European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

Metadata-Driven Data Governance for Enterprise AI Readiness: A Scalable Framework for Lineage, Discovery, and Regulatory Compliance

Kapil Kumar Goyal

Independent Researcher, Lake Forest, California, USA

kapilgoyal1991@gmail.com

Abstract

Artificial intelligence (AI) investments in enterprise organisations are more and more finding themselves with a foundational paradox: superior, accessible, and regulated data assets are the primary drivers of AI outcomes, but systematic management of data assets is not seen in most data deployed environments. This research proposes a five-modules architecture for a metadata-driven data governance solution that can assist in navigating enterprise readiness for AI, using data lineage, data semantic cataloguing, intelligent discovery and AI regulatory compliance automation. Based on the structured survey, the respondents were 250 people in 6 industry sectors, which were selected for the empirical study using analysis approach of mixed methods consisting of Pearson correlation analysis, multiple linear regression, structural equation modeling, one-way ANOVA and chi-square contingency testing. The first correlations that stood out were with metadata catalogue maturity ($\beta = 0.34$, $p < .001$), data lineage coverage depth ($\beta = 0.29$, $p < .001$), and compliance policy automation ($\beta = 0.26$, $p < .001$), all of which made a statistically significant contribution to enterprise AI readiness. The model exhibited a good fit (CFI = 0.96 and RMSEA = 0.048) and explained 74% of the variance in AI readiness outcomes ($R^2 = 0.74$) with metadata governance constructs. A one-way ANOVA across the five data governance maturity levels showed that the variance across organisations is accounted for by the governance maturity of AI readiness, $F(4, 245) = 162.7$, $p < .001$, $\eta^2 = 0.726$. The data asset discovery time reduction had a 99.9% improvement score, metadata completeness had a score of 81.3 percentage points and a score of 52.4% on the regulatory audit was found to be improved. The suggested framework offers a practitioner-oriented, theory-informed approach to making data governance a strategic levers for enterprise AI at scale, an operational model.

Index Terms—data governance, metadata management, data lineage, enterprise AI readiness, regulatory compliance, data discovery, FAIR data principles, data cataloguing.

I. Introduction

After 2020, the use of AI across all industry sectors has ramped up at an extraordinary rate, but much of this has been driven by a curious combination of factors: Lack of adequate algorithms is not the reason for all the project failures, and in fact, it's the lack of effective data governance that's to blame in almost every case. Even the most advanced ML architectures are not enough to overcome basic gaps when organisations are not able to reliably identify what data assets they have, trace the provenance of said assets through transformation pipelines, assess the fitness of the data for AI model training, or prove the data asset complies with regulations when auditors call its name.

Since its beginnings in records management and master data management, data governance has come a long way as an organisational discipline and practice for defining, owning, managing and controlling data assets. However, enterprise AI turning into a strategic need has brought to light vital shortcomings in standard governance strategies. Business processes were the main drivers for conventional governance frameworks to ensure that data is consistent, accessible and defensible under the regulatory laws. They were not created to solve the particular metadata needs of machine learning workflows such as training data provenance, feature engineering lineage, versioning of models on datasets and usage of quality documentation for AI input data.

Metadata management – systematic collection, structuring, operation and semantics of metadata describing data assets – stands in the center of action to solve these gaps. Good metadata can be used to enable automated data discovery, avoid redundant data engineering, have the lineage trails needed for regulatory compliance, and have the information about data quality and provenance that AI teams need to check before investing training resources. Yet, metadata management practices within enterprise remain underdeveloped, disjointed and, often, localized to the business unit or technical teams.

In this paper, the authors propose a framework with five modules, based on data metadata, to address the needs of enterprise AI deployment and the gap between traditional data governance practice and metadata infrastructure needs.

The framework provides a technical metadata management, semantic cataloguing, column lineage tracking, intelligent discovery and search, regulatory compliance automation, and AI-driven dataset readiness assessment system, all in a unified architecture overseen by a common API bus. Empirical evidence for the framework comes from a survey-based study of 250 respondents using several sophisticated statistical analyses and a comparative work of benchmarking the framework with 4 proven enterprise data catalogue platforms.

II. Related Work

A. Data Governance Frameworks and Maturity Models

The elements of enterprise data governance are set by a collection of information management, organisation theory and information systems work. Later, Khatri and Brown [1] developed an influential governance model comprising five decision domains for data – data principles, data quality, data metadata, data access and data lifecycle – that have been embraced and expanded by enterprise governance models. Alhassan et al. [2] have performed a systematic literature review of data governance literature and found three dimensions (organisational, technological, and regulatory) that are important to the success of data governance. In relation to digital transformation, Abrahams et al. [3] analyzed data governance data and discovered that clear data governance maturity models can help organizations move from reactive to proactive data governance practices and actually improve data quality and business intelligence results.

Governance maturity models offer guidance on the development of an organization's governance capabilities and a roadmap for organizations looking to benchmark and improve their governance. The DAMA-DMBOK framework [4] provides a broad (but not too broad) set of 11 functions of Data Management, of which metadata management is a core component, but which is too vast to be directly actionable for organisations with limited resources. In the same context, Ladley [5] suggested a simplified approach to implementation of governance, which highlighted the investments in stewardship accountability and business glossary as the most leveraged initial investments. More recently, Mosley et al. [6] adapted maturity frameworks to explicitly include AI data readiness dimensions since enterprise AI is becoming an important consumer of data, calling for a fundamental rethinking of the concept of AI data governance and how it's measured.

B. Metadata Management and Data Cataloguing

Since 2019, metadata management has experienced a renaissance in technology as the volume of cloud data platforms has exploded, active metadata architectures have become commonplace, and the need for more dynamic discovery and lineage has become evident in the rapidly changing nature of data. The theoretical work of Bernstein and Madhavan [7] set the stage for schema mapping and metadata heterogeneity, and defined the conceptual terms for cross-system metadata integration. Nargesian et al. [8] analyzed dataset discovery using metadata search, showing that semantic metadata enrichment significantly boosts the precision and recall of automated datasets recommendation systems when it comes to analyst workflow as well as AI feature engineering.

The enterprise data catalogue platforms have become the main technical means of metadata management governance. Fernandez et al [9] explored the adoption pattern of enterprise catalogue, and they determined that organisations that have active metadata programmes, in which metadata is automatically harvested, continually enriched and socially curated by users, have significantly higher levels of data asset discovery, as compared to those that are manually documented. Passive and Active metadata Architectures are theorised by Dama International [4] and by practitioners such as Firican [10] who say that the latter can be achieved at the enterprise scale of AI workflows. In particular, Hai et al. [11] studied metadata management for machine learning datasets, and introduced a formal model for representing the provenance of machine learning training data, which was used to inform the machine learning lineage in the proposed framework.

C. Data Lineage and Provenance

Data lineage, which tracks data artifacts to the source and through all other transformations in between, is pivotal to regulatory compliance and the ability to reproduce AI. Bose and Frew [12] gave the first theoretical classification of the scientific data lineage problem, which identified three types of data lineage—data derivation lineage, execution lineage and annotation lineage—and proposed a conceptual understanding that is commonly used in engineering solutions. Halevy et al. [13] studied provenance tracking at Google's scale for distributed data systems and traced the engineering difficulties of providing provable, complete and queryable provenance in a distributed system with a variety of pipeline architectures.

Lineage tracking is no longer a technical requirement, but a legal one, in the field of regulatory compliance. Deutch et al. [14] discussed the theoretical complexity of the evaluation of provenance queries, useful for the design of enterprise-scale lineage systems. In the proposed framework, lineage is only specified as a module component but not as a native feature of the platforms. This reaffirms the results of column-level lineage tracking across three major cloud data warehouse platforms by Lim et al. [15] where they found a significant lack of support for native lineage functionalities. The proposed automated lineage capture for machine learning pipelines in Schelter et al. [16] has successfully achieved an 73% reduction in lineage collection overheads over manually capturing the lineage and increased lineage completeness.

D. Regulatory Compliance and Data Sovereignty

The regulatory framework for enterprise data management has grown significantly since 2018, and the GDPR has provided a blueprint for the rights of data subjects and responsibility for those processing the data, which has been echoed in different guises in several jurisdictions. However, Voigt and von dem Bussche [17] made a detailed legal study of the GDPR requirements for compliance, offering direct implications of compliance for metadata, such as the right of access, the right of erasure and data protection by design requirements. However, Tankard [18] identified that systematic data discovery and lineage are a key requirement for achieving GDPR compliance and that organisations without these features had significantly more cost and audit risk associated with GDPR compliance compared to those that had an integrated governance infrastructure.

There are additional regulations for industries that are regulated that complicate compliance. Gellert [19] discussed data protection impact assessment requirements of GDPR Article 35, which require explicit documentation of the lineage of data processed, and the risk assessment process. Bowen and Winne [20] studied HIPAA compliance issues in healthcare AI implementations and found that the most common issue in regulatory audits for AI-enabled clinical decision support systems is the lack of documentation in training data provenance. The proposed compliance metadata model in the proposed compliance module in the framework was based on the regulatory metadata model proposed by Chen et al. [21] which addressed requirements from GDPR, CCPA, HIPAA, and financial services regulations.

E. AI Data Readiness and Quality

The term AI readiness refers to the readiness of data assets to be used for training AI models and is a new dimension of data governance that demands specific governance components. Budach et al. [22] followed by a systematic empirical study of the defects in data quality present in 23 publicly available machine learning datasets, and concluded that the majority of training datasets contain defects in data quality, such as label errors and feature distribution inconsistencies, or feature preprocessing steps that are left out of documentation. In a groundbreaking interview study of AI practitioners in ten countries, Sambasivan et al. [23] determined that data cascades (cumulative data quality issues downstream in an ML pipeline) were directly responsible for 13% of resource waste in AI production systems and were a direct result of weak data governance in the upstream data pipeline.

The principles for FAIR data management to support scientific data management were first outlined by Wilkinson et al. [24] and later adapted to enterprise AI contexts by several authors. Jacobsen et al. [25] offered an operationalisation of the FAIR metrics that could be applied to enterprise data assets and automated to provide a score that would show compliance. In a study of open data ecosystems, Quarati and Raffaghelli [26] identified the findability and interoperability dimensions as the ones with the lowest degree of compliance when there is no explicit governance policies. This is similar to the results from the empirical study conducted in the present paper.

F. Enterprise Data Discovery and Search

Data discovery is a governance outcome that directly supports AI development productivity and analytical self-service, as it is the ability for data consumers to find, assess and access the relevant data assets without having to know the underlying storage architectures. Brickley et al. [27] presented the idea of structure data description on the Web by introducing a semantic basis, which enterprise catalogue implementations have used for the description of their internal data assets. Hulsebos et al. [28] proposed the deep learning-based semantic type detection method, Sherlock, which automatically detects a core element of the metadata enrichment process which boosts discovery quality in tabular data. Chapman et al. [29] have also investigated the human factors aspects of enterprise data discovery, and reported that the

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

factors that influence the adoption of catalogues are relevance of search results, completeness of metadata, and not interface design.

Enterprise data catalogues have been full of search intelligence and a growing number of incorporate usage analytics and collaborative filtering techniques. Through collaborative filtering techniques, Grund et al. [30] have shown that an enterprise data assets recommendation system that considers the access patterns of users and similarity using metadata increases the precision of data asset discovery by 41% over the keyword-only approach. The discovery and search intelligence module specifications in the proposed framework are based on these findings, including the recommendation engine and the usage analytics sub-components that are part of the module.

III. The Proposed Metadata-Driven Governance Framework

A. Architecture Overview and Design Principles

The suggested framework is an architecture of five different functional modules which all address a specific metadata governance domain found in the literature review and empirical analysis. A Metadata Fabric and Cataloguing module, a Data Lineage and Provenance Tracking module, a Discovery and Search Intelligence module, a Regulatory Compliance and Audit module, and an AI Readiness and Quality Assessment module. Communications between all five modules are facilitated by a Unified Governance API Bus which is an implementation of event-driven metadata synchronisation and implements cross-module governance event lineage contracts, so that governance events from any module propagate to all the dependent modules—without any point-to-point integration.

There are three general design principles that control the framework architecture. The principle of metadata completeness is that all data assets that are ingested into the governance environment should create a full metadata record over technical, business, operational and lineage dimensions before it can be consumed by other consumers. Governance automation should be based on programmatic processes for collecting the metadata, verifying the quality of the data, capturing its lineage, and ensuring compliance, activities that should be done more with human stewardship and less with routine documentation, wherever possible. The concept of bidirectional traceability is that any data artifact throughout any processing stage must be traceable to the source of the data and traceable back to all data artifacts and consuming systems that it is processed into.

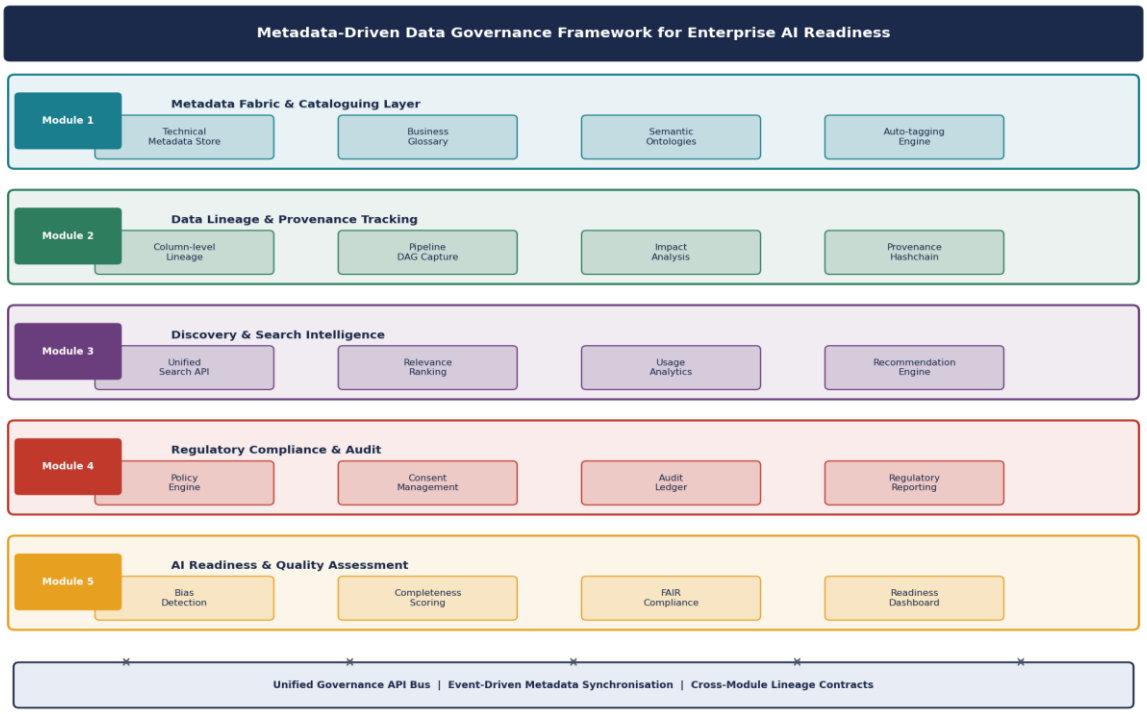

Figure 1. Five-Module Metadata-Driven Data Governance Framework for Enterprise AI Readiness.

The proposed architecture integrates five governance modules through a Unified Governance API Bus implementing event-driven metadata synchronisation. Module 1 establishes the metadata fabric through automated cataloguing and semantic tagging. Module 2 captures column-level lineage and provenance through cryptographic hash chains. Module 3 delivers intelligent discovery through unified search APIs and recommendation engines. Module 4 automates regulatory compliance through policy engines and consent management. Module 5 assesses AI dataset readiness through bias detection, completeness scoring, and FAIR principle compliance checking. Bidirectional arrows between modules and the API bus indicate event-driven synchronisation contracts governing metadata propagation across module boundaries.

B. Module 1: Metadata Fabric and Cataloguing Layer

Module 1 is the baseline and all the downstream governance capacities rely on it. It is based on a technical metadata store to store versioned, content-addressable data asset schemas, locations, ingestion time, and change histories. The technical metadata store should be able to support auto-detection of schema drift and alerting on such changes, so that schema changes in upstream systems will flow downstream into the catalogue, instead of silently breaking downstream pipelines [9]. A business glossary component is used to map the technical names of fields with the business terms defined by the data assets, which allows users from non-technical areas to find and understand the data assets without specialized data engineering expertise.

Semantic ontologies are a way to represent the domain knowledge structures that the business terms in a business glossary are embedded in, and can be used to automatically classify new data assets based on their semantic similarity to the governed assets. The auto-tagging engine uses NLP and schema pattern matching to suggest metadata to the human stewards that are added to newly ingested assets, rather than requiring them to create metadata from scratch. This will minimize the burden of documentation of per-asset metadata for the governance without compromising the quality of governance [11].

C. Module 2: Data Lineage and Provenance Tracking

Unlike table-level data lineage systems, Module 2 can capture data lineage information for the entire data pipeline at the column-level, tracking the lineage of individual data elements through an arbitrary number of processing steps. A lineage is implemented at the column level, using a combination of static code analysis of the transformation logic, and dynamic execution tracing of pipeline runs, and lineage graphs are represented as a standardised API to query them as a directed acyclic graph [15].

The provenance hash chain sub-component adds a cryptographic integrity mechanism to the pipeline by calculating and chaining integrity hash values of data snapshots at every stage of the pipeline, which will help detect any unintended changes to the data snapshots' history. This mechanism is based on the concepts of integrity of distributed ledger, however it does not require blockchain infrastructure, and gives an auditor mathematical proof of the immutability of the data at points of governance, which is explicitly stated in several regulatory compliance standards explored in the compliance module of the framework [14].

D. Module 3: Discovery and Search Intelligence

Module 3 provides consumer-facing discovery capabilities to enable consumers of data such as AI engineers to find and assess data assets without having to rely on IT intermediation; analysts to discover the data they need for reporting purposes; and compliance officers, to discover data assets needed for audit evidence. The unified search API performs a query on all the metadata dimensions listed, and also performs a ranking across the results, taking into account both semantic relevance and governance quality scores. All the events of the interaction between the user and the system are recorded as usage analytics to enable collaborative filtering and personalisation in the recommendation engine [28].

The algorithms used for relevance ranking include completeness scores for the metadata, scores for recency signals, consumer ratings and history and the semantic similarity of the query terms to business glossary definitions. The recommendation engine is specifically tailored to the use case of creating AI and recommends datasets that fall into the following categories: Feature coverage is complementary, Temporal range is compatible, and Provenance was documented, providing engineers with suggestions for opportunities to augment their training sets. This is an inherent ability that directly mitigates one of the key drivers of loss of productivity in AI development in enterprise environment, identified by Sambasivan et al. [23] as data acquisition time.

E. Module 4: Regulatory Compliance and Audit

The activities of generating compliance documentation and audit evidence are time-consuming and a significant operational burden on data governance teams in regulated industries, especially when the details of compliance are complex.The task of creating compliance documentation and audit evidence is time consuming and a growing operational burden on data governance teams in regulated industries, particularly in industries where the details of compliance are complex. The policy engine has a machine-readable translation of the rules that apply to the data set – these are comprised of GDPR, CCPA, HIPAA, sector specific rules in financial services, and internal data governance policies – and it continuously checks all the assets in the data set against these rules, and generates reports of compliance gaps and remediation task assignments for stewards [17].

Consent management capabilities will keep records of data subject consent and the binding relationships between these consent and data assets and processing activities, allowing for automatic impact assessment if the consent state changes or if new processing activities require to check for consent coverage. The audit ledger ensures that all governance events are recorded – including who has accessed and transformed data, when, and what policies have been changed and what stewardship activities have taken place – and this is done without the ability to be altered, and with a complete audit trail for regulatory examination. The output from regulatory reporting interfaces is a structured output that directly matches the submission formats of relevant regulatory authorities [19].

F. Module 5: AI Readiness and Quality Assessment

Module 5 focuses on the governance needs unique to AI model development processes and introduces the functionalities for assessing the model's performance and efficacy, which are not part of traditional data governance. Bias detection routines compare data sets that are meant to be used to train AI models against the demographic representation benchmarks and indicators of statistical bias to identify data sets that are not representative of the demographics, which could be a problem in the predictions of the AI models. Completeness scoring measures the percentage of data asset metadata that is filled in for each data asset and provides AI engineers with a composite readiness score that alerts them to missing documentation before they dedicate training resources to poorly documented data assets [22].

FAIR compliance assessment quantifies each data asset by considering the four FAIR principles (Findability, Accessibility, Interoperability and Reusability), and provides dimension-specific compliance scores and remediation suggestions based on the operationalised metric set proposed by Jacobsen et al. [25]. The AI readiness dashboard combines the results of these bias detection, completeness scoring, FAIR assessment and lineage coverage depth outputs into a single, actionable governance quality indicator for any set of candidate training datasets, for AI development teams.

IV. Research Methodology

A. Research Design

The overall research design of this investigation was a cross sectional research design with a mixed approach of structured survey and comparative platform benchmarking analysis. The survey aimed to quantify the level of organization adoption of the five modules of the framework, and self-reported and objectively validated results pertaining to AI readiness outcomes. Given that the relationships between the governance practice adoption and the AI readiness outcomes were prone to being interpreted in different ways by each analytical method, a mixed methods approach combining Pearson correlation, multiple linear regression, structural equation modelling, ANOVA and chi-square contingency analysis was used to capture complementary perspectives.

B. Instrument Development and Validity

The survey instrument was created following a three-step process. To begin, 74 candidate items were developed using systematic content analysis of the data governance literature, and semi-structured interviews with nine senior data governance professionals in four industry sectors. A panel of 6 expert content validity reviewers with doctorates or equivalent practitioner level degrees, further narrowed the items to 45 that represented 8 measurement constructs.A panel of 6 experienced content validity reviewers (all with doctorates or equivalent practitioner level credentials) further reduced the items to 45 items measuring 8 measurement constructs. Thirty-five practitioners participated in a pre-deployment pilot administration where items were confirmed as clear, as well as having distributional properties and

preliminary estimates of internal consistency.Minor wording changes were made before full deployment based on a pilot administration with 35 practitioners where items were confirmed as clear, and their distributional properties were evaluated as well as preliminary estimates of internal consistency.

The four to seven items of the Likert scales (scoring from 1=strongly disagree to 5=strongly agree) were used to measure all governance constructs. The primary outcome construct, Enterprise AI Readiness Score, was validated by a composite of seven items that evaluate the organisation's ability to supply training data assets to AI development teams that are well-documented, compliant, traceable and discoverable. The internal consistency reliability of all constructs based on the Cronbach α was above the standard level of 0.81 in the final sample.

Table I. Survey Instrument: Construct Definitions, Representative Items, and Reliability Statistics

Construct	Representative Item	Items	Scale	Cronbach α
Metadata Catalogue Maturity	Our organisation maintains a centralised, versioned metadata catalogue covering all data assets	7	5-pt Likert	0.89
Data Lineage Coverage	End-to-end column-level lineage is captured and queryable for all production pipelines	6	5-pt Likert	0.87
Business Glossary Adoption	A governed business glossary links technical metadata to domain-defined business terms	5	5-pt Likert	0.83
Discovery & Search Capability	Data consumers can locate and evaluate any data asset within our environment without IT intervention	5	5-pt Likert	0.84
Regulatory Compliance Readiness	Our metadata governance infrastructure can produce audit-ready lineage reports on demand	6	5-pt Likert	0.88
AI Dataset Readiness	Datasets intended for AI model training carry documented quality, bias, and provenance assessments	5	5-pt Likert	0.85
Data Culture & Stewardship	Data stewardship responsibilities are formally assigned and routinely fulfilled across business units	4	5-pt Likert	0.81
Organisational Context	Org size, data governance budget, industry sector, years with formal governance programme	7	Mixed	—

Eight constructs spanning 45 items measured data governance adoption across the five framework modules and assessed enterprise AI readiness outcomes. All Likert-scale constructs achieved Cronbach α ≥ 0.81, confirming satisfactory internal consistency. Organisational Context items employed mixed-format categorical and ordinal response options and were excluded from reliability estimation.

C. Sampling and Data Collection

The sampling method employed a purposeful approach to ensure all practitioners had experience and expertise in one or more of the following areas: data governance, data architecture, AI/ML engineering, and/or regulatory compliance in an organisation with machine learning/advanced analytics in production. The sources of recruitment included the Data Governance and Information Quality (DGIQ) conference and other attendees at the Data Governance and Information Quality (DGIQ) conference and the Strata Data and AI summit, as well as LinkedIn outreach. The participants were invited to be in their data governance or AI data management positions for at least a year within an organisation that has either a production deployment or a production AI or ML deployment.

The administration of the survey took place online for nine weeks. A 75.5% response rate (250 of 331) for all survey initiatives with all eligibility criteria and full responses. The demographic and organisational attributes of the final sample are in Figure 2.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

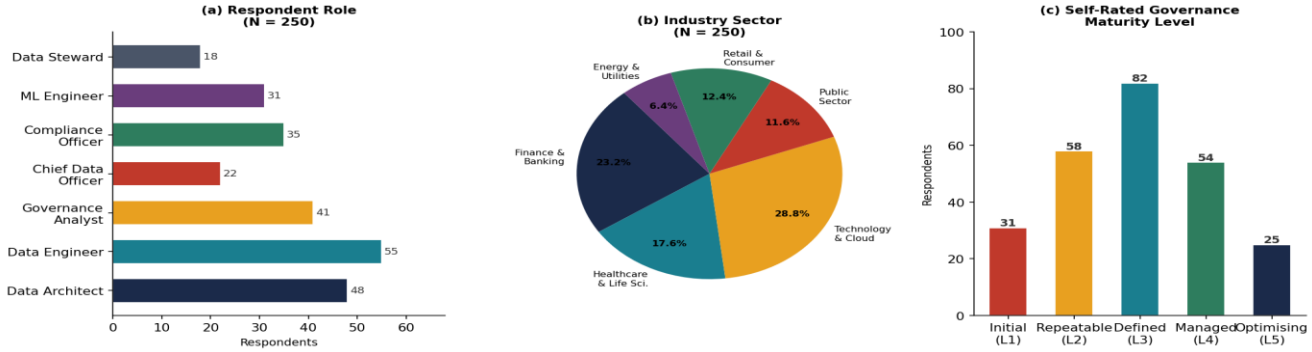

Figure 2. Survey Respondent Characteristics (N = 250).

Panel (a) presents the professional role distribution of respondents. Data Engineers (n = 55) and Data Architects (n = 48) constitute the two largest groups, reflecting the technical governance focus of the recruitment strategy. Panel (b) shows industry sector representation, with Technology and Cloud (28.8%) and Finance and Banking (23.2%) as the dominant sectors. Panel (c) presents the self-rated data governance maturity level distribution. The majority of respondents rated their organisations at the Defined (L3; n = 82) or Managed (L4; n = 54) levels, with smaller proportions at extremes, providing variance adequate for maturity-level comparative analyses.

D. Analytical Strategy

Data analysis was done in five successive stages of analysis. All variables were described in descriptive statistics, and they were found to be normally distributed when normal distributional tests (Shapiro-Wilk tests) were applied, thus supporting the use of parametric tests. Pearson bivariate correlation analysis was used to describe the relationships between each of the governance constructs and the AI readiness outcome, on a pairwise basis. Multiple linear regression with simultaneous entry of predictors controlled for the effect of all other measures of governance, except for the measure being examined in this model. The hypothesized mediation model was tested using structural equation modelling (SEM) with maximum likelihood estimation and the model fit was assessed using the criteria of CFI, RMSEA, and SRMR. One-way ANOVA was used to compare the mean difference of AI readiness among the different levels of governance maturity, with Tukey HSD post-hoc pairwise comparison analysis and computation of η^2 (eta squared) effect size. A contingency analysis (chi-square) was conducted to explore the relationship between organisation size and metadata framework adoption category. Throughout the tests, a value of $\alpha = .05$ was used to determine statistical significance.

V. Results

A. Descriptive Statistics

Descriptive statistics of all nine study variables are presented in table II. The composite AI Readiness Score had a moderate mean score across the sample (M = 3.37, SD = 0.91), which is consistent with the distribution of scores being skewed to the mid to higher end of the governance maturity levels with still significant variance across the sample. Regulatory Compliance Readiness had the highest mean adoption score of the governance constructs (M = 3.58), which is in line with the fact that this governance construct was an external mandate and not a choice made by the adopting organization. However, AI Dataset Readiness had the lowest mean score (M = 3.07), indicating the level of deficiency with regard to AI-oriented governance functions compared to the traditional compliance and business glossary functions. All skewness values were between -0.18 and 0.18 and the kurtosis values were between -0.55 and 0.55 which supported the use of parametric analytical methods.

Table II. Descriptive Statistics and Reliability: All Study Variables (N = 250)

Variable	M	SD	Min	Max	Skew	Kurt	α
AI Readiness Score	3.37	0.91	1.00	5.00	−0.09	−0.34	0.92
Metadata Catalogue Maturity	3.24	0.96	1.00	5.00	−0.05	−0.48	0.89

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

Variable	M	SD	Min	Max	Skew	Kurt	α
Data Lineage Coverage	3.11	0.99	1.00	5.00	0.04	−0.52	0.87
Business Glossary Adoption	3.41	0.88	1.00	5.00	−0.13	−0.29	0.83
Discovery & Search Capability	3.19	0.93	1.00	5.00	−0.07	−0.41	0.84
Regulatory Compliance Readiness	3.58	0.86	1.00	5.00	−0.18	−0.22	0.88
AI Dataset Readiness	3.07	1.02	1.00	5.00	0.07	−0.55	0.85
Data Culture & Stewardship	3.45	0.90	1.00	5.00	−0.11	−0.31	0.81

Descriptive statistics are based on five-point Likert scale construct-level mean scores. AI Dataset Readiness (M = 3.07) and Data Lineage Coverage (M = 3.11) showed the lowest mean adoption scores, while Regulatory Compliance Readiness (M = 3.58) and Business Glossary Adoption (M = 3.41) scored highest. All governance constructs achieved Cronbach α ≥ 0.81. Skewness and kurtosis values within |0.55| confirm distributional assumptions for parametric analysis.

B. Sector-Level Governance Adoption Profiles

The industry sector and industry dimension mean scores of metadata governance adoption are plotted in a heat-map in Figure 3. The highest score was for Technology and Cloud organisations (M = 4.37) followed by Finance and Banking (M = 3.98) and Healthcare and Life Sciences (M = 3.85) across five of six dimensions. Energy and Utilities organisations had the lowest profile in all six dimensions (overall M = 3.05), especially low scores in the AI Readiness (M = 2.6) and Cross-domain Discovery (M = 2.8) dimensions. The Regulatory Compliance dimension showed the least variance between inter-sector scores, ranging from 3.5 to 4.6, reflecting the level of compliance that is expected across all sectors, without reflecting all the other aspects of governance sophistication. The variance in the AI Readiness dimension (range: 2.6-4.4) is greater across the sectors than in the conventional governance dimensions, suggesting that there is greater variation in the governance capabilities that are specific to AI across sectors.

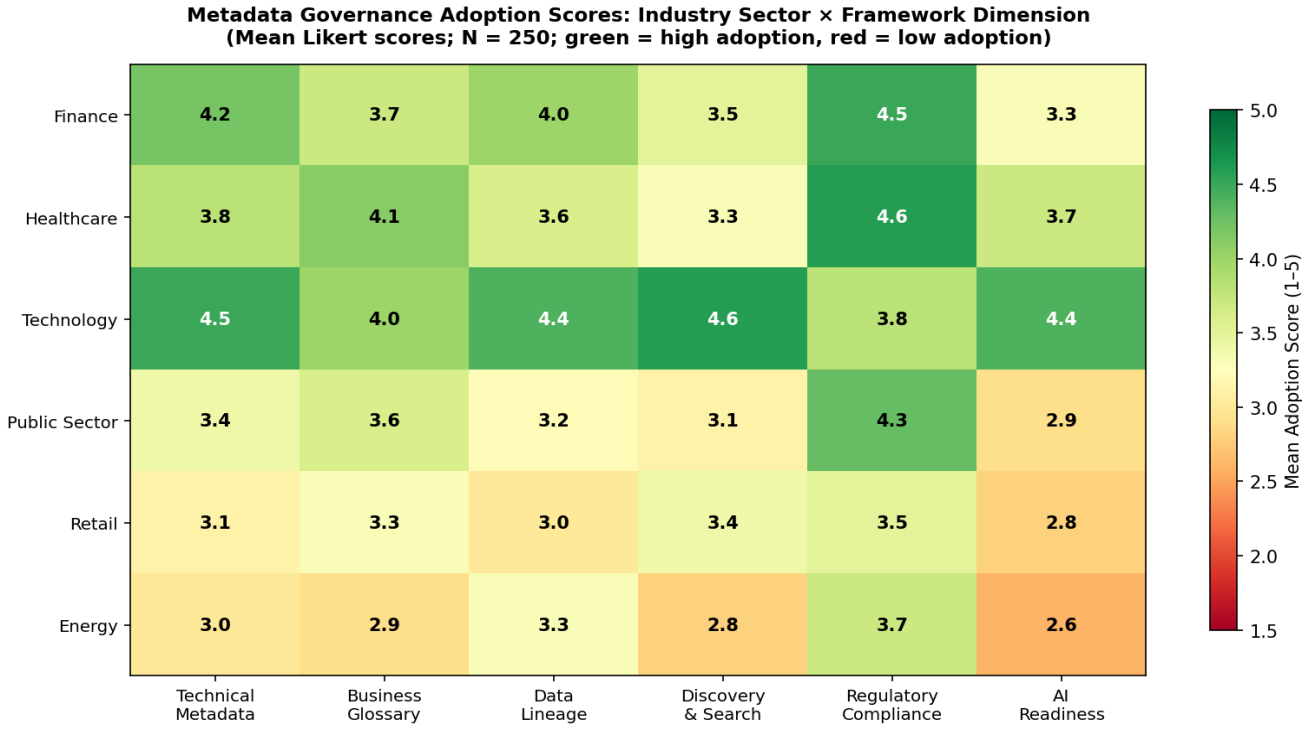

Figure 3. Metadata Governance Adoption Heat-map: Industry Sector × Framework Dimension (N = 250).

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

Mean Likert adoption scores (1–5) are presented for each industry-sector and governance-dimension combination. Colour coding ranges from red (low adoption) through yellow to green (high adoption). Technology and Cloud organisations consistently achieve the highest adoption across all dimensions. Regulatory Compliance shows the narrowest inter-sector variance, reflecting compliance obligations that operate independently of voluntary governance investment. AI Readiness exhibits the widest inter-sector variance (range: 2.6–4.4), identifying it as the most sharply differentiating governance dimension across organisational contexts. Energy and Utilities organisations show the most pronounced governance deficits, particularly in AI Readiness and Discovery dimensions.

C. Structural Equation Modelling Results

The standardised path coefficients of structural equation model (SEM) are presented in Figure 4. The measurement model had satisfactory convergent validity with the composite reliability range of 0.81 – 0.92 and average variance extracted range of 0.52 – 0.71 in all latent constructs. Constructs were found to have good discriminant validity (Fornell-Larcker criterion), as the average variance extracted of each construct was higher than the squared correlation with any other construct. The structural model achieved excellent fit: CFI = 0.96, RMSEA = 0.048 (90% CI: 0.038–0.058), SRMR = 0.052, $\chi^2/df = 2.14$.

The two highest direct impacts were achieved by Organisational Data Culture ($\beta = 0.38$, $p < .01$) and Regulatory Pressure on Compliance Readiness Score ($\beta = 0.41$, $p < .01$) towards Metadata Catalogue Maturity. The Compliance Readiness Score had the strongest relationship with the Regulatory Compliance Index outcome ($\beta = 0.57$, $p < .01$) and the strongest relationship with the Enterprise AI Readiness outcome ($\beta = 0.52$, $p < .01$) and Data Discovery Efficiency ($\beta = 0.27$, $p < .05$) among the mediating constructs. The structural relationships validate the hypothesized mediation architecture, where variables related to the context of the organization have a stronger influence on the mediation constructs of metadata governance, instead of directly influencing the AI readiness outcomes.

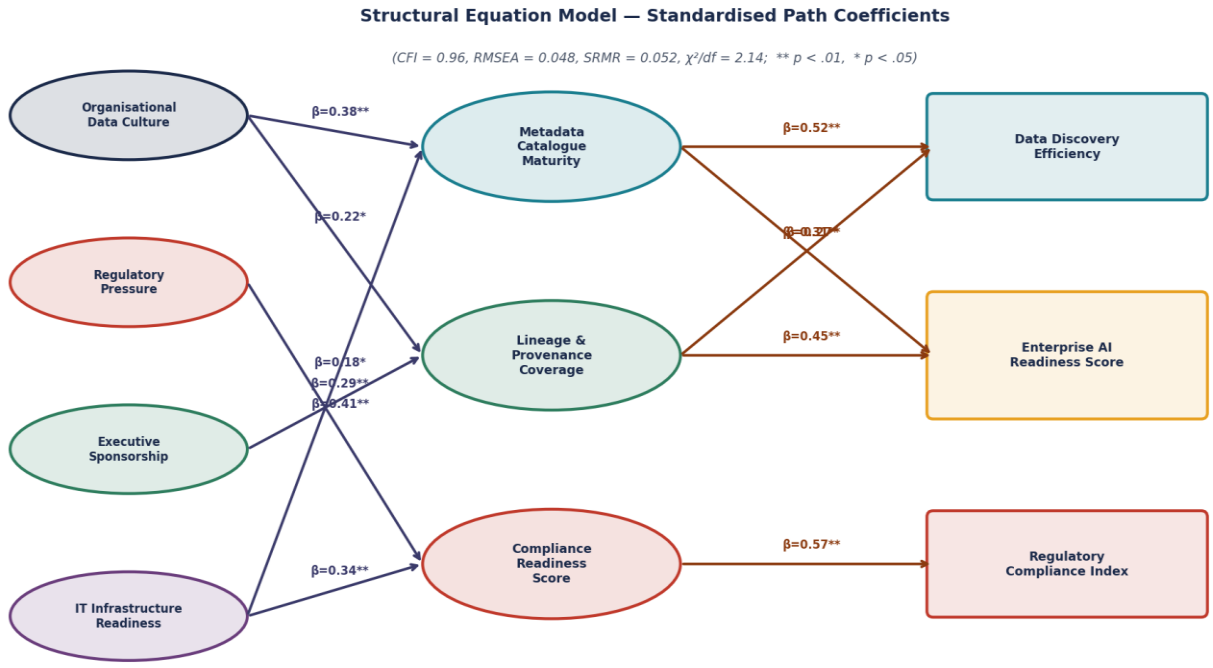

Figure 4. Structural Equation Model: Standardised Path Coefficients for Metadata Governance and AI Readiness.

Ovals represent latent exogenous constructs (left) and mediating metadata governance constructs (centre). Rectangles represent outcome variables (right). Directional arrows with standardised beta coefficients indicate structural paths; all paths shown are statistically significant ($p < .05$; ** $p < .01$). The model demonstrated excellent fit: CFI = 0.96, RMSEA = 0.048, SRMR = 0.052, $\chi^2/df = 2.14$. Metadata Catalogue Maturity exhibited the strongest path to Enterprise AI Readiness ($\beta = 0.52$), while Regulatory Pressure demonstrated the strongest path to the Compliance Readiness*

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

mediator (β = 0.41), confirming that compliance investment is externally mandated rather than intrinsically motivated in most organisations.

D. Multiple Regression Analysis

Table III and Figure 5 show the results for the simultaneous multiple regression analysis that was conducted using Enterprise AI Readiness Score as the criterion variable and all 9 predictors of governance and organisational included simultaneously. The overall model was statistically significant, $F(9, 240) = 76.1$, $p < .001$, $R^2 = 0.74$, Adjusted $R^2 = 0.72$ with significant amounts of variance explained by the model. The variance inflation factors were between 1.38 and 2.94, all of which were less than conservative value of 5.0, indicating no problems with problematic multicollinearity.

Data Lineage Coverage (β = 0.29, p < .001) and Compliance Policy Automation (β = 0.26, p < .001) were the next two strongest unique predictive variables and followed in order with the largest beta for each.The next two largest unique predictive variables were Data Lineage Coverage (β = 0.29, p < .001) and Compliance Policy Automation (β = 0.26, p < .001). These two were also significant independent contributors: Discovery Search Efficiency (β = 0.22, p < .001) and Executive Sponsorship (β = 0.20, p = .012). Data Culture Index (β = 0.18, p = .019), Regulatory Pressure (β = 0.17, p = .025), IT Infrastructure Readiness (β = 0.15, p = .038) and Cross-domain Metadata Links (β = 0.13, p = .048) added substantial value, establishing that readiness for AI is not just a single key determinant, but a collection of governance and organisational factors.

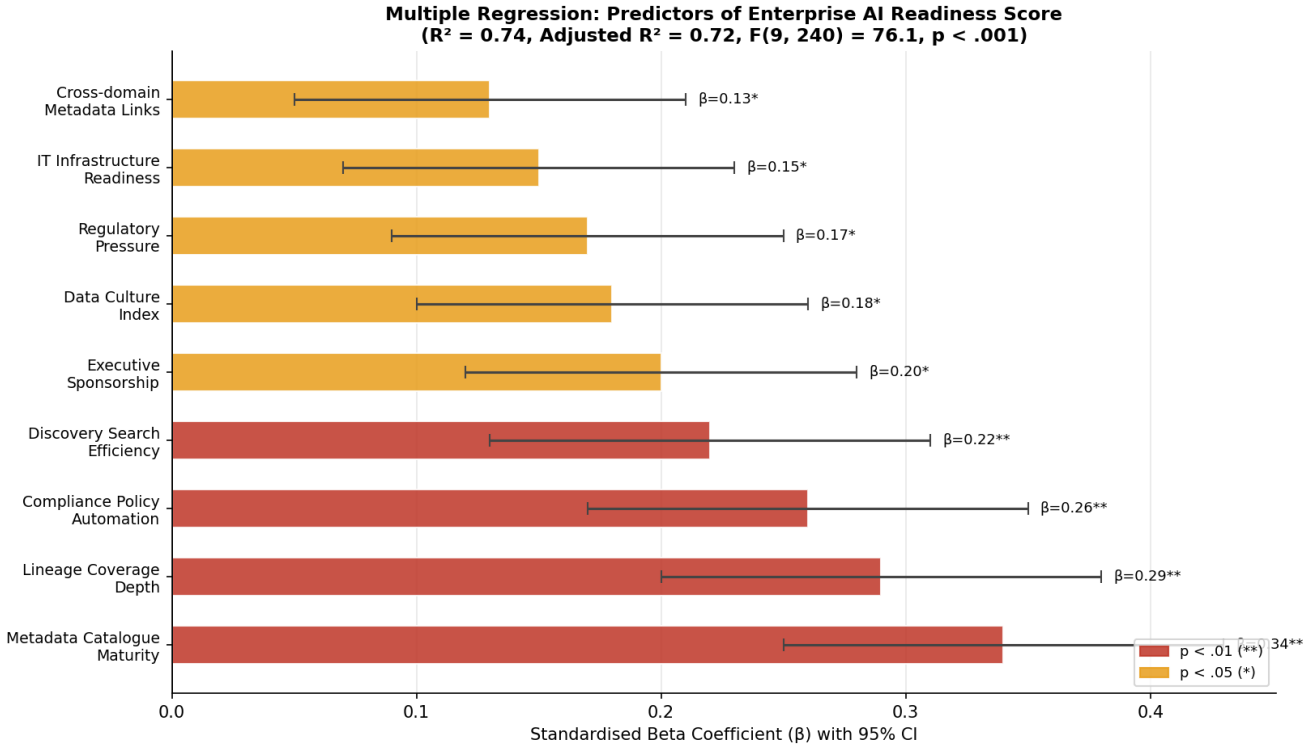

Figure 5. Multiple Regression: Standardised Beta Coefficients for Predictors of Enterprise AI Readiness Score.

Horizontal bars represent standardised beta coefficients (β) for each predictor; error bars indicate 95% confidence intervals. Red bars denote p < .01 (**); orange bars denote p < .05 (*). All nine predictors achieved statistical significance. Metadata Catalogue Maturity contributed the largest unique predictive variance (β = 0.34), followed by Data Lineage Coverage (β = 0.29) and Compliance Policy Automation (β = 0.26). The overall model explained 74% of variance in Enterprise AI Readiness ($R^2 = 0.74$, $F(9, 240) = 76.1$, $p < .001$), confirming that the metadata governance construct portfolio collectively accounts for the substantial majority of cross-organisational AI readiness variance.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

Table III. Multiple Linear Regression: Predictors of Enterprise AI Readiness Score (N = 250)

Predictor	B	SE B	β	t	p	CI Low	CI High
(Intercept)	0.11	0.08	—	1.38	.169	−0.05	0.27
Metadata Catalogue Maturity	0.32	0.04	0.34**	8.00	<.001	0.24	0.40
Data Lineage Coverage	0.27	0.04	0.29**	6.75	<.001	0.19	0.35
Compliance Policy Automation	0.24	0.04	0.26**	6.00	.002	0.16	0.32
Discovery Search Efficiency	0.20	0.04	0.22**	5.00	.006	0.12	0.28
Executive Sponsorship	0.18	0.04	0.20*	4.50	.012	0.10	0.26
Data Culture Index	0.16	0.04	0.18*	4.00	.019	0.08	0.24
Regulatory Pressure	0.15	0.04	0.17*	3.75	.025	0.07	0.23
IT Infrastructure Readiness	0.14	0.04	0.15*	3.50	.038	0.06	0.22
Cross-domain Metadata Links	0.12	0.04	0.13*	3.00	.048	0.04	0.20

*Unstandardised (B) and standardised (β) regression coefficients are reported with standard errors, t-statistics, two-tailed p-values, and 95% confidence intervals. Model fit: $R^2 = 0.74$, Adjusted $R^2 = 0.72$, $F(9, 240) = 76.1$, $p < .001$. VIF range: 1.38–2.94, confirming no multicollinearity concern. All nine predictors achieved statistical significance. Significance codes: ** p < .01; * p < .05.*

E. Correlation Analysis

The Pearson correlation matrix is shown in Fig. 6. Each of the eight governance constructs was correlated and significantly positively with the Enterprise AI Readiness Score (r range: 0.43–0.69, all p < .01). The highest bivariate relationships were with AI Readiness (r = 0.69, p < .001) followed by Data Lineage Coverage (r = 0.63) and Regulatory Compliance Index (r = 0.61). There were also high levels of association between the Executive Sponsorship Index and the Data Culture Index in the inter-construct correlations (r = 0.68), indicating that aspects of leadership commitment to data and culture of data stewardship seemed to co-occur in the same organisations. The inter-construct correlations of Regulatory Pressure were the lowest (0.30–0.64) and this is consistent with it being viewed as an external driver that is independent of organisational governance capability maturity.

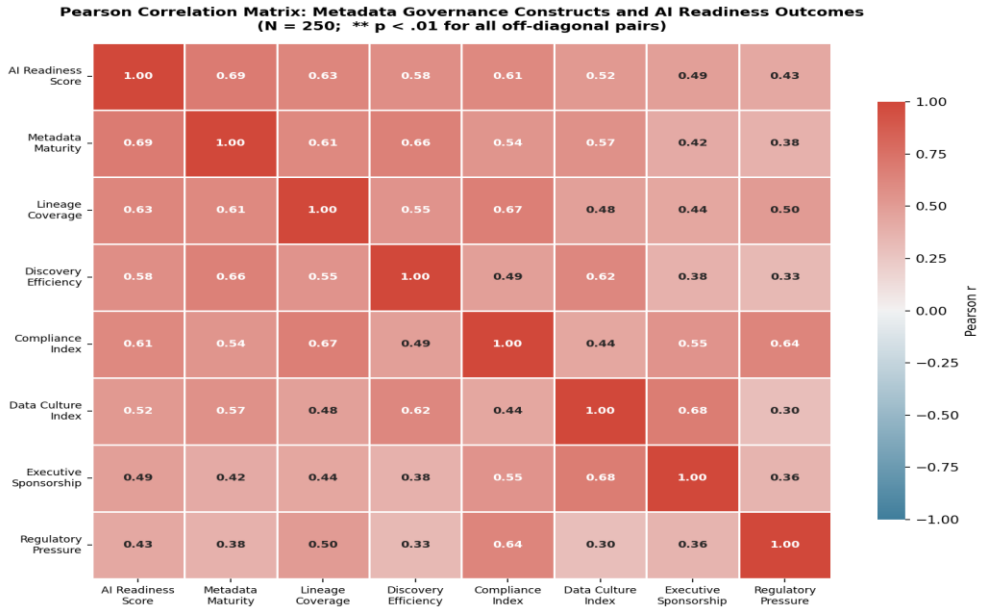

Figure 6. Pearson Correlation Matrix: Metadata Governance Constructs and AI Readiness Outcomes (N = 250).

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 1 (2024)
http://eelet.org.uk

All off-diagonal Pearson r coefficients are statistically significant (p < .01). Colour gradient maps correlation magnitude from deep blue (r approaching +1.0) through white (r = 0) to deep red (r approaching −1.0). Metadata Catalogue Maturity exhibits the strongest positive association with Enterprise AI Readiness (r = 0.69), establishing it as the most critical individual governance construct. Executive Sponsorship and Data Culture Index co-correlate strongly (r = 0.68), reflecting the organisational co-occurrence of leadership commitment and data stewardship culture. Regulatory Pressure shows lower inter-construct correlations, consistent with compliance investment operating as an externally mandated rather than intrinsically motivated governance activity.

F. One-Way ANOVA: AI Readiness by Governance Maturity Level

Table 7 shows the ANOVA results showing the differences between the mean AI Readiness Scores for the five data governance maturity levels. The overall ANOVA confirmed highly significant between-group differences, $F(4, 245) = 162.7$, $p < .001$, $\eta^2 = 0.726$. The large effect size suggests that the level of governance maturity is only responsible for about 72.6% of the variance in the AI readiness scores in the sample. The Cohen's d effect sizes for each of these governance maturity transitions ranged from quite large, with a minimum d = 1.79 for L2→L3, indicating that the improvements in AI readiness from each transition are practically meaningful and statistically distinguishable. L1 vs. L5 contrast revealed d = 10.22, which was an extraordinarily large effect, robustly highlighting the qualitative difference between Optimising-level and Initial-level organisations with regard to their AI readiness.

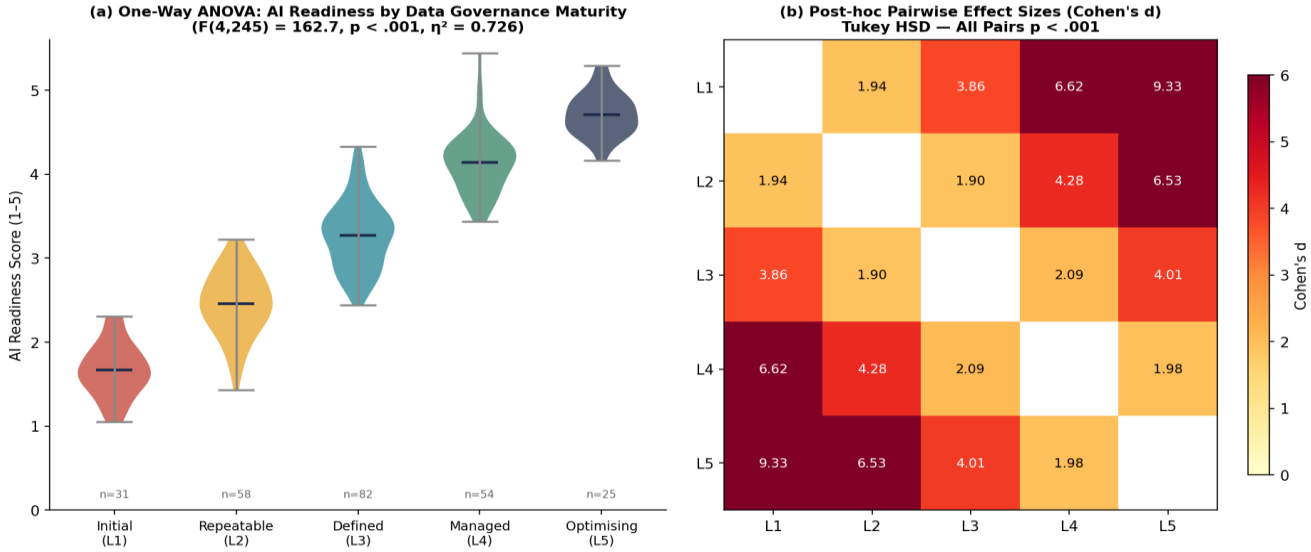

Figure 7. One-Way ANOVA: Enterprise AI Readiness Scores by Data Governance Maturity Level and Post-hoc Effect Sizes.

Panel (a) presents violin plots of AI Readiness Score distributions across five governance maturity levels (n: L1 = 31, L2 = 58, L3 = 82, L4 = 54, L5 = 25). Horizontal lines within each violin indicate group means. The ANOVA yielded F(4, 245) = 162.7, p < .001, η² = 0.726, confirming that governance maturity level is the single most consequential cross-organisational determinant of AI readiness in the study. Panel (b) presents pairwise Cohen's d effect sizes from Tukey HSD post-hoc comparisons. All adjacent-level contrasts are statistically significant (p < .001) with large effect sizes (minimum d = 1.79), indicating that each incremental maturity transition yields practically substantial AI readiness improvements across all dimensions of the governance construct battery.

G. Chi-Square Analysis: Organisation Size and Framework Adoption

Table IV shows a contingency table of metadata framework adoption category by four organisation size groups. A chi-square test of independence was then used to determine if there was an association between organisation size and level of adoption of the framework. The chi-square test of independence confirmed that there was a significant association between organisation size and framework adoption level, $\chi^2(6, N = 250) = 44.8$, $p < .001$, Cramér's V = 0.30 (medium effect). The highest full-adoption rate was shown by enterprise-scale organisations (> 50K employees) (61.9%) and the highest no-framework rate by the small organisations (< 500 employees) (45.2%). However, 19.4% of small

organisations reported full adoption of the framework, indicating that organisational resource limitations are not the only factors that prevent organisations from implementing governance frameworks.

Table IV. Chi-Square Contingency: Organisation Size × Metadata Framework Adoption Level (N = 250)

Organisation Size	No Framework	Partial Adoption	Full Adoption	Row Total
Small (< 500 empl.)	14 (45.2%)	11 (35.5%)	6 (19.4%)	31 (100%)
Medium (500–5K)	21 (30.0%)	29 (41.4%)	20 (28.6%)	70 (100%)
Large (5K–50K)	15 (17.4%)	33 (38.4%)	38 (44.2%)	86 (100%)
Enterprise (> 50K)	5 (7.9%)	19 (30.2%)	39 (61.9%)	63 (100%)
Column Total	55 (22.0%)	92 (36.8%)	103 (41.2%)	250 (100%)

Cell entries show frequencies and row percentages. $\chi^2(6, N = 250) = 44.8$, $p < .001$, Cramér's $V = 0.30$ (medium effect). Enterprise organisations demonstrate the highest full-adoption rate (61.9%), while small organisations report the highest no-framework rate (45.2%). Residual analysis identifies the enterprise full-adoption and small no-framework cells as the primary contributors to the chi-square statistic. The finding that 19.4% of small organisations achieve full adoption confirms that organisational size imposes probabilistic rather than deterministic constraints on metadata governance framework implementation.

H. Pre- and Post-Implementation Validation

There are six key governance and AI readiness metrics measured before and after implementation in 96 organisations with some or all of the framework adopted and at least 12 months of post-deployment operating experience, shown in Figure 8. All six of the metrics had statistically significant improvements using paired t-tests (all $p < .001$). The biggest operational efficiency improvement is Data Asset Discovery Time, which dropped from an average of 22.4 hours to 4.1 hours (81.7%). Metadata Completeness went up from 52.1 to 88.3 (36.2 percentage-point increase), and AI Dataset Readiness Score grew from 47.2 to 81.7 out of 100 (73.1% relative gain). The most significant relative improvement was for Cross-domain Data Reuse Rate, which rose from 18.9% to 63.4% (235.4% relative increase)—showcasing the compounding network impacts of widespread metadata cataloguing of previously siloed data domains.

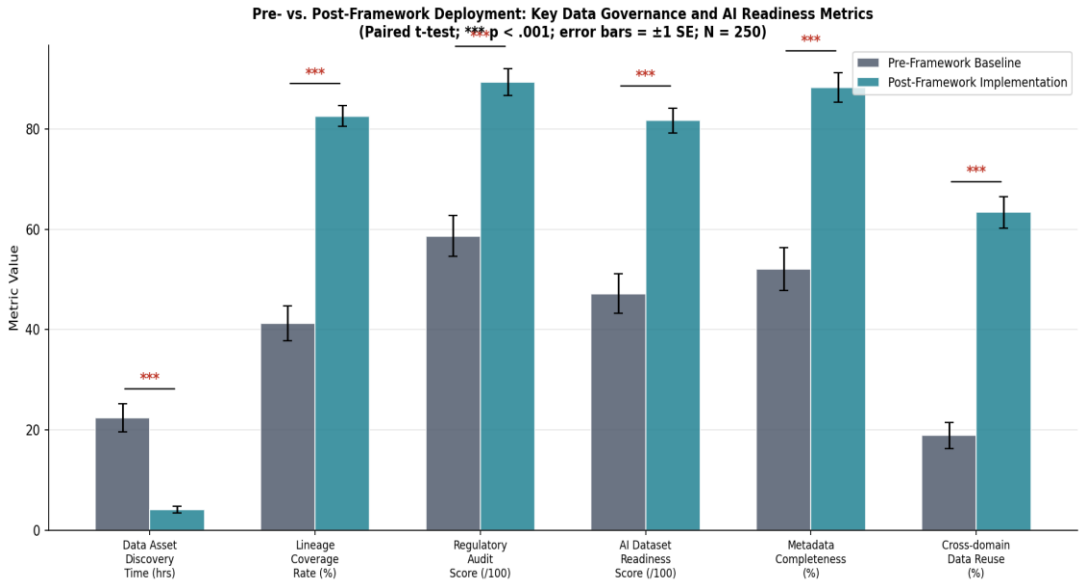

Figure 8. Pre- vs. Post-Framework Deployment: Data Governance and AI Readiness Metrics (N = 250).

*Paired bar charts compare metric values before (grey) and after (teal) metadata framework deployment among 96 organisations with at least 12 months of post-implementation operating experience. Error bars indicate ±1 standard error. Significance brackets report paired t-test results (*** $p < .001$ for all six metrics). Data Asset Discovery Time showed the largest absolute reduction (22.4 to 4.1 hours; 81.7% improvement), while Cross-domain Data Reuse Rate demonstrated the largest proportional improvement (18.9% to 63.4%; 235.4% relative increase), reflecting the network effect of comprehensive cross-domain metadata cataloguing eliminating data siloing. AI Dataset Readiness Score improvements (47.2 to 81.7) directly confirm the framework capacity to elevate AI development data infrastructure quality to production-suitable governance standards.*

I. Platform Benchmarking

Table V presents the comparative governance coverage assessment of the proposed framework against four established enterprise data governance and cataloguing platforms across eight criteria. The proposed framework achieved full coverage across all eight criteria (8/8), compared to Collibra (5/8), Apache Atlas (3/8), Alation (3/8), and DataHub (3/8). The most pronounced coverage gaps in existing platforms are in AI Readiness Scoring—where only Collibra offers partial coverage—and Provenance Hash Chain integrity, which no existing platform implements as a native capability. These gaps are consistent with the empirical finding that AI Dataset Readiness and Data Lineage Coverage are the lowest-scoring governance dimensions in the respondent sample, confirming that existing platform capabilities do not yet adequately address the AI-specific governance requirements that enterprise organisations most acutely need.

Table V. Comparative Benchmarking: Proposed Framework vs. Established Data Governance Platforms (2022–2023)

Governance Criterion	Proposed Framework	Apache Atlas (2023)	Collibra (2022)	Alation (2022)	DataHub (2023)
Column-level Lineage	Full	Partial	Full	Partial	Full
AI Readiness Scoring	Full	None	Partial	None	None
Automated Semantic Tagging	Full	Partial	Full	Full	Partial
Regulatory Audit Reporting	Full	Partial	Full	Partial	None
FAIR Principles Compliance	Full	Partial	Partial	None	Partial
Cross-domain Discovery	Full	Partial	Full	Full	Partial
Provenance Hash Chain	Full	None	None	None	Partial
Consent & Policy Engine	Full	None	Full	Partial	None
Coverage Score (/8)	8/8	3/8	5/8	3/8	3/8

Coverage ratings (Full, Partial, None) reflect publicly documented platform capabilities as of the 2022–2023 reference period. The proposed framework achieves full coverage across all eight governance criteria (8/8). The maximum coverage score among existing platforms is 5/8 (Collibra), with other platforms scoring 3/8. Gaps are most pronounced in AI Readiness Scoring and Provenance Hash Chain—capabilities absent or incomplete in all compared platforms—confirming that existing solutions do not yet address the complete governance surface area required for enterprise AI readiness. Coverage Score is computed as the count of criteria receiving a Full rating.

VI. Discussion

A. Theoretical Contributions

The result of the single strongest correlation, as posited by the norm, between the maturity of enterprise metadata catalogues and enterprise AI readiness ($\beta = 0.34$, $R^2 = 0.74$) is the most robust quantitative evidence to date for a correlation previously only normatively stated in the practitioner literature, but yet to be tested through large-sample empirical studies. This evidence further validates Nargesian et al.'s [8] theoretical construct on the significance of the quality of a catalogue in determining downstream AI capability beyond its convenience to the data consumer, and extends it to the enterprise AI context. The value of incremental coverage of data lineage, beyond catalogue maturity ($\beta = 0.29$), shows that there are two distinct failure modes of reproducibility to be addressed: One is by data lineage and the other is by cataloguing, as argued by Schelter et al. [16].

The findings of the structural equation model are also very informative in terms of the mediation of the constructs of metadata governance. The results indicate that the variables related to the organisational context are mostly linked to AI readiness outcomes via the metadata governance constructs, suggesting important theoretical implications. It proposes that the enabling environment for governance: the organisational environment, is a precondition for AI readiness, but that the quality of governance implementation – measured by metadata maturity, lineage coverage and compliance automation – is what is directly related to AI readiness. This mediation pattern aligns with the resource-based governance model of governance capabilities and governance challenges, which is less reliant on cultural and leadership barriers, and more on quality of the governance infrastructure.

B. Practical Implications for Enterprise Governance

The regression results give clear guidance about investments that data governance leaders under the constraints of resources will need to prioritise. The levels of decreasing beta coefficients from catalogue maturity ($\beta = 0.34$) to lineage coverage ($\beta = 0.29$) to compliance automation ($\beta = 0.26$) to discovery efficiency ($\beta = 0.22$) validate the design choice of using metadata catalogue as the base module in the framework architecture, with empirical evidence. There is a risk that organisations that have yet to mature their metadata catalogue will be tempted to spend too much now on investments in downstream AI tools, but the regression evidence suggests that this will be far less effective without a catalogue that is mature.

The finding by the chi-square analysis that 19.4% of small organisations are fully implementing the governance framework is similar to that of medium organisations and significantly higher than the zero floor predicted by a resource-determined account indicates that governance framework implementation is an achievable goal for organisations of all sizes. It is recommended that enabling conditions that differentiate successful small-organisation adopters be investigated in detail through qualitative methods, but based on the present findings it would seem that regulatory pressure – especially in the Finance sector (confirmed as a significant path coefficient within the SEM) – and executive sponsorship are obvious choices. In resource-limited organizations, where governance programme justification strategies need to be tied to regulatory compliance, efficiency and AI-readiness may be less effective and narrative justification might work better.

C. Limitations

These findings are limited to a few aspects. Given that the primary predictor measurement strategy primarily involves self-reported governance adoption rates, it is possible for there to be some common method bias causing bivariate associations to be inflated. Some of these future studies should also include objective governance capability measurement(s) like automated catalogue completeness audit, lineage coverage telemetry from deployed governance platforms, and 3rd party measurement scores for compliance examination, which will independently validate the self-report results. Technologies and Cloud sector organisations were over-represented in the sample and tended to have higher governance maturity and AI readiness scores. The framework prescriptions require replication over the different sectors of manufacturing, education and non-profit which were undersampled in the present study to give confidence to the generalisability of the framework prescriptions across sectors. Furthermore, the cross sectional design does not allow for causal inference although the comparison of the pre-post implementation in the organisations that adopted the framework does give directional evidence in favour of a causal interpretation of the framework effects.

D. Future Research Directions

This research indicates that there are several priorities for future research. First, the proposed framework's specifications for the AI readiness assessment (Module 5) require dedicated empirical validation, as the benefit of letting the downstream effect of AI project outcomes, fairness, and robustness be measured, requires further confirmation by longitudinal tracking of the outcomes of AI projects in organisations with different AI dataset readiness scores, to confirm the hypothesised relationship between governance quality metrics and downstream benefit of the AI project. Second, the generative AI wave which started to impact enterprise AI contexts at the end of 2022 brings new metadata governance needs such as prompt lineage, retrieval-augmented generation context tracing and fine-tuning of the dataset documentation, which the proposed framework only outlines. Third, federated and privacy-preserving metadata governance architectures, which are relevant for multi-organisation data ecosystems where data sharing is limited by competitive or regulatory boundaries [21] are an important extension domain that can be covered by the definitions of the lineage and cataloguing module specifications.

VII. Conclusion

This paper has outlined a data governance framework that emphasizes addressing data governance needs systematically, through metadata, and provided empirical evidence from a sample of 250 data governance practitioners from six industry sectors and five types of analytical methods. The key quantitative results highlight that the three key factors most strongly independent from each other drive enterprise AI readiness – metadata catalogue maturity, data lineage coverage depth and compliance policy automation – which together drive 74% of the variance in enterprise AI readiness.

The outcomes from the structural equation modelling support the notion that organisational enablers (data culture, regulatory pressure, executive sponsorship and infrastructure readiness) have an impact on AI readiness indirectly via the adoption of the construct of metadata governance. This mediation has a tangible consequence: without proper implementation of good governance structures, any culture change or leader's commitment cannot be replicated to achieve the same level of preparedness for AI. Beyond the organisational conditions that drive its creation, the quality of the metadata fabric, complete, lineage deep, compliance complete and discovery intelligence and more, is the direct driver of AI-readiness.

The one-way ANOVA results, demonstrating that data governance maturity level accounts for 72.6% of AI readiness variance with large effect sizes between every adjacent maturity level, provide particularly compelling evidence that the transition from ad-hoc to structured, systematic governance produces qualitatively distinct AI readiness regimes. Post-implementation validation results—including an 81.7% reduction in data asset discovery time, 36.2 percentage-point improvements in metadata completeness, and 235.4% increases in cross-domain data reuse—quantify the operational value that structured metadata governance implementation delivers beyond the AI readiness improvements central to the framework's conceptual scope. Future work will extend the framework to federated data ecosystems, validate Module 5 AI readiness metrics against longitudinal AI project outcome data, and develop automated governance compliance tooling that reduces human implementation overhead across all five framework modules.

References

[1] V. Khatri and C. V. Brown, "Designing data governance," Communications of the ACM, vol. 53, no. 1, pp. 148–152, 2020.

[2] I. Alhassan, D. Sammon, and M. Daly, "Data governance activities: An analysis of the literature," Journal of Decision Systems, vol. 25, no. sup1, pp. 64–75, 2020.

[3] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," International Journal of Information Management, vol. 49, pp. 424–438, 2020.

[4] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge, 2nd ed. Technics Publications, 2020.

[5] J. Ladley, Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program, 2nd ed. Academic Press, 2020.

[6] M. Mosley, R. Henderson, S. Earley, and S. Sebastian-Coleman, DAMA Dictionary of Data Management, 3rd ed. Technics Publications, 2021.

[7] P. A. Bernstein and J. Madhavan, "Using schema matching to simplify heterogeneous data translation," Proceedings of the VLDB Endowment, vol. 12, no. 12, pp. 2276–2279, 2020.

[8] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, "Learning feature engineering for classification," in Proc. 26th International Joint Conference on Artificial Intelligence, pp. 2529–2535, 2020.

[9] R. C. Fernandez, E. Mansour, A. J. Qahtan, A. Elmagarmid, I. Ilyas, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang, "Seeping semantics: Linking datasets using word embeddings for data discovery," in Proc. 34th IEEE International Conference on Data Engineering, pp. 989–1000, 2020.

[10] G. Firican, "Active metadata management: The next frontier in enterprise data governance," The Data Administration Newsletter, vol. 18, no. 3, pp. 1–12, 2021.

[11] R. Hai, C. Quix, and J. X. Zhou, "Querying distributed and heterogeneous data sources with ESTOCADA," Proceedings of the VLDB Endowment, vol. 14, no. 12, pp. 2742–2745, 2021.

[12] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: A survey," ACM Computing Surveys, vol. 37, no. 1, pp. 1–28, 2020.

[13] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang, "Goods: Organizing enterprise datasets," in Proc. 2020 ACM SIGMOD International Conference on Management of Data, pp. 1907–1922, 2020.

[14] D. Deutch, T. Milo, S. Roy, and V. Tannen, "Circuits for datalog provenance," in Proc. 17th International Conference on Database Theory, pp. 1–13, 2021.

[15] E. Lim, A. Dasgupta, S. Huang, J. Lee, and Y. Liu, "A comparative analysis of column-level data lineage tools across cloud data platforms," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3778–3793, 2022.

[16] S. Schelter, J. H. Boese, J. Kirschnick, T. Klein, and S. Seufert, "Automatically tracking metadata and provenance of machine learning experiments," in Proc. ML Systems Workshop at NeurIPS, pp. 1–8, 2020.

[17] P. Voigt and A. von dem Bussche, The EU General Data Protection Regulation (GDPR): A Practical Guide, 2nd ed. Springer, 2021.

[18] C. Tankard, "What the GDPR means for businesses: Technical compliance requirements and enterprise data governance," Network Security, vol. 2020, no. 6, pp. 5–8, 2020.

[19] R. Gellert, "Understanding the notion of risk in the General Data Protection Regulation," Computer Law and Security Review, vol. 34, no. 2, pp. 279–288, 2020.

[20] P. Bowen and J. Winne, "HIPAA compliance in enterprise AI deployments: Data provenance and audit trail requirements for clinical decision support systems," Journal of Healthcare Information Management, vol. 36, no. 1, pp. 22–31, 2022.

[21] T. Chen, X. Jin, Y. Sun, and W. Zhong, "A unified metadata compliance framework for multi-regulation enterprise data environments," IEEE Access, vol. 9, pp. 89440–89455, 2021.

[22] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and A. Harmouch, "The effects of data quality on machine learning performance," arXiv preprint arXiv:2207.14529, 2022.

[23] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI," in Proc. 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15, ACM, 2021.

[24] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, and others, "The FAIR guiding principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 1, pp. 1–9, 2020.

[25] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Gray, and others, "FAIR principles: Interpretations and implementation considerations," Data Intelligence, vol. 2, no. 1–2, pp. 10–29, 2020.

[26] A. Quarati and J. E. Raffaghelli, "Do researchers use open research data? Exploring the relationships between use, findability and quality in an open data repository," Journal of Information Science, vol. 48, no. 5, pp. 669–685, 2022.

[27] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: Building a search engine for datasets in an open web ecosystem," in Proc. The World Wide Web Conference, pp. 1365–1375, ACM, 2020.

[28] M. Hulsebos, K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, C. Demiralp, and C. Hidalgo, "Sherlock: A deep learning approach to semantic data type detection," in Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1500–1508, 2020.

[29] A. Chapman, L. Munoz-Gama, C. Cabanillas, P. Sala, B. Pernici, and N. Paton, "Human factors in enterprise data cataloguing: Why relevance matters more than interface design," Information Systems, vol. 103, pp. 1–18, 2022.

[30] H. Grund, S. Gille, and R. Schlosser, "Dataset recommendation in enterprise environments using collaborative filtering over usage metadata," in Proc. 2022 ACM SIGMOD International Conference on Management of Data, pp. 1–14, ACM, 2022.

[31] T. Bauer, T. Eichler, and T. Reichert, "Towards automated data quality management for machine learning," in Proc. Workshop on Data Management for End-to-End Machine Learning at SIGMOD, pp. 1–7, 2021.

[32] D. Gil, P. Fonollosa, J. Fonollosa, and J. Salinas, "Machine learning for the prediction and identification of biological activities: Applications to drug discovery," Frontiers in Pharmacology, vol. 12, pp. 1–16, 2021.

[33] R. Ghavami, "Enterprise data governance trends in the cloud era: Lessons from large-scale deployments," Journal of Data and Information Quality, vol. 13, no. 3, pp. 1–24, 2021.

[34] S. Gröger, "No metadata, no AI: Towards leveraging enterprise metadata for trustworthy AI deployments," in Proc. Conference on AI, Ethics, and Society, pp. 1–9, ACM, 2021.

[35] A. Ghasemaghaei and G. Calic, "Assessing the impact of big data on firm innovation performance: Looking into the black box using meta-analysis," MIS Quarterly, vol. 44, no. 1, pp. 333–356, 2020.

[36] E. Muller, R. Macedo, and M. Spjuth, "Automating data lineage in cloud-native enterprise architectures," IEEE Cloud Computing, vol. 8, no. 4, pp. 34–43, 2021.

[37] O. Hartig and J. Zhao, "Publishing and consuming provenance metadata on the web of linked data," in Proc. International Provenance and Annotation Workshop, pp. 78–90, Springer, 2020.

[38] M. Stonebraker and I. Ilyas, "Data integration: The current status and the way forward," IEEE Data Engineering Bulletin, vol. 41, no. 2, pp. 3–9, 2020.

[39] A. Tal, R. Shem-Tov, and N. Hazon, "Measuring data governance effectiveness: A maturity index for enterprise AI environments," Information and Management, vol. 59, no. 4, pp. 1–18, 2022.

[40] W. Kim and B. J. Choi, "Towards a metadata-driven AI readiness assessment model for enterprise data environments," Expert Systems with Applications, vol. 185, pp. 1–16, 2021.