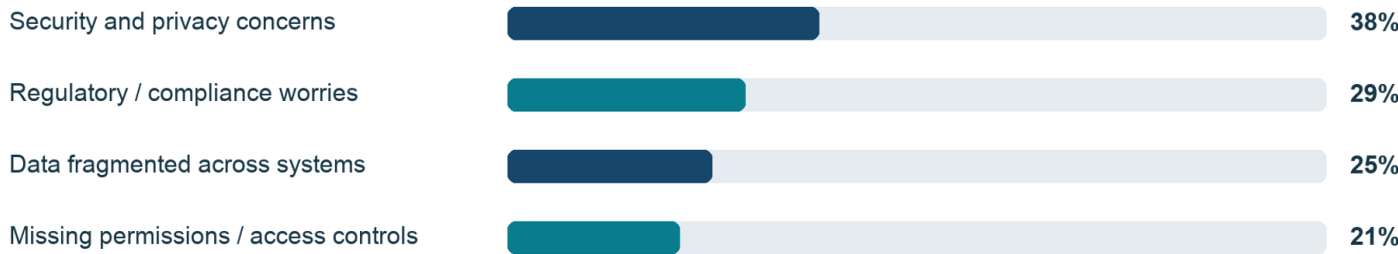
Box 2026 • Different questions and respondent bases; the two arcs are not an adoption funnel  
Colored arc = respondents reporting the stated finding; gray = remaining respondents.

### Unstructured content already used as an AI advantage



Box 2026 • Documents, contracts and reports; self-reported maturity groups

### Barriers to giving agents access to organizational content



Box 2026 • Respondents citing barriers; not verified control-failure rates

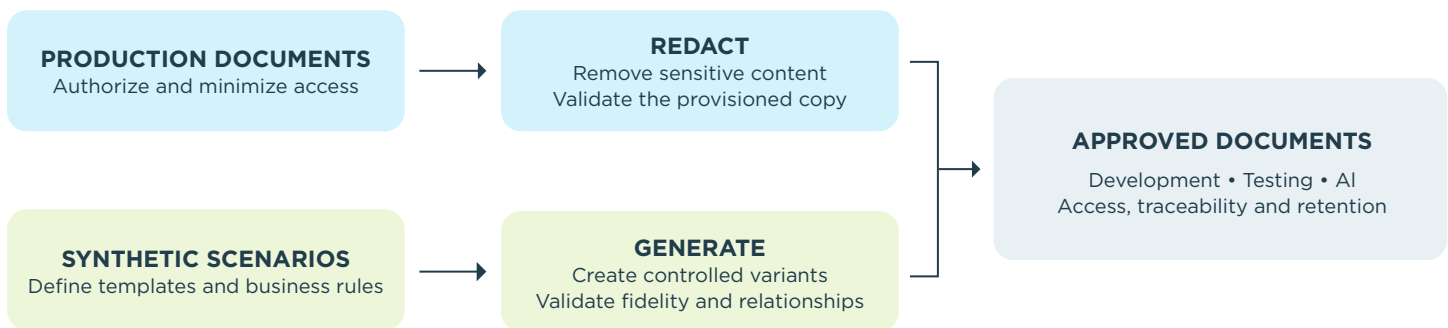
Box / Harris Poll: 1,640 IT decision-makers in the US, UK, France and Japan, April-May 2026. Maturity comparisons are cross-sectional, not year-over-year growth measures. Agent content-access barriers include organizational content more broadly than documents alone.

## What the Research Reveals: An Enterprise Data Protection Gap

The findings establish incomplete visibility and protection, sensitive content in AI file uploads, and demand for enterprise knowledge in agentic systems. **They support a clear requirement: make useful documents available through an approved preparation and provisioning process.**

## Two Complementary Approaches to Secure Document Provisioning

Secure provisioning gives engineering teams and AI initiatives a dependable way to obtain useful documents. Redact authorized production documents when authentic characteristics matter, generate synthetic documents for controlled scenarios, or combine both approaches. Each route requires validation and lifecycle controls.



Two complementary routes governed by the same enterprise controls

## Redact When Authentic Document Characteristics Matter

Production documents can contain complex layouts, annotations, handwriting, and image-quality variations that are difficult to reproduce. Controlled redaction removes sensitive content while preserving features needed for the task. Protection must address every relevant occurrence: a patient's name may appear in a header, a record number in a table, and identifying details in a clinical narrative.

## Generate When Controlled Scenarios Can Replace Production Records

Synthetic documents use designed values, templates, and business scenarios without incorporating actual consumer or patient records. A lending team can create mortgage packages across income levels, employment histories, and loan products; a healthcare team can create representative documentation for classification, extraction, and workflow testing. Controlled variations also support missing information, negative tests, edge cases, and repeatable failure analysis.

## Accountability and Responsible AI Start with Provisioning

Responsibility crosses organizational boundaries. Business owners authorize the use of information; governance teams establish classification and lifecycle standards; security and privacy teams determine controls and permissible uses. Application owners manage repositories, while engineering teams define representative layouts, exceptions, and related-document scenarios. **A legitimate testing need should lead to an approved provisioning process with accountable ownership, rather than an informal document export.**

A repeatable provisioning process should address six requirements:

- **Classify.** Identify PII, PHI, confidential financial information, and other sensitive content, including material outside predictable form fields.
- **Authorize.** Name the accountable owner and define who may approve production-derived document access for each purpose.
- **Protect.** Apply and validate redaction, de-identification, or other appropriate controls before distribution.
- **Offer synthetic alternatives.** Provide representative scenarios that satisfy engineering needs without production records whenever practical.
- **Trace.** Record sources, transformations, validation results, approvals, and provisioning activities.
- **Manage the lifecycle.** Enforce access restrictions, retention periods, and disposal across development, testing, vendor, and AI environments.

These requirements extend existing enterprise security and governance practices. [NIST SP 800-53](#) offers a broader control framework for access, audit, and information handling, while the [NIST AI Risk Management Framework](#) emphasizes privacy, security, reliability, and accountability across the AI lifecycle.

## Govern AI Data Flows Alongside Model Development

Define what the task must demonstrate before choosing its document inputs. Classification, extraction, workflow testing, retrieval evaluation, and agent development may require different layouts, scenarios, and permissions. Synthetic documents can provide controlled coverage; authorized redacted samples can preserve real-world complexity. A combined workflow can offer both.

## Business Value and the GenRocket Approach

Controlled document provisioning can reduce unnecessary copies of sensitive production material, shorten preparation and approval cycles, and improve coverage of unusual conditions and exceptions. It gives security and engineering teams a shared process for delivering useful data. Measure progress through provisioning turnaround time, scenario coverage, protection effectiveness, and the number of nonproduction environments containing sensitive documents.

## How GenRocket Supports Secure Document Provisioning

GenRocket's *Unstructured Data Accelerator* supports controlled document preparation. UDA-Redact detects PII and PHI in supported PDFs and images, with operator review and permanent redaction before provisioning. Batch processing, offline deployment, and audit capabilities support a controlled workflow. UDA-Generate provides synthetic documents using templates and designed data for document-centric testing and AI use cases.

### Two Paths to Compliant Unstructured Data

Pick based on what you have and what you need — or combine both for an end-to-end workflow that evolves production data into scalable synthetic data.

Claims Document – Page 1 • PROCESSED

|                   |               |
|-------------------|---------------|
| Social Security # | [REDACTED]    |
| Patient Name      | [REDACTED]    |
| Claim ID          | CL-48201-2025 |
| Address           | [REDACTED]    |
| Provider NPI      | 1750398421    |

#### UDA-REDACT

##### Start with Real Data

Use the fidelity of existing production documents while permanently removing the PII/PHI that prevents them from being safely used.

- ✓ Maximum fidelity to production documents
- ✓ Existing layouts, formats and naturally occurring edge cases
- ✓ Auditable, on-prem PII/PHI removal
- ✓ Production-derived data for testing or AI training

Synthetic Template • UNLIMITED VARIANTS

|                       |                     |
|-----------------------|---------------------|
| Volume                | On demand           |
| Scenarios             | Positive + Negative |
| Edge cases            | Fully controlled    |
| Referential integrity | Preserved           |

#### UDA-GENERATE

##### Create the Data You Need

Move beyond what happens to exist in production – generate documents, images and unstructured content with fully controlled conditions.

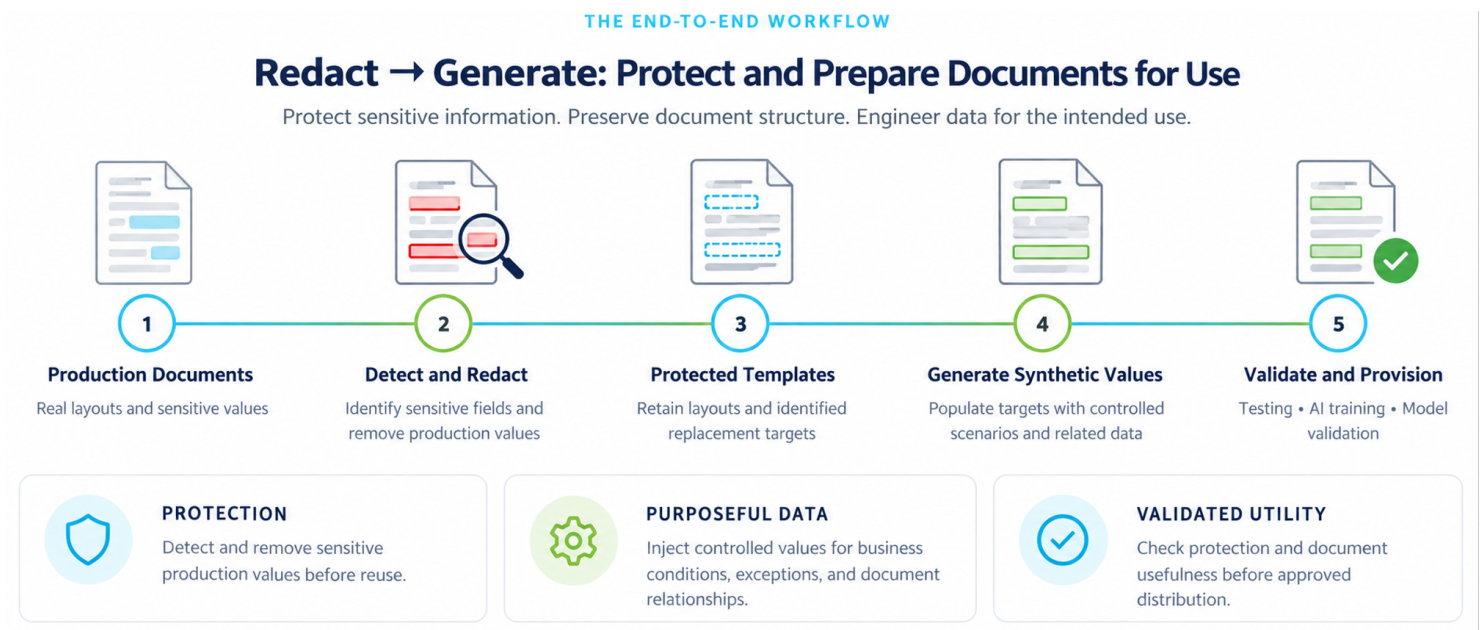
- ✓ Unlimited data volume on demand
- ✓ Positive and negative scenarios
- ✓ Rare edge cases that don't exist in production
- ✓ Structured and unstructured data stay logically connected

UDA-Redact prepares document collections selected for an authorized use. Its detection, review, and redaction capabilities work alongside the organization's discovery, access, and retention controls. Prepared documents can then be used directly or serve as templates for controlled synthetic generation.

## Redact and Generate in a Combined Workflow

UDA-Redact automates the identification and removal of sensitive information from production documents. The identified fields become replacement targets in protected templates. UDA-Generate populates those targets with controlled synthetic values while preserving document layout and structure.

Teams design business conditions, exceptions, and relationships across documents to create purposeful variants for testing, AI training, and model validation. Check protection and document usefulness before approved distribution.



## Five Questions for Enterprise Leaders

- **Where are documents reused?** Can the organization locate sensitive document collections across development, testing, vendor, and AI environments?
- **Who authorizes access?** Are business owners, privacy, and security involved in decisions about production-derived samples?
- **How is protection validated?** Do controls address sensitive content throughout the document and confirm that removed information cannot be recovered?
- **Could synthetic documents meet the need?** Have teams assessed whether designed scenarios satisfy engineering and AI evaluation requirements?
- **Can compliance be demonstrated?** Are provisioning activities traceable and governed by access, retention, and disposal policies?

## Data Protection Must Follow the Data

Sensitive information remains sensitive when embedded in a PDF, extracted from a scanned form, or copied into testing. Clear ownership, consistent controls, and validated alternatives allow enterprises to make realistic documents available while reducing unnecessary exposure. Integrating redaction and synthetic generation into enterprise data provisioning supports software quality and responsible AI without treating security as a separate downstream task.

### About GenRocket

GenRocket provides enterprise test data generation and provisioning technology across structured and unstructured data for software development, quality engineering, and AI initiatives.

Learn more at [www.genrocket.com](https://www.genrocket.com).

