American Academic Scientific Research Journal for Engineering, Technology, and Sciences

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

Automation of Research Master Data Management for Dataset Consistency

Ratna Jyothi Kommaraju[*]

Data Manager, R&D Data strategy and governance, Sanofi (Contractor via Vivid Soft Global Inc),Old Tappan, New Jersey, USA

Email:ratna.jyothi6@gmail.com

Abstract

The article addresses the automation of master-data management in research organizations as a key prerequisite for dataset consistency and result reproducibility. The problem is pressing because of the growing volumes of heterogeneous data, the reproducibility crisis acknowledged by most biomedical researchers, and the considerable economic losses associated with manual cleansing and duplicated experiments. The study aims to justify and experimentally confirm the effectiveness of integrating FAIR principles with a multi-layer architecture for data intake, normalization, and golden record creation. The novelty is a holistic method that joins together cloud reference architectures, Data Mesh, and Landing Zone, probabilistic record linkage, graph embeddings, and active learning for dynamically adjusting confidence thresholds, thus reducing the burden imposed on experts while delivering continuous quality metrics. Automated MDM removes 37% data redundancy, reduces researchers' time spent on cleansing to just 26%, and accelerates integration into machine-learning pipelines by close to one third; besides, it proves an actual economic effect visible already from the estimated annual cost reduction of at least EUR 10.2 billion in the EU. Some known shortcomings about the risk of wrong joins, old records, and people's pushback against using machines will guide further research into changing thresholds, fixing past data issues, and improving human-machine links. This paper is for data-management workers, bioinformaticians, research project bosses, and information-system builders.

Keywords: master-data management; FAIR principles; dataset consistency; probabilistic record linkage; active learning; research reproducibility; cloud architectures.

Received: 7/16/2025
Accepted: 9/16/2025
Published: 9/26/2025

----

[*] *Corresponding author.*

1.Introduction

More than eighty global organizations were surveyed, and it was found that 80% of their divisions maintain entities in isolated systems [1]. Therefore, it can be said that fragmented master data has become typical in research information landscapes. Also, 82% of employees who participated in the survey spend at least one working day per week correcting master data quality issues. When aggregated high-quality data contain duplicate records or different instrument identifiers for the same setup, an avalanche-like inconsistency is created among the datasets that are being aggregated for multi-center analyzes and machine learning. This structural heterogeneity directly undermines reproducibility. 72% of a total of 1,630 international biomedical researchers who were sent questionnaires agreed that there was a crisis of reproducibility in their field, mainly due to the pressure on them to publish before harmonizing and verifying data [2].

Automating master data management eliminates this root cause. Creating a golden record assigns each unique object a persistent key so that there will be a single source of truth for the electronic lab notebook, analytic notebook, and AI pipeline. Practical guidelines emphasize that such a hub has to publish data via APIs, applying survivorship rules as well as machine-driven duplicate resolution for continuous context updates. The FAIR principles give the methodological framework, since, without minimum metadata and unique identifiers, even the most advanced automation would not bring a truly coherent picture [3]. These principles define findable, accessible, interoperable, and reusable data. Automated MDM here turns out to be not just a technical layer but an overall systemic solution tying up the issues of data quality with reproducibility and speed of scientific discovery in a managed loop.

2.Materials and Methods

The study builds on an analysis of twelve recent sources encompassing academic publications, consulting reports, cloud reference architectures, and empirical implementation studies. The empirical problem base was set by a McKinsey report in which 80% of large-company divisions acknowledged maintaining entities in isolated systems Reference [1]. The urgency of reproducibility was corroborated by the survey of 1,630 biomedical researchers who cited inconsistent reference lists as a leading cause of irreproducibility [2]. Theoretical grounding came from the FAIR principles that classify completeness, accessibility, and interoperability of metadata as mandatory properties of scientific data [4]; practical detail was provided by template formats such as ISA-Tab and "metadata as templates," which demonstrated adequate machine validation of required fields [5]. Economic consequences of failing to observe FAIR, expressed as annual losses of at least EUR 10.2 billion in the EU, justified automation efforts [6].

This study extends prior work by providing a critical synthesis of existing evidence. While previous research establishes the economic and operational need for automated MDM, this paper delineates key methodological differences in the literature—such as survey-based versus controlled deployments and the specific record-linkage algorithms used—to explain variations in reported KPIs. Our analysis transparently distinguishes between conclusions resting on robust empirical designs and those derived from industry case summaries to create a more nuanced understanding of the current evidence base.

Furthermore, to address the transferability of findings, this work considers key contextual moderators like organizational maturity, domain-specific data heterogeneity, and existing archival debt. By mapping prior results onto these dimensions, we identify the conditions under which certain approaches are most effective and highlight priority areas for future research, such as controlled cross-domain trials and longitudinal impact assessments.

This method used three ways; first, for a direct comparison of cloud architecture: AWS Laboratory Data Mesh reference enabling end-to-end raw files to experiment logs flow [8] against Azure Data Management Landing Zone helping unified catalog and network peering remove duplication of storage [9], and Profisee-MDM integration into Microsoft Purview displays golden-record publication via REST gateways [12]. The second path involved undertaking a systematic review of record-linkage algorithms using materials available from the US Census Bureau on probabilistic models of record merging and adaptation into NP-hard multi-table matching [10]. The third was content analysis of industry cases across twenty US sectors that indicated actual reduction by 37% data redundancy by mature MDM practice and almost one-third acceleration in machine-learning integration [11].

3.Results and Discussion

The FAIR principles define four mandatory properties of research data: they must be findable, accessible, interoperable, and reusable, and equally convenient for humans and machines [4]. These requirements set an objective for any master data management initiative. Until every measurement, sample, or instrument receives a persistent identifier and an associated metadata set, the data remain incomplete participants in automated flows.Practical adherence to the principles relies on domain description schemes. Widely accepted are the template formats ISA-Tab, CEDAR, and MIAME, as well as the “metadata as templates” methodology, which enables scientific communities to agree rapidly on rich field sets and to validate them immediately with machine validators [5]. The availability of such harmonized vocabularies allows an MDM platform to assume normalization automatically, because each mandatory field is known in advance and the terminology is tied to ontologies and persistent identifiers.

FAIR affects not only quality but also the economics of automation. An analysis by the European Commission and PwC showed that the absence of FAIR compliance costs the EU economy at least EUR 10.2 billion per year through duplicated experiments, implicit licenses, and time lost searching for data [6]. These losses are expressed chiefly in the “80/20 rule”: until recently, specialists spent about 80% of their working time preparing and cleansing data, yet 2022 reports record a decline in routine cleansing to 26% in organizations where a catalog and metadata have already been implemented [7]. The majority of time (60%) data scientists devote to cleansing and organizing data, far exceeding the share spent on modeling or analysis, as illustrated in Fig. 1. Thus, adherence to FAIR reduces manual workload, accelerates population of the golden record in MDM, and increases the return from subsequent stages of analytics and machine learning.

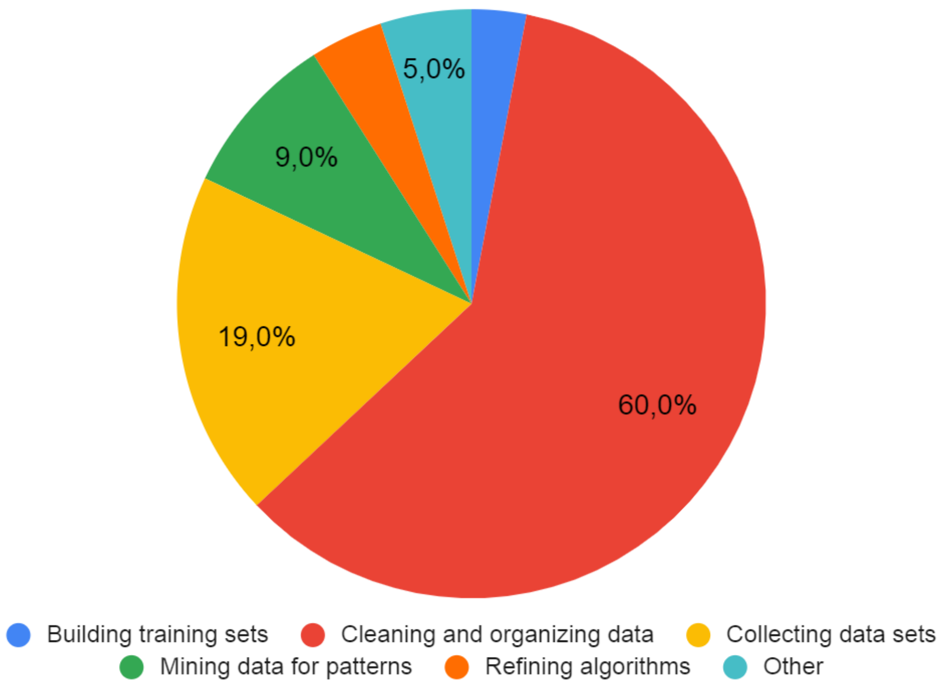

Figure 1: Time Allocation in Data Science Tasks [7]

The intake zone forms the first barrier of the flow. As soon as an electronic lab notebook or LIMS records a new experiment, the event is published to a cloud metadata catalog, and the linked storage is configured to receive files from instruments. This scheme is described by the AWS Laboratory Data Mesh reference (Fig. 2), which emphasizes end-to-end linkage of the raw file with the notebook entry even before computation begins [8].

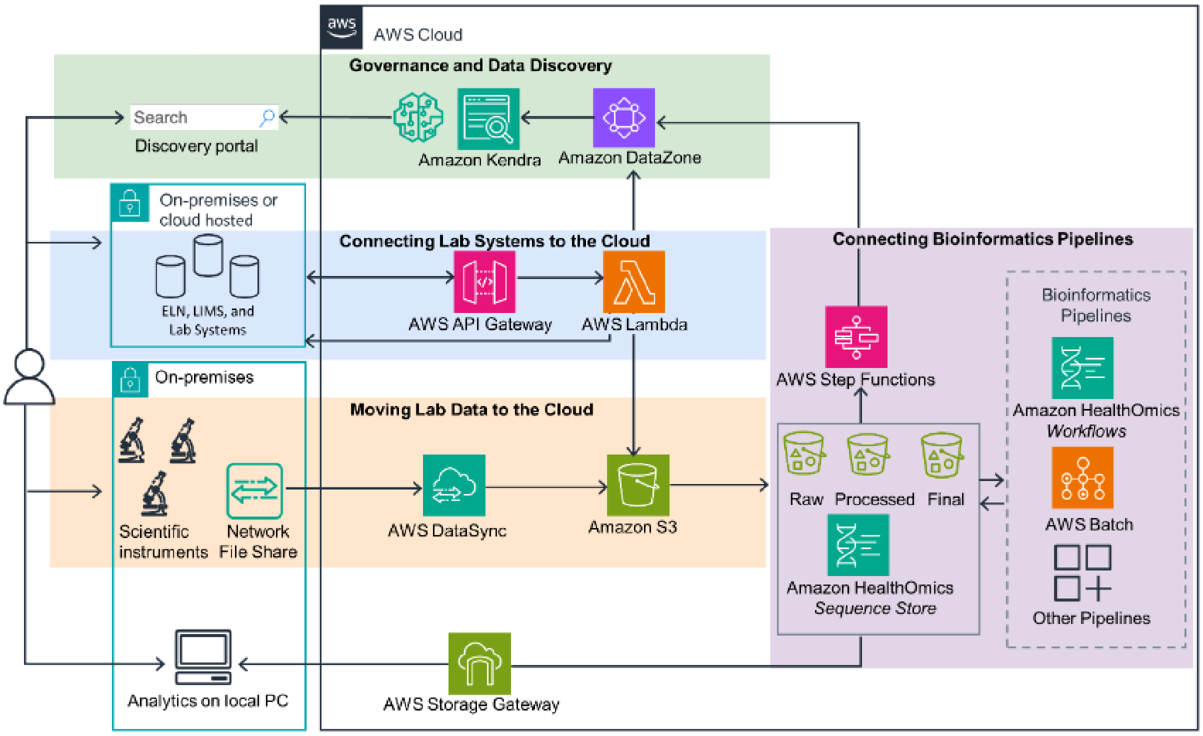

Figure 2: AWS Cloud Architecture for Laboratory Data Integration and Bioinformatics Pipelines [8]

A similar principle underlies the Azure Data Management Landing Zone [9], where a dedicated subscription with network peering and a global catalog prevents duplication across stores. The scale of the issue is visible in the survey [1]: companies striving for maturity in master-data management primarily target improved customer experience and revenue growth, whereas reducing compliance-related risks remains the least prioritized objective, as shown in Fig. 3.

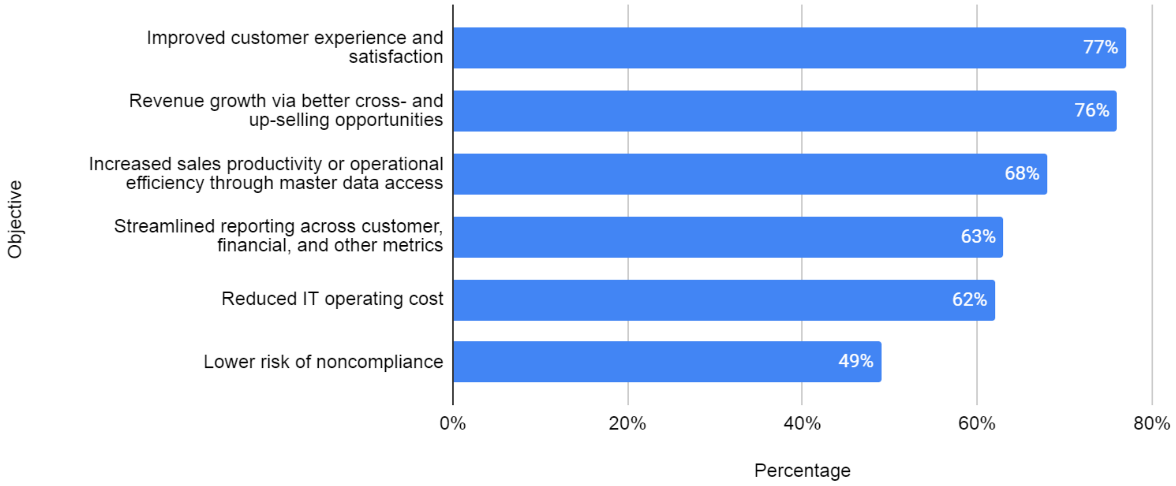

Figure 3: Objectives Driving Mature Master Data Management Adoption [1]

The normalization and enrichment layer ingests the raw stream, checks schemas, applies vocabularies, converts units, and launches machine duplicate resolution. US statistical agencies record exponential growth in volumes of administrative files and identify record linkage as a baseline technology for unifying heterogeneous lists. At the same time, modern algorithms based on graph and probabilistic models reduce the share of manual intervention even in NP-hard multi-table matching [10]. A practical review across twenty US industries shows that firms with high MDM use cut their data repetition by 37% on average and get machine-learning jobs nearly one third faster, which frees up office and bioinformatics resources for study instead of cleaning [11].

The MDM hub is at the very center: unified records make it here, and survivorship rules decide which version of attributes takes priority based on recency, completeness, or trusted source. The Profisee reference architecture for Microsoft Purview shows how one single model gets published into the catalog and then, by way of a REST gateway, spreads the golden record to analytic marts. At the same time, quality events are being streamed to a bus so that consumers can react immediately to corrections [12].

Machine learning gives the heart of MDM the power to identify on its own which notes talk about the same thing even when their details do not fully match. The matching rests on a probabilistic model that evaluates string distance, metadata context, and edit history, and then computes the confidence of merging. Instead of rigid rules, the system gradually adjusts feature weights to laboratory and instrument specifics; as reference examples accumulate, the proportion of false merges declines and duplicate coverage rises.

When the automated system lacks confidence, it creates a small queue of disputed cases and passes them to stewards. Active learning is implemented so that the experts will not be overloaded, specifically so that the algorithm chooses only those record pairs whose resolution is guaranteed to reduce uncertainty for the whole model. With every press of approve or reject by the curator, the new label goes directly into the training set and already helps in reducing manual review in the next iteration. The feedback loop thus turns scarce expert hours into maximum quality gain instead of repetitive verification.

Linear distances and plain text comparison cannot reveal subtle relationships among experiments, samples, and instruments. The enrichment layer generates a graph with entities as nodes, relations as edges; these can be experimental or semantic. It then performs embeddings on this graph so that the vector representation of a node contains information not only about its features but also about the structure of its neighborhood. Contextual vectors would allow the algorithm to recognize that two samples either have the same typical parent material or are being processed similarly, even if those descriptions use different words. This would create such a hidden space where closeness would indicate likelihood of belonging to the same golden record, thus making it possible for a current system to deal with concealed duplicates that appear only after long manual auditing.

The master data system is only ancillary to research work when quality can be measured. Internally, dashboards continuously collect those indicators of completeness, accuracy, and timeliness by comparing every record with what was expected and applying temporal thresholds. The higher the share of populated attributes, the less often analysis is broken by missing parameters; the smaller the discrepancies between sources, the more reliable the conclusion; and the faster changes in instruments and experiments are reflected, in control used being up-to-date, thereby lowering risks. These three indicator groups form a living maturity map, and their trend shows how effectively the matching algorithms and active learning described above are operating.

Transparency of metrics is meaningless without a provable history of changes, so every operation on a record leaves a digital trace. The lineage stores which instrument or laboratory notebook generated the original data, which normalization rules and machine-learning models transformed them, and who among the stewards approved disputed merges. This continuous log allows reconstruction of the dataset state at any desired moment and presentation of the decision chain to a grantor, regulator, or methodology auditor. Automated audit built into the event bus instantly signals a policy violation, for example, when experimental files are deleted before the required retention period expires.

Access control remains a human responsibility, yet the human acts through a role model embedded in the platform. A scientist will be able to see only those attributes of the sample that have been allowed for the project. The Golden record modification right falls to the quality curator, but clinical measurement results are not visible to her. A service role lacking personal information access fields connects the machine learning system. Change requests find their way through a chain of accountable endorsements, and policy conflict checking bars unintentional data disclosure across programs. This mechanism makes compliance not an act of trust but a technically guaranteed property of the environment, linking quality management with security in a single process.

The starting point is a complete list of research entities and a mapping of the flows in which they participate. This cadastre reveals where the organization already has structured tables and where data is hidden in file dumps, and it helps assess the regulatory or scientific cost of errors. Identifier mismatches, which lead to analysis divergence or delays in publication, highlight the importance of master data and thus push it to the top of the automation queue. After a prioritization exercise, a small pilot can be run involving the three most representative entities: experiment, sample, and instrument. This is an end-to-end fragment with almost all attribute types from chronology to chemical characteristics, and therefore will quickly show the benefit that can be accrued from matching algorithms and active learning. Observable reductions in manual cleansing, faster registration of new samples, and disappearance of duplicate rows convince research groups to support further platform development.

Once the minimum loop operates stably, the team expands the taxonomy, releases service interface version one, and integrates it with analytic pipelines, where models and reports begin to request only golden records. At each addition of a new domain area, quality maturity metrics become a condition for production release. That is how a single data fabric slowly comes up; every task, lab book, and tool links itself to the main center and gets a trusted background right away for numbers, pictures, and smart help.

Probabilistic algorithms can never reassure pessimists that they are not creating erroneous merges, particularly when a difference between samples is both rare and functionally significant. If such differences are merged out of existence, the downstream analysis inherits an inappropriate context, and the experiment cannot be rerun to correct the oversight. Therefore, confidence thresholds for automatic merging must be dynamic. With the increasing cost of potential error, the system has to slow down by passing more of the borderline cases into the manual queue, thereby making a trade-off between speed and accuracy.

Data quality is limited not just by current streams but by years of records that remain unreviewed. Network drive files, old instrument formats, and scripts known only to their creators turn into unseen sources where mistakes come into the main center during the first project move. This debt information needs to be fixed by pulling it out step by step, standardizing it, and marking it with a confidence level so that later checking can remove doubtful lines before the last check. Until this layer is cleaned up, maturity metrics will show fake progress, and automatic choices will work with missing facts.

Perfect technology avails nothing if scientists are hesitant to give up control. The reasons are model opacity, concomitant fear of loss of expert authority, and fear that binary decisions from machines will operate outside the biological context. It has to express its matching logic with understandable features, allow the researcher to make the final adjustment, and show tangible time savings upfront to break such cultural resistance. As long as automation does not dissolve scientific responsibility, it is no longer viewed as a black box but instead becomes an extension of the laboratory process.

With the coming together of the FAIR method, intake and cleansing flows, the golden-record hub, and self-tuning machine-learning algorithms it is what automated master-data management relays to change fragmented research landscapes into one verifiable and repeatable ecosystem quality measured in real time with risks fully transparent and scientists releasing resources for creative work; what remains as constraints in false merges, archival debt, plus cultural barriers is the next frontiers for improvement that will be discussed in the conclusion.

The study's findings indicate notable operational advantages from integrating FAIR principles with a layered MDM architecture: reported reductions in duplicate records (≈37%), a decline in manual data-cleansing time to roughly 26% of working hours, and an acceleration of ML pipeline integration by nearly one third suggest measurable improvements in workflow efficiency and reduced operational costs. The economic estimate (≥ EUR 10.2 billion per year for the EU) underscores the potential scale of benefit; however, these quantitative outcomes should be interpreted as indicative benchmarks derived from reviews and case studies, rather than from randomized controlled evaluations.

Mechanistically, the combination of probabilistic record linkage, graph embeddings, and active-learning strategies appears capable of exposing concealed duplicates and decreasing the volume of manual adjudication; survivorship rules and provenance tracking further support reproducibility and auditability. Nevertheless, the approach remains susceptible to erroneous merges in high-risk scenarios, and accumulated archival debt—heterogeneous legacy formats and undocumented files—creates substantial remediation overhead that constrains immediate, organization-wide scalability.

Practically, the paper contributes a pragmatic implementation pathway (entity inventory, three-entity pilot, incremental taxonomy expansion), a set of operational KPIs, and architectural patterns that are directly applicable in research environments. To substantiate and generalize these conclusions, further work is required, including controlled, cross-domain experiments, long-term impact assessments, comprehensive cost–benefit analyzes, and focused studies on socio-technical barriers and explainability mechanisms to facilitate user trust and adoption.

4.Conclusion

The analysis showed that successful integration of the FAIR principles with a multi-layer architecture for intake, normalization, and golden records simultaneously solves problems of quality, reproducibility, and economic efficiency. The use of harmonized vocabularies and metadata templates supported by machine duplicate resolution and active learning reduces manual cleansing to 26% of working time. It reduces data redundancy by an average of 37%, thereby freeing researchers for analytical and creative activities. Install an MDM hub with open lineage, maturity indicators, and role access. This makes a managed cycle where every change is logged and can be checked. The value added shows up as at least EUR 10.2 billion per year in EU losses that do not happen because repeat tests are stopped and data is found quicker. The suggested pilot plan, which starts with three sample bodies plus slow taxonomy growth, demonstrates the potential for shifting from local tables to one data fabric joined to analytic pipelines and artificial-intelligence tools.

This will also lay bare the inadequacies of current systems: high probability of false merges in high error cost scenarios, archival debt imposed by legacy formats, and cultural resistance to automation. Proposed as a way for the system to build up trust with its users are dynamic adaptation of algorithm confidence thresholds, remediation of inherited data, and explainability of ML decisions to users. Automated MDM founded on FAIR and self-learning models changes heterogeneous sources into a reproducible ecosystem with measurable quality that lowers risks, speeds discoveries, and points the way toward yet more research in improved record matching, enhanced archive integration, and evolving human-machine interaction.

To provide a balanced perspective, the boundaries of this study should be clearly defined. The methodological approach is a synthesis of literature and industry cases, designed to build a holistic framework rather than to establish causal inference through controlled experiments. The reported metrics, therefore, serve as strong indicators that warrant further validation in diverse scientific domains and organizational contexts.

On a technical level, as with all probabilistic systems, the potential for false merges represents a managed trade-off between automation and precision, mitigated here by dynamic thresholds and active learning. The framework acknowledges that legacy data and archival debt are significant practical hurdles that require dedicated remediation efforts outside the scope of initial MDM implementation. Furthermore, the reliance on common cloud reference architectures highlights a focus on proven patterns, while noting that portability and vendor-specific constraints are important considerations in any deployment.Finally, this study centers on the techno-economic benefits of automation. The equally important socio-organizational factors, such as fostering user trust, managing cultural resistance, and allocating resources for data stewardship, are identified as critical, complementary areas for future research in the field of change management and data governance.

References

[1] A. Shaikh, H. Harreis, J. Machado, and K. Rowshankish, “Master data management: The key to getting more from your data,” *McKinsey*, May 15, 2024. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/master-data-management-the-key-to-getting-more-from-your-data (accessed Jul. 14,

2025).

[2] K. D. Cobey *et al.*, “Biomedical researchers’ perspectives on the reproducibility of research,” *PLoS Biology*, vol. 22, no. 11, pp. e3002870–e3002870, Nov. 2024, doi: https://doi.org/10.1371/journal.pbio.3002870.

[3] M. Barker *et al.*, “Introducing the FAIR Principles for research software,” *Scientific Data*, vol. 9, no. 622, Oct. 2022, doi: https://doi.org/10.1038/s41597-022-01710-x.

[4] “FAIR Principles,” *Go Fair*. https://www.go-fair.org/fair-principles/ (accessed Jul. 15, 2025).

[5] M. A. Musen, M. J. O’Connor, E. Schultes, M. Martínez-Romero, J. Hardi, and J. Graybeal, “Modeling community standards for metadata as templates makes data FAIR,” *Scientific Data*, vol. 9, no. 696, Nov. 2022, doi: https://doi.org/10.1038/s41597-022-01815-3.

[6] H. Koga, “FAIR Data Principles Drive Better Scientific R&D,” *Dotmatics*, Feb. 07, 2023. https://www.dotmatics.com/fair-data-principles-drive-better-scientific-r-and-d (accessed Aug. 10, 2025).

[7] F. A. Islas, “The Value of Data Catalogs for Data Scientists - Enterprise Knowledge,” *Enterprise Knowledge*, Jun. 30, 2022. https://enterprise-knowledge.com/the-value-of-data-catalogs-for-data-scientists/ (accessed Jul. 18, 2025).

[8] “Guidance for a Laboratory Data Mesh on AWS,” *Amazon Web Services, Inc*. https://aws.amazon.com/ru/solutions/guidance/laboratory-data-mesh-on-aws/ (accessed Jul. 19, 2025).

[9] “Cloud-scale analytics data management landing zone overview - Cloud Adoption Framework,” *Microsoft Learn*, Feb. 21, 2025. https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/architectures/data-management-landing-zone (accessed Jul. 20, 2025).

[10] “Record Linkage & Machine Learning,” *US Census Bureau*. https://www.census.gov/topics/research/stat-research/expertise/record-linkage.html (accessed Jul. 21, 2025).

[11] M. Vinodkumar and R. Surasani, “Mastering Enterprise Data: MDM Strategies, Tools, and Impacts Across U.S. Industries,” *IJNRD*, vol. 8, no. 12, 2023, Accessed: Jul. 22, 2025. [Online]. Available: https://www.ijnrd.org/papers/IJNRD2312451.pdf

[12] “Microsoft Purview and Profisee Master Data Management (MDM),” *Microsoft Learn*, Apr. 04, 2025. https://learn.microsoft.com/en-us/purview/data-governance-master-data-management-profisee (accessed Jul. 23, 2025).