iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** X **Month of publication:** October 2025

DOI: https://doi.org/10.22214/ijraset.2025.74518

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

The Big Data Turn in Libraries: Innovation, Ethics, and Practice

Nandkishor Balu Gosavi[1], Sr. Lect. Sheetal Gosavi[2]
[1]*Department of Library and information Science, Sant Gadge Baba University Amravati, Maharashtra*
[2]*Department of Computer Technology, SNJB's Shri Hiraral Hastimal (Jain Brothers) Polytechnic, Chandwad, Nashik Maharashtra*

Abstract: ***Information is arriving faster and in more varied forms than ever before. Libraries repositories of cultural and research assets now hold both legacy and born-digital collections whose size and complexity position them in the "big data" domain. This paper reviews the role of big data in LIS, highlights opportunities (collection development, discovery, preservation, administrative analytics), outlines key technical, ethical and capacity challenges, and presents recent case studies (2023–2025) from India and major national libraries demonstrating practical approaches and pitfalls. The paper concludes with actionable recommendations for LIS institutions to develop big-data strategies, strengthen metadata practices, build skills, and protect user privacy.***

Keywords: Big data, libraries, Metadata, Digital libraries, NDLI, Data Privacy.

I. INTRODUCTION

Big data denotes datasets whose volume, velocity, and variety exceed the handling capacity of conventional tools, requiring specialized storage, processing, and analytics. Libraries accumulate large volumes of structured and unstructured material (catalog records, digitized texts, images, audio, usage logs, institutional repositories) that are valuable both for users and for organizational planning. Recent advances in distributed computing, cloud services, and open-source analytical tools make many big-data techniques accessible to LIS practitioners; simultaneously, expanding user bases and digitization programs are producing large data volumes that warrant strategic attention.

II. PROBLEM STATEMENT

Although many libraries possess large digital collections and usage logs, they frequently lack coherent strategies, standards, and technical capacity to convert raw data into actionable knowledge. Without intentional metadata practices, analytics pipelines, privacy safeguards, and staff training, libraries risk under-utilizing assets, exposing user privacy, or outsourcing essential functions to vendors without adequate oversight.

III. RESEARCH QUESTIONS

1) What concrete opportunities does big data present for core LIS activities (collection development, discovery, preservation, and service evaluation)?
2) What are the technical, ethical, and organizational barriers preventing libraries especially in India from leveraging big data?
3) How have leading library initiatives (2023–2025) applied big-data tools and what lessons can be drawn?
4) What practical roadmap can LIS institutions adopt to become data-aware and data-driven while protecting privacy?

IV. SCOPE & LIMITATIONS

Scope: Review of literature, market and initiative reports (2020–2025), and documented case studies of national and large research libraries (with emphasis on India, British Library, Library of Congress).

Limitations: This is a review paper. Some institutional usage statistics vary by source and change rapidly; figures used here reflect the latest published values at the time of writing (2023–2025). The review focuses on institutional (library) data applications and does not cover commercial media/platform big-data practices in depth.

V. METHODOLOGY

This review synthesizes: (a) recent institutional reports and blog posts from major libraries and national initiatives (2023–2025); (b) market and national digital statistics relevant to library digitization and internet adoption; and (c) peer-reviewed and gray literature addressing metadata, privacy, and library analytics. Case studies were selected for being recent, well-documented, and representative of diverse strategic responses to big-data challenges.

©IJRASET: All Rights are Reserved

VI. LITERATURE REVIEW

1) Three (and four) Vs: Volume, Variety, Velocity (and often Veracity/Value) remain the conceptual basis for big-data work in LIS.
2) Opportunities: analytics for collection planning; recommendation systems; automated metadata enrichment (NLP, entity extraction); and preservation triage.
3) Challenges: inconsistent metadata, limited IT infrastructure, vendor dependence, staff skill gaps, and privacy/ethical risks when usage data are collected or shared.

VII. RECENT CASE STUDIES (2023–2025)

A. *National Digital Library Of India (Ndli)*

The National Digital Library of India (NDLI) has rapidly expanded its resources and user base in recent years, emerging as a cornerstone of digital education infrastructure. As of 2025, NDLI hosts over 125 million learning resources and serves more than 94 million registered users across the country. Its multilingual interface supports search in 14 major Indian languages, enhancing accessibility and inclusivity. These vast user and content volumes present a unique opportunity to apply advanced analytics for personalization, demand-driven digitization, and educational service design. NDLI's federated search architecture, built on open-source technologies, efficiently handles high query loads, while its tailored educational modules support diverse learner segments from school and college students to job aspirants and researchers. The platform's success illustrates that centralized, government-backed digital libraries can rapidly achieve scale, enabling the deployment of recommendation systems and usage analytics. However, to be truly effective, such systems require robust metadata harmonization and thoughtful federated search design.

B. *British Library*

The British Library has actively embraced AI and machine learning (ML) in its digital transformation efforts, as reflected in its 2024–2025 year-end reviews and project outputs. Notable initiatives include the *Recovered Pages* crowdsourcing project, which engaged the public in reconstructing lost web content via the Internet Archive's Wayback Machine, contributing to the Digital Scholarship Training Programme. The library has experimented with ML tools such as Transkribus and eScriptorium for automated text recognition across diverse scripts, including Arabic and Bengali, enhancing computational description and accessibility of digitized collections. It also partnered with Sheffield University on the FRAIM project (Framing Responsible AI Implementation & Management), helping shape its AI Strategy and Ethical Guide. These efforts demonstrate how national libraries can pilot ML applications, crowdsource labeled data, and build internal capacity through targeted training programs and collaborative research. The British Library's participation in the AI4LAM community and its hosting of the *Fantastic Futures 2025* conference further underscore its commitment to advancing responsible AI in cultural heritage sectors.

C. *Library of Congress (LOC)*

The Library of Congress (LOC) has undertaken significant initiatives to integrate AI into its cataloging and metadata workflows, exemplified by its *Exploring Computational Description* experiments. These include projects such as generating MARC records from eBooks and visualizing historical ship logs, which support catalogers in metadata creation and enhance access to complex archival materials. LOC's 2023–2027 Digitization Strategy marks a formal institutional shift toward data-driven accessibility, aiming to systematically digitize rare, distinctive, and rights-restricted materials. As of 2025, the Library has digitized over 9 million items, with particular strengths in newspapers, manuscripts, and pictorial collections. The opening of a new Digital Scan Center in 2021 significantly increased throughput and postproduction capabilities. LOC's strategy also emphasizes open access via APIs and technical documentation, enabling external researchers to build tools, conduct analysis, and contribute to a broader digital ecosystem. These efforts illustrate how national libraries can align digitization with research outputs and infrastructure modernization to maximize public value.

D. *Privacy Incident Context*

The Adobe Digital Editions telemetry incident, widely reported in 2014, remains a cautionary example of the privacy risks posed by unsecured vendor software. Investigations revealed that Adobe Digital Editions version 4.0 was transmitting detailed user data including book titles, publishers, pages read, and reading timestamps in plain text over unsecured HTTP to Adobe servers2. This data collection extended beyond books opened in the application, scanning metadata from all EPUB files stored on connected devices and uploading it without encryption or user consent.

The breach sparked widespread concern across library and privacy communities, prompting Adobe to release version 4.0., which introduced HTTPS encryption and limited telemetry to DRM-protected content1. The incident underscores the critical need for libraries to audit vendor telemetry practices, enforce contractual privacy guarantees, and demand encrypted data transmission, minimal retention policies, and transparent user consent mechanisms before adopting third-party platforms.

VIII. FINDINGS & DISCUSSION

A. *Opportunities*

The integration of advanced data analytics and machine learning presents significant opportunities for enhancing library operations and services. Collection development and budget optimization can be achieved through usage analytics and demand prediction, enabling more strategic acquisitions and effective de-duplication. Improved discovery and user engagement are facilitated by natural language processing (NLP) and recommender systems, which enhance search relevance and service personalization. Preservation efforts can be prioritized using metadata on usage and material condition, guiding decisions on digitization and conservation. Furthermore, operational efficiency is strengthened through the analysis of workflows, circulation patterns, and resource utilization, informing staff deployment and space management strategies.

B. *Key Barriers*

Libraries face several critical challenges in leveraging data-driven innovations. Metadata fragmentation, stemming from heterogeneous schemas, hinders effective analytics across collections and complicates interoperability. Infrastructure limitations and insufficient funding further constrain the adoption of big-data storage and processing capabilities, which are essential for modern library operations. A pronounced skills gap exists, as Library and Information Science (LIS) curricula often lack comprehensive data science training, and opportunities for continuing professional development remain scarce. Moreover, the collection and analysis of user interaction data raise significant privacy and ethical concerns, necessitating the implementation of privacy-by-design principles and transparent governance policies. Finally, reliance on third-party platforms introduces vendor-related risks, particularly around telemetry and data governance, underscoring the need for explicit contractual safeguards to protect institutional and user interests.

IX. RECOMMENDATIONS

To harness the potential of big data in libraries, institutions must develop a comprehensive strategy encompassing policy formulation for data collection, metadata standards, storage, access, retention, and privacy. Standardizing and enriching metadata through frameworks like Dublin Core, BIBFRAME, and schema.org, alongside automated techniques such as Named Entity Recognition (NER) and OCR correction, enhances data quality and interoperability. Infrastructure should be built incrementally, starting with cloud-based storage and serverless analytics or scalable distributed tools like Hadoop or Spark, with consortia models offering cost-effective solutions for smaller libraries. Investing in staff training—covering data science fundamentals, SQL, Python, and ethical data practices—and collaborating with academic data science departments is essential to bridge the skills gap. Privacy protection and vendor governance must be prioritized through negotiated telemetry limits, robust encryption, clear service-level agreements (SLAs), and transparent user consent and anonymization protocols. Pilot initiatives, such as recommendation engines, digitization triage, and cataloging assistance, should be launched and scaled based on documented outcomes. Finally, fostering collaboration through federated search systems, shared repositories, and open APIs will enable cross-institutional analytics and expand the research utility of library collections.

X. CONCLUSION

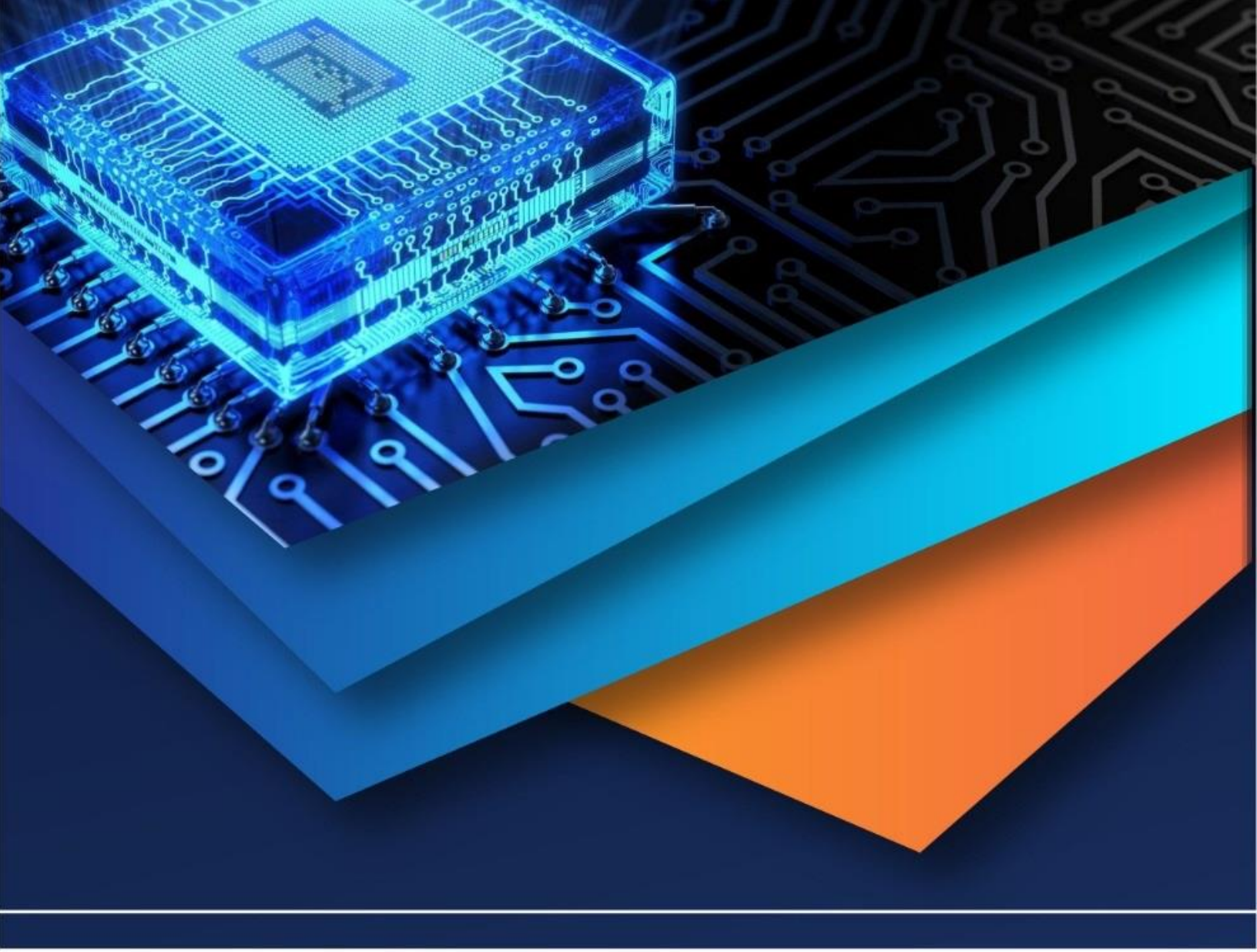
Libraries are well positioned to harness big data for improving collections, discovery, preservation, and institutional decision-making. Recent initiatives (NDLI, British Library, Library of Congress) demonstrate practical ways to combine digitization, AI/ML pilots, crowdsourcing, and policy to generate value. However, technical infrastructure, metadata standardization, staff capability, and privacy protections are essential preconditions for success. A staged, policy-backed approach beginning with pilots and extending to institutional strategies—offers the safest and most effective path for LIS institutions to become data-driven while protecting users and collections.

REFERENCES

[1] British Library. (2025). *AI and machine learning projects*. Retrieved from https://www.bl.uk/research
[2] Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. META Group Research Note.
[3] Library of Congress. (2024). *Digitization strategy 2023–2027*. Retrieved from https://loc.gov
[4] Mashey, J. (1998). *Big data and the next wave of infraStress*. Usenix Conference.
[5] National Digital Library of India. (2025). *NDLI milestones*. Retrieved from https://ndl.iitkgp.ac.in

[6] Nair, S. (2023). Infrastructure challenges in Indian libraries. *Journal of LIS Development, 39*(2), 45–58.
[7] Shiri, A. (2022). *Discoverability and personalization in digital libraries.* Springer.
[8] Tripathi, A. (2025). Data science training needs in LIS education. *Library Trends, 73*(3), 321–338.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)