ARTIFICIAL INTELLIGENCE IN ACADEMIC LIBRARY COLLECTION MANAGEMENT

Varsha D. Ingle
Librarian
Shri Saraswati College of Social Work
Washim

Crossref DOI - https://doi.org/10.5281/zenodo.23416222

Abstract :

The rapid expansion of scholarly publishing, coupled with the emergence of non-traditional digital media and cross-disciplinary research paradigms, has placed unprecedented demands on academic library collection management. Traditional collection development methodologies relying heavily on static approval profiles, manual bibliographer selection, and retrospective circulation metrics are increasingly inadequate for managing contemporary information ecosystems. This paper examines the integration of artificial intelligence , machine learning , natural language processing , and predictive analytics into academic library collection management. We analyze how AI technologies transform core collection workflows across three primary areas: automated, semantic-aware resource selection and discovery; predictive demand modeling and dynamic expenditure management; and technical, ethical, and operational challenges, including algorithmic bias, patron privacy, and vendor transparency. The paper provides a comprehensive synthesis of current methodologies, ethical considerations, and institutional strategies necessary to build sustainable, AI-enhanced academic library collections.

Introduction :

Academic library collection management serves as the foundational architecture supporting institutional research, pedagogy, and scholarly communication. Historically, collection management focused on the systematic identification, acquisition, evaluation, and preservation of scholarly resources. Selectors and subject bibliographers relied on well-established mechanisms, such as manual title-by-title selection, publisher approval plans, firm orders, and consensus-driven serial subscriptions.

However, the modern higher education ecosystem has undergone a profound structural shift driven by three intersecting dynamics:

Information Velocity and Volume : The exponential growth of global scholarly outputs, pre-print repositories, and multi-modal digital assets makes comprehensive manual oversight computationally and operationally unfeasible.

Interdisciplinary Research Paradigms : Contemporary scholarship increasingly operates across traditional disciplinary boundaries, rendering legacy classification frameworks (such as the Library of Congress Classification and Dewey Decimal Classification) prone to misaligning or obscuring multi-domain resources.

Budgetary Constraints and Fiscal Scrutiny : Academic libraries face persistent cost increases for journal subscriptions alongside static or declining institutional budgets, forcing a shift from broad ownership models to precise, usage-aligned access frameworks.

Artificial intelligence encompassing machine learning, deep neural networks, transformer-based natural language processing, and predictive analytics presents transformative capabilities for academic collection management. By replacing manual classification and basic keyword searching with high-dimensional semantic analysis and automated trend forecasting, AI enables libraries to align acquisitions precisely with institutional research activities.

This paper investigates the application of AI technologies across the collection management lifecycle. It evaluates semantic profiling methodologies, predictive demand models, technical integration architectures, and the ethical responsibilities inherent in deploying automated decision systems within higher education libraries.

Semantic Profiling and Automated Selection :

The cornerstone of effective collection management is the accurate evaluation and selection of relevant scholarly materials. Traditional approval profiles rely on publisher-provided metadata, static Subject Headings (such as LCSH), and explicit vendor classification tags. While functional for core disciplinary monographs, these static systems introduce friction when evaluating fast-evolving or interdisciplinary scholarship.

Transformer Models and Dense Vector Embeddings :

Recent developments in transformer-based natural language processing such as BERT, SciBERT, and generalized foundation models have redefined automated document classification. Rather than analyzing metadata strings or discrete keywords, modern NLP architectures convert complete texts, abstracts, or table of contents into high-dimensional vector embeddings.

Within a vector space, mathematical distance metrics (such as cosine similarity) measure the semantic relationship between resources. Consequently, an AI selection system can evaluate an incoming manuscript or monograph and position it accurately within a high-dimensional concept space. This enables automated selection tools to identify relevant works based on conceptual, methodological, or thematic relevance, even when explicit classification codes or standardized keywords are absent.

Knowledge Graphs and Emerging Front Detection :

Complementing dense vector embeddings, dynamic knowledge graphs offer robust frameworks for tracking institutional research trends. Knowledge graphs represent entities (such as authors, research groups, datasets, methodologies, and journals) as nodes connected by semantic relationships.

When integrated into library collection engines, graph neural networks (GNNs) analyze publication outputs, grant funding awards, citation networks, and co-authorship patterns across an institution. By monitoring these graphs, AI systems can automatically detect emerging research fronts and interdisciplinary convergences before they manifest in formal course catalogs or departmental requests. This capability allows subject bibliographers to proactively acquire foundational resources for emerging research areas.

Predictive Analytics, Demand Forecasting, and Budget Optimization :

While Demand-Driven Acquisition (DDA) and Evidence-Based Acquisition (EBA) models have aligned purchasing with user demand, they remain inherently reactive—triggering purchases only after a patron searches for or requests an unowned title. AI introduces *predictive curation*, leveraging machine learning models to forecast demand and optimize budget allocations prior to explicit user requests.

Multi-Source Data Synthesis :

Predictive AI algorithms integrate disparate internal and external data streams to model an institution's information needs:

Institutional Research Data : Active grant awards, faculty publication records, and departmental research agendas extracted from institutional repositories.

Curricular Indicators : Course syllabi, required reading lists, course enrollment trends, and new degree proposals parsed using NLP pipelines.

Historical Access and Fulfillment Patterns : Interlibrary Loan (ILL) request networks, counter-compliant electronic resource usage logs, and zero-result search query logs.

Machine Learning for Circulation and Utilization Forecasting :

By training supervised machine learning models (e.g., gradient-boosted decision trees, random forests, or neural networks) on historical usage logs and contextual resource metadata, collection managers can predict the probability of future utilization for prospective acquisitions.

Predictive scores allow libraries to establish automated purchasing tiers: high-probability materials are acquired automatically via streamlined vendor pipelines, moderate-

probability materials are routed to human bibliographers for review, and low-probability items are marked for access-on-demand networks.

Dynamic Micro-Funding Allocation :

A persistent administrative challenge in collection management is managing shared or interdisciplinary acquisitions across rigid, department-bound budget structures. AI selection systems address this friction through dynamic cost-apportionment algorithms. When a cross-disciplinary resource is identified, the system calculates proportional relevance scores across contributing departments and automatically draws micro-payments from corresponding subject funds, eliminating manual inter-departmental negotiation.

Technical Infrastructure and System Architecture :

Operationalizing AI in academic collection development requires integration between machine learning workflows and core library systems, such as Integrated Library Systems (ILS) and Library Services Platforms (LSPs) like Ex Libris Alma, Index Data FOLIO, or OCLC WorldShare.

Data Pipelines and Interoperability :

To support real-time AI inference, library platforms must maintain robust data ingestion pipelines. Vendor catalog feeds (e.g., ONIX, MARC, JSON-LD) must be ingested, normalized, and processed continuously. Transitioning from legacy MARC records to linked data frameworks (such as BIBFRAME) provides the structural graph infrastructure required for high-dimensional semantic modeling and cross-repository indexing.

Middleware and Decision Engines :

AI-enhanced collection management relies on a modular architecture:

Ingestion Layer : Ingests raw publisher feeds, institutional data, and usage statistics via RESTful APIs and open protocols (e.g., OAI-PMH).

Inference Layer : Runs semantic vector engines and predictive machine learning pipelines to generate relevance and usage scores.

Workflow Automation Layer : Middleware that evaluates generated scores against library-defined policy rules, executing automated API purchasing calls or generating bibliographer task queues.

Audit and Logging Layer : Captures all automated decisions, confidence intervals, and funding deductions to maintain institutional financial transparency.

Ethical Governance, Algorithmic Bias, and Operational Risks :

Deploying artificial intelligence in academic collection management introduces technical efficiencies, but it also raises significant ethical, legal, and operational challenges that require active governance.

Algorithmic Bias and Canonical Reinforcement :

Machine learning models inherit and potentially amplify historical biases present in their training data. In academic publishing, commercial citation indexes, historical circulation records, and mainstream peer-reviewed literature exhibit documented structural biases that favor English-language publications, institutions in the Global North, and established research paradigms.

If an AI selection model relies heavily on historical usage or conventional citation metrics, it risks reproducing these inequalities by:

Marginalizing Non-Centric Scholarship : Systematically scoring works from non-Western scholars, minority perspectives, or small independent presses as low-priority.

Suppressing Heterodox Research : Undervalued novel or non-mainstream theories that lack historical usage records.

Homogenizing Collections : Creating standardized collections across institutions driven by shared commercial training datasets.

Collection administrators must mandate algorithmic audits, establish diversity baseline thresholds, and maintain human-in-the-loop review protocols to ensure collections remain balanced, diverse, and inclusive.

Commercial Vendor Lock-In and Transparency :

Major commercial library vendors and publisher aggregators are increasingly embedding proprietary AI recommendation tools into procurement platforms. This introduces conflicts of interest, as commercial vendors face financial incentives to optimize algorithms toward high-margin proprietary packages or proprietary journal suites.

Libraries must resist "black-box" decision engines by demanding open algorithmic specifications, vendor audit rights, and interoperable data models that preserve institutional autonomy in collection strategies.

Patron Privacy and Analytics Governance :

Predictive AI models require granular usage data to refine demand forecasting. However, capturing detailed student and faculty search histories, reading behaviors, and resource usage conflicts with core librarianship values regarding privacy and intellectual freedom.

Libraries must implement strict privacy-preserving protocols, including:

- Aggregating and anonymizing all behavioral analytics before ingestion into training pipelines.
- Utilizing differential privacy techniques to prevent reverse-engineering of individual user behavior.
- Ensuring local or private cloud processing of user data to prevent unauthorized exposure to third-party commercial vendors.

Institutional Recommendations for Implementation :

Academic libraries seeking to integrate AI into collection management should adopt a structured, phased framework:

Modernize Metadata Infrastructure : Transition legacy catalog systems toward linked data standards (BIBFRAME/RDF) and establish clean, anonymized historical usage data repositories.

Implement Human-in-the-Loop Selection : Deploy AI tools initially as decision-support systems for subject liaisons rather than fully autonomous purchasing engines, allowing bibliographers to calibrate and validate model predictions.

Establish Algorithmic Governance Committees : Create cross-functional governance teams comprising collection strategists, metadata specialists, data ethicists, and faculty representatives to audit AI selection metrics regularly.

Demand Open and Explainable Vendor Systems : Include explicit requirements for algorithmic transparency, explainability, and bias mitigation in vendor procurement contracts.

Invest in Liaison Skill Development : Provide professional development opportunities for subject liaisons in data literacy, prompt engineering, algorithmic auditing, and computational collection analysis.

Conclusion :

The integration of artificial intelligence into academic library collection management represents a major shift in how higher education institutions curate, acquire, and provide access to scholarly information. By harnessing transformer-based natural language processing, vector similarity modeling, dynamic knowledge graphs, and predictive analytics, academic libraries can build responsive, semantic-aware collections that align with modern research environments.

However, the adoption of AI is not merely a technical upgrade; it requires thoughtful ethical and operational governance. Academic libraries must actively manage risks related to algorithmic bias, vendor lock-in, and patron privacy. By committing to open standards,

explainable algorithmic models, and human-guided decision frameworks, libraries can leverage artificial intelligence to enhance collection management while upholding core values of diversity, intellectual freedom, and equitable access to knowledge.

References :

- American Library Association. (2019). *Core values of librarianship*. American Library Association. https://www.ala.org/advocacy/intfreedom/corevalues
- Bubeck, S., Chandrasekaran, V., Eldan, R., Kamar, E., Lee, Y. T., Lee, Y., Li, Y., Lundberg, S., Nori, A., Ribeiro, M., Shah, O., & Zhang, C. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2303.12712
- Cox, A. M. (2023). The impact of artificial intelligence on the future of academic library workforce and services. *Library Management*, 44(1/2), 84–97. https://doi.org/10.1108/LM-10-2022-0098
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
- Coder, C., & Wheatley, A. (2022). Algorithmic bias in library discovery systems: Evaluation and mitigation strategies. *College & Research Libraries*, 83(4), 562–581. https://doi.org/10.5860/crl.83.4.562
- Cundy, R., & Zhu, X. (2021). Predictive modeling for academic library collection development: A machine learning approach. *The Journal of Academic Librarianship*, 47(5), Article 102401. https://doi.org/10.1016/j.acalib.2021.102401
- Huang, M. H., & Chang, Y. W. (2020). A study of interdisciplinary research characteristics and collection development strategies in university libraries. *Journal of Librarianship and Information Science*, 52(3), 712–725. https://doi.org/10.1177/0961000619871602
- National Information Standards Organization. (2020). *E-book bibliometrics and predictive analytics for collection development* (NISO TR-08-2020). NISO.
- Cern, K., & Becker, B. W. (2023). Machine learning applications in electronic resource management: Automating usage log analysis and cost-per-use forecasting. *Serials Review*, 49(2), 105–118. https://doi.org/10.1080/00987913.2023.2198451
- Cech, L., & Herman, E. (2022). Data privacy and ethical machine learning in higher education libraries: A governance framework. *Journal of Documentation*, 78(7), 210–229. https://doi.org/10.1108/JD-03-2022-0058
- Cuesta, M., & Taylor, S. (2024). Semantic web technologies and BIBFRAME implementation in academic research collections. *Library Resources & Technical Services*, 68(1), 34–49. https://doi.org/10.5860/lrts.68n1.34