



WHITE PAPER · DATA GOVERNANCE

Governing the AI Data Lifecycle

Training and retrieval data, prompt logs, synthetic data and agent access

White Paper 1 · Data Governance series · Version 1.0 · October 2026

Bassirah Technology — Innovation & R&D

Governing the AI Data Lifecycle

White Paper 1 · Data Governance series · Bassirah Technology — Innovation & R&D · 28 October 2026

Document control

Field	Value
Series	Data Governance, paper 1 of 6
Author	Bassirah Technology — Innovation & R&D
Version	1.0 · 28 October 2026
Scope	A technology and regulatory survey. It describes published research, official legal texts, standards and publicly documented systems; it does not report Bassirah benchmarks or customer results, and it is not legal advice.

Contents

1. Abstract.....	4
2. Introduction	4
2.1 Why AI changes data governance.....	4
2.2 About the Data Governance series	4
2.3 Scope.....	5
2.4 Terms used in this paper	5
3. A lifecycle model for AI data	5
4. The regulatory and standards baseline	6
4.1 The EU AI Act and the Digital Omnibus on AI.....	6
4.2 The GDPR and the EDPB opinion on AI models.....	7
4.3 International standards.....	7
4.4 Voluntary frameworks	8
5. Training data: provenance, rights and documentation	8
5.1 Provenance cannot be inherited	8
5.2 Documentation as a governed artifact.....	9
5.3 Recording lineage and content provenance	9
6. Data quality for machine learning	9
7. Retrieval knowledge bases	10
8. Prompts, responses and logs	10
9. Synthetic and AI-generated data.....	11
10. Agents and access to data	12
11. What governance can and cannot fix	12
12. Making it operational	13
12.1 Lineage from data to model to answer.....	13
12.2 Quality expectations as agreements	13
12.3 Controls that adapt to risk.....	13
12.4 Assets and components.....	13
13. Comparative analysis.....	13

14. Open challenges and future opportunities.....	14
14.1 Provenance that stays accurate	14
14.2 Measuring representativeness and bias	14
14.3 Integrity of knowledge bases	14
14.4 Permission parity in retrieval and agents.....	15
14.5 Policies that follow data	15
14.6 Privacy of synthetic data	15
14.7 Telemetry standards for AI	15
14.8 Evidence that governance works.....	15
15. What this means for platform decisions	15
References	16

1. Abstract

Data governance was built for tables, reports and the people who read them. AI systems use data in more places and in more forms: corpora collected for training, labeled sets for evaluation, documents indexed for retrieval, prompts and responses captured in logs, synthetic data generated by models, and data that agents read and write through tools. Each of these is governed data, and each brings risks that classic governance programs were not designed for. This paper follows data through the AI lifecycle and asks, at each stage, what must be recorded, checked and controlled. It sets out the regulatory and standards baseline as of October 2026: the EU AI Act, including the dates changed by the Digital Omnibus on AI, the GDPR and the European Data Protection Board's opinion on AI models, ISO/IEC 42001, the ISO/IEC 5259 data quality series, ISO/IEC 8183, and the NIST AI Risk Management Framework. It then examines training data provenance and documentation, data quality for machine learning, retrieval knowledge bases, prompt and response logs, synthetic data, and agent access to data, and it is explicit about what data governance can and cannot do for hallucination and bias. It compares the maturity of current practice by stage, sets out open challenges and closes with notes for platform decisions.

2. Introduction

2.1 Why AI changes data governance

Research on data governance has grown quickly, and its focus has moved from privacy and regulatory compliance toward AI, international governance and health data [1]. The shift is practical as well as academic. A conventional governance program asks who owns a data set, what it means, who may see it and whether it is accurate. An AI program must also ask where training data came from and under what terms, whether an evaluation set is itself correct, which documents a retrieval system may show to which user, how long prompts and responses may be kept, which data was generated by a model, and what an agent may do with the data it reaches.

Governance programs already find implementation hard. A systematic review of empirical studies groups the barriers into organizational, cultural, technical, quality and regulatory dimensions, and finds that formal frameworks give guidance without removing the barriers, which reinforce each other [2]. A review of data governance frameworks finds that most concentrate on data quality, security and privacy, and give less attention to accessibility, transparency and alignment with business goals [3]. AI raises the stakes on all three neglected areas: data must be accessible to the teams that build models, transparent enough to explain where a model's behavior comes from, and tied to the purpose for which the model is built.

2.2 About the Data Governance series

This is the first of six papers. This paper covers the data that AI systems consume and produce. Paper 2 covers continuous, risk-adaptive governance and policy as code. Paper 3 covers data products and data contracts in mesh and fabric architectures. Paper 4 covers data sovereignty and cross-border compliance in multi-cloud. Paper 5 covers privacy-enhancing computation. Paper 6 covers data sharing ecosystems: intermediaries, data trusts and data spaces. Several topics introduced here, such as risk-based controls, contracts and privacy techniques, are treated in depth in those papers.

2.3 Scope

This paper covers the governance of data across the AI lifecycle: collection and acquisition, preparation and labeling, training and evaluation, retrieval knowledge bases, prompts and logs, synthetic data, agent access and retirement. It concentrates on official legal texts, international standards, peer-reviewed research and official documentation, and cites foundational work where it introduced an idea still in use. Statements about laws, regulations and standards are general information as of October 2026. They are not legal advice: obligations depend on the facts of each case, on national implementation and on guidance that continues to change, and organizations should take qualified legal advice. Statements about products reflect documentation accessed in October 2026. The paper does not quote performance, prevalence or error figures from its sources; where a source reports results, we describe them qualitatively and attribute them to the authors.

2.4 Terms used in this paper

Training data is the data used to fit a model, including fine-tuning data. **Evaluation data** is held-out data used to validate or test a model. A **knowledge base** is the corpus a retrieval-augmented generation (RAG) system searches at inference time to ground a model's answer. **Logs** or **traces** are records of prompts, responses, retrieved content, tool calls and their metadata. **Synthetic data** is data generated by a model or algorithm rather than collected from the world. An **agent** is an AI system that plans and calls tools to act on data or systems. **Provenance** is the record of where data came from and under what terms; **lineage** is the record of how it moved and was transformed.

3. A lifecycle model for AI data

ISO/IEC 8183 defines the stages of data processing across the AI system life cycle, from acquisition and creation through development, deployment and maintenance to decommissioning, and is deliberately neutral about tools and platforms [4]. A risk-adaptive view adds that controls should be mapped to each stage, rather than applied uniformly, and adjusted as risk changes [5]. Table 1 applies both ideas to the data that AI systems use. It is the frame for the rest of this paper.

Stage	Data involved	Main governance questions	Typical controls
Collection and acquisition	Web, licensed, internal and user-contributed data	Origin, rights, consent and purpose	Provenance records, license review, datasheets
Preparation and labeling	Cleaned, annotated and enriched sets	Assumptions, label quality, representativeness	Validation rules, labeling guidelines, bias examination
Training and evaluation	Training, validation and test sets	Suitability, gaps, contamination of test data	Versioned data sets, lineage to models, model documentation
Retrieval	Indexed documents and embeddings	Who may retrieve what; freshness; integrity	Permission-aware retrieval, source authentication, retrieval logs
Use and logging	Prompts, responses, retrieved chunks, tool calls	Personal data in logs, retention, reuse	Masking at source, retention and deletion policies
Synthetic and generated data	Generated records and model outputs reused as data	Privacy, quality, feedback into training	Labeling as synthetic, privacy testing, share limits

Stage	Data involved	Main governance questions	Typical controls
Agent access	Data read and written through tools	Scope, identity, attribution	Least-privilege scopes, consent, action logs
Retirement	Data sets, models, indexes and logs	Deletion, archiving, obligations that outlive the system	Deletion that follows data into derived stores

The stages are not a pipeline that runs once. Logs feed evaluation sets, model outputs return as training data, and retrieval corpora change daily. Much of the governance difficulty lies in these loops, where data leaves a governed store and enters another without its policy (Section 12).

4. The regulatory and standards baseline

4.1 The EU AI Act and the Digital Omnibus on AI

The EU Artificial Intelligence Act, Regulation (EU) 2024/1689, contains the most specific data governance obligations in current AI law [6]. Article 10 requires that the training, validation and testing data of high-risk AI systems be subject to data governance and management practices appropriate to the system's purpose. These cover design choices; data collection processes and data origin, and for personal data the original purpose of collection; preparation operations such as annotation, labeling, cleaning, updating, enrichment and aggregation; the assumptions about what the data measure and represent; the availability, quantity and suitability of the data; examination for possible biases, including feedback loops in which outputs influence future inputs, with measures to detect, prevent and mitigate them; and the identification of data gaps [6]. The data sets must be relevant, sufficiently representative and, to the best extent possible, free of errors and complete in view of the intended purpose, and must account for the geographical, contextual, behavioral or functional setting of use [6].

Article 12 requires high-risk systems to allow the automatic recording of events over their lifetime, so that risk situations and substantial modifications can be identified and post-market and deployer monitoring are supported, and Article 19 requires providers to keep such logs for a period appropriate to the purpose, of at least six months unless other law, notably data protection law, provides otherwise [6]. Article 53 applies to providers of general-purpose AI models. It requires technical documentation including the training and testing process, information for downstream providers, a policy to comply with EU copyright law including text-and-data-mining opt-outs, and a publicly available summary of the content used for training on a template provided by the AI Office [6]. The Commission has issued that template, which asks providers to describe their main data sources, including public data sets, licensed data, web-scraped data with the top domains, user data and synthetic data, and their data processing, and which applies to open-source models as well [7].

The application dates have changed. Regulation (EU) 2026/1744, the Digital Omnibus on AI, published on 24 July 2026, postponed the high-risk requirements: Chapter III Sections 1 to 3, which include Articles 10 and 12, now apply from 2 December 2027 for high-risk systems listed in Annex III and from 2 August 2028 for those covered by Annex I product legislation [8]. It moved the former Article 10(5), which allowed special categories of personal data to be processed for bias detection, into a new Article 4a, which sets strict safeguards and extends

the legal basis beyond high-risk systems without creating an obligation to detect bias [8]. The obligations for general-purpose AI models in Article 53 were not postponed and have applied since 2 August 2025, with models placed on the market before that date given until 2 August 2027 [6], [8]. As of October 2026, Article 53 therefore applies, while Articles 10 and 12 do not yet apply. Organizations building high-risk systems have time, but the data practices Article 10 describes take longer to build than the remaining period suggests, and the data used to train systems that will be on the market in 2027 is being collected now.

4.2 The GDPR and the EDPB opinion on AI models

Where AI data includes personal data, the GDPR applies in full [9]. Its principles in Article 5 map directly onto the lifecycle: lawfulness, fairness and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality; and accountability, which requires the controller to be able to demonstrate compliance and is the legal basis for much of the documentation and lineage discussed below [9]. Article 25 requires data protection by design and by default, so that only the data necessary for each purpose is processed, kept and made accessible [9]. Article 22 restricts decisions based solely on automated processing that have legal or similarly significant effects, which matters at the use stage rather than in training [9].

The European Data Protection Board's Opinion 28/2024 addresses AI models directly [10]. It finds that models trained on personal data cannot in all cases be considered anonymous: a model is anonymous only if the likelihood of extracting personal data about the people in its training data, directly or through queries, is insignificant, and controllers must document that assessment. It sets out the three-step test for relying on legitimate interest, applied separately to development and deployment. And it states that a controller deploying a model developed by another party must assess, as part of its accountability, whether the model was developed lawfully [10]. For governance, the consequence is that a model is not a clean boundary: personal data and its obligations may travel with the model, and provenance records are needed by the deployer as well as the developer. A Commission proposal of November 2025 to amend the GDPR, including a provision on AI training, was still under negotiation when this paper was written and is not treated here as law.

4.3 International standards

ISO/IEC 42001:2023 specifies an AI management system: requirements to establish, implement, maintain and continually improve the management of AI in an organization that develops, provides or uses it, following the harmonized structure of ISO management system standards [11]. Its Annex A includes a control objective on data for AI systems; secondary sources describe controls for documented data management processes, data acquisition, defined data quality requirements, recording of data provenance over the data and AI system life cycles, and documented data preparation methods [11]. The ISO/IEC 5259 series adds the data quality detail: terminology, quality measures, requirements for a data quality management system, a process framework that includes labeling, a governance framework for governing bodies, and, since 2026, a technical report on visualizing data quality results [12]. Together with ISO/IEC 8183 [4], these standards give an organization a management system, a quality model and a lifecycle, which are the three things a regulator or auditor is likely to ask about.

4.4 Voluntary frameworks

The NIST AI Risk Management Framework organizes AI risk work into four functions: govern, map, measure and manage [13]. Its data-related outcomes include documenting how data was collected and selected, including its availability, representativeness and suitability; managing risks from third-party data, including intellectual property risks; and examining and documenting privacy risk and bias [13]. It notes that maintaining the provenance of training data supports transparency and accountability [13]. The generative AI profile adds risks specific to generative systems, including data privacy, intellectual property, harmful bias and homogenization, and information integrity, and suggests actions such as documenting provenance with source, signature, version and watermark information and verifying that fine-tuning and retrieval data are of known provenance [14]. The OECD AI Principles, revised in May 2024, ask AI actors to ensure traceability of data sets, processes and decisions across the lifecycle [15]. These frameworks are voluntary, but they describe the practices that regulators and auditors increasingly expect, and they translate well into internal policy.

Table 2 summarizes the baseline. It is a simplified orientation, not a compliance checklist.

Instrument	Type	Data governance focus	Status as of October 2026
EU AI Act, Art. 10	Regulation (EU)	Training, validation and test data practices for high-risk systems	Applies from 2 December 2027 (Annex III) or 2 August 2028 (Annex I)
EU AI Act, Arts. 12 and 19	Regulation (EU)	Automatic logging and log retention for high-risk systems	Same dates as Art. 10
EU AI Act, Art. 53	Regulation (EU)	Documentation, copyright policy and training content summary for general-purpose models	Applies since 2 August 2025; earlier models by 2 August 2027
EU AI Act, Art. 4a	Regulation (EU)	Safeguarded processing of special-category data for bias detection	Introduced by Regulation (EU) 2026/1744
GDPR	Regulation (EU)	Principles, data protection by design, automated decisions	Applies
EDPB Opinion 28/2024	Supervisory opinion	Anonymity of models, legitimate interest, deployer due diligence	Adopted December 2024
ISO/IEC 42001	International standard	AI management system, including data for AI systems	Published 2023
ISO/IEC 5259 series	International standards	Data quality for analytics and machine learning	Parts 1 to 5 published 2024 to 2025; Part 6 in 2026
ISO/IEC 8183	International standard	AI data life cycle stages	Published 2023
NIST AI RMF and AI 600-1	Voluntary frameworks	Data documentation, third-party data, privacy, bias, provenance	Published 2023 and 2024

5. Training data: provenance, rights and documentation

5.1 Provenance cannot be inherited

The most direct evidence on training data provenance comes from audits. A joint legal and machine learning team traced the lineage of a large collection of widely used fine-tuning data sets back to their sources and found that license information on popular hosting

platforms was frequently missing or miscategorized, and that restricted data was concentrated in particular areas, including low-resource languages, creative tasks and synthetic data [16]. A companion audit of the web domains behind major training corpora found that consent signals, in robots.txt files and terms of service, were becoming more restrictive quickly and were often inconsistent with each other [17]. Two lessons follow. Provenance metadata copied from a hosting platform is a claim to be checked, not a fact. And provenance is not fixed at collection: the terms that applied when data was gathered may differ from those that apply when a model is retrained.

A provenance record for AI data therefore needs more than a source URL. It should hold the source and its owner, the license or contract and its version, the consent or opt-out signals observed and when, the purpose of collection, and the transformations applied since. The AI Act's training content summary [7] and copyright policy [6] for general-purpose models, and the provenance actions in the NIST profile [14], all depend on such records existing.

5.2 Documentation as a governed artifact

Datasheets for datasets proposed that every data set be accompanied by a document describing its motivation, composition, collection process, preprocessing and labeling, recommended uses, distribution and maintenance [18]. Data Cards developed the idea into structured, comparable documentation built from practice [19], and model cards apply the same discipline to trained models, reporting intended uses and evaluation across relevant conditions and groups [20]. The questions in a datasheet correspond closely to the items Article 10 asks about: origin, collection purpose, preparation steps, assumptions and gaps [6]. Documentation of this kind is most useful when it is treated as governed data in its own right: versioned with the data set, owned, reviewed and linked to the models trained on it.

5.3 Recording lineage and content provenance

Two open standards help to capture provenance automatically rather than by hand. OpenLineage records lineage as events for runs, jobs and data sets, with extensible facets that can carry schemas, versions, source code and data quality results [21]; it covers pipelines rather than model weights or prompts. The C2PA specification attaches tamper-evident provenance to media assets through signed manifests that are preserved as an asset is edited [22]. It is relevant both to labeling AI-generated content and to tracking the origin of media used for training or retrieval, but it records provenance; it does not establish that content is true [22]. The specification has been revised several times in 2025 and 2026, so implementations should track its current version.

6. Data quality for machine learning

Data-centric AI moves attention from model design to the data, and frames data work as continuing across training data development, inference data development and data maintenance rather than as a cleaning step before training [23]. Interviews with practitioners in high-stakes domains describe data cascades: compounding harms that start with data issues, surface late and are hard to see, because data work is undervalued compared with model work [24]. Even benchmark test sets contain mislabeled examples, enough in some cases to change which model appears best [25], so evaluation data needs the same quality controls as training data.

The most mature part of the AI data lifecycle is automated validation. Production systems validate data entering machine learning pipelines against expected schemas and distributions, and detect skew between training and serving data [26]. Declarative constraints, in effect unit tests for data, can be run at scale and incrementally as data sets grow [27]. The ISO/IEC 5259 series provides a common vocabulary and quality model for such checks, and a governance framework for those who oversee them [12]. Sector studies, for example in banking, describe governance practices such as standardization, cleansing and monitoring as a basis for reliable models, though the evidence in such single studies is limited [28].

Quality for AI also has dimensions that schema checks do not reach. Article 10 asks whether data is representative of the people and settings the system will meet, and whether biases and gaps have been examined [6]. Those questions need documented analysis and human review, supported by tooling but not replaced by it (Section 11).

7. Retrieval knowledge bases

Retrieval-augmented generation moves part of what a model knows from its weights into a corpus that is searched at inference time. That corpus is governed data, with risks of its own. The OWASP Top 10 for LLM Applications lists vector and embedding weaknesses, including unauthorized retrieval, leakage of context between tenants, embedding inversion and poisoning of the knowledge base, and recommends permission-aware vector stores with logical partitioning, validation and authentication of sources, classification of data, and immutable retrieval logs [29]. It lists data and model poisoning separately, at pre-training, fine-tuning and embedding stages, and recommends provenance tracking, data version control and vetting of data suppliers [29].

Research supports both concerns. Injecting a small number of crafted texts into a knowledge base can steer a model to answers chosen by an attacker, and the defenses the authors evaluated were insufficient [30]. RAG systems can be induced to reveal the contents of their private retrieval corpus, even though retrieval may reduce leakage of the model's own training data [31]. Retrieved content can also carry instructions: indirect prompt injection plants prompts in documents, web pages or email that an application is likely to retrieve, blurring the line between data and instructions [32].

The governance response is to treat ingestion into a knowledge base as a controlled publication step. Each chunk should carry its source, owner, classification, access rights and the time it was last verified; retrieval should enforce the same permissions as the source system for the user making the request; and retrieval events should be logged so that an answer can be traced to the documents that grounded it. Model-side defenses help, but they cannot substitute for controlling what enters the corpus and who can see it [30].

8. Prompts, responses and logs

LLM applications produce a new kind of data at volume. Tracing tools record each request's prompt, response, token usage, latency and cost, together with tool calls and retrieval steps and their inputs and outputs, across multi-step agent workflows [33]. These traces are needed for debugging, evaluation and, for high-risk systems, the logging that Article 12 will require [6]. They also contain whatever users typed and whatever the system retrieved, which can include

personal and confidential data. Trace stores must therefore be governed like any other store of sensitive data, under the GDPR's principles of minimization and storage limitation [9].

Three controls matter most. The first is capturing content by explicit decision. The OpenTelemetry semantic conventions for generative AI, still at Development status, make content-bearing attributes such as output messages and retrieved documents opt-in and note that some attributes may contain sensitive information [34]. The second is masking at source. Langfuse documents masking of trace inputs, outputs and metadata in its SDKs before data leaves the application, or centrally in an OpenTelemetry Collector, and documents the limits: masking applies only to span attributes, other exporters receive unmasked copies, and Collector-side masking happens after data has left the application, so the Collector must sit inside the trust boundary [35]. The third is retention and deletion that follow the data. Langfuse documents per-project retention windows after which traces and related data are deleted, while audit logs and data set items are not affected [36], and deletion of traces that does not remove copies made elsewhere, such as in evaluation data sets [37].

That last point generalizes. Evaluation workflows build data sets from production examples and score them with automated, human and model-based judges [38]. When a production trace becomes an evaluation example, it leaves the trace store's retention and deletion policy [36], [37]. A request to erase personal data, or a decision to stop retaining it, must follow the data into every derived store. Logs feed evaluation, and evaluation feeds training; without lineage between them, deletion is incomplete.

9. Synthetic and AI-generated data

Synthetic data is attractive for AI because it can fill gaps, balance classes and reduce exposure of real personal data. It is not automatically private. An evaluation of synthetic data generated by models found that it either failed to prevent privacy inference attacks or lost substantial utility, did not clearly outperform traditional anonymization, and had an unpredictable trade-off between privacy and utility [39]. The UK Information Commissioner's Office states that whether synthetic data is anonymous depends on whether source personal data can be inferred from it, that some generation methods are vulnerable to model inversion and membership inference, and that creating synthetic data still involves processing real personal data [40]. NIST's guidelines on differential privacy add that synthetic data techniques without differential privacy generally offer only informal privacy guarantees, and that differentially private synthetic data is the route to formal ones [41]. An explainer commissioned by the Royal Society reaches the same conclusion for a general audience: privacy and utility must be evaluated, not assumed [42].

Synthetic data also affects model quality. Training indiscriminately on content generated by models causes irreversible defects in which the rare tails of the original distribution disappear over successive generations, a result shown across language models and other generative models [43]. The finding concerns indiscriminate recursive use, not every use of synthetic data, but its practical lesson is clear: knowing which data was generated by a model becomes more valuable as such content spreads [43]. Lineage should mark synthetic and AI-generated data as such, record the generator and its version, and let teams limit and monitor its share in training sets.

The Digital Omnibus on AI makes synthetic data part of a legal test. Under the new Article 4a, special categories of personal data may be processed for bias detection only where that cannot be done effectively with other data, including synthetic or anonymized data [8]. Organizations relying on Article 4a will need to show that they considered synthetic data and why it was not sufficient. Paper 5 of this series examines synthetic data alongside other privacy-enhancing techniques.

10. Agents and access to data

Agents change the access question from what a person may see to what an AI system may do on someone's behalf. OWASP's list for agentic applications highlights tools used in unintended ways that lead to data exfiltration, and agents that inherit or accumulate more privilege than their task needs [44]; its LLM list names excessive agency as a risk [29]. The Model Context Protocol's security guidance, in its current draft revision, sets normative rules relevant to data access: servers must not accept tokens that were not issued for them, partly because passing tokens through breaks accountability and audit trails; proxy servers need per-client consent; and scopes should start minimal and be elevated step by step, with elevation events logged and wildcard scopes avoided [45].

Prompt injection makes these controls necessary rather than optional. Data an agent retrieves or a tool returns can carry instructions that steer the agent [32], and benchmark environments for tool-using agents show that defenses remain partial [46]. Governance cannot prevent every injection, but it can limit the damage: give each agent an identity of its own, grant scopes per task rather than per agent, require confirmation for actions that write, send or delete, and log every tool call with enough context to attribute it to a user, an agent and a request. These are the same principles of least privilege and accountability that apply to people, applied to software that acts faster and on more data.

11. What governance can and cannot fix

Data governance is sometimes presented as a remedy for hallucination and bias. The evidence supports a narrower claim. Surveys of hallucination identify data-related causes, such as divergence between sources and references in training data, alongside causes in training and decoding [47], and organize the causes of hallucination in large language models into data, training and inference stages [48]. Governed data, with known provenance, freshness and quality, can reduce hallucinations that originate in the data and makes failures traceable to their sources. It does not remove hallucination, which remains an open research problem [48].

Bias is similar. A survey of data set development and use finds that common collection and use practices produce biased and poor outcomes, reviews technical mitigations, and argues for careful qualitative and quantitative documentation of data sets during creation and use [49]. Article 10 frames bias in the same way, as something to be examined, documented and mitigated rather than removed by a tool [6]. Governance provides the evidence and the accountability; judgment about what is fair in a given use remains human.

12. Making it operational

12.1 Lineage from data to model to answer

The controls in Sections 5 to 10 depend on one capability: knowing how data, models, indexes, logs and evaluation sets relate. Catalogs can now hold much of this. DataHub records table-level and column-level lineage, pipeline lineage and impact analysis, collected automatically from supported sources, edited by hand or written through an API, and documents caveats: manual and programmatic lineage can overwrite each other, and the view shows current rather than historical lineage [50]. Its metadata model for machine learning models can record training and evaluation data sets, training runs with their inputs and outputs, versions, features, deployments, model card documentation and governance metadata such as owners, terms and domains [51]. OpenLineage provides a vendor-neutral event format for capturing lineage from pipelines [21]. Linking those records to trace stores and evaluation data sets closes the loop described in Section 3.

12.2 Quality expectations as agreements

Data contracts express quality expectations as agreements between the producers and consumers of a data set. DataHub's documentation describes a contract as a set of assertions on schema, freshness and data quality, one per physical asset, whose status depends on assertion results that can be produced by DataHub or by external tools and published back [52]. For AI data, contracts give a place to state, test and monitor the quality requirements that ISO/IEC 42001 and ISO/IEC 5259 ask organizations to define [11], [12]. Paper 3 examines data contracts and their emerging open standards in depth.

12.3 Controls that adapt to risk

Not every data set or AI use needs the same controls. A risk-adaptive framework combines static policy with sensitivity classification and contextual risk scoring, maps lifecycle stages to mitigation mechanisms in a governance-risk matrix, and uses a rules engine to tighten or relax controls as risk changes; its validation so far is by simulation [5]. The idea fits AI well, where the same data may be low risk in an internal analytics model and high risk in a system that makes decisions about people. Paper 2 develops it, with policy as code.

12.4 Assets and components

Finally, a governance program needs to know what it is trying to achieve. Separating the outcomes a framework should deliver from the components that must be managed to reach them helps to show where a program is thin [3]. For AI data, the outcomes include lawful and documented provenance, fit-for-purpose quality, controlled access, traceability and timely deletion. Each maps to components described in this paper, and gaps are easier to see when the mapping is written down.

13. Comparative analysis

Table 3 compares the maturity of current practice at each stage. The judgments are qualitative, drawn from the sources cited, and describe the state of practice, not any single product.

Stage	Most mature practice	Standards and guidance	Main open problems
Collection and acquisition	Data set documentation templates	AI Act Art. 53 and training content template; NIST AI 600-1	License and consent provenance that is accurate and current
Preparation and quality	Automated validation and declarative quality checks	ISO/IEC 5259; AI Act Art. 10	Representativeness, label quality and bias examination at scale
Training and evaluation	Lineage in catalogs; model documentation	ISO/IEC 42001; ISO/IEC 8183	Lineage across organizations; evaluation data quality
Retrieval	Partitioned indexes; source metadata	OWASP LLM Top 10	Poisoning and leakage defenses; permission parity with sources
Use and logging	Tracing with masking, retention and deletion	AI Act Arts. 12 and 19; GDPR; OpenTelemetry conventions (in development)	Policies that follow data into derived stores
Synthetic data	Differentially private generation	NIST SP 800-226; ICO guidance	Privacy guarantees outside differential privacy; share in training
Agent access	Scoped tokens and consent	MCP security guidance; OWASP agentic list	Prompt injection; attribution across chains of agents

Validation and lineage for structured data are proven and have open standards. Documentation, model metadata and tracing are well defined but depend on people and processes to keep them complete. The open problems cluster where data crosses a boundary: between organizations, between a source system and a retrieval index, between a trace store and an evaluation set, and between a human and an agent.

14. Open challenges and future opportunities

14.1 Provenance that stays accurate

Challenge. Inherited license metadata is often wrong, and consent signals change after data is collected [16], [17]. **Research direction.** Provenance tracing tools, signed provenance for content [22] and lineage events that carry rights information [21]. **Opportunity.** Provenance records that are checked and refreshed automatically would make training content summaries and copyright policies easier to evidence.

14.2 Measuring representativeness and bias

Challenge. Article 10 asks for representative data and examined biases, but representativeness depends on context and is hard to measure [6], [49]. **Research direction.** Quality measures and process frameworks for machine learning data [12] and documentation practices [18], [19]. **Opportunity.** Shared methods for evidencing representativeness would reduce the cost of preparing for the high-risk requirements.

14.3 Integrity of knowledge bases

Challenge. Small injections can corrupt a RAG system's answers, and evaluated defenses were insufficient [30]. **Research direction.** Source authentication, provenance per chunk and anomaly detection at ingestion [29]. **Opportunity.** Knowledge bases with verifiable integrity would let organizations ground AI in more of their own content.

14.4 Permission parity in retrieval and agents

Challenge. Retrieval and agents can expose data beyond what a user could see in the source system [29], [44]. **Research direction.** Permission-aware retrieval and progressive, logged scope elevation [45]. **Opportunity.** AI access that matches source permissions by construction would remove a common blocker to deployment.

14.5 Policies that follow data

Challenge. Retention and deletion stop at the boundary of the store that defines them, while traces become evaluation data and outputs become training data [36], [37], [43]. **Research direction.** Lineage across trace stores, catalogs and model registries [50], [51]. **Opportunity.** Deletion and retention that propagate along lineage would make erasure requests and storage limits enforceable for AI.

14.6 Privacy of synthetic data

Challenge. Synthetic data outside differential privacy offers only informal guarantees, and its privacy-utility trade-off is unpredictable [39], [41]. **Research direction.** Differentially private generation and standard privacy tests for each release [40], [41]. **Opportunity.** Synthetic data with stated, tested guarantees would support the Article 4a test and wider data sharing [8].

14.7 Telemetry standards for AI

Challenge. The shared vocabulary for AI telemetry is still at Development status [34]. **Research direction.** Stable conventions that distinguish content from metadata and support masking. **Opportunity.** Stable, privacy-aware telemetry would let organizations change tools without losing governance controls.

14.8 Evidence that governance works

Challenge. Frameworks give guidance without removing implementation barriers, and much of the evidence on effectiveness is limited [2], [28]. **Research direction.** Empirical studies of governance programs for AI and design-science evaluation in real settings [5], [3]. **Opportunity.** Evidence about which controls matter most would let organizations spend governance effort where it reduces the most risk.

15. What this means for platform decisions

Start from the lifecycle in Table 1 and name an owner for each stage. Keep provenance records for training and evaluation data that include rights and consent, and re-check them before each retraining. Document data sets and models with datasheets and model cards, and version that documentation with the data. Put automated quality checks and data contracts on the pipelines that produce AI data, and treat evaluation data as carefully as training data. Treat ingestion into a knowledge base as publication, with source, owner, classification and permissions on every chunk, and enforce source permissions at retrieval. Capture prompt and response content only by decision, mask it at source, and set retention and deletion that follow the data into evaluation and training sets. Label synthetic and AI-generated data in lineage and test each synthetic release for privacy. Give agents their own identities and least-privilege scopes, and log their actions. Use lineage to connect all of these, and map the controls to ISO/IEC 42001, ISO/IEC 5259 and the NIST AI RMF so that the same

evidence serves several regimes. Plan against the AI Act dates in Table 2, which may change again, and take legal advice on how they apply.

For LLM observability, Bassirah works with Langfuse as a Strategic Technology Partner; Langfuse documents tracing, evaluation, masking, retention and deletion with the limits described in Section 8. For metadata, lineage and catalog functions, Bassirah provides independent support for DataHub Core, the open-source edition of DataHub; this support is Bassirah's own and is not vendor support from DataHub's maintainers, and features documented for DataHub Cloud, such as managed data contracts, should be checked against the open-source edition before relying on them. Other products, projects and standards named in this paper are ecosystem references. In each case, validate the design against your own data, risks, legal advice and versions before relying on it.

This paper opens the Data Governance series. The next five papers examine continuous, risk-adaptive governance; data products and contracts; sovereignty and cross-border compliance; privacy-enhancing computation; and data sharing ecosystems. One idea runs through all of them: governance has to move with the data, through every copy and every use, rather than stay with the store where the data first arrived.

References

1. Hassani, H., Huang, X. and MacFeely, S. (2025). Mapping the evolution of data governance scientific research. *Data & Policy*, 7, e51. (<https://doi.org/10.1017/dap.2025.10014>)
2. Acuña, I., Cravero, A. and Zorrilla, M. E. (2026). From Compliance to Value: Data Governance Implementation Challenges to Turn Data into a Strategic Asset—A Systematic Literature Review. *Applied Sciences*, 16(11), 5289. (<https://doi.org/10.3390/app16115289>)
3. Acev, D., Biyani, S., Rieder, F., Aldenhoff, T. T., Blazevic, M., Riehle, D. M. and Wimmer, M. A. (2025). Systematic analysis of data governance frameworks and their relevance to data trusts. *Management Review Quarterly*, 76(3), 3045–3098. (<https://doi.org/10.1007/s11301-025-00545-1>)
4. ISO/IEC (2023). ISO/IEC 8183:2023 Information technology — Artificial intelligence — Data life cycle framework. Edition 1, July 2023. ISO/IEC JTC 1/SC 42. (<https://www.iso.org/standard/83002.html>)
5. Gharib, J. and Gahi, Y. (2026). DRADG: A Dynamic Risk-Adaptive Data Governance Framework for Modern Digital Ecosystems. *Information*, 17(1), 102. (<https://doi.org/10.3390/info17010102>)
6. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 2024/1689, 12 July 2024. (<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>)
7. European Commission (2025). Communication from the Commission: Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models required by Article 53(1)(d) of Regulation (EU) 2024/1689 (AI Act). C(2025) 8311 final, Brussels, 5 December 2025. (<https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models>)

8. European Parliament and Council of the European Union (2026). Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, L series, 2026/1744, 24 July 2026. (<https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>)
9. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. (<https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>)
10. European Data Protection Board (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted 17 December 2024. (https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en)
11. ISO/IEC (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. Edition 1, December 2023. ISO/IEC JTC 1/SC 42. (<https://www.iso.org/standard/42001>)
12. ISO/IEC (2024–2026). ISO/IEC 5259 series, Artificial intelligence — Data quality for analytics and machine learning (ML). ISO/IEC JTC 1/SC 42. Parts 1 (Overview, terminology, and examples), 2 (Data quality measures), 3 (Data quality management requirements and guidelines) and 4 (Data quality process framework), 2024; Part 5 (Data quality governance framework), 2025; ISO/IEC TR 5259-6 (Visualization framework for data quality), 2026. (<https://www.iso.org/standard/81088.html>)
13. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST, January 2023. (<https://doi.org/10.6028/NIST.AI.100-1>)
14. National Institute of Standards and Technology (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: NIST, July 2024. (<https://doi.org/10.6028/NIST.AI.600-1>)
15. OECD (2019, amended 2024). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. Adopted 22 May 2019; amended 8 November 2023 (AI system definition) and 3 May 2024 (Ministerial Council revision). (<https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>)
16. Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K. et al. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8), 975–987. (<https://doi.org/10.1038/s42256-024-00878-8>)
17. Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C. et al. (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024, Datasets and Benchmarks Track)*, 108042–108087. (<https://doi.org/10.52202/079017-3431>)
18. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. (<https://doi.org/10.1145/3458723>)

19. Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)*, 1776–1826. (<https://doi.org/10.1145/3531146.3533231>)
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, Atlanta, GA, ACM, pp. 220–229. (<https://doi.org/10.1145/3287560.3287596>)
21. OpenLineage Project. (2026). Object Model. OpenLineage documentation (site version 1.54.0). Official documentation, accessed October 2026. (<https://openlineage.io/docs/spec/object-model>)
22. Coalition for Content Provenance and Authenticity (C2PA). (2026). Content Credentials: C2PA Technical Specification, Version 2.5. Official documentation, accessed October 2026. (https://c2pa.org/specifications/specifications/2.5/specs/C2PA_Specification.html)
23. Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric Artificial Intelligence: A Survey. *ACM Computing Surveys*, 57(5), 1–42. (<https://doi.org/10.1145/3711118>)
24. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–15. (<https://doi.org/10.1145/3411764.3445518>)
25. Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS 2021)*. (<https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html>)
26. Polyzotis, N., Zinkevich, M., Roy, S., Breck, E., & Whang, S. (2019). Data Validation for Machine Learning. *Proceedings of Machine Learning and Systems 1 (MLSys 2019)*. (https://proceedings.mlsys.org/paper_files/paper/2019/hash/928f1160e52192e3e0017fb63ab65391-Abstract.html)
27. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12), 1781–1794. (<https://doi.org/10.14778/3229863.3229867>)
28. Singamsetty, S. (2024). Efficacy of Data Governance a Cutting Edge Approach to Ensuring Data Quality in Machine Learning for Banking Industry. In *2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, IEEE, pp. 1–7. (<https://doi.org/10.1109/scopes64467.2024.10991944>)
29. OWASP Foundation, GenAI Security Project. (2025). OWASP Top 10 for LLM Applications 2025. Official documentation, accessed October 2026. (<https://genai.owasp.org/llm-top-10/>)
30. Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. *34th USENIX Security Symposium (USENIX Security 25)*, 3827–3844. (<https://www.usenix.org/conference/userixsecurity25/presentation/zou-poisonedrag>)

31. Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y. et al. (2024). The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). Findings of the Association for Computational Linguistics: ACL 2024, 4505–4524. (<https://doi.org/10.18653/v1/2024.findings-acl.267>)
32. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISeC '23), 79–90. (<https://doi.org/10.1145/3605764.3623985>)
33. Langfuse. (2026). Observability Overview (LLM Observability & Application Tracing). Langfuse documentation. Official documentation, accessed October 2026. (<https://langfuse.com/docs/observability/overview>)
34. OpenTelemetry. (2026). Semantic conventions for generative AI systems (GenAI spans, events, metrics, agent spans, MCP). OpenTelemetry GenAI Semantic Conventions repository. Official documentation, accessed October 2026. (<https://github.com/open-telemetry/semantic-conventions-genai>)
35. Langfuse. (2026). Masking sensitive LLM data. Langfuse documentation. Official documentation, accessed October 2026. (<https://langfuse.com/docs/observability/features/masking>)
36. Langfuse. (2026). Data Retention. Langfuse documentation. Official documentation, accessed October 2026. (<https://langfuse.com/docs/administration/data-retention>)
37. Langfuse. (2026). Data Deletion. Langfuse documentation. Official documentation, accessed October 2026. (<https://langfuse.com/docs/administration/data-deletion>)
38. Langfuse. (2026). Evaluation Overview. Langfuse documentation. Official documentation, accessed October 2026. (<https://langfuse.com/docs/evaluation/overview>)
39. Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic Data – Anonymisation Groundhog Day. 31st USENIX Security Symposium (USENIX Security 22), 1451–1468. (<https://www.usenix.org/conference/usenixsecurity22/presentation/stadler>)
40. Information Commissioner's Office (ICO). (2023). Privacy-enhancing technologies (PETs): Synthetic data. UK GDPR guidance and resources. Official documentation, accessed October 2026. (<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/what-pets-are-there/synthetic-data/>)
41. Near, J. P., Darais, D., Lefkovitz, N., & Howarth, G. (2025). Guidelines for Evaluating Differential Privacy Guarantees (NIST Special Publication 800-226). National Institute of Standards and Technology. (<https://doi.org/10.6028/NIST.SP.800-226>)
42. Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic Data -- what, why and how? arXiv preprint 2205.03257. (<https://arxiv.org/abs/2205.03257>)
43. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755–759. (<https://doi.org/10.1038/s41586-024-07566-y>)
44. OWASP Foundation, GenAI Security Project. (2025). OWASP Top 10 for Agentic Applications for 2026. Official documentation, accessed October 2026. (<https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>)
45. Model Context Protocol. (2026). Security Best Practices. Model Context Protocol specification (draft). Official documentation, accessed October 2026. (https://modelcontextprotocol.io/specification/draft/basic/security_best_practices)

46. Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024, Datasets and Benchmarks Track)*, 82895–82920. (<https://doi.org/10.52202/079017-2636>)
47. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. (<https://doi.org/10.1145/3571730>)
48. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. et al. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55. (<https://doi.org/10.1145/3703155>)
49. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 100336. (<https://doi.org/10.1016/j.patter.2021.100336>)
50. DataHub. (2026). Lineage. DataHub documentation. Official documentation, accessed October 2026. (<https://docs.datahub.com/docs/features/feature-guides/lineage>)
51. DataHub. (2026). ML Model entity (metadata model). DataHub documentation. Official documentation, accessed October 2026. (<https://docs.datahub.com/docs/generated/metamodel/entities/mlmodel>)
52. DataHub. (2026). Data Contracts. DataHub documentation. Official documentation, accessed October 2026. (<https://docs.datahub.com/docs/managed-datahub/observe/data-contract>)

Product and project names are trademarks of their respective owners. Their mention describes publicly documented technology and does not imply a business relationship, except where Bassirah names a Strategic Technology Partner.