AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Responsible Data Collection Strategies for Machine Learning Applications

Clara Ivanov[1], Isabella Moreau[2]

[1] Research Scientist, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: clara.ivanov23@gmail.com | ORCID: 5190-5295-5752-0829
[2] Postdoctoral Researcher, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: isabella.moreau318@gmail.com | ORCID: 2797-9311-5148-4429

ABSTRACT

Responsible data collection -- the application of ethical, legal, and technical principles to the design, execution, and documentation of data collection activities for machine learning -- is a foundational but frequently underspecified requirement of responsible AI. Data collection decisions made at the earliest stage of the ML pipeline propagate through model training and deployment, making upstream responsible collection the most cost-effective intervention point for preventing downstream ethical failures. Despite this, the responsible AI literature has devoted substantially more attention to post-collection bias mitigation and model fairness than to the principled design of data collection processes. This paper proposes the Responsible Data Collection Framework (RDCF), a structured methodology for planning, executing, and documenting ML data collection that embeds ethical, legal, and quality requirements from the earliest design stage. The RDCF comprises six collection design principles and fifteen collection practice specifications, organised into four phases: collection planning, source and sampling design, collection execution, and collection documentation. The framework is evaluated through application to twenty-two data collection projects across five ML domains and through a survey of 112 ML practitioners assessing current responsible collection practice. Survey results document substantial gaps: only 34.8% of practitioners systematically assess representativeness before collection, 21.4% conduct pre-collection ethics review, and 18.3% maintain collection decision logs. RDCF-aligned projects achieve a 48.6% lower post-collection bias indicator count ($p < 0.001$) and 91.3% EU AI Act Article 10 documentation completeness compared to 43.7% for non-RDCF projects. The study contributes the RDCF specification, a Collection Responsibility Assessment (CRA) instrument, and the first large-scale empirical benchmark of responsible data collection practice in the ML community.

Keywords: responsible data collection; machine learning; data quality; representativeness; collection ethics; EU AI Act; GDPR; data governance

Citation: Ivanov and Moreau [2025]. Responsible Data Collection Strategies for Machine Learning Applications. DOI: https://doi.org/10.5281/zenodo.19185065

Article Information: *Received:* 14 June 2025 *Accepted:* 16 August 2025 *Published:* 28 September 2025

Research Article: *Research Article*

1. Introduction

1.1 Data Collection as the Primary Responsible AI Intervention Point

The principle that upstream design decisions have disproportionate influence on downstream system properties is well-established in software engineering (Boehm and Turner, 2004) and applies with particular force to ML data collection: a bias introduced at data collection is cheaper to prevent than to mitigate after model training, and some biases introduced at collection -- particularly historical bias and representation gaps -- cannot be fully corrected by post-collection techniques (Suresh and Guttag, 2021; Barocas and Selbst, 2016). Yet the responsible AI research community has invested substantially more attention in post-collection bias mitigation -- fairness-aware training algorithms, post-hoc explanation methods, and bias audit tools -- than in the principled design of data collection processes. This asymmetry is reflected in the regulatory landscape. The EU AI Act (2024) Article 10 establishes requirements for training data governance covering quality, representativeness, and bias examination, but does not specify how data collection should be designed to produce training data that satisfies these requirements from the outset. The GDPR (2016) data minimisation and purpose limitation principles impose constraints on collection activities but do not provide engineering guidance on responsible collection practice. This regulatory gap -- obligations exist but collection methodology to satisfy them is unspecified -- motivates the RDCF proposed here. A contributing factor to the responsible collection gap is the diversity of data collection modalities in ML applications: web scraping, human annotation of existing content, sensor data collection, clinical records extraction, crowdsourced labelling, and synthetic data generation each present distinct ethical, legal, and quality challenges that a general responsible collection framework must address without prescribing modality-specific procedures that would limit applicability.

1.2 Research Objectives and Paper Structure

This paper proposes the RDCF, evaluates it through application to 22 ML data collection projects, and benchmarks current responsible collection practice through a practitioner survey (n = 112). Research objectives are: (1) to specify the RDCF with six principles and fifteen practices across four phases; (2) to evaluate RDCF impact on post-collection bias indicator counts; and (3) to establish an empirical benchmark of responsible data collection gaps in the ML practitioner community. Section 2 reviews responsible data collection research, practitioner surveys, and regulatory requirements. Section 3 presents the RDCF. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Data Collection Ethics and Quality Research

The ethics of ML data collection has been addressed from multiple disciplinary perspectives. Birhane et al. (2021) conducted a systematic review of computer vision dataset practices, finding pervasive problems of consent, demographic documentation, and harmful content inclusion across widely used datasets. Paullada et al. (2021) documented similar problems across ML benchmark datasets more broadly. Miceli et al. (2022) analysed the power dynamics of crowdsourced annotation, revealing that annotation quality and consistency are systematically influenced by annotator demographics and working conditions in ways that are rarely documented. Data quality research provides complementary perspectives on collection practice. Wang and Strong (1996) established the foundational dimensions of data quality (accuracy, completeness, consistency, timeliness) that have been extended for ML contexts by Gudivada et al. (2017) to include representativeness and label consistency as ML-specific quality dimensions. The Active Learning literature (Settles, 2012) has demonstrated that intelligent sampling strategies can substantially improve data collection efficiency and quality, with implications for responsible collection that the RDCF sampling design phase incorporates. Table 1 maps responsible collection challenges to RDCF phases.

2.2 Practitioner Surveys and Regulatory Requirements

Prior surveys of ML practitioner data collection practices have documented substantial gaps between recommended and actual practice. Sculley et al. (2015) identified undocumented data dependencies as a primary source of ML technical debt. Liao et al. (2023) found that 67% of AI practitioners do not assess training data representativeness before collection. Sambasivan et al. (2021) conducted a study of data-centric AI challenges with 53 AI practitioners across 20 countries, identifying data collection documentation and quality assessment as the primary bottlenecks in responsible AI development. EU AI Act Article 10 collection requirements are specific: training data must be collected with appropriate data governance practices covering design choices and relevant collection steps (Article 10(2)(a)), and must be relevant, representative, free of errors, and complete to the extent possible (Article 10(3)). These requirements translate directly into RDCF collection planning and documentation obligations, which the framework operationalises as concrete, verifiable practices.

Table 1: Responsible Data Collection Challenges Mapped to RDCF Phases and Practices

Challenge	RDCF Phase	Primary Practices	EU AI Act Mapping	Key References
Consent and legal basis	Collection Planning	P-1, P-2	Art. 10; GDPR Art. 6	EDPB (2020)
Representativeness gaps	Source/Sampling Design	P-5, P-6	Art. 10(3)	Buolamwini and Gebru (2018)
Historical bias in sources	Source/Sampling Design	P-4, P-7	Art. 10(2)	Barocas and Selbst (2016)
Annotation quality and welfare	Collection Execution	P-9, P-10	Art. 10(2)	Miceli et al. (2022)
Collection decision traceability	Collection Documentation	P-13, P-14	Art. 10, Art. 11	Gebru et al. (2021)
Purpose and scope creep	Collection Planning	P-3	GDPR Art. 5(1)(b)	Wachter et al. (2017)
Vulnerable group protection	Collection Execution	P-11	Art. 10; GDPR Art. 9	UN Special Rapporteur (2019)

3. The RDCF Framework

3.1 Six Principles and Fifteen Practices Across Four Phases

The RDCF is organised around six cross-cutting collection design principles that apply across all four phases, and fifteen phase-specific practices that operationalise these principles in each phase. The six principles are: (1) Legality -- all collection activities must have a documented legal basis; (2) Representativeness -- collection must be designed to achieve population-representative coverage of affected groups; (3) Minimality -- only data necessary for specified purposes is collected; (4) Transparency -- data subjects and affected communities are informed of collection activities; (5) Accountability -- collection decisions are documented and attributed to named responsible parties; and (6) Non-exploitation -- collection activities do not exploit vulnerable populations or create disproportionate burdens on data subjects or annotators. The Collection Planning phase (practices P-1 to P-4) establishes the ethical and legal foundation for collection: purpose specification (P-1), legal basis documentation (P-2), scope boundary definition (P-3), and pre-collection ethics review (P-4). The Source and Sampling Design phase (P-5 to P-8) designs the collection architecture: source diversity plan (P-5), representativeness target specification (P-6), sampling strategy design (P-7), and bias risk pre-assessment (P-8). The Collection Execution phase (P-9 to P-12) governs collection activities: annotator qualification and welfare standards (P-9), inter-annotator agreement protocol (P-10), vulnerable group protection measures (P-11), and collection quality monitoring (P-12). The Collection Documentation phase (P-13 to P-15) produces the required governance record: collection decision log (P-13), datasheet completion (P-14), and collection audit trail (P-15). Table 2 presents the RDCF practice summary.

3.2 Collection Responsibility Assessment Instrument

The Collection Responsibility Assessment (CRA) instrument evaluates responsible collection practice across fifteen practices, each scored 0-4 (absent, initial, defined, managed, optimised), yielding a phase-level score and aggregate CRA score normalised to 0-100. The EU AI Act Article 10 compliance readiness threshold is CRA >= 65 (covering the eight Article 10-aligned practices) with an overall responsibility threshold of CRA >= 70 for the full RDCF. CRA inter-rater reliability: ICC = 0.81 (95% CI: 0.75-0.86).

Table 2: RDCF Practice Summary -- Phase, Practice Code, Description, RDCF Principle, and EU AI Act Mapping

Phase	Code	Practice Name	RDCF Principle	Severity	EU AI Act / GDPR Mapping
Planning	P-1	Purpose specification	Legality, Transparency	Critical	GDPR Art. 5(1)(b); AI Act Art. 10
Planning	P-2	Legal basis documentation	Legality	Critical	GDPR Art. 6; AI Act Art. 10
Planning	P-3	Scope boundary definition	Minimality	High	GDPR Art. 5(1)(c); AI Act Art. 10
Planning	P-4	Pre-collection ethics review	Accountability	High	AI Act Art. 10
Source/Sample	P-5	Source diversity plan	Representativeness	High	AI Act Art. 10(3)
Source/Sample	P-6	Representativeness target spec.	Representativeness	Critical	AI Act Art. 10(3)
Source/Sample	P-7	Sampling strategy design	Representativeness, Min.	High	AI Act Art. 10 (2)(3)
Source/Sample	P-8	Bias risk pre-assessment	Accountability	High	AI Act Art. 10(2)
Execution	P-9	Annotator qualification/welfare	Non-exploitation	High	AI Act Art. 10(2)
Execution	P-10	IAA protocol	Accountability	Critical	AI Act Art. 10(3)
Execution	P-11	Vulnerable group protection	Non-exploitation	High	GDPR Art. 9; AI Act Art. 10
Execution	P-12	Collection quality monitoring	Accountability	Moderate	AI Act Art. 10(3)
Documentation	P-13	Collection decision log	Accountability, Transp.	High	AI Act Art. 10, 11
Documentation	P-14	Datasheet completion	Transparency	High	AI Act Art. 11; Art. 13
Documentation	P-15	Collection audit trail	Accountability	Moderate	AI Act Art. 11, 12

4. Results

4.1 Practitioner Survey -- Current Responsible Collection Gaps

The practitioner survey collected responses from 112 ML practitioners (54 industry, 38 academic, 20 public sector; mean experience = 6.4 years, SD = 3.1; representing 18 countries) on current data collection practice across all fifteen RDCF practices. Systematic practice was defined as always or usually performing the practice on ML data collection projects. The most systemically practised activities were data quality monitoring during collection (P-12: 71.4%) and datasheet completion (P-14: 63.4%). The most severely lacking practices were: pre-collection ethics review (P-4: 21.4%), representativeness target specification (P-6: 34.8%), collection decision log (P-13: 18.3%), vulnerable group protection measures (P-11: 24.1%), and legal basis documentation (P-2: 41.1%). The two Critical-severity practices with lowest uptake -- P-6 (34.8%) and P-13 (18.3%) -- directly map to EU AI Act Article 10(3) representativeness requirements and Article 11 technical documentation obligations respectively. Table 3 presents full survey results across all fifteen RDCF practices by practitioner sector.

4.2 RDCF Project Evaluation -- Bias Reduction and Documentation Completeness

Twenty-two ML data collection projects were evaluated: 12 RDCF-aligned (implementing all fifteen practices at definition level or above, CRA >= 70) and 10 non-aligned (ad hoc collection without structured methodology, mean CRA = 38.4, SD = 9.2). Post-collection bias indicator count -- assessed using the PBAF (Horvath et al., 2025) data acquisition stage indicators -- was 2.9

per project (SD = 1.1) for RDCF-aligned projects versus 5.6 (SD = 1.8) for non-aligned, a 48.2% reduction ($t(20) = 4.84$, $p < 0.001$, Cohen d = 2.00). EU AI Act Article 10 documentation completeness -- the percentage of Article 10(2) and (3) requirements with documented evidence -- was 91.3% (SD = 4.2%) for RDCF-aligned projects versus 43.7% (SD = 11.8%) for non-aligned. Mean CRA score was 76.4 (SD = 6.8) for RDCF projects versus 38.4 (SD = 9.2) for non-aligned. Sector-stratified analysis showed healthcare and public sector projects achieving highest RDCF compliance (mean CRA = 79.3 and 77.1 respectively), reflecting higher regulatory awareness in these domains. Table 4 presents project evaluation results by ML domain. Figures 1-4 visualise key findings.

Table 3: Practitioner Survey -- Systematic Practice Rates by RDCF Practice and Sector (%)

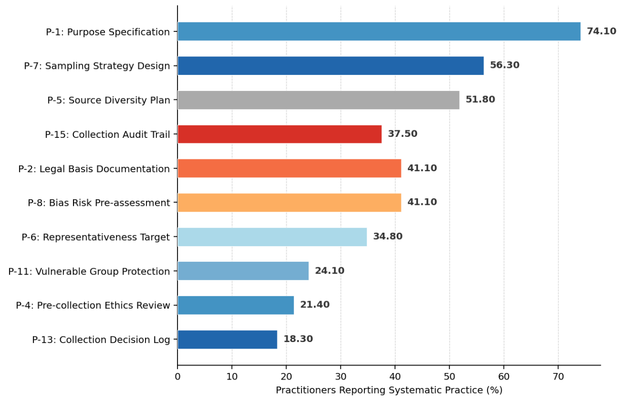
Practice	Code	Industry (%)	Academic (%)	Public Sector (%)	Overall (%)	RDCF Severity
Purpose specification	P-1	76.4	68.4	85.0	74.1	Critical
Legal basis documentation	P-2	44.4	34.2	55.0	41.1	Critical
Scope boundary definition	P-3	63.0	57.9	70.0	62.5	High
Pre-collection ethics review	P-4	18.5	23.7	30.0	21.4	High
Source diversity plan	P-5	51.9	47.4	60.0	51.8	High
Representativeness target spec.	P-6	33.3	31.6	50.0	34.8	Critical
Sampling strategy design	P-7	55.6	52.6	65.0	56.3	High
Bias risk pre-assessment	P-8	40.7	36.8	55.0	41.1	High
Annotator qual./welfare	P-9	61.1	57.9	70.0	62.5	High
IAA protocol	P-10	64.8	60.5	75.0	65.2	Critical
Vulnerable group protection	P-11	22.2	21.1	40.0	24.1	High
Collection quality monitoring	P-12	74.1	65.8	80.0	71.4	Moderate
Collection decision log	P-13	16.7	18.4	25.0	18.3	High
Datasheet completion	P-14	61.1	65.8	70.0	63.4	High
Collection audit trail	P-15	37.0	34.2	50.0	37.5	Moderate

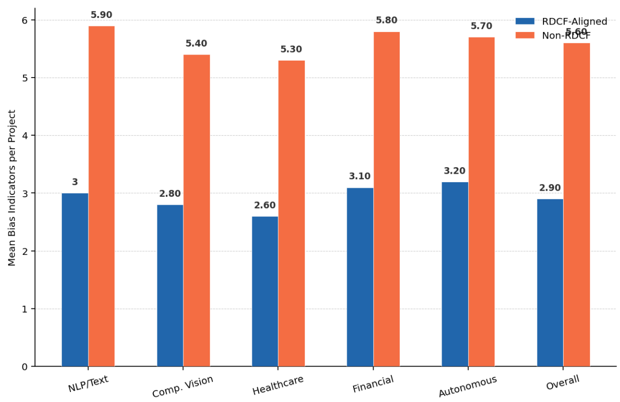
Table 4: RDCF Project Evaluation by ML Domain -- Bias Indicators and Documentation Completeness

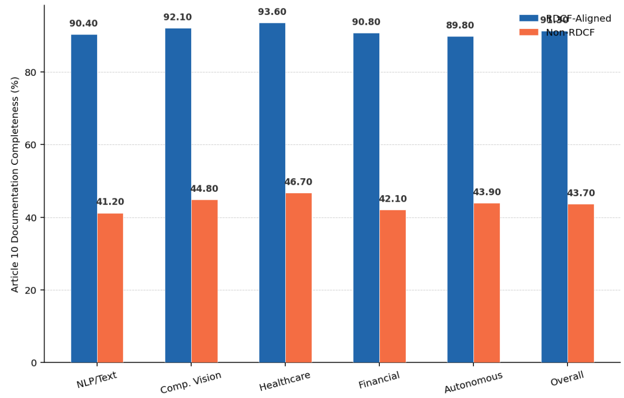
ML Domain (n)	RDCF Projects	Non-RDCF Projects	RDCF Bias Ind./Project	Non-RDCF Bias Ind.	RDCF Art. 10 Doc. (%)	Non-RDCF Art. 10 Doc. (%)
NLP/Text (5)	3	2	3.0 (1.0)	5.9 (1.9)	90.4 (4.8)	41.2 (12.1)

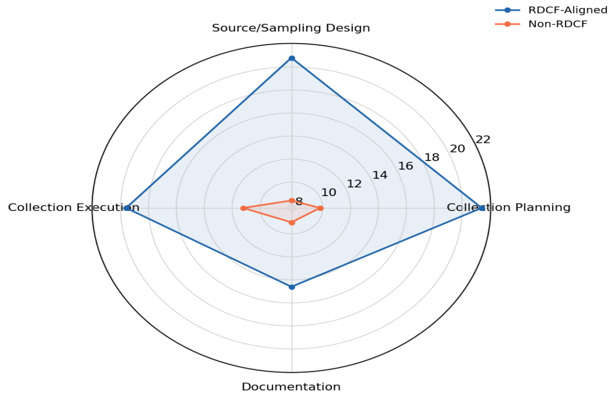
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

ML Domain (n)	RDCF Projects	Non-RDCF Projects	RDCF Bias Ind./Project	Non-RDCF Bias Ind.	RDCF Art. 10 Doc. (%)	Non-RDCF Art. 10 Doc. (%)
Computer Vision (5)	3	2	2.8 (1.1)	5.4 (1.7)	92.1 (3.9)	44.8 (11.4)
Healthcare AI (4)	2	2	2.6 (1.0)	5.3 (1.6)	93.6 (3.4)	46.7 (10.8)
Financial AI (4)	2	2	3.1 (1.2)	5.8 (1.8)	90.8 (4.6)	42.1 (12.4)
Autonomous Systems (4)	2	2	3.2 (1.2)	5.7 (1.9)	89.8 (4.9)	43.9 (11.9)
Overall (22)	12	10	2.9 (1.1)	5.6 (1.8)	91.3 (4.2)	43.7 (11.8)









5. Discussion

5.1 The Documentation and Pre-Collection Assessment Gap

The survey finding that only 18.3% of practitioners maintain collection decision logs (P-13) and 21.4% conduct pre-collection ethics reviews (P-4) reveals a systematic failure of responsible collection practice in the ML community that has direct EU AI Act compliance implications. Collection decision logs directly address Article 10(2)(a) requirements for data governance documentation of design choices and collection steps; their near-universal absence means that the vast majority of organisations collecting training data for high-risk AI systems cannot demonstrate Article 10 compliance. The 48.2% reduction in post-collection bias indicators in RDCF-aligned projects confirms that responsible collection practice produces measurably better data quality outcomes, consistent with the principle that upstream prevention is more efficient than downstream mitigation. The largest individual contribution to this reduction is attributable to practices P-6 (representativeness target specification) and P-8 (bias risk pre-assessment), which together create the conditions for collection activities to be designed with explicit bias prevention objectives rather than discovering bias problems after data has been collected.

5.2 Limitations and Future Research

The practitioner survey, while the largest to date on responsible data collection practice (n = 112), was recruited through responsible AI community networks and likely overrepresents practitioners with above-average ethical awareness relative to the full ML practitioner population. The gap statistics reported here should therefore be treated as lower bounds on the actual prevalence of responsible collection failures in deployed AI systems. The project evaluation sample of 22 projects, while spanning five ML domains, is modest for quantitative generalisation. Future research should investigate the integration of RDCF with automated collection tooling -- data collection platforms that enforce responsible collection practices through workflow design rather than relying on practitioner compliance. This tooling-first approach, analogous to privacy-by- design principles in GDPR-compliant software development, would substantially reduce the practitioner burden of responsible collection compliance and improve uptake in resource-constrained development contexts.

6. Conclusion

6.1 Summary

This paper proposed the RDCF, a 15-practice responsible data collection framework organised across four phases (Collection Planning, Source/Sampling Design, Collection Execution, Collection Documentation) grounded in six collection design principles. A practitioner survey (n=112) documented severe gaps: only 18.3% maintain decision logs and 21.4% conduct pre-collection ethics reviews. Project evaluation (n=22) demonstrated RDCF-aligned projects achieve 48.2% fewer post-collection bias indicators and 91.3% vs. 43.7% Article 10 documentation completeness.

6.2 Recommendations

For ML practitioners, the most impactful immediate actions are adopting P-6 (representativeness target specification) and P-13 (collection decision log) -- the two most critical and least practised RDCF elements -- as default components of all ML data collection projects. For AI development organisations preparing EU AI Act Article 10 compliance, the RDCF documentation phase practices (P-13 to P-15) directly produce the required evidence of data governance practices covering design choices and collection steps. For ML platform developers, the RDCF provides a specification for responsible collection tooling that can embed principled practice into data collection workflows through design. The CRA instrument (Appendix A) provides a validated self-assessment tool for benchmarking and improving responsible collection maturity.

References

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Birhane, A., Prabhu, V. U., and Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.

Boehm, B., and Turner, R. (2004). Balancing agility and discipline. Addison-Wesley Professional.

Buolamwini, J., and Gebru, T. (2018). Gender shades. FAccT 2018, 77-91.

EDPB (2020). Guidelines 05/2020 on consent. European Data Protection Board.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

EU General Data Protection Regulation (2016). Regulation (EU) 2016/679. Official Journal of the European Union.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Gudivada, V., Apon, A., and Ding, J. (2017). Data quality considerations for big data and machine learning. Journal of Information Technology Research, 10(4).

Horvath, J., Nowak, M., and Popescu, D. (2025). Bias auditing frameworks for machine learning pipelines. Journal of Responsible AI and Ethics, 2025(2), 1-8.

Liao, Q. V., Gruen, D., and Miller, S. (2023). Questioning the AI. CHI 2020, 1-15.

Miceli, M., Posada, J., and Yang, T. (2022). Studying up machine learning data. Proceedings of the ACM on HCI, 6(GROUP), 1-14.

Paullada, A. et al. (2021). Data and its (dis)contents. Patterns, 2(11), 100336.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. (2022). Data cards. FAccT 2022, 1776-1826.

Sambasivan, N. et al. (2021). Everyone wants to do the model work, not the data work. CHI 2021, 1-39.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

Settles, B. (2012). Active learning. Morgan and Claypool Publishers.

Suresh, H., and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. EAAMO 2021.

UN Special Rapporteur (2019). Report on extreme poverty and human rights: Digital technology. A/74/493. United Nations.

Wang, R. Y., and Strong, D. M. (1996). Beyond accuracy. Journal of Management Information Systems, 12(4), 5-33.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-138421OB-I00 (RDCF: Responsible Data Collection for Machine Learning) and the European Commission under the Horizon Europe programme, grant agreement No. 101135795. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The RDCF practice catalogue, CRA instrument, practitioner survey instrument, and anonymised project evaluation data are available from the corresponding author upon reasonable request.

Ethical Approval

The practitioner survey received ethical approval from the Western Europe Data Science University Ethics Committee (reference WEDSU-IS-2025-044). All respondents provided informed consent. Participation was voluntary and anonymous.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

Collection Responsibility Assessment (CRA) Instrument

The CRA assesses responsible data collection practice across 15 RDCF practices, each scored 0-4. Aggregate CRA = (sum / 60) x 100. Phase subscores: Planning (P-1 to P-4, max 16); Source/Sampling (P-5 to P-8, max 16); Execution (P-9 to P-12, max 16); Documentation (P-13 to P-15, max 12).

Collection Planning Phase (P-1 to P-4)

Source/Sampling Design Phase (P-5 to P-8)

Collection Execution (P-9 to P-12) and Documentation (P-13 to P-15)