

Unstructured Data & AI Governance
Survey 2026 by GDPR Tech

Information quality deserves a place in strategic AI risk oversight

EXECUTIVE INSIGHT REPORT



Enterprise AI brings the condition of underlying information into sharper focus. This expert sample points to gaps in visibility, information quality and the controls that make information dependable.

AI governance depends on the information underneath

The findings are based on an exploratory international expert survey of 86 professionals across Europe, North America and Asia-Pacific, as well as respondents representing globally operating organisations. Participants worked in privacy, information governance, compliance, technology, business and executive roles.

58% of respondents consider unstructured data highly important or critical to their organisation. Yet the survey reveals substantial gaps in visibility, retention and governance maturity. The results suggest that information quality deserves a place in executive AI risk oversight, alongside visibility, access and lifecycle controls.

18 %

Only 18% have good or accurate visibility

Only 15 of 85 valid responses report fairly good or accurate, continuously maintained visibility of unstructured data volumes.

73 %

Sensitive information leads the risk agenda

Uncontrolled storage of sensitive or personal data is the most selected risk. Access controls and approved information locations belong in the same discussion.

53 %

Outdated information and incorrect or low-quality information were each selected as significant risks by 53% of respondents.

Fewer respondents select explicitly quality-focused development priorities. The two questions are not directly comparable.

65 %

Unstructured data governance is immature

Unstructured data practices are not established or initial / ad hoc. This also applies to 54% of production AI or analytics users.

7 %

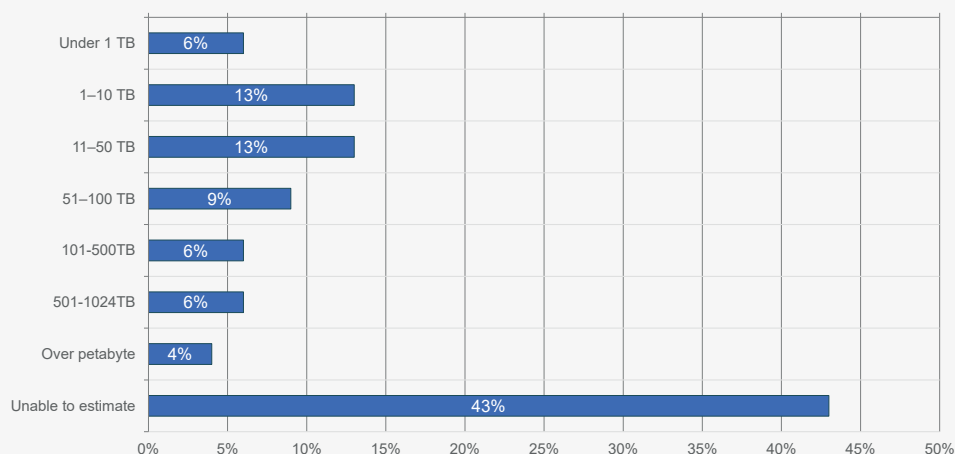
Only 7% report implemented and monitored retention and deletion practices

*Percentages use valid responses to each question.
Multiple-response handling and question-specific bases are explained in the methodology.*

THE INFORMATION LANDSCAPE

Information scale calls for software-assisted analysis

Estimated total amount of unstructured data



The distribution combines a large visibility gap with substantial scale among those able to estimate. These are self-reported volume bands, not a measured inventory.

The information is spread across everyday repositories, including email, collaboration spaces and file shares.

ILLUSTRATIVE SCALE REFERENCE

At an assumed average size of 1 MB per document, one terabyte corresponds to approximately one million documents using decimal units. This illustrative assumption is informed by GDPR Tech's experience in data assessments over the years.

An illustrative scale reference, not a survey finding. Actual file sizes vary substantially.

EXECUTIVE IMPLICATION

At tens or hundreds of terabytes, manual review is not a realistic primary method. Data management and information

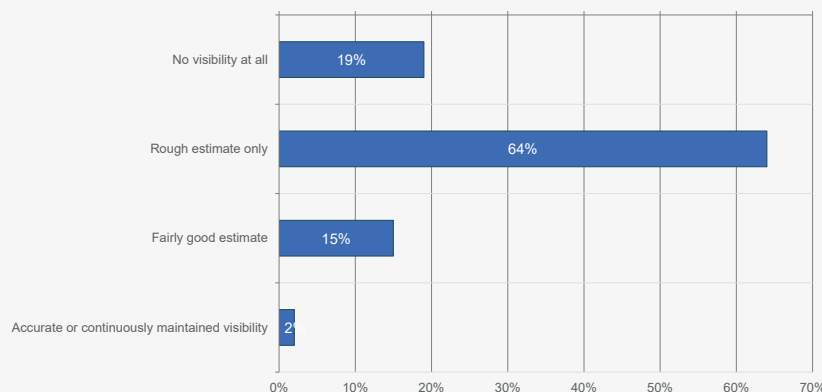
governance tools support discovery, metadata and permission analysis, and content classification at scale. Humans validate and interpret the results, assess risk and make governance decisions.

VISIBILITY

The information estate is still an estimate

Limited visibility makes it difficult to set a defensible scope for AI use.

Visibility of unstructured data volumes



Only
18 %

have good or accurate
visibility of unstructured
data volumes

Only **7 %** report broadly mapping, analysing or assessing unstructured data in the past two years.

In a separate question, **43 %** are unable to estimate total volume.

These measures describe visibility and assessment coverage, not the accuracy of a technical inventory. Together, they suggest that selecting trustworthy sources is difficult when the information landscape itself is only partly known.

OPEN TEXT RESPONSES

Question: "What is the single biggest challenge in managing unstructured data in your organisation?"

"Inventorying and understanding usage by business users."

"No clear understanding of how much of the data is still relevant."

EXECUTIVE IMPLICATION

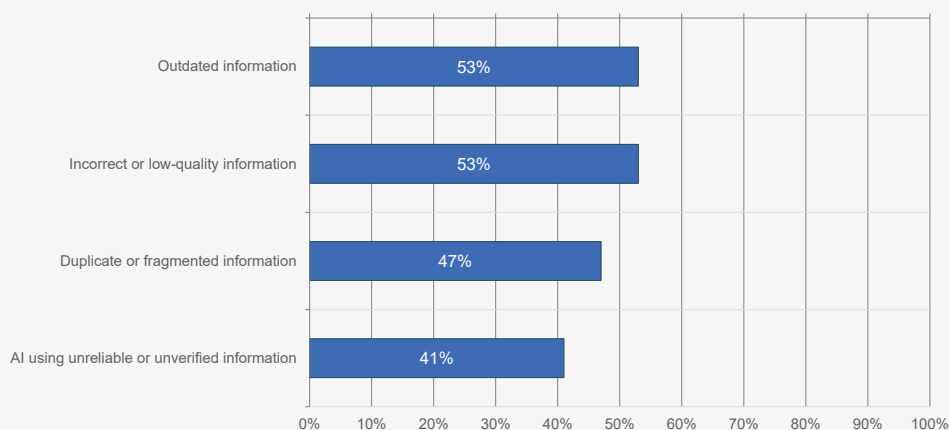
Start with the repositories needed for a specific business use case. Record location, ownership, sensitivity, currency and access, then expand coverage.

INFORMATION QUALITY

Quality is recognised as a risk but less explicit as a priority

The gap is a signal for executive scrutiny, not proof that quality work is absent.

Information quality risks selected



Multiple responses allowed; percentages do not sum to 100%. Selected answer options.

15 % Poor or inconsistent data quality Explicit priority for the next 12 months

10 % AI using outdated or obsolete sources Explicit priority for the next 12 months.

The risk and priority questions use different option lists and are answered by partly different groups. Their percentages should not be treated as a like-for-like measure of underinvestment. Quality work may also sit within mapping, classification or AI readiness programmes.

The strategic concern is nevertheless clear: making information easier to retrieve does not establish its authority or currency. An AI governance programme needs explicit decisions about trusted sources and the handling of stale or conflicting content. This is an interpretation, not a measured AI outcome.

EXECUTIVE IMPLICATION

Give information quality a named owner and measurable acceptance criteria within each AI use case.

ACCESS AND SENSITIVE INFORMATION

Uncontrolled sensitive data is the leading risk

Risk perceptions connect information location, permissions and AI access.

73 %

selected sensitive or personal data in uncontrolled locations as a significant risk

This was the most frequently selected risk. These responses describe perceived risks, not confirmed incidents or measured exposure.

48% say classification is not defined for unstructured data. Without a consistent view of content sensitivity and purpose, access decisions are harder to apply and review.

For connected AI services, verify the effective access boundary in practice and test what different users can retrieve.

OPEN TEXT RESPONSES

Question: "What is the single biggest challenge in managing unstructured data in your organisation?"

"Security and access control. Data is fragmented across multiple systems and locations, making it difficult to find, access, and effectively use relevant information."

"Raise user awareness about the use, management and security of unstructured data, particularly transfers of unstructured data outside the organisation."

EXECUTIVE IMPLICATION

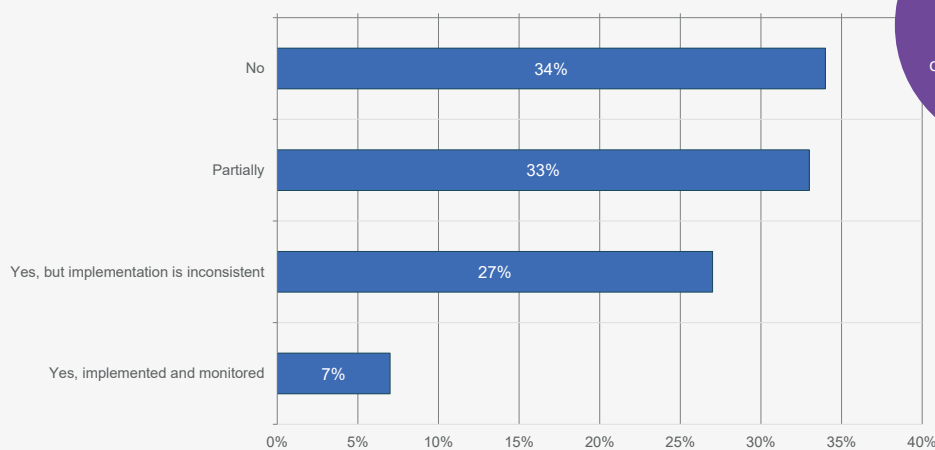
Review effective access and sensitive content together. Test retrieval with realistic user roles and correct inappropriate sharing before extending AI scope.

RETENTION AND DELETION

Lifecycle policy needs operational follow-through

Old information needs an explicit decision about continued use, retention or deletion.

Retention periods and deletion practices



Only
7 %

report retention and deletion practices that are implemented and monitored.

Information can remain accessible after its business purpose has changed or its content has been superseded. Availability alone does not establish whether it remains appropriate to use as an AI source. Retention obligations and holds must be

considered before removal. Among the respondents reporting production use of unstructured data in AI or analytics, only one reported implemented and monitored retention and deletion practices.

OPEN TEXT RESPONSES

Question: "What is the single biggest challenge in managing unstructured data in your organisation?"

"The staffing resources, time, and/or technology cost in doing the work to get everything in order. While we prioritize this work to an extent, it is competing with very limited resources against many other high-priority work items."

"Senior leadership reluctance to delete what isn't needed"

EXECUTIVE IMPLICATION

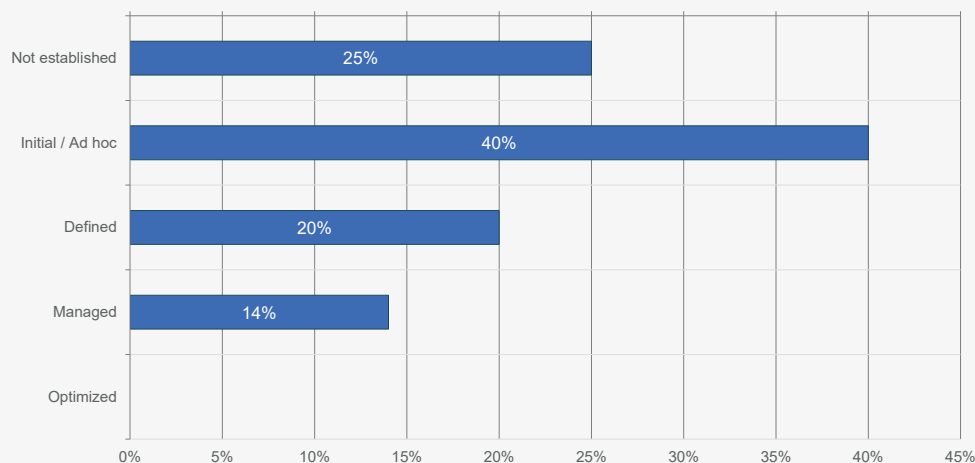
Connect lifecycle rules to repository controls. Check that decisions are executed, exceptions are handled and obsolete material is not treated as current guidance.

UNSTRUCTURED DATA GOVERNANCE

Production use of AI or analytics coexists with immature unstructured data governance

The maturity measure concerns unstructured data. Use includes AI or analytics.

Current unstructured data governance maturity



47 % report production use of unstructured data in AI or analytics

Limited or wide production use combined.

54 % of these production users report unstructured data governance that is not established or initial / ad hoc

Production-use subgroup; question-specific bases are reported in the methodology.

This does not measure AI governance maturity or demonstrate that these deployments are unsafe. It shows that production use of unstructured data in AI or analytics can coexist with relatively immature governance of the underlying information.

In a separate question, only **20 %** rate current data governance as supporting safe and reliable AI usage well or very well.

EXECUTIVE IMPLICATION

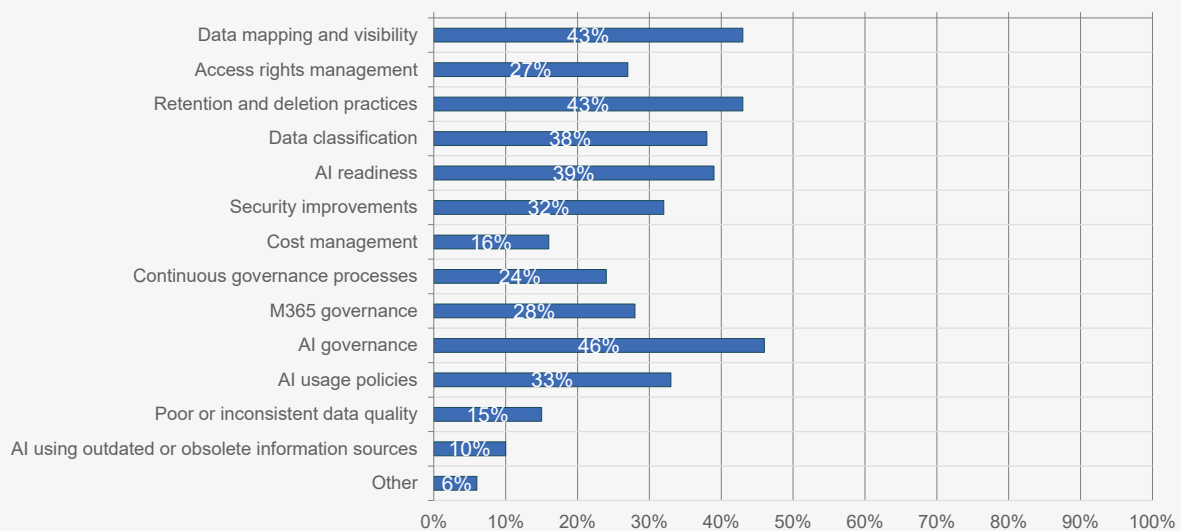
Require owners, approved sources, access checks and ongoing review for the information in scope.

DEVELOPMENT PRIORITIES

AI governance and information foundations share the agenda

The next step is to connect the work programmes and their measures of success.

Development priorities for the next 12 months



Multiple responses allowed; percentages do not sum to 100%.

AI governance is the most selected priority, closely followed by mapping and visibility and by retention and deletion. The pattern does not support a simple claim that participants are focusing only on AI-specific controls. Foundational information work remains prominent.

Continuous governance processes are selected by 24%. This suggests an executive question: will today's assessment and clean-up work become a sustained operating responsibility?

EXECUTIVE IMPLICATION

Set shared outcomes for AI and information governance: approved source coverage, timely content review, access exceptions resolved and lifecycle decisions executed.

Make information quality part of AI accountability

Open-text responses add an operational explanation. Participants describe limited time and resources, weak executive sponsorship, unclear ownership, fragmented repositories and difficulty turning rules into everyday practice.

Themes overlap and are not ranked.

"The biggest challenge is turning fragmented, text-heavy information from different projects, reports and communication channels into reliable, searchable and GDPR-compliant knowledge, without creating additional administrative burden for staff."

OPEN-TEXT RESPONSE

01. Set ownership and purpose

Name a business owner for the information used by each material AI use case. Define the decisions it supports and who can approve a change of scope.

Practical output: Accountable owner and approved information scope

02. Validate source quality and suitability

Analyse the content, sensitivity, currency, necessity and effective access of information used by the AI use case. Verify that the source is appropriate for its intended purpose, not merely available.

Practical output: Validated information scope with identified content, access and quality risks

03. Verify effective access

Review permissions, shared links and connector scope. Test retrieval for different user roles and resolve exposure before extending availability.

Practical output: Access review and recorded retrieval tests

04. Operationalise lifecycle decisions

Apply retention rules, holds and deletion decisions. Distinguish records that must be retained from content appropriate for routine AI use.

Practical output: Implemented rules and an exception process

05. Monitor information and AI together

Review content changes, retrieval quality and unresolved exceptions. Fund ongoing ownership and maintenance alongside the deployment.

Practical output: Joint measures reviewed by an accountable forum

An exploratory international expert sample

Participants were recruited through several international privacy, data management and information governance professional networks. The survey was also openly available on LinkedIn and distributed through GDPR Tech customer communications and newsletters, as well as direct invitations to professionals in the field.

The objective was to collect informed professional perspectives, not a statistically representative sample. The survey was open from 11 June to 17 September 2026. The publication dataset contains 86 completed responses; the final response was received on 2 September 2026. We aim to repeat this survey annually.

Operating region / n=86

Europe	40 %
North America	22 %
Finland	19 %
Global	8 %
Asia-Pacific	7 %
Nordic countries	5 %

Organization size / n=86

Under 50 employees	16 %
50–250	26 %
251–1,000	20 %
1,001–5,000	20 %
5,001–10,000	7 %
More than 10,000	12 %

Industry / n=85

Public sector	34 %
IT / Technology	25 %
Other	19 %
Finance / Insurance	7 %
Manufacturing / Industry	6 %
Services	6 %
Healthcare	4 %

Professional role / n=86

Data protection / Security	37 %
Compliance / Legal	19 %
Other	16 %
Business / Operations	12 %
IT / Technology	9 %
Executive management	7 %

How to read the results. Closed-question valid bases range from 79 to 86 because of missing answers. The open question received 52 responses. The production-use subgroup findings on governance maturity and retention and deletion practices each use 39 valid responses. Multiple-response bases count participants with at least one selection; their percentages do not sum to 100%. Figures are unweighted and rounded to whole percentages.

Limits. This exploratory expert sample is not statistically representative. Responses are self-reported and do not verify practices, incidents or unique organisations. No causal or time-trend conclusions are drawn. Open-text quotations add context; themes are interpretative and unranked.



Information governance and AI governance belong together

The findings point to a practical executive priority: make the condition of enterprise information visible in AI decisions. Ownership, quality, access and lifecycle management provide the evidence needed to decide which information can be used and how it should be maintained.

ABOUT GDPR TECH

GDPR Tech is a Finnish consultancy specialising in information governance, privacy and AI governance. Since 2016, we have helped Finnish, Nordic and international customers manage information risk, strengthen privacy and data governance, and prepare information environments for responsible AI use.

Our work covers unstructured data and Microsoft 365 governance, information lifecycle management, privacy consulting and outsourced DPO services, DPIA and TIA assessments, privacy engineering, and practical governance controls.

Professional training and digital learning

GDPR Tech also provides professional training, e-learning and online programmes in privacy, information governance and AI governance.

Read and share
the research:

[gdprtech.com/en/
unstructured-data-ai-
survey-2026/](https://gdprtech.com/en/unstructured-data-ai-survey-2026/)

GDPR TECH RESEARCH

GDPR Tech and Tutkimusvoima conducted the annual Data Protection Survey from 2021 to 2024: 56 responses in 2021, 102 in 2022, 51 in 2023 and 41 in 2024 - 250 responses across the series. The 2026 survey expands this research activity into an international expert study of unstructured information, information governance and AI.

The answer may be 42. The harder question is whether you can trust the source.