

Article

Examining the Challenges of Data Curation

Abdullahi Ayegba¹, Salawu Fatima Ayomide², Adekemi Olagunoye³, Ekene Carol Ajana⁴, Memunat Lawal⁵, Aisha Ibrahim⁶

^{1,2,3,4,5,6}Management of Information Technology Department, African University of Science and Technology, Abuja, Nigeria

Article information:

Manuscript received: 9 May 2026; **Accepted:** 10 Jun 2026; **Published:** 1 Jul 2026

Abstract: In today's digital world, data plays a major role in almost every sector, including healthcare, education, business, government, and scientific research. Organizations generate and store large amounts of data daily, making proper data management increasingly important. This paper examines the major challenges affecting data curation and discusses their impact on organizations, researchers, and information systems, using explanatory research method approach. It also explores possible solutions such as the adoption of data governance policies, improved technological infrastructure, staff training, and collaboration among stakeholders. It was observed from the results that the common challenges facing data curation in the society include inadequate storage infrastructure, lack of standard data management practices, data security and privacy concerns, insufficient funding, and shortage of skilled professionals. The rapid growth of big data and emerging technologies has also increased the complexity of handling large volumes of structured and unstructured data. The study concludes that effective data curation is essential for ensuring the long-term value, reliability, and accessibility of data in the digital age.

Keywords: Data curation, Data quality and integrity, Insufficient funding, skilled professionals.

1. Introduction

The rapid growth of digital technology has led to a significant increase in the amount of data generated daily across different sectors such as healthcare, education, business, and scientific research. As data continues to grow in volume and importance, organizations now place greater emphasis on proper data management and preservation. This has made data curation an essential aspect of modern information management. Data curation refers to the process of collecting, organizing, managing, preserving, and maintaining data to ensure that it remains accurate, secure, accessible, and useful over time. According to the Digital Curation Centre [1], data curation involves maintaining and adding value to digital data throughout its lifecycle. Proper data curation helps organizations improve decision-making, enhance research quality, and ensure easy retrieval of information when needed.

Despite its importance, many organizations face several challenges in managing data effectively. Common issues include poor data quality, inadequate storage systems, lack of standardization, data security and privacy concerns, insufficient funding, and shortage of skilled professionals. In addition, the rapid rise of big data and emerging technologies has increased the complexity of handling and preserving large volumes of information.

This study therefore examines the major challenges associated with data curation, their effects on organizations and research activities, and possible solutions for improving data management practices in the digital age.

2. Literature Review

2.1 Conceptualizing Data Curation

The concept of data curation has evolved significantly over the past two decades, moving from a largely archival practice to a multidisciplinary enterprise encompassing information science, computer science, and domain-specific knowledge management. Yakei [2] defined digital curation as a field of practice focused on the active management and appraisal of digital objects, emphasizing that curation extends well beyond mere storage to encompass the processes of selection, transformation, and long-term stewardship of data assets. Building on this foundation, the Digital Curation Centre articulates data curation as maintaining, preserving, and adding value to digital research data throughout its lifecycle.

Borgman [3] offers a broader theoretical lens, arguing that data does not speak for itself; its value is contingent on the social, institutional, and technical contexts in which it is embedded. This perspective foregrounds the human and organizational dimensions of curation, highlighting that technical infrastructure alone is insufficient to ensure data quality and usability. Researchers and practitioners must therefore attend to governance structures, metadata standards, and community norms alongside tool-based solutions.

2.2 The Data Curation Lifecycle

Several lifecycle models have been proposed to guide data curation practice. The DCC Curation Lifecycle Model [4] is among the most widely cited, presenting curation as a series of sequential and iterative actions—from conceptualization and creation through access, use, appraisal, reappraisal, disposal, ingest, preservation, storage, and transformation. This model underscores that curation is not a one-time activity but an ongoing commitment requiring coordinated policies, resources, and expertise at every stage.

More recent literature has extended this lifecycle thinking to accommodate big data environments and cloud-based repositories. Tenopir [5] found that while researchers broadly recognize the importance of data sharing and preservation, institutional support structures frequently lag behind the aspirations of individual scientists, creating practical gaps in curation practice. These gaps manifest in uneven metadata quality, fragmented repository systems, and ad hoc approaches to data documentation.

2.3 Identified Challenges in the Literature

The literature consistently identifies a cluster of interrelated challenges that impede effective data curation. Corti [6] catalogue barriers including the heterogeneity of data formats, the absence of discipline-wide metadata standards, and the tension between open-access mandates and data privacy obligations. Similarly, Wallis, Rolker, and Borgman [7] highlight the challenge of incentivizing researchers to invest time in data documentation and curation tasks that yield limited immediate professional reward. Data quality emerges as a particularly persistent concern. Pipino, Lee, and Wang [8] argue that data quality is multidimensional, encompassing accuracy, completeness, consistency, timeliness, and accessibility, and that organizations frequently lack systematic frameworks for assessing and improving these dimensions. More recently, machine learning and AI-driven analytics have intensified the demand for high-quality training datasets, making poor curation practices an obstacle not only to archival integrity but also to downstream analytical validity [9]. Security and privacy challenges have grown more salient in the era of data protection legislation. Ohm [10] and subsequent scholars have argued that anonymization techniques frequently offer weaker privacy guarantees than commonly assumed, complicating the curation of sensitive datasets. The enactment of the General

Data Protection Regulation (GDPR) in the European Union and analogous frameworks globally has imposed new compliance obligations on data custodians, raising the stakes for inadequate curation practice[11].

3. Methodology

This study adopted the explanatory research method. The explanatory research design was considered suitable because it enabled the researchers to critically explain and analyze the various challenges facing data curation in Nigeria and beyond. The study relied mainly on secondary sources of data collection, such as information obtained from textbooks, scholarly journals, conference papers, and online publications on data curation, its challenges and the related research topics. [12][13] The explanatory research method enabled the authors to critically identify the major challenges facing data curation in Nigeria and beyond. Explanatory research identifies root causes instead of symptoms, as it tests whether these outcomes result from funding gaps, lack of training, weak policy, or outdated technology. This causal clarity helps prioritise interventions by revealing which factors have the strongest impact on regarding the success and failure of data curation.

4. Results and Discussion

The following are some challenges of Data Curation.

4.1 Data Quality and Integrity

Maintaining high-quality data is one of the most persistent and consequential challenges in data curation. Data quality is multidimensional, encompassing accuracy (the degree to which data correctly represents real-world values), completeness (whether all required data elements are present), consistency (the absence of contradictory values across datasets or systems), and timeliness (whether data is sufficiently current for its intended use). When any of these dimensions is compromised, the reliability of research findings, managerial decisions, and public records may be seriously undermined.

Common sources of data quality degradation include human error during data entry, format inconsistencies arising from data migration across systems, incomplete documentation by original data creators, and the accumulation of duplicate records over time. In large organizations and research institutions, these issues are compounded by the distributed nature of data production, where multiple teams or departments generate data independently without shared standards or validation protocols. The result is often a heterogeneous and difficult-to-reconcile data landscape that curators must laboriously clean, harmonize, and document before the data can be reliably used or preserved.

4.2 Technological Infrastructure and Format Obsolescence

Rapid technological change poses significant challenges for the long-term preservation of digital data. As software platforms, file formats, and hardware systems evolve, data encoded in older formats may become inaccessible without specialized tools or migration procedures. This problem—commonly referred to as format obsolescence—is particularly acute for complex file types such as proprietary databases, multimedia files, and software-dependent datasets that cannot be straightforwardly converted to open or archival formats without loss of functionality or fidelity.

Inadequate storage infrastructure compounds these challenges. Many organizations, particularly those in resource-constrained environments, rely on aging hardware, insufficient backup systems, or fragmented storage architectures that lack the redundancy and geographic distribution necessary for reliable long-term preservation. The emergence of cloud storage has offered new possibilities for scalable and cost-effective data management, but cloud migration introduces its own challenges, including vendor lock-in, data sovereignty concerns, network dependency, and the need for ongoing financial commitment to subscription-based services.

4.3 Standardization and Interoperability

The absence of widely adopted metadata standards and data documentation practices creates significant barriers to data discovery, exchange, and reuse. Metadata—the descriptive, structural, and administrative information that gives context to raw data—is essential for enabling users to understand what a dataset contains, how it was collected, under what conditions it can be accessed, and how it relates to other datasets [14]. Without consistent and comprehensive metadata, data assets become difficult to locate, interpret, and integrate with other sources.

Interoperability challenges arise when data systems employ incompatible schemas, controlled vocabularies, or technical formats, making it difficult to aggregate or compare data across institutional or disciplinary boundaries. This is a particular challenge in multidisciplinary research environments, where investigators from different fields may use domain-specific terminologies and data structures that are difficult to reconcile. Efforts to develop cross-domain metadata standards—such as the Dublin Core Metadata Initiative, the Data Documentation Initiative (DDI), and the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable)—have made important progress, but adoption remains uneven across sectors and institutions[15].

4.4 Data Security and Privacy

Protecting the security and privacy of curated data is a critical challenge, particularly for datasets containing sensitive personal, medical, financial, or governmental information. Data breaches, unauthorized access, and inadequate access control mechanisms can expose individuals and institutions to significant legal, reputational, and financial risks. Curators must therefore implement robust security measures—including encryption, access authentication, audit logging, and network security protocols—while simultaneously ensuring that legitimate users can access and use the data they need.

Privacy challenges are further complicated by evolving legal and regulatory frameworks. The enactment of the General Data Protection Regulation (GDPR) in the European Union, Nigeria's Nigeria Data Protection Regulation (NDPR), and analogous instruments globally has imposed substantial compliance obligations on organizations that collect, process, or preserve personal data [16]. Data curators must navigate these requirements carefully, balancing the imperative of data openness and sharing—increasingly mandated by research funders and government policy—against the obligation to protect individuals' privacy rights. Anonymization and de-identification techniques offer partial solutions but are not infallible, and the residual risk of re-identification in richly attributed datasets remains a significant concern.

4.5 Resource and Capacity Constraints

Effective data curation requires sustained investment in financial resources, skilled personnel, and institutional infrastructure. Many organizations, particularly universities, public sector agencies, and non-governmental organizations in developing countries, lack the dedicated funding streams necessary to support professional curation staff, invest in appropriate storage and software systems, or maintain curation activities over the long term. In research institutions, data curation is frequently treated as an administrative afterthought rather than a core component of the research process, resulting in inadequate documentation, poor data archiving practices, and high rates of data loss following the conclusion of funded projects.

The shortage of trained data curation professionals is a related and equally pressing challenge. Data curators require a distinctive combination of skills spanning information science, domain-specific subject knowledge, data management technology, and legal and ethical literacy. Academic programs offering systematic training in these competencies remain limited relative to the growing demand, particularly in sub-Saharan Africa and other regions where digital research infrastructure is expanding rapidly. Without investment in capacity development at both individual and institutional levels,

organizations will continue to struggle to implement the data governance policies and technical systems necessary for effective curation.

5. Result and Discussion

The challenges identified above are not independent phenomena but form an interconnected system in which weaknesses in one area reinforce difficulties in others. Poor data quality, for instance, is both a cause and a consequence of inadequate metadata standards: without robust documentation practices, curators cannot effectively identify and remediate quality problems, while poor quality data undermines the credibility of the metadata that describes it. Similarly, resource constraints limit the ability of organizations to invest in the technological infrastructure and professional expertise needed to address quality, security, and interoperability challenges, creating a cycle of underfunding that is difficult to break without deliberate institutional commitment.

Several intersecting solutions emerge from the literature. The adoption of formal data governance frameworks—including clearly defined roles and responsibilities for data stewardship, documented data management plans, and institution-wide policies for data quality assessment—has been identified as a foundational intervention that enables more systematic approaches to curation [17]. The implementation of the FAIR Data Principles provides a practical framework for organizations seeking to improve the findability, accessibility, interoperability, and reusability of their data assets, and has been increasingly adopted as a benchmark by research funders and institutional repositories.

Investment in staff training and capacity development is equally essential. Organizations should prioritize the development of in-house data curation expertise through professional development programs, collaboration with library and information science schools, and participation in national and international networks for research data management. In the Nigerian context, the growth of university library systems and the emergence of national data repositories offer institutional anchors for capacity development efforts, but these initiatives require sustained funding and policy support from government agencies and academic leadership.

Technological solutions, including automated metadata generation tools, AI-assisted data quality monitoring, and cloud-based preservation platforms, offer significant potential for reducing the labour intensity of curation at scale. However, these tools must be implemented within a broader governance framework that addresses the organizational and human dimensions of the curation challenge; technology alone cannot substitute for institutional commitment, clear accountability structures, and sustained investment in skilled personnel.

6. Conclusion

This paper has examined the major challenges confronting data curation practice in the contemporary digital environment. Drawing on a systematic review of scholarly and practitioner literature, the study identified five principal categories of challenge: data quality and integrity, technological infrastructure and format obsolescence, standardization and interoperability, data security and privacy, and resource and capacity constraints. The analysis reveals that these challenges are deeply interrelated and that effective responses require coordinated action across technical, organizational, governance, and capacity dimensions.

The study underscores that data curation is not merely a technical problem amenable to software solutions but a fundamentally social and institutional challenge that requires sustained commitment from organizational leadership, adequate funding, skilled professionals, and enabling policy environments. Institutions that invest in robust data governance frameworks, staff training, and standards adoption are better positioned to

realize the long-term value of their data assets and to contribute meaningfully to the broader ecosystem of open and reproducible research.

Looking ahead, several areas merit further investigation. Primary empirical research—including surveys of data management practice in Nigerian universities and public institutions, case studies of successful curation initiatives in low-resource settings, and longitudinal studies of the impact of data governance interventions—would substantially enrich the evidence base available to practitioners and policymakers. As data continues to grow in volume, variety, and velocity, the imperative to address data curation challenges with urgency and intentionality will only intensify.

REFERENCES

- [1] Digital Curation Centre, “What is digital curation?,” 2024. [Online]. Available: <https://www.dcc.ac.uk/digital-curation/what-digital-curation>
- [2] E. Yakei, “Digital curation,” *OCLC Systems & Services*, vol. 23, no. 4, pp. 335–340, 2007.
- [3] C. L. Borgman, **Big Data, Little Data, No Data: Scholarship in the Networked World**. Cambridge, MA, USA: MIT Press, 2015.
- [4] S. Higgins, “The DCC curation lifecycle model,” *International Journal of Digital Curation*, vol. 3, no. 1, pp. 134–140, 2008.
- [5] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, “Data sharing by scientists: Practices and perceptions,” *PLOS ONE*, vol. 6, no. 6, Art. no. e21101, 2011.
- [6] L. Corti, V. Van den Eynden, L. Bishop, and M. Woollard, **Managing and Sharing Research Data: A Guide to Good Practice**. London, U.K.: SAGE Publications, 2014.
- [7] J. C. Wallis, C. Rolker, C. L. Borgman, M. S. Mayernik, and A. Pepe, “The challenges of data management infrastructure in research,” *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–10, 2010.
- [8] L. L. Pipino, Y. W. Lee, and R. Y. Wang, “Data quality assessment,” *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.
- [9] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [10] P. Ohm, “Broken promises of privacy: Responding to the surprising failure of anonymization,” *UCLA Law Review*, vol. 57, no. 6, pp. 1701–1777, 2010.
- [11] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.
- [12] European Commission, **General Data Protection Regulation (GDPR): Regulation (EU) 2016/679**. Official Journal of the European Union, 2018.
- [13] R. J. Torraco, “Writing integrative literature reviews: Guidelines and examples,” *Human Resource Development Review*, vol. 4, no. 3, pp. 356–367, 2005.
- [14] B. Lavoie and R. Gartner, **Preservation Metadata**, 2nd ed. DPC Technology Watch Report. York, U.K.: Digital Preservation Coalition, 2013.
- [15] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, *et al.*, “The FAIR guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, Art. no. 160018, 2016.
- [16] National Information Technology Development Agency (NITDA), **Nigeria Data Protection Regulation (NDPR)**. Abuja, Nigeria: Federal Republic of Nigeria, 2019.
- [17] P. P. Tallon, “Corporate governance of big data: Perspectives on value, risk, and cost,” *Computer*, vol. 46, no. 6, pp. 32–38, 2013.