Asian Journal of Information Science and Technology
ISSN: 2231-6108 (P); 3049-2661 (E)
Vol.15 No.2, 2025, pp.62-66

www.crijournals.org
DOI: https://doi.org/10.70112/ajist-2025.15.2.4393

Transforming Library Services with Metadata Lakes: Leveraging Big Data and AI for Enhanced Resource Management

Arti Sawale* **and Paramjeet Kaur Walia**
Department of Library and Information Science, University of Delhi, Delhi, India
E-mail: pkwalia2002@gmail.com
*Corresponding Author: artisawale@gmail.com
(Received 3 November 2025; Revised 24 November 2025; Accepted 3 December 2025; Available online 10 December 2025)

Abstract **- This study explores the transformative role of metadata in the management of library resources and services, particularly within the context of big data. Metadata enhances data discoverability, indexing, and automation, which, in turn, improves searchability, analytics, and decision-making processes. Libraries are facing an explosion of big data, diverse formats, and evolving user expectations, necessitating a shift in how metadata is managed. In response, this paper proposes a theoretical framework for implementing a "metadata lake" within libraries operating in a big data environment. The concept of a metadata lake is a novel approach to metadata management, and this study aligns it with the characteristics of big data to demonstrate its potential in library systems. Metadata management in libraries has evolved to fit into the big data ecosystem, creating opportunities for knowledge discovery, AI-driven automation, and predictive analytics. By adopting big data technologies, libraries can enhance access, efficiency, data transfer and research impact. The metadata lake concept offers scalability, AI-driven enrichment, and interoperability, allowing libraries to leverage big data and semantic technologies to improve the storage, processing, and retrieval of metadata. The integration of AI-driven recommendations will further enhance content discovery and user engagement. This paper highlights the importance of adopting advanced analytics, linked open data, and semantic technologies in library metadata management. It presents the metadata lake as the future of library metadata, ensuring libraries remain adaptable and capable of contributing to the big data landscape. Furthermore, future research could explore the use of machine learning (ML) and natural language processing (NLP) to automate metadata creation, enrich bibliographic records, and improve classification accuracy, ensuring that libraries continue to evolve in line with technological advancements.**
Keywords: **Metadata Management, Big Data, Metadata Lake, Library Systems, Artificial Intelligence (AI)**

I. INTRODUCTION

Library metadata has traditionally been used for organizing and managing library resources. Libraries are now facing significant challenges in metadata creation and management due to the explosion of big data, diverse formats, evolving standards, and increasing user expectations. Therefore, libraries need to evolve their metadata management strategies to handle vast and diverse datasets efficiently. Libraries depend on metadata for content management (Boukraa *et al.,* 2024). Along with traditional library metadata management (MARC, Dublin Core, BIBFRAME), there is a distinct need to meet the extensive demands of digital libraries, research repositories, and large-scale knowledge systems. While traditional library metadata management and processing might not always meet the extreme scale of big data in domains like social media or IoT, modern digital library systems and global cataloguing networks increasingly embody big data characteristics. This paper introduces the concept of leveraging a "metadata lake" in libraries. While the exact origin of the metadata lake is not well-documented, it has gained attention in libraries as a centralized repository to manage the increasing volume and complexity of metadata generated by modern and diverse data systems and sources. Metadata lakes are evolving into artificial intelligence (AI)-generated catalogues that manage metadata and integrate with AI and machine learning workflows (Fendy Feng & Shri Varsheni R, n.d.). The technical architecture framework introduced in this paper provides a detailed breakdown of layers, including technology choices and workflows, which certainly improves metadata management for a big data environment. This paper highlights that by leveraging big data tools, AI, and semantic technologies, libraries can ensure efficient metadata management for better searchability, interoperability, and user experience.

II. LITERATURE REVIEW

Rousidis *et al.,* (2014) emphasized the importance of metadata for the long-term sustainability of research data repositories and explained the data quality problems associated with metadata. Similarly, Löffler *et al.,* (2021) highlighted that existing metadata currently poorly reflects information needs and leads to poor retrieval of relevant data from a large set of heterogeneous data. To overcome this scenario, the usage of big data, deep learning, and natural language processing (NLP) types of artificial intelligence (AI) technologies is increasing. So far, libraries are positively approaching the implementation of these technologies for automatic text understanding and metadata creation (Wang *et al.,* 2023). Sawadogo *et al.,* (2019) proposed architecture for textual metadata management using text mining and information retrieval techniques to extract, store, and reuse metadata. Attanasio (2022) discussed the awareness of the importance of quality and richness of descriptive metadata. Oladokun & Gaitanou (2024) carried out research on shifting linked open data to library metadata for reformatting and reusing large and complex data. The study by Boukraa *et al.,* (2024) provided valuable insights to the digital library

community, offering them a technological outlook on metadata management**.**

III. OBJECTIVES

1. To introduce the leveraging of a metadata lake in libraries.
2. To understand libraries' metadata alignments with Big Data characteristics.
3. To propose a theoretical architecture for implementing a metadata lake in libraries in a Big Data environment.

A. Metadata Lake in Libraries

A metadata lake is an evolution of traditional metadata management, designed to handle big data scalability, AI-driven enrichment, and advanced analytics. While traditional metadata is structured metadata stored in fixed schemas for organizing and managing data, a metadata lake is a centralized repository (Himpe C., 2024) that stores, manages, and processes large volumes of metadata from various sources in a scalable big data environment. A metadata lake is a big data-powered, scalable repository that ingests, processes, and analyzes metadata from multiple sources. Adoption of a metadata lake in the library field is gaining traction as library centers seek to manage and integrate diverse metadata sources more effectively. It allows libraries to store raw metadata, catalogue records, digital archives, research data, and multimedia resources without predefined schemas. Unlike traditional databases, a metadata lake can handle structured, semi-structured, and unstructured metadata from various sources, making it ideal for libraries and information centers in the era of big data. Table I lists the functional aspects of traditional metadata and metadata lakes.

TABLE I DETAILED COMPARISON OF FUNCTIONS OF TRADITIONAL META DATA AND META DATA LAKE

Functions	Traditional Metadata	Metadata lake
Predefined Schemas	Yes	No
Datatype	Structured	Structured, Semi-structured/Unstructured
Scalability	Limited	Highly
Storage & analysis	In relational databases like MSSQL, MYSQL catalogues, or meta data repositories with strict schema constraints.	Big data storage like Hadoop, AWSS3, Google Cloud, Google Big Query environment, No SQL databases, and distributed file systems.
Purpose	Data discovery, indexing, classification, and interoperability	Scalable metadata storage, AI-driven enrichment, and big data analytics for large and diverse metadata sets
Processing Approach	Manual or rule-based processing with pre-defined metadata standards.	Uses big data frame works (Apache Spark, Kafka, AI, NLP) for real-time metadata processing and enrichment and intelligent discovery.
Data Sources	Primarily structured metadata from books, journals, and research databases.	Ingests metadata from structured, semi- structured, and unstructured sources (e.g., user-generated tags, AI-generated metadata, linked data).
Metadata Enrichment	Manual tagging and classification by librarians and metadata specialists.	Uses AI, NLP, and Machine Learning to automatically enrich and classify metadata.
Search &Discovery	Traditional keyword-based search within predefined metadata fields.	Uses semantic search, AI-driven recommendations, and knowledge graphs for intelligent discovery.
Analytics & Insights	Basic analytics for metadata completeness and usage statistics.	Uses big data analytics, AI-driven insights, and meta data quality analysis to track trends, user behaviour, and data gaps.
Use Cases	Library catalogues, digital repositories, institutional repositories, archival systems research databases.	AI-powered metadata management, large- scale metadata analytics, metadata standardization across platforms, smart search & discovery platforms.
Examples	MARC record, metadata tags and Dublin Core metadata for digital assets.	Storing & analyzing millions of metadata records using a big data ecosystem.
Summary	Structured, rule-based, manually curated, and stored in fixed relational databases.	Scalable, AI-powered, bigdata-driven, & optimized for real-time processing, discovery, and insights.

IV. METADATA LAKE ALIGNMENTS WITH BIGDATA CHARACTERISTICS IN LIBRARIES

Metadata often involves interrelations between resources, e.g., citations, references, and related works. Handling multilingual metadata, cross-referencing across databases, and ensuring interoperability add layers of complexity, especially with big data. Library Metadata exhibits the fundamental characteristics of big data, often referred to as the '5 Vs'. Additionally, library metadata can be considered a form of big data under certain conditions, depending on the '5 Vs'-volume, variety, velocity, veracity, and value-explained as follows:

1. *Volume*: Libraries manage vast amounts of metadata for books, journals, articles, digital collections, and more. Modern library systems generate and store massive datasets of metadata records, including bibliographic records, digital repositories, subject classifications, and usage logs.
2. *Variety*: Metadata comes in various formats and describes diverse resources like text, images, videos,

and e-resources. Libraries include structured data (catalogue entries), semi-structured data (Linked Open Data), and unstructured data (annotations, reviews, and full-text indexing).

3. *Velocity:* Digital libraries update their metadata frequently, especially with the increasing shift to online and real-time indexing of new materials. Automated systems continuously ingest data from publishers, aggregators, and other sources. This includes continuous updates in library catalogues, real-time indexing of digital content, and automated metadata generation using AI.
4. *Veracity:* Data quality concerns, such as duplicate records, inconsistent cataloguing, and variations in metadata standards.
5. *Value:* Enhances resource discoverability, supports academic research, enables predictive analytics, and improves library services.

TABLE II EXAMPLES OF LIBRARY META DATA ALIGNMENTS WITH BIG DATA CHARACTERISTICS

Big Data Characteristics	Library Metadata Example
Volume	Millions of bibliographic records in systems like OCLC World Cat.
Variety	MARC records, Dublin Core, subject headings, keywords, user tags, and linked data.
Velocity	Real-time updates in digital library systems for new acquisitions or changes.
Veracity	Need for high accuracy and reliability in metadata for scholarly use.
Value	Metadata enriches discoverability, enhances research, and supports analytics.

V. TECHNICAL ARCHITECTURE FOR IMPLEMENTING METADATA LAKE IN LIBRARIES IN BIG DATA ENVIRONMENT

A. Prerequisites

Implementing a Metadata Lake in libraries requires careful strategic planning, infrastructure, and technology selection. The key prerequisites are mostly related to technical, infrastructural, and organizational aspects. It includes certain parameters such as storage, data ingestion, metadata sources, processing, standardization, indexing, analytics & visualization, access control and security, and expertise.

B. Architecture Overview

A Metadata Lake in a big data environment is designed to handle, process, and analyze large volumes of metadata efficiently. The proposed theoretical architecture for implementing a metadata lake in libraries in a big data environment includes five layers. This architecture ensures that metadata is ingested, stored, processed, indexed, and analyzed using big scalable data technologies. Figure.1 is a detailed breakdown of each layer, including technological choices and workflows.

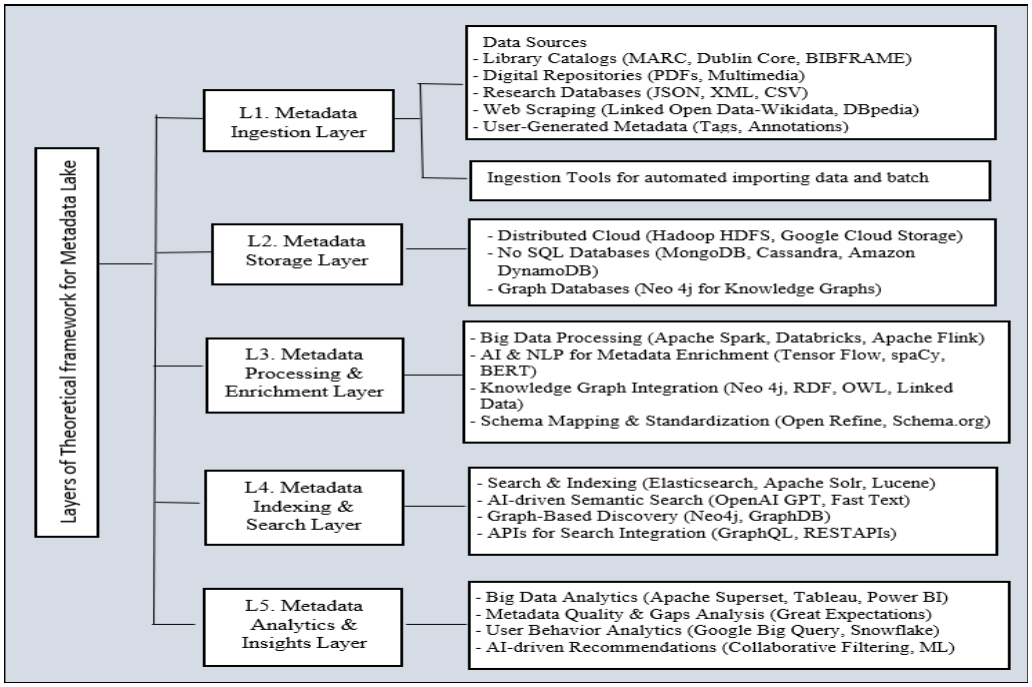

Fig.1 Framework Including Big Data Technology Choices and Workflows

This architecture layout enables scalable metadata management by leveraging distributed cloud storage like Hadoop Distributed File System (HDFS), AI-powered enrichment, and real-time analytics. The key benefits include real-time structured and unstructured metadata ingestion from diverse sources. The flexible and scalable storage can handle massive volumes of data. NLP and AI-driven metadata enrichment for indexing enables fast retrieval and discovery and extracts insights using big data analytics.

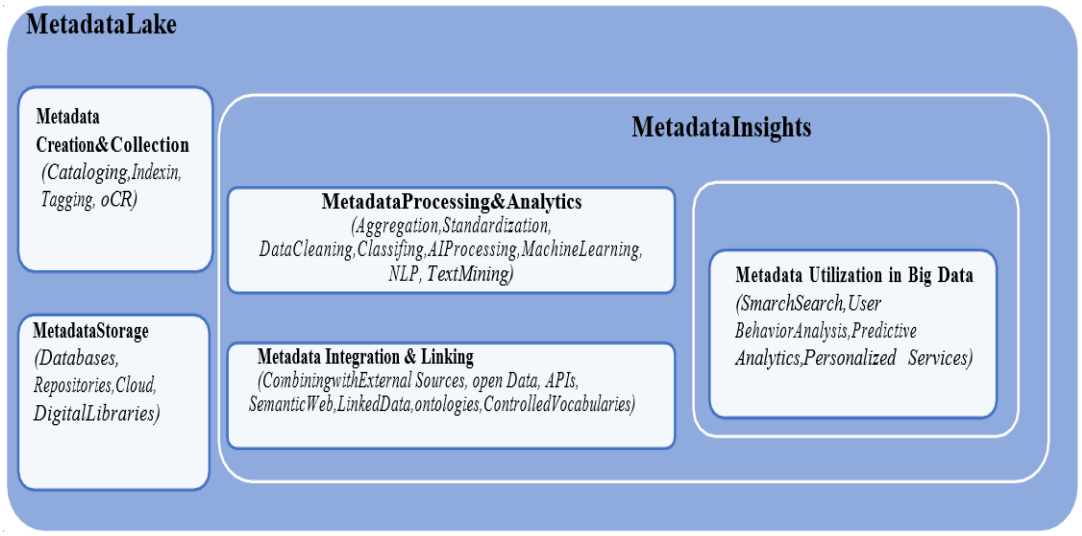

Fig.2 How Big Metadata is Stored, Processed, and Retrieved

Above Figure 2, shows the conceptual technical architecture assists to overcome the existing system struggles with metadata inconsistencies, slow search speeds, and lack of AI-driven recommendations. AI-driven enrichment enhanced metadata accuracy and improve the metadata quality (Oladokun & Gaitanou, 2024) (Tani *et al*., 2013). Big data indexing enabled instant search and faster information retrieval for library users. Link data integration provides the seamless metadata interoperability across platforms.

VI. CONCLUSION

The term "metadata lake" is a relatively recent concept in the field of data management, and its application within libraries is still emerging. A metadata lake is the future of library metadata management, enabling scalability, AI-driven enrichment, interoperability, and intelligent search. Libraries must adopt big data storage, AI metadata processing, and semantic technologies to transform how metadata is stored, processed, retrieved, and support decision-making for acquisitions and resource allocation. Practicing AI-driven library automation and big data analytics to extract insights from metadata can improve reading trends, citation networks, and knowledge graph development. Personalized AI-driven recommendations boost content discovery and increase user engagement. As libraries embrace advanced analytics, linked open data, and interoperability, their metadata not only resembles but also actively contributes to the big data landscape. The term "metadata lake" is a relatively recent concept in the field of data management, and its application within libraries is still emerging.

A metadata lake is the future of library metadata management, enabling scalability, AI-driven enrichment, interoperability, and intelligent search. Libraries must adopt big data storage, AI metadata processing, and semantic technologies to transform how metadata is stored, processed, retrieved, and support decision-making for acquisitions and resource allocation. Practicing AI-driven library automation and big data analytics to extract insights from metadata can improve reading trends, citation networks, and knowledge graph development. Future research can be carried out for a case study on machine learning (ML) and natural language processing (NLP) applications to automate metadata creation and processing, enrich bibliographic records, and improve classification accuracy.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation

The authors confirm that no AI-assisted technologies were used in the preparation or writing of the manuscript, and no images were altered using AI.

ORCID

Arti Sawale https://orcid.org/0000-0001-7070-8272

Paramjeet Kaur Walia https://orcid.org/0000-0003-4174-6311

REFERENCES

Attanasio, P. (2022). New challenges in metadata management between publishers and libraries. *DOAJ: Directory of Open Access Journals*, *13*(1).

Boukraa, D., Bala, M., & Rizzi, S. (2024). Metadata management in data lake environments: A survey. *Journal of Library Metadata*, *24*(4), 215–274. https://www.researchgate.net/publication/382277284_Metadata_Management_in_Data_Lake_Environments_A_Survey.

Feng, F., & Shri Varsheni, R. (n.d.). *How metadata lakes empower next-gen AI/ML applications*. Zilliz. Retrieved January 23, 2025, from https://zilliz.com/blog/how-metadata-lakes-empower-next-gen-ai-ml-apps.

Himpe, C. (2024). DatAasee: A metadata-lake as metadata catalog for a virtual data-lake. [This entry is incomplete. A journal title, conference proceedings, or publisher is needed for a complete citation.]

Löfler, F., Wesp, V., König-Ries, B., & Klan, F. (2021). Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PLOS ONE*, *16*(3), e0246099. https://doi.org/10.1371/journal.pone.0246099.

Oladokun, B. D., & Gaitanou, P. (2024). Leveraging open data for reference services delivery in academic libraries. *Library Hi Tech News*, *41*(4), 12–14. https://doi.org/10.1108/LHTN-07-2023-0112.

Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.-A. (2014). Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. *Information Services & Use*, *34* (3-4), 279–286. https://doi.org/10.3233/ISU-140746.

Sawadogo, P., Kibata, T., & Darmont, J. (2019). Metadata management for textual documents in data lakes. *Proceedings of the 21st International Conference on Enterprise Information Systems*, 72–83. https://www.scitepress.org/Papers/2019/77063/77063.pdf.

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, *49*(6), 1194–1205. https://www.sciencedirect.com/science/article/abs/pii/S0306457313000526.

Wang, Z., Chen, H., Zhang, C., Lu, W., & Wu, J. (2023). JCDL2023 Workshop: Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 303–305. https://ieeexplore.ieee.org/document/10266056.