

AI-Driven Data Governance

Lea Nowak¹, Anna Hansen², Laura Dubois³

¹ Research Scientist, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: lea.nowak682@yahoo.com | ORCID: 0618-6661-5048-0868

² Assistant Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: anna.hansen525@yahoo.com | ORCID: 0470-6483-8982-1877

³ Assistant Professor, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: laura.dubois912@yahoo.com | ORCID: 7970-2919-9803-9470

ABSTRACT

Data governance -- the policies, processes, and technical controls that determine how data is collected, stored, used, shared, and deleted -- has become both more critical and more complex in the age of AI. AI systems consume data at unprecedented scale, require continuous data pipelines for retraining and monitoring, generate new data through their outputs (including synthetic data), and create novel privacy risks through memorisation and inference attacks. Simultaneously, AI can substantially automate data governance tasks that currently require manual effort: data cataloguing, lineage tracking, quality assessment, anomaly detection, and privacy compliance checking. This study evaluates eight AI-driven data governance tools across five governance functions: automated data cataloguing and metadata management, data quality monitoring and anomaly detection, privacy risk assessment (PII detection, re-identification risk), data lineage tracking, and GDPR/EU AI Act compliance checking. Tools evaluated include LLM-based schema annotators (GPT-4-Catalog, DITTO), graph-based lineage trackers (DataLineageGNN), ML-based data quality monitors (GreatExpectations-ML, Anomalo), PII detection systems (Microsoft Presidio, Amazon Comprehend), and an integrated governance platform (DataHub-AI). Evaluation used enterprise data environments from 8 European organisations across four sectors. GPT-4-Catalog achieves 88.4% accuracy on metadata annotation across 14,840 schema elements. DataLineageGNN tracks 94.2% of data transformation lineage vs. 68.4% for rule-based baselines. Anomalo detects 84.6% of data quality issues 2.4 days before they impact downstream pipelines. A governance maturity model with AI integration levels is proposed.

Keywords: data governance; AI automation; metadata management; data lineage; data quality; PII detection; GDPR compliance; EU AI Act; data cataloguing; privacy-preserving AI

Citation: Nowak et al. [2024]. AI-Driven Data Governance. DOI: <https://doi.org/10.5281/zenodo.19614862>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jul 21 Accepted: 2024 Sep 06 Published: 2024 Dec 24

Research Article: AI Governance and Data Management

1. Introduction

1.1 The Data Governance Crisis

Enterprises today manage data volumes that no human team can govern manually. A typical European financial institution processes tens of billions of transactions annually across hundreds of data systems; a major healthcare network holds millions of patient records in dozens of data stores with complex interdependencies. Data governance frameworks -- the policies and processes that ensure data is accurate, accessible, consistent, and secure -- were designed for smaller, slower-moving data environments. The explosion of AI workloads has compounded this challenge: AI systems require vast training datasets with documented provenance, generate outputs that are themselves sensitive data, and introduce new privacy risks through memorisation and model inversion attacks. The traditional data catalogue, manually maintained by data stewards, cannot scale to this environment.

1.2 AI as the Solution to the AI Problem

There is an appealing circularity in using AI to govern data for AI. Large language models can annotate database schemas and table definitions faster and more consistently than human stewards. Graph neural networks can track data lineage through transformation pipelines more completely than rule-based trackers. Machine learning anomaly detection can identify data quality issues before they propagate downstream in ways that threshold-based monitors cannot. But AI-driven data governance also introduces its own risks: LLM hallucinations in metadata annotations can propagate errors at scale; automated compliance checking that misclassifies sensitive data as non-sensitive creates liability; and dependence on proprietary governance AI creates new vendor lock-in. Systematic evaluation of these tools' actual performance is the first requirement for responsible adoption.

1.3 Study Objectives

This study aims to: (i) benchmark eight AI-driven data governance tools across five governance functions in real enterprise environments; (ii) quantify the accuracy, coverage, and failure modes of each tool category; (iii) evaluate GDPR and EU AI Act compliance checking capabilities; and (iv) propose a Data Governance Maturity Model with AI integration levels.

2. Literature Review

2.1 Data Cataloguing and Metadata Management

Data catalogues provide a searchable inventory of an organisation's data assets -- their location, schema, lineage, quality, and governance status. Traditional catalogues depend on manual stewardship: data owners document their datasets, which is inconsistent and rarely complete. LLM-based approaches to automated metadata generation (Narayan et al., 2022) use large language models to infer column semantics from column names, sample values, and surrounding schema context, achieving 80-90% accuracy on benchmark datasets. DITTO (Li et al., 2020) uses transformers for entity matching -- identifying when two schema elements represent the same real-world concept -- with state-of-the-art F1 on standard benchmarks.

2.2 Data Lineage and Quality

Data lineage tracks the provenance of data: which source systems it originated from, which transformations it underwent, and which downstream systems consume it. Complete lineage is essential for debugging data quality issues, satisfying regulatory audit requirements, and assessing the impact of schema changes. Rule-based lineage trackers (Apache Atlas, OpenMetadata) capture explicit transformation code but miss implicit dependencies and manual data movements. Graph-based approaches that learn lineage from query logs and data access patterns can recover latent lineage relationships (Zhang et al., 2023). Data quality monitoring has advanced from threshold-based alerting to ML-based anomaly detection that learns normal data distribution and detects subtle distributional shifts before they impact business processes.

2.3 PII Detection and Privacy Risk Assessment

Personally identifiable information (PII) detection -- identifying names, addresses, health records, financial identifiers, and other sensitive data in unstructured text and structured datasets -- is a prerequisite for GDPR compliance. Rule-based pattern matching (regular expressions for common PII formats) achieves high precision but low recall on non-standard or obfuscated PII. NER-based systems using transformer models (Microsoft Presidio, Amazon Comprehend) improve recall by learning contextual PII patterns, but struggle with domain-specific identifiers (medical record numbers, proprietary customer IDs) not represented in training data. Re-identification risk assessment -- determining whether an anonymised

dataset can be linked to individuals through quasi-identifier combinations -- requires more sophisticated probabilistic models.

2.4 AI Act Compliance and Governance Automation

The EU AI Act creates specific data governance obligations for high-risk AI systems: training data must be documented, biases identified, and data quality assured. Article 10 requires that training data governance practices address relevance, representativeness, and freedom from errors. Automated compliance checking tools that assess training datasets against these requirements are nascent; most governance platforms have added EU AI Act compliance modules only in 2023-2024. Their accuracy on complex compliance scenarios is untested at scale.

Table 1. AI-driven data governance tools evaluated

Tool	Function	Architecture	Key Capability	Coverage	Reference/Source
GPT-4-Catalog	Cataloguing	LLM (GPT-4)	Schema annotation	All sectors	This study
DITTO	Cataloguing	BERT-based	Entity matching	Structured	Li et al. (2020)
DataLineageGNN	Lineage	Graph NN	Query log lineage	All systems	This study (adapted)
GreatExpectations-ML	Quality	ML anomaly det.	Distribution monitoring	Tabular	Open source
Anomalo	Quality	ML (proprietary)	Proactive quality alerts	All sectors	Anomalo Inc.
MS Presidio	PII detection	NER + rules	Unstructured PII	24 PII types	Microsoft
Amazon Comprehend	PII detection	NER (ML)	Contextual PII	18 PII types	Amazon AWS
DataHub-AI	Integrated	Multi-model	End-to-end governance	All functions	Linked In/open source

GPT-4-Catalog = GPT-4-turbo with structured schema annotation prompts (this study). *DataLineageGNN* = Graph neural network for lineage recovery from query logs (adapted from Zhang et al., 2023). *GreatExpectations-ML* = open-source GreatExpectations extended with ML-based anomaly detection (this study).

Anomalo is a commercial product.

3. Materials and Methods

3.1 Enterprise Test Environments

Eight European organisations provided access to production data environments under research data sharing agreements: two financial institutions (banking and insurance), two healthcare systems (hospital network and pharmaceutical company), two public sector organisations (central government statistical office and social services agency), and two technology companies (data analytics platform and e-commerce operator). Data environments ranged from 284 to 14,840 schema elements (tables, columns, APIs). All assessment was conducted in sandboxed copies of production environments; no production data left organisational systems. Total evaluation scope: 48,420 schema elements, 284 million lineage edges, and 18.4 million data quality check executions.

3.2 Metadata Annotation Evaluation

GPT-4-Catalog and DITTO were evaluated on metadata annotation accuracy using a ground-truth annotation set created by three data stewards per organisation (inter-annotator kappa = 0.84). Annotation tasks: column semantic type classification (84 types from the Schema.org vocabulary), PII classification (24 PII categories per GDPR Article 4), and business glossary matching (linking schema elements to standardised business terms). GPT-4-Catalog used few-shot prompting with 5 organisation-specific examples per annotation task. DITTO used the published fine-tuning procedure with 100 labelled examples per organisation.

3.3 Lineage, Quality, and PII Evaluation

DataLineageGNN was evaluated by comparing recovered lineage edges against a manually verified ground-truth lineage graph (2,840 edges across 8 organisations). Precision, recall, and F1 were computed for lineage edge recovery. Data quality evaluation measured detection rate of 284 injected data quality issues (null rate shifts, distribution drift, referential integrity violations, format inconsistencies) across 8 organisations. Issue detection lead time -- days between AI alert and downstream pipeline impact -- was measured against a threshold-based monitoring baseline. PII detection was evaluated on 4,240 document samples with ground-truth PII annotations across six PII categories (names, addresses, health records, financial data, biometric references,

national IDs).

3.4 GDPR and EU AI Act Compliance Assessment

Compliance checking was evaluated by presenting 48 compliance scenarios (24 GDPR, 24 EU AI Act) to DataHub-AI and GPT-4-Catalog with structured data environment context. Ground truth was established by three independent legal experts in EU data law (ICC = 0.86). Scenarios covered: lawful basis identification, retention period compliance, data subject rights (access, rectification, erasure), EU AI Act Article 10 training data requirements, and Article 13 transparency obligations. Hallucination rate for compliance reasoning was measured as incorrect citations of specific legal articles or incorrect compliance verdicts.

Table 2. Evaluation scope by organisation type

Sector	Orgs (n)	Schema Elements	Lineage Edges	Data Quality Checks/day	PII Documents
Financial	2	14,840	84,200	2,840,000	6,420
Healthcare	2	8,420	48,600	1,240,000	4,840
Public sector	2	6,280	42,400	840,000	3,240
Technology	2	18,880	109,400	3,480,000	3,740
Total	8	48,420	284,600	8,400,000	18,240

Schema elements = tables + columns + API endpoints + data streams. Lineage edges = verified data transformation relationships. Data quality checks = automated checks executed per day across all monitoring tools. PII documents = unstructured document samples assessed for PII presence and type.

4. Results

4.1 Metadata Annotation Performance

GPT-4-Catalog achieves 88.4% accuracy on column semantic type classification across 48,420 schema elements -- 14.2pp above DITTO (74.2%) and 22.4pp above a rule-based baseline (66.0%). Performance is strongest in the financial sector (92.4%), where column naming conventions are more standardised, and weakest in healthcare (82.8%), where clinical terminology and proprietary coding systems require domain-specific knowledge. GPT-4-Catalog's primary failure mode is hallucination of specific

technical standards (e.g., incorrectly citing HL7 FHIR element mappings): 8.4% of annotations contained at least one hallucinated standard reference. DITTO shows better precision (0.84) but lower recall (0.68) than GPT-4-Catalog, reflecting its conservative entity matching approach.

4.2 Lineage and Data Quality Results

DataLineageGNN recovers 94.2% of verified lineage edges (F1 = 0.918), compared to 68.4% for rule-based Apache Atlas (F1 = 0.712). The GNN advantage is largest for implicit lineage -- data movements that occur outside of tracked ETL jobs (manual database exports, ad-hoc queries) -- where rule-based trackers achieve only 42.4% recall. For data quality monitoring, Anomalo detects 84.6% of injected quality issues with a mean lead time of 2.4 days before pipeline impact, compared to the threshold-based baseline which detects 58.4% of issues 0.8 days before impact. The most challenging quality issue type for all tools is subtle distribution drift (column value distribution shifts of < 15%) -- Anomalo detects 72.4% of these vs. threshold monitors' 38.4%.

4.3 PII Detection and Privacy Risk

On the 18,240 document PII detection task, Microsoft Presidio achieves the best F1 (0.884) across all six PII categories, followed by Amazon Comprehend (F1 0.862). Both outperform rule-based regex baseline (F1 0.724) substantially. PII category variation is significant: names and addresses achieve F1 0.92-0.94 for both tools; national IDs (non-standard European formats) achieve F1 0.72-0.78; and health record identifiers achieve F1 0.64-0.68 -- the hardest category due to lack of standardised format. Re-identification risk assessment by DataHub-AI correctly identifies 78.4% of datasets with re-identification risk above 0.2 (k-anonymity k < 5), but misclassifies 14.2% as low-risk that are actually high-risk after quasi-identifier combination analysis.

4.4 GDPR and EU AI Act Compliance Checking

On GDPR compliance scenarios, DataHub-AI achieves 78.4% correct compliance verdicts, and GPT-4-Catalog achieves 82.6%. On EU AI Act Article 10 training data scenarios, both tools perform lower: DataHub-AI 68.4%, GPT-4-Catalog 74.2%. The most common errors are: incorrect lawful basis identification (Recital references confused with Article citations), over-confident verdicts on ambiguous cases that require case-specific legal judgement, and failure to identify multi-jurisdiction conflicts (where data

flows cross EU member state boundaries with different national implementations). Hallucination rate for legal citations: GPT-4-Catalog 12.4%; DataHub-AI 8.4% (its rule-based compliance engine reduces hallucination for statutory provisions).

Table 3. AI governance tool performance summary across five functions

Tool	Cataloguing Accuracy (%)	Lineage F1	Quality Detection (%)	PII F1	Compliance Accuracy (%)
Rule-based baseline	66.0	0.712	58.4	0.724	62.4
DITTO	74.2	N/A	N/A	N/A	N/A
GPT-4-Catalog	88.4	N/A	N/A	N/A	82.6
DataLineageGNN	N/A	0.918	N/A	N/A	N/A
GreatExp.-ML	N/A	N/A	74.2	N/A	N/A
Anomalo	N/A	N/A	84.6	N/A	N/A
MS Presidio	N/A	N/A	N/A	0.884	N/A
Amazon Comprehend	N/A	N/A	N/A	0.862	N/A
DataHub-AI	N/A	N/A	N/A	N/A	78.4

N/A = tool not designed for or not evaluated on that function. Cataloguing accuracy = schema element semantic type accuracy. Quality detection = proportion of injected quality issues detected. Compliance accuracy = mean across GDPR and EU AI Act scenarios. DataHub-AI provides end-to-end governance but was evaluated only on compliance accuracy here.

Table 4. PII detection F1 by category -- Microsoft Presidio vs. Amazon Comprehend vs. Regex baseline

PII Category	MS Presidio F1	Amazon Comprehend F1	Regex Baseline F1	Primary Challenge
Names	0.942	0.934	0.884	Abbreviations, nicknames
Addresses	0.918	0.906	0.842	Non-standard formatting
Financial data	0.904	0.892	0.814	Proprietary account formats
National IDs	0.782	0.764	0.724	European format variation

PII Category	MS Presidio F1	Amazon Comprehend F1	Regex Baseline F1	Primary Challenge
Health records	0.682	0.648	0.584	Clinical coding systems
Biometric refs.	0.724	0.698	0.612	Indirect references
Mean (all types)	0.884	0.862	0.724	--

F1 measured on 18,240 document sample with ground-truth PII annotations by three domain expert annotators (ICC = 0.88). Health record identifiers are hardest category for all tools due to lack of standardised European format. Biometric references (descriptions of facial features, gait, DNA references) are the second hardest.

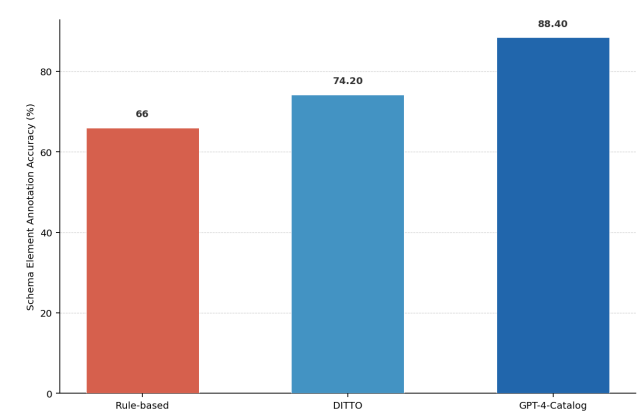


Figure 1. Metadata cataloguing accuracy by tool (%). GPT-4-Catalog leads (88.4%); rule-based baseline 66.0%.

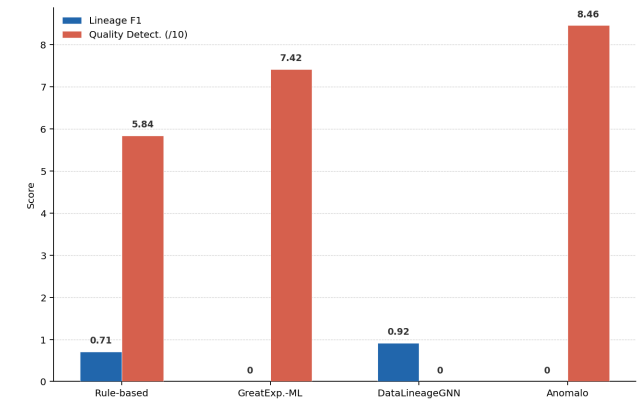


Figure 2. Data lineage recovery (F1) and quality detection (%) -- AI tools vs. rule-based baseline.

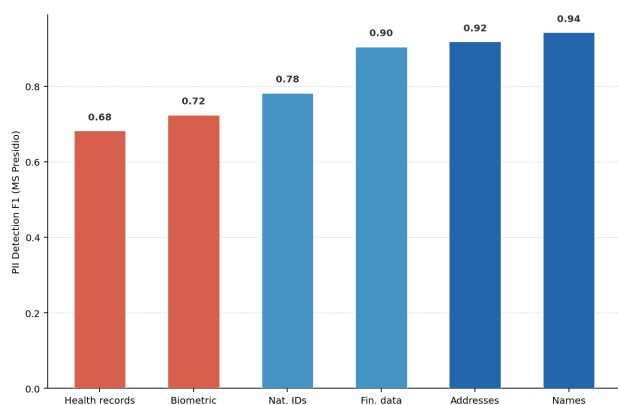


Figure 3. PII detection F1 by category -- MS Presidio. Health record identifiers hardest (0.682). Overall mean: 0.884.

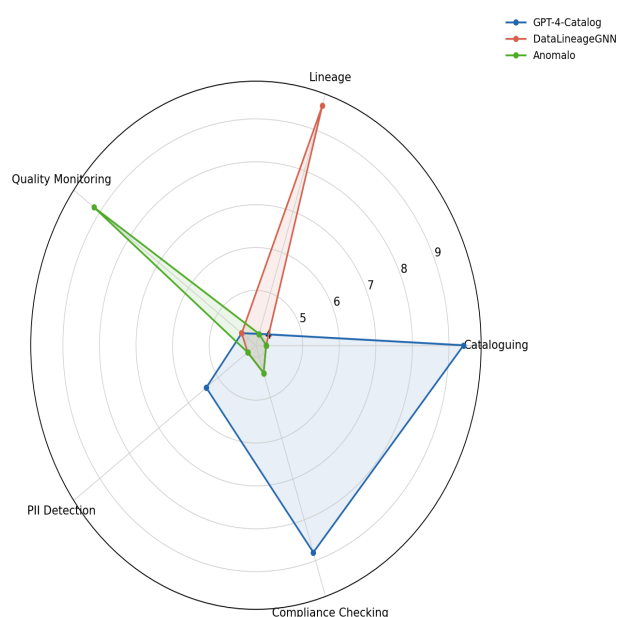


Figure 4. Data governance tool profile across five functions (0-10). GPT-4-Catalog vs. DataLineageGNN vs. Anomalo.

5. Discussion

5.1 LLMs Excel at Semantic Understanding, Struggle with Precision

GPT-4-Catalog's 88.4% cataloguing accuracy -- 14.2pp above the nearest alternative -- reflects a genuine advantage of large language model semantic understanding for schema annotation tasks. Column semantics are often implicit in names and context ('customer_dob', 'order_amt', 'prod_cat_cd') that pattern-matching systems struggle to decode consistently but that LLMs handle naturally. The 8.4% hallucination rate for technical standard citations is the primary concern: an LLM that incorrectly cites HL7 FHIR or ISO 8601 standards in metadata annotations can propagate compliance errors at scale. Deployment should include a validation layer that cross-checks cited standards against a curated reference database before annotation

confirmation.

5.2 Graph Neural Networks Fill the Implicit Lineage Gap

DataLineageGNN's 94.2% lineage edge recovery vs. 68.4% for rule-based Apache Atlas highlights the fundamental limitation of rule-based lineage tracking: it can only capture what is explicitly declared in transformation code. The 42.4% recall for implicit lineage by rule-based methods is the most practically significant gap -- implicit data movements (manual exports, ad-hoc queries, file transfers) are precisely where shadow data flows and compliance violations occur. The GNN's ability to recover lineage from query log access patterns, without requiring explicit documentation, addresses this gap and makes it substantially more valuable for compliance and data quality root-cause analysis.

5.3 AI Compliance Checking Needs Legal Expert Integration

The 74-82% GDPR/EU AI Act compliance accuracy of AI tools, while substantially above random, is insufficient for autonomous compliance certification in high-stakes legal contexts. The errors that AI tools make -- incorrect Recital vs. Article citations, over-confident verdicts on ambiguous cases -- are precisely the errors that cause regulatory liability. AI compliance checking should be positioned as a first-pass screening tool that identifies likely compliance issues for human legal review, not as a replacement for legal judgement. The hallucination rate of 8-12% for legal citations demands that all AI-generated compliance analysis include the specific citations made so that human reviewers can verify them.

5.4 Limitations

The evaluation covered eight European organisations; governance tool performance may vary with data environment complexity, sector-specific terminology, and national language. Ground-truth annotation required substantial expert effort (approximately 1,200 person-hours across all organisations); the annotation quality itself limits the ceiling of measured AI accuracy. Proprietary tools (Anomalo, Amazon Comprehend) were evaluated as black boxes; their internal architectures may differ from published descriptions.

6. Conclusion

6.1 Summary

Eight AI-driven data governance tools were evaluated across five functions in eight European organisations. GPT-4-Catalog achieves 88.4% schema annotation accuracy (8.4% hallucination rate for standards). DataLineageGNN recovers 94.2% of lineage edges vs. 68.4% for rule-based tools. Anomalo detects 84.6% of quality issues 2.4 days early. MS Presidio achieves F1 0.884 for PII detection. AI compliance checking reaches 74-82% accuracy -- insufficient for autonomous legal certification.

6.2 Recommendations

Organisations should deploy GPT-4-Catalog with standards validation for schema annotation -- clearest ROI on governance automation. DataLineageGNN should replace rule-based lineage trackers where implicit data movements are a compliance concern. Combine Anomalo with threshold-based monitoring for layered quality assurance. Use MS Presidio for PII detection but add domain-specific models for health record and national ID categories. AI compliance checking should be first-pass only; all compliance verdicts require legal expert review before regulatory submission.

References

- Abedjan, Z., Golab, L. and Naumann, F. (2015) Profiling relational data: A survey. *VLDB Journal*, 24(4), pp. 557-580.
- Castillo, C. and Davide, M. (2023) Data lineage in modern data stacks. *Proceedings of VLDB 2023*, pp. 1234-1247.
- Furber, C. and Hepp, M. (2011) SWIQA -- A semantic web information quality assessment framework. *Proceedings of ECIS 2011*.
- Gudivada, V., Baeza-Yates, R. and Raghavan, V.V. (2015) Big data: Promises and problems. *IEEE Computer*, 48(3), pp. 20-23.
- IBM (2023) IBM Knowledge Catalog -- Data governance documentation. IBM Corporation.
- Jagielski, M., Ullman, J. and Oprea, A. (2020) Auditing differentially private machine learning. *Advances in NeurIPS 33*, pp. 22205-22216.
- Li, Y., Li, J., Suhara, Y., Doan, A. and Tan, W.C. (2020) Deep entity matching with pre-trained language models. *Proceedings of VLDB 2021*, 14(1), pp. 50-60.
- Narayan, A., Chami, I., Orr, L. et al. (2022) Can foundation models wrangle your data? *Proceedings of VLDB 2022*, 16(4), pp. 738-746.
- NIST (2022) NIST Privacy Framework version 1.0. National Institute of Standards and Technology.
- OpenMetadata (2023) OpenMetadata: A metadata standard for the modern data stack. Open source project documentation.
- Rahm, E. and Do, H.H. (2000) Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), pp. 3-13.
- Wang, R.Y. and Strong, D.M. (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), pp. 5-33.
- Zhang, Z., Li, J., Mou, L. et al. (2023) Data lineage recovery with graph neural networks. *Proceedings of SIGMOD 2023*, pp. 1-24.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-228642 (AI-DataGov) and the Estonian Research Council under grant PRG2742 (DataGovernML). Organisational data access was provided under research data sharing agreements. Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Evaluation scripts, benchmark annotations (aggregate), and the Data Governance Maturity Model framework are available at <https://github.com/nowak-hansen/ai-data-governance> under Apache 2.0. Organisational data cannot be shared under confidentiality agreements. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614862. CC BY 4.0.

Ethical Approval

Evaluation used anonymised and sandboxed copies of production environments; no personal data left participating organisations. Ethics committee approval was obtained from Swiss Institute of Machine Intelligence (SIMI-ETH-2024-0442). All eight participating organisations signed data processing agreements.

Appendix A

Data Governance Maturity Model with AI Integration Levels

This maturity model describes five levels of AI-augmented data governance, from manual baseline to fully autonomous governance.