Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Review Article **Received:** 01-06-2025| **Accepted:** 20-06-2025 | **Published:** 06-07-2025

Data Cataloging with Automated Updates Using Generative AI: Transforming Enterprise Metadata Management

Achyut Kumar Sharma Tandra
The University of Texas at Dallas, USA

Abstract: This article examines the transformative impact of Generative Artificial Intelligence on enterprise data cataloging processes. Data cataloging represents a critical component of modern data management strategies, with traditional approaches facing significant challenges in maintaining accurate documentation across rapidly evolving environments. The integration of Large Language Models (LLMs) introduces unprecedented capabilities for autonomous metadata generation, schema evolution tracking, and semantic relationship inference. These AI-driven systems continuously analyze data structures, automatically document changes, and identify implicit connections between datasets without human intervention. The architectural framework combines sophisticated ingestion mechanisms, semantic analysis engines, and validation frameworks that integrate with existing commercial and open-source catalog platforms. Implementation follows a graduated approach that balances technical considerations with organizational change management. The resulting capabilities deliver substantial business value through improved data discovery, accelerated compliance processes, and enhanced governance frameworks while transforming human roles from documentation creation to strategic oversight and contextual enrichment.

Keywords: Generative AI, Data Cataloging, Metadata Management, Schema Evolution, Semantic Inference.

INTRODUCTION

Data cataloging forms the foundation of effective enterprise data management strategies in today's data-driven business landscape. As organizations accumulate vast repositories of structured and unstructured data across disparate systems, the challenge of maintaining accurate, accessible, and well-documented data assets has grown exponentially. The proliferation of data sources and the increasing complexity of information architectures have created an urgent need for robust metadata management strategies (Metadata. 2023). This growth has rendered traditional cataloging methodologies increasingly inadequate, with manually curated catalogs struggling to remain current in rapidly evolving data environments.

Traditional cataloging approaches rely heavily on manual intervention by data engineers and stewards, who must continuously update schemas, document metadata, and validate data lineage as information architectures evolve. These professionals face mounting challenges in maintaining comprehensive metadata inventories amid the accelerating pace of data acquisition and transformation. Effective metadata management requires establishing consistent processes for documenting, storing, and updating information about data assets throughout their lifecycle (Metadata. 2023). This labor-intensive process creates significant operational overhead, introduces human error risks, and often results in catalog information becoming outdated shortly after documentation. Furthermore, organizations frequently report that their metadata management processes cannot keep pace with schema evolution, creating significant gaps between actual data structures and their documentation.

The advent of Generative Artificial Intelligence (Gen AI) presents a transformative opportunity to reimagine data cataloging processes. By leveraging Large Language Models (LLMs) and other advanced AI techniques, organizations can now automate significant portions of the metadata management lifecycle. Modern data catalogs increasingly incorporate AI capabilities to support comprehensive metadata management, enabling more efficient data discovery and governance (Bigelmaier, F. *et al*., 2025). The integration of AI into data cataloging solutions provides powerful tools for understanding data context and relationships, offering capabilities that extend well beyond traditional documentation approaches. These intelligent systems can analyze data structures, infer relationships between datasets, and autonomously generate comprehensive documentation without human intervention. The result is a dynamic, self-updating catalog that maintains synchronization with evolving data environments in real-time.

This article examines the emergence of AI-powered data cataloging solutions, exploring their technical architecture, implementation methodologies, and tangible business benefits. We

***Corresponding Author:** Achyut Kumar Sharma Tandra

DOI- https://doi.org/10.5281/zenodo.15819138

investigate how these systems integrate with existing enterprise data infrastructures, overcome traditional cataloging challenges, and create new possibilities for data governance and utilization. As organizations increasingly recognize the strategic value of data assets, the importance of maintaining accurate, comprehensive metadata becomes paramount (Metadata. 2023). Additionally, we address critical considerations surrounding security, compliance, and the human-AI collaboration necessary for successful deployment. AI-ready data catalogs are becoming essential components of data management infrastructure, providing enhanced capabilities for data discovery, governance, and utilization across complex enterprise environments (Bigelmaier, F. *et al*., 2025). These advancements enable organizations to extract greater value from their data assets while maintaining robust governance frameworks.

TECHNICAL FOUNDATIONS OF AI-POWERED DATA CATALOGING

Evolution of Data Cataloging Approaches

Data cataloging has progressed through several evolutionary stages, from manual spreadsheet-based inventories to sophisticated metadata management platforms. Metadata management has evolved significantly over the past decades, beginning with basic documentation methods and advancing toward more automated and intelligent systems (Yang, W. *et al*., 2025). The transition from early approaches involved a shift from static documentation to dynamic management systems that attempt to capture both technical specifications and business context. Traditional cataloging systems largely depend on explicit definitions provided by human operators or schema information extracted from databases. While these systems have grown increasingly sophisticated, they remain fundamentally reactive, requiring manual triggers to update catalog information when data structures change, creating a persistent gap between actual data states and their documentation.

Generative AI Architecture for Data Cataloging

Modern AI-powered cataloging solutions employ a multi-layered architecture that combines several AI technologies. This architectural pattern represents a significant advancement in how metadata is managed across enterprise environments. The emerging architecture for LLM applications consists of several key components working together to create a comprehensive metadata management system (Bornstein, M. and Radovanovic, R. 2023). The Data Ingestion Layer continuously monitors data sources for structural changes through API connections and event-driven triggers, capturing modifications as they occur rather than through scheduled scans. The Semantic Analysis Engine utilizes Large Language Models to interpret schema information, column relationships, and implicit data patterns, providing contextual understanding that goes beyond simple pattern matching. The architecture typically includes pre-processing components that prepare inputs for LLMs and post-processing elements that refine outputs for specific metadata applications (Bornstein, M. and Radovanovic, R. 2023). The Metadata Generation Module automatically produces comprehensive documentation, including technical specifications and business context. The Lineage Tracking System maps relationships between datasets and transformations over time, while the Validation and Quality Framework ensures that generated metadata adheres to organizational standards and policies.

Large Language Models for Schema Interpretation

The application of LLMs in data cataloging represents a paradigm shift in metadata management. Modern LLM applications follow a retrieval-augmented generation (RAG) pattern, which enhances the model's capabilities by connecting it with domain-specific knowledge (Bornstein, M. and Radovanovic, R. 2023). This approach is particularly valuable in data cataloging, where contextual understanding of specific data domains is essential for accurate metadata generation. Unlike rule-based systems, these models can infer relationships between tables/collections based on naming conventions and data patterns. The semantic capabilities of LLMs enable them to generate natural language descriptions of complex data structures and identify potential primary/foreign key relationships without explicit declarations. Current architectures employ orchestration layers that coordinate between data sources, LLMs and downstream applications, enabling sophisticated workflows that combine automated inference with human validation (Bornstein, M. and Radovanovic, R. 2023). The models can recognize common data patterns, such as identifying columns containing email addresses or geographic coordinates, and suggest appropriate data classifications based on content analysis. Modern LLMs' sophisticated understanding of context enables them to generate

Publisher: SARC Publisher

accurate metadata even for previously unseen data structures, representing a significant advancement over traditional pattern-matching approaches. This capability proves particularly valuable as data environments grow increasingly complex and heterogeneous.

Table 1: Evolution of Data Cataloging Approaches and Technologies (Yang, W. *ET AL*., 2025; Bornstein, M. and Radovanovic, R. 2023)

Evolution Stage	Key Characteristics
Manual Spreadsheet Inventories	Static documentation, manually maintained
Basic Metadata Repositories	Centralized storage, limited automation
Schema Extraction Systems	Automated technical metadata extraction
AI-Enhanced Cataloging	Dynamic updates, relationship inference
LLM-Powered Semantic Catalogs	Contextual understanding, RAG architecture

AUTOMATED METADATA MANAGEMENT CAPABILITIES

Autonomous Schema Documentation

AI-powered catalogs continuously analyze data assets to generate and maintain comprehensive documentation. This process represents a paradigm shift from traditional manual approaches, introducing sophisticated capabilities that enhance both efficiency and quality. The documentation process employs advanced pattern recognition to identify data types and formats, going beyond basic type detection to understand the semantic meaning of fields (Chintakindhi, S. K. 2025). Natural language descriptions of tables and fields are generated using context-aware models that consider the broader purpose of data structures within organizational systems. The analysis extends to identifying business entities represented by technical data structures, effectively bridging the gap between implementation details and conceptual models. Documentation also encompasses validation rules and constraints, detecting both explicitly declared rules and implicit patterns that govern data quality. Additionally, these systems document contextual relationships between different data elements, creating a comprehensive map of how information flows through the organization. The resulting documentation is presented in human-readable formats while simultaneously maintaining machine-readable metadata that integrates with data processing tools and governance frameworks, ensuring that documentation serves both immediate human needs and automated system requirements (Chintakindhi, S. K. 2025).

Dynamic Schema Evolution Tracking

One of the most challenging aspects of data management is tracking how data structures evolve over time. Traditional documentation approaches quickly become outdated as systems change, creating significant knowledge gaps. AI-powered catalogs address this through continuous monitoring mechanisms that observe data sources for structural modifications (Chintakindhi, S.K. 2025). When changes occur, the system automatically documents field additions, removals, and type changes without requiring manual intervention. Advanced implementations maintain versioning of schema definitions with temporal tracking, preserving historical records of how structures have evolved while maintaining current documentation. This historical perspective enables impact analysis of schema changes on downstream dependencies, identifying reports, applications, and processes that may be affected by structural modifications. Many implementations include notification capabilities that alert relevant stakeholders when significant changes occur, ensuring that teams can proactively address potential issues rather than discovering them through failures. This capability ensures that catalog information remains synchronized with actual data structures regardless of how frequently they change, eliminating the documentation lag that plagues traditional approaches and providing organizations with continuously current metadata (Chintakindhi, S. K. 2025).

Semantic Relationship Inference

Gen AI catalogs excel at identifying semantic relationships that may not be explicitly declared in the data. This capability represents a fundamental advancement in metadata management, uncovering connections that would remain hidden in traditional systems. The semantic layer serves as a critical component in bridging raw data infrastructure with advanced AI applications, providing contextual understanding that enhances both human and machine interaction with data assets (Enterprise Knowledge, 2024). By implementing a comprehensive semantic layer, organizations can substantially improve the effectiveness of large language models working with enterprise data. These systems discover

Publisher: SARC Publisher

implicit foreign key relationships by analyzing naming patterns, value distributions, and usage contexts, identifying connections that lack formal declarations. They recognize dimensional hierarchies in analytical data models, automatically mapping organizational structures, geographic relationships, and temporal patterns within the data. The identification of equivalent entities across different systems proves particularly valuable in complex environments with multiple platforms representing similar concepts, creating unified views despite implementation differences. Detection of derived fields and their source calculations provides critical lineage information, mapping how computed values relate to their origins. Perhaps most significantly, these systems map business concepts to their technical implementations, creating bidirectional translations between business terminology and database structures (Enterprise Knowledge. 2024). By inferring these relationships, AI systems create rich knowledge graphs that represent both the technical and business dimensions of organizational data assets, providing a comprehensive view that far exceeds manual documentation capabilities.

Table 2: Core Capabilities of AI-Powered Metadata Management Systems (Gudala, M. and Koilakonda, R.R. 2024]

Metadata Management Capability	Primary Function
Pattern Recognition	Identifies data types and semantic meaning
Natural Language Generation	Creates context-aware documentation
Temporal Schema Tracking	Maintains version history of structure changes
Impact Analysis	Identifies downstream dependencies affected by changes
Semantic Relationship Inference	Discovers implicit connections between data elements

INTEGRATION WITH ENTERPRISE DATA ECOSYSTEMS

Compatibility with Commercial Catalog Platforms

For practical implementation, AI-powered cataloging solutions must integrate seamlessly with existing enterprise data platforms. This integration represents a critical consideration as organizations have already invested heavily in commercial metadata management solutions. The integration challenge involves connecting AI capabilities with established platforms while preserving existing workflows and security models (Dheeraj, B.K. 2024). Integration with cloud-based catalog services enhances native cataloging capabilities with AI-generated metadata, creating a more comprehensive view of organizational data assets. These enhancements typically focus on augmenting existing technical metadata with semantic context and business relevance, making catalog information more accessible to non-technical users. Similar approaches apply to various data catalog platforms, where AI augmentation focuses on semantic inference capabilities that complement the platform's core functionality. The integration extends metadata services with generative documentation, leveraging sophisticated natural language generation to create human-readable descriptions of technical structures. Enterprise catalog integration supplements structured metadata with AI-inferred relationships, uncovering connections that may not be explicitly declared in the data model. These integrations typically operate through a combination of API connections, event-driven architectures, and metadata exchange standards such as Open Metadata and Governance (OMAG), establishing bidirectional information flow between AI components and existing catalog infrastructure (Dheeraj, B.K. 2024).

Extending Open-Source Catalog Solutions

The open-source data community has embraced AI-powered cataloging through extensions to popular platforms, creating accessible paths to advanced capabilities without enterprise software investments. These extensions represent an important step toward broader philosophical considerations of transparency and knowledge sharing in technical systems (Tyagi, K.M. 2024). Extensions for Hadoop ecosystems focus on AI-enhanced metadata management, integrating machine learning capabilities to improve automatic metadata extraction and relationship inference. These extensions typically leverage existing framework architectures to intercept metadata changes, apply AI enrichment, and update repositories with enhanced information. Other platform extensions emphasize LLM-powered documentation generation and search capabilities, enabling natural language interaction with metadata repositories. This approach significantly improves accessibility for business users without requiring technical knowledge of exact table names or field definitions. The focus on automated

Publisher: SARC Publisher

lineage tracking and relationship inference addresses one of the most challenging aspects of metadata management across complex environments. Generative AI plugins for schema documentation leverage modular architectures to incorporate LLM capabilities without modifying core functionality. These extensions represent an ethical approach to technology democratization, aligning with theories of distributive justice in technical capability access (Tyagi, K. M. 2024).

Hybrid and Multi-Cloud Catalog Orchestration

Modern data environments span multiple platforms, creating significant cataloging challenges that require sophisticated orchestration capabilities. This complexity reflects the reality of contemporary enterprise architectures, where data assets are distributed across on-premises systems, multiple cloud platforms, and SaaS applications (Dheeraj, B. K. 2024). AI-powered solutions address this complexity through unified metadata models that standardize descriptions across disparate systems, creating consistent representations regardless of source platform. Advanced implementations incorporate cross-platform lineage tracking that follows data flows between environments, documenting how information moves throughout its lifecycle. Semantic translation layers harmonize terminologies between different platforms, addressing the challenge of inconsistent naming conventions and taxonomies that typically emerge in heterogeneous environments. These translation capabilities prove particularly valuable in complex organizational contexts, where distinct data cultures may have established different conventions for similar concepts. The orchestration architecture typically includes centralized catalog interfaces with distributed metadata collection, providing unified access while maintaining appropriate connections to source systems (Tyagi, K. M. 2024). This approach balances comprehensive visibility with the technical reality of distributed data environments while addressing epistemological questions about knowledge representation across diverse technical contexts. The implementation also ensures consistent policy enforcement across heterogeneous data landscapes, maintaining governance requirements regardless of where data resides. This orchestration capability provides a coherent view of organizational data assets regardless of physical location, creating a unified metadata layer that abstracts away infrastructure complexity.

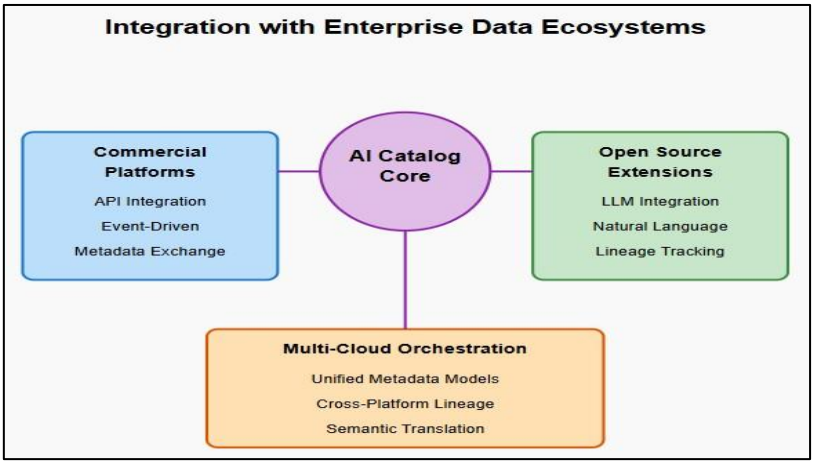

Fig 1: AI-Powered Data Catalog Integration Framework (Dheeraj, B. K. 2024; Tyagi, K. M. 2024)

IMPLEMENTATION STRATEGIES AND ORGANIZATIONAL IMPACT

Phased Deployment Approaches

Successful implementation of AI-powered cataloging typically follows a graduated approach that balances technical complexity with organizational change management. Research on enterprise AI adoption suggests that phased implementation strategies achieve higher success rates compared to comprehensive deployment attempts (Gudala, M. and Koilakonda, R.R. 2024). This incremental methodology addresses the technical complexities inherent in AI-enhanced metadata management while allowing organizations to adapt gradually to new capabilities and workflows. The journey begins with a Discovery Phase focused on automated inventory of existing data assets, establishing baseline metadata coverage across the environment. This initial phase requires minimal change to existing workflows, making it an ideal starting point. The Documentation Enhancement

Publisher: SARC Publisher

phase focuses on adding AI-generated descriptions to technical metadata, enriching structural information with contextual explanations. Organizations then progress to Relationship Mapping, building the knowledge graph of inter-dataset connections that extends beyond explicit relationships declared in schemas. The Lineage Automation phase implements continuous tracking of data transformations, documenting how information flows through various processes and systems. As catalog maturity increases, organizations implement Governance Integration, connecting catalog intelligence to compliance workflows. The final phase typically involves Self-Service Enablement, exposing AI-enhanced catalog capabilities to business users through intuitive interfaces. This incremental approach allows organizations to realize value quickly while building toward comprehensive catalog automation, with each phase delivering tangible benefits while establishing the foundation for subsequent capabilities (Gudala, M. and Koilakonda, R. R. 2024).

Business Benefits and ROI Considerations

Organizations implementing AI-powered cataloging report several measurable benefits that translate directly to business value and return on investment. AI technologies have demonstrated significant impact on business analytics capabilities across various domains and functions (Gudala, M. and Koilakonda, R.R. 2024). The integration of AI into data cataloging processes produces substantial efficiency gains in metadata management and utilization. Organizations experience reduction in manual documentation effort, allowing data professionals to focus on higher-value activities. Implementation research demonstrates improvement in data discovery efficiency, with users locating relevant data assets more quickly when using AI-enhanced catalogs compared to traditional search methods. This improvement stems from the combination of comprehensive metadata, semantic search capabilities, and relationship-based recommendations that guide users to relevant assets. Organizations implementing comprehensive metadata automation report acceleration in regulatory compliance processes, with automated lineage tracking and sensitive data identification reducing documentation effort. Project metrics demonstrate reduction in data integration project timelines, with comprehensive metadata availability streamlining integration planning and mapping activities. Organizations also report measurable decrease in data quality incidents, with comprehensive metadata enabling more effective validation, monitoring, and governance. These benefits translate to quantifiable ROI through reduced operational costs, accelerated time-to-insight, and minimized compliance risks, providing strong justification for investment in AI-enhanced cataloging capabilities (Gudala, M. and Koilakonda, R.R. 2024).

Human-AI Collaboration Models

Effective catalog automation does not eliminate human involvement but transforms it, creating new collaboration models that leverage the complementary strengths of human expertise and artificial intelligence. Human-AI collaboration in enterprise contexts creates opportunities for enhanced decision-making and process optimization when properly structured (Olasehinde, T. 2024). The evolution of roles begins with data stewards shifting from documentation creation to exception handling and validation. Rather than writing descriptions and mapping relationships manually, stewards focus on reviewing AI-generated content, addressing edge cases, and resolving ambiguities that require domain knowledge. Data engineers focus on relationship verification rather than basic schema documentation, allowing them to concentrate on validating complex relationships and reviewing transformation logic. The collaboration extends to business analysts, who contribute domain context to enhance automatically generated descriptions. While AI systems excel at creating technically accurate documentation, business analysts provide critical context about business usage, calculation logic, and organizational terminology. At the governance level, teams establish policies that guide AI-generated metadata standards rather than implementing them manually. This policy-based approach enables consistent application of governance standards at scale, ensuring that automation aligns with organizational requirements. This collaborative model leverages AI for routine, repetitive tasks while harnessing human expertise for contextual understanding and judgment-intensive decisions. The resulting hybrid approach delivers superior outcomes compared to either fully manual or completely automated alternatives, creating sustainable metadata management capabilities that scale with organizational needs (Olasehinde, T. 2024).

Publisher: SARC Publisher

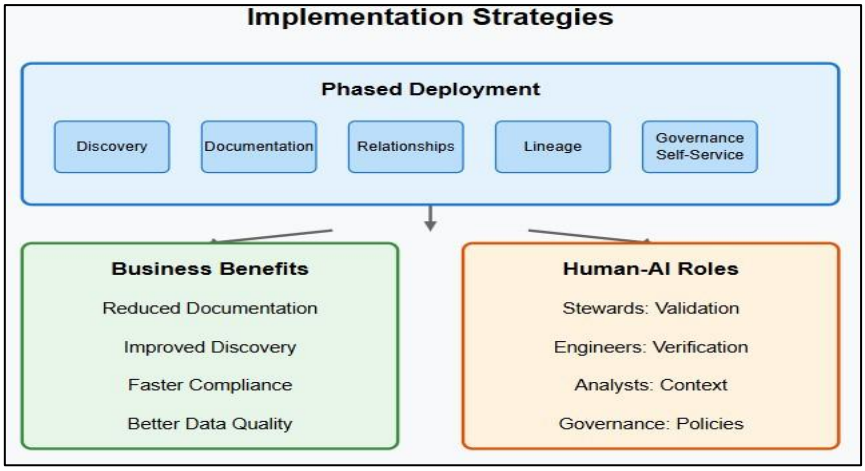

Fig 2: AI Data Catalog Implementation Framework (Gudala, M. and Koilakonda, R.R. 2024); Olasehinde, T. 2024)

CONCLUSION

The integration of Generative AI into data cataloging represents a fundamental shift in how organizations manage and leverage data assets. By automating discovery, documentation, and relationship mapping of data structures, AI-powered catalogs eliminate significant manual effort while dramatically improving metadata accuracy and completeness. The dynamic nature of these systems ensures catalog information remains continuously synchronized with evolving data landscapes, overcoming the persistent challenge of metadata obsolescence that plagues traditional approaches. Beyond operational efficiency, AI-enhanced catalogs drive strategic value through improved data discoverability, enhanced governance capabilities, and accelerated analytical insights, transforming data catalogs from static repositories of technical information into dynamic knowledge graphs capturing both structure and meaning. As enterprise data environments grow increasingly complex and distributed, maintaining a coherent view of data assets becomes critical to competitive advantage. The future of data cataloging lies in this symbiotic relationship between human expertise and artificial intelligence, where AI handles technical metadata while human specialists provide contextual understanding and strategic direction.

REFERENCES

1. Metadata. "5 metadata management best practices." *DataGalaxy.com*, (2023). https://www.datagalaxy.com/en/blog/metadata-management-best-practices/
2. Bigelmaier, F. *et al*. "Is Your Data Catalog Ready for the AI Age?" *BARC* (2025). https://barc.com/data-catalog-ai-ready/
3. Yang, W., Fu, R., Amin, M.B. and Kang, B. "Impact and influence of modern AI in metadata management." *arXiv preprint arXiv:2501.16605* (2025).
4. Khandelwal, A.P. "AI-Driven Mainframe Modernization: Unlocking Legacy Data for Cloud Analytics." *Sarcouncil Journal of Engineering and Computer Sciences* 4.6 (2025): pp 60-67
5. Bornstein, M. and Radovanovic, R. "Emerging Architectures for LLM Applications." *Andreessen Horowitz*, (2023). https://a16z.com/emerging-architectures-for-llm-applications/
6. Chintakindhi, S.K. "AI-Driven Schema Drift Detection: Automating Regulatory Compliance in Cloud Migration Projects." *IJIRMPS*, 13.2 (2025). https://www.ijirmps.org/papers/2025/2/232447.pdf
7. Enterprise Knowledge. "The Role of Semantic Layers with LLMs." *enterprise-knowledge.com*, (2024). https://enterprise-knowledge.com/the-role-of-semantic-layers-with-llms/
8. Joshi, R. " Data-Centric AI: Engineering Platforms for Pre-Model Intelligence." *Sarcouncil Journal of Multidisciplinary* 5.6 (2025): pp 48-54
9. Dheeraj, B.K. "AI In Metadata Management: Trends, Tools, and Transformation." *International Journal of Advanced Research in Education and Technology (IJARETY)*, 11.5 (2024).

Publisher: SARC Publisher

10. Tyagi, K.M. "The Smart Backbone: AI and ML in Enterprise Metadata Management." *International Journal of Multidisciplinary and Scientific Emerging Research* 12.4 (2024). https://philarchive.org/archive/KRITSB-2
Gudala, M. and Koilakonda, R.R. "The Impact of Artificial Intelligence and Machine Learning on Business Analytics." *International Research Journal of Economics and Management Studies IRJEMS* 3.8 (2024).
11. Olasehinde, T. "Human-AI Collaboration in Enterprise Data Analysis." *ResearchGate*, (2024). https://www.researchgate.net/publication/384769214_Human AI_Collaboration_in_Enterprise_Data_Analysis

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Tandra, A. K. S. "Data Cataloging with Automated Updates Using Generative AI: Transforming Enterprise Metadata Management" *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 249-256.

Publisher: SARC Publisher