A Pratibha Spandan's Journal

CONFLUENCE OF KNOWLEDGE
(A Multidisciplinary International Peer-Reviewed/Refereed Journal)
ISSN 2320–5237 | Volume 12, Issue 02, July-December, 2025
http://cok.pratibha-spandan.org | © The Author(s) 2025

THE ROLE OF AI-ENABLED MULTILINGUAL METADATA GENERATION USING LARGE LANGUAGE MODELS FOR LIBRARY CATALOGUES

Srinivas Sambari
Librarian, Government Degree College, Thorrur, Mahabubabad, Telangana, India. sri.wgl@gmail.com
Chandra Chary Sreeramoj
University Librarian, Malla Reddy University, Hyderabad. Telangana State, India. Chandrachary81@mail.com
Dr. Raja Suresh Kumar Pitla
Librarian, Koneru Lakshmaiah Education Foundation (Deemed to be University), K L University, Bowrampat, Hyderabad - 500043, dr.prajasureshkumar@klh.edu.in.

ABSTRACT
Artificial intelligence (AI), primarily Large Language Models (LLMs), is increasingly altering library cataloging and metadata development via attractive effectiveness, consistency, and scalability. This study investigates the use of AI-enabled tools such as ChatGPT, Gemini, Copilot, retrieval-augmented generation (RAG) systems, and open-source LLMs in metadata creation, with a particular emphasis on their implications for the multilingual library environment. The article, based on recent experiments, discusses how AI may automate MARC21- and RDA-compliant metadata creation, subject assignment, indexing, and mistake correction at speeds significantly faster than traditional human cataloging. Systems like CATMELK illustrate how generative AI can turn textual and image-based bibliographic material into structured records, while Annif and TRANSLIB highlight breakthroughs in automatic subject indexing and multilingual access. Despite these advantages, the review identifies significant challenges, including variable accuracy across languages, intellectual property concerns, bias, formatting inconsistencies, and limitations in handling non-Latin scripts. Ethical considerations, data privacy, environmental impact, and the risk of de-skilling library professionals are also emphasized. The findings suggest that while AI and LLMs hold substantial promise for multilingual metadata generation and enhanced discovery in library catalogues, they are best deployed as assistive technologies rather than fully autonomous solutions. The paper concludes that responsible implementation, human oversight, adherence to bibliographic standards and further research into multilingual robustness are essential to ensure equitable, reliable, and sustainable integration of AI in library metadata practices.

Keywords: *Artificial Intelligence; Large Language Models; Metadata Generation; Library Cataloguing; Multilingual Access; MARC21*

Introduction

The paper discusses the potential of language models like ChatGPT for generating accurate MARC records in library cataloging, which could extend to multilingual metadata generation. By utilizing standards such as RDA and Dublin Core, ChatGPT can streamline record creation, enhancing efficiency in libraries. However, the implementation of AI-generated records raises concerns about intellectual property rights and bias, indicating a need for further research to ensure responsible use in multilingual contexts within library settings (Brzustowicz, 2023). The paper evaluates AI tools, specifically Large Language Models (LLMs) like ChatGPT and Gemini, for automating book cataloguing in libraries. It highlights that these AI applications can significantly enhance metadata generation by being 183 times faster and capable of cataloguing 187 times more books than human cataloguers. While the study focuses on cataloguing efficiency, it implies potential for multilingual metadata generation, although specific multilingual capabilities are not directly addressed in the research (Chisaba et al., 2025). The paper focuses on the use of Open-Source Large Language Models (LLMs) for generating high-quality metadata specifically for Open Data portals, rather than library catalogues. It highlights the integration of LLMs into the metadata generation process, improving consistency and categorization. However, it does not specifically address multilingual metadata generation or its application in library catalogues. The challenges noted include precise temporal assignment and scalability across different data formats, which may also apply to library contexts (Eger et al., 2025). The paper discusses the TRANSLIB system, which integrates AI-based methods to provide multilingual access to library catalogues. While it does not specifically address AI-enabled multilingual metadata generation using large language models, it highlights the use of bilingual dictionaries, terminology lexica, and intelligent thesauri to support multilingual search and user interface localization. The system has shown improved search processes and user friendliness, demonstrating the potential for enhanced multilingual access in library automation systems (Michos et al., 1999).

CONFLUENCE OF KNOWLEDGE
(A Multidisciplinary International Peer-Reviewed/Refereed Journal)
ISSN 2320–5237 | Volume 12, Issue 02, July-December, 2025
http://cok.pratibha-spandan.org | © The Author(s) 2025

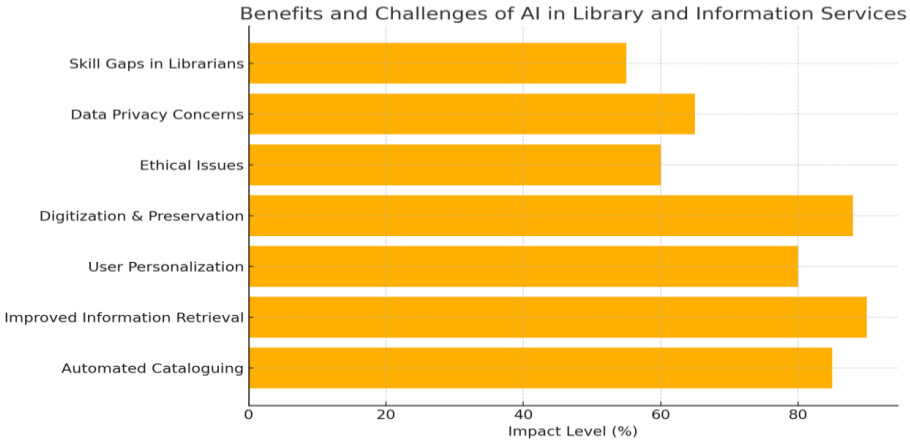

Figure 1.1: Opportunities and Concerns of AI in Libraries

Applications of AI in Metadata Generation:

Artificial Intelligence (AI) is gradually more applied in library and information science to hold metadata creation, authority control, subject indexing, and excellence assessment of bibliographic records. Techniques such as natural language processing, machine learning, and computer vision are being used to extract metadata from textual and visual sources like title pages and tables of contents, thereby improving efficiency and consistency in cataloguing workflows.

The research paper discusses the application of ChatGPT in library processes, including cataloging and indexing, highlighting its potential for multilingual metadata generation. It notes differences in response accuracy across languages and emphasizes the importance of fact-checking to ensure the reliability of generated metadata. Recommendations for formulating queries are provided to enhance the effectiveness of ChatGPT in creating multilingual metadata for library catalogues, ultimately aiming to streamline traditional bibliographic work (Stepanov ey al., 2024). The paper does not specifically address AI-enabled multilingual metadata generation for library catalogues. Instead, it focuses on a retrieval augmented generation system that enhances publication management through chat-based large language models (LLMs). It streamlines the organization of personal publication libraries, automating metadata and tag addition, but does not explicitly cover multilingual capabilities or library catalogues. The primary use cases are explorative search and retrieval, and cataloguing and management within platforms like BibSonomy (Volker et al., 2024).

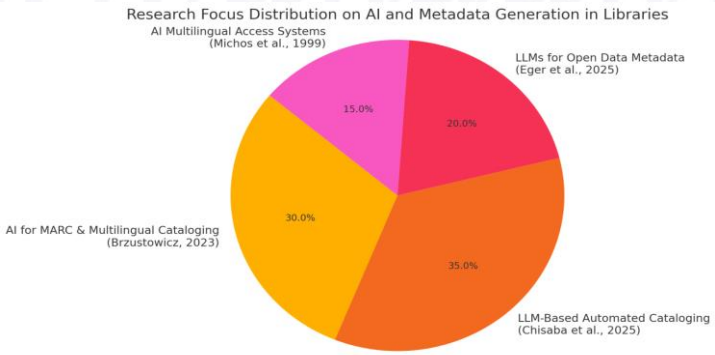

Figure 2.1: Distribution of Research Focus on AI and Metadata Generation in Libraries

The paper does not specifically address AI-enabled multilingual metadata generation using large language models for library catalogues. However, it discusses the construction of a library subject information intelligent perception system that integrates LLMs to enhance library subject services, including the efficient integration of multi-source subject data and automatic generation of subject knowledge graphs. This could potentially support multilingual applications, but the focus is primarily on subject services rather than cataloguing specifically (Xu, 2024)). The paper does not specifically address AI-enabled multilingual metadata generation using large language models for library catalogues. However, it discusses the integration of retrieval-augmented generation (RAG) systems in academic libraries, focusing on enhancing search precision and user experience. RAG's capabilities in natural language understanding and contextual processing could potentially support multilingual

ISSN 2320–5237 | Volume 12, Issue 02, July-December, 2025
http://cok.pratibha-spandan.org | © The Author(s) 2025

applications, but the study emphasizes the need for careful implementation regarding technical architecture, data protection, and ethical compliance in library systems (Bevara et al., 2025). The paper does not specifically address AI-enabled multilingual metadata generation using large language models for library catalogues. However, it discusses the development of Seamless M4T, an AI-based application that enhances communication between librarians and users from diverse linguistic backgrounds. This application utilizes natural language processing techniques for real-time translation and transcription, which could indirectly support multilingual metadata generation by improving overall communication and service delivery in university libraries (Kusumaningtiyas et al., 2024). The paper explores the capabilities of Large Language Models (LLMs) like ChatGPT and Copilot in performing subject cataloging tasks, specifically in assigning subject headings and class numbers. - It highlights the significance of LLMs in automating cataloging processes, which could enhance efficiency in libraries and information management. - However, the paper identifies several shortcomings of LLMs, including lack of specificity, unauthorized term usage, and incorrect formatting of MARC records. - The research aims to assess the potential applications of LLMs in cataloging, suggesting they may serve better as aides and teaching tools rather than fully automated solutions. - It also addresses concerns related to environmental impact, de-skilling, intellectual theft, and bias associated with the use of LLMs in cataloging tasks (Holstrom, 2025). The paper discusses the transformative role of artificial intelligence (AI) in library management, highlighting its potential to revolutionize traditional practices in information management and user engagement. - It emphasizes the importance of AI in automating library operations, improving information retrieval, personalizing user experiences, and optimizing resource management. The review identifies existing challenges, such as data privacy concerns and the need for staff training, which must be addressed for successful AI integration. - The paper aims to synthesize literature on AI applications in libraries from 2010 to 2024, providing insights into current practices and future implications for library services. - It calls for further research into the long-term effects of AI on user trust and information equity, advocating for a cautious and deliberate approach to AI adoption in libraries (Patil et al., 2025). The Annif system ranked 1st in both the overall quantitative evaluation and the qualitative evaluation of Subtask 2 at GermEval-2025. - The main quantitative metric used was nDCG20, with Annif achieving a score of 0.8787 in Case 1 (considering both correct and technically correct predictions) and 0.7699 in Case 2 (considering only correct predictions). - The bilingual predictions using the M24 LLM ranked 1st with an nDCG20 score of 0.5697. - The new LLM ranking ensemble approach improved nDCG scores of topic suggestions by 0.01-0.03 compared to simple ensembles. - The Bonsai model achieved the best scores in all cases, followed by XTransformer and MLLM, indicating a strong performance of traditional XMTC algorithms enhanced by LLMs (Suominen et al., 2025). The paper highlights the potential of ChatGPT in automating and enhancing the cataloging and description processes of library resources, which can significantly streamline metadata creation and save time for library staff and information technology professionals. This capability allows for the efficient generation of metadata from textual descriptions, adhering to established standards such as Dublin Core, MARC 21, and UNIMARK. - It also discusses the challenges and limitations of using ChatGPT for metadata generation, particularly concerning the accuracy and quality of the generated text, emphasizing the need for careful consideration when implementing this technology in library settings (Sokół & Andrukhiv, 2024).The study concluded that Generative Artificial Intelligence (GAI) can significantly enhance cataloging efficiency, demonstrating high satisfaction in correcting errors within existing MARC21 records. This indicates GAI's potential to improve the overall quality of bibliographic descriptions in library catalogs. - The custom GPT model, CATMELK, effectively converted images from book title, copyright, and cover pages into RDA compliant MARC21 records, showcasing GAI's capability as a prospective educational tool for novice librarians in understanding and applying MARC21 and RDA standards. However, challenges were noted in converting Sinhala and Tamil data from images to MARC21 records (Gamage et al., 2024). The paper introduces a novel iterative retrieval-generation collaborative framework that enhances large language models by integrating both parametric and non-parametric knowledge, which is crucial for effectively addressing knowledge-intensive tasks that require multi-step reasoning. - Through experiments conducted on four question answering datasets, including both single-hop and multi-hop QA tasks, the proposed method demonstrates a significant improvement in the reasoning capabilities of large language models, outperforming previous baseline approaches (Gamage & Wanigasooriya 2024).

CONFLUENCE OF KNOWLEDGE
(A Multidisciplinary International Peer-Reviewed/Refereed Journal)
ISSN 2320–5237 | Volume 12, Issue 02, July-December, 2025
http://cok.pratibha-spandan.org | © The Author(s) 2025

Table 2.1: Comparative Impacts of AI Technologies Metadata Generation.

AI Technology Type	Primary Function	Key Benefits	Performance Impact	Implementation Complexity
Large Language Models (LLMs) – ChatGPT, Gemini, Copilot	Automated cataloguing, metadata generation, subject assignment, indexing	Faster record creation, support for standards (MARC21, RDA, Dublin Core), reduced staff workload	Up to 183× faster cataloguing; can process 187× more books than humans; quality varies by language	Medium–High: Requires prompt engineering, validation workflows, bias and IP safeguards
Generative AI for MARC/RDA Conversion (e.g., CATMELK)	Conversion of bibliographic data (including images) into structured MARC21 records	Error correction, training support for novice librarians, improved metadata quality	High accuracy for English-language records; effective MARC21 error correction	High: OCR integration, multilingual limitations (e.g., Sinhala, Tamil), quality control needed
Retrieval-Augmented Generation (RAG) Systems	Enhanced search, metadata enrichment, contextual retrieval	Improved search precision, contextual understanding, reduced hallucinations	Better relevance in discovery systems; improved user experience	High: Requires technical infrastructure, data governance, and system integration
Open-Source LLMs for Metadata Generation	Automated metadata generation for Open Data portals	Consistency, scalability, cost efficiency	Improved categorization and consistency across datasets	Medium: Challenges with temporal precision and cross-format scalability
AI-Based Multilingual Access Systems (TRANSLIB)	Multilingual search and interface localization	Improved access for multilingual users, enhanced usability	Demonstrated improvement in search effectiveness and user friendliness	Medium: Depends on dictionaries, thesauri, and linguistic resources
AI for Subject Knowledge Graph Generation	Integration of multi-source subject data, knowledge graph creation	Better subject services, automated data integration	Enhanced subject discovery and knowledge organization	High: Complex data modeling and integration requirements
AI Translation & Communication Tools (e.g., Seamless M4T)	Real-time translation and transcription in libraries	Improved librarian–user communication across languages	Enhanced service delivery in multilingual environments	Medium: Language coverage and accuracy management required
Automated Subject Indexing Systems (Annif + LLMs)	Topic suggestion and subject indexing	High accuracy, improved bilingual subject predictions	nDCG20 up to 0.8787; top-ranked in GermEval-2025	Medium: Model tuning and assembly optimization needed
AI-Driven Library Management Systems	Automation of operations, personalization, resource optimization	Improved efficiency, personalized user services	Long-term improvements in service delivery and management	High: Staff training, data privacy, and ethical compliance essential
Wearable AI, Computer Vision & Biomechanical AI (Education-focused)	Training, evaluation, and performance analysis	Personalized, data-driven learning and assessment	High accuracy in training feedback	High: Specialized hardware, limited relevance to cataloguing

Conclusion

The incorporation of Artificial Intelligence, mostly Large Language Models (LLMs), into library metadata creation and cataloguing practice represents a significant advancement in library and information science. The studies reviewed in this paper express that AI-driven tools such as ChatGPT, Gemini, RAG-based systems, Annif, and specialized generative models can substantially improve the speed, scalability, and consistency of metadata creation while supporting established standards like MARC21, RDA, and Dublin Core. These technologies show strong potential in automating subject assignment, indexing, and error correction, and in enhancing multilingual access to library resources.

However, the findings also show that AI-enabled metadata creation is not without limitations. Variations in truth across languages, challenges in handling non-Latin scripts, risks of bias, intellectual property concerns, and inconsistencies in record formatting highlight the need for cautious and responsible adoption. Additionally, ethical issues such as data privacy, environmental impact, and the potential de-skilling of library professionals must be carefully addressed. Current evidence suggests that AI systems perform most effectively when used as assistive tools that complement human expertise rather than as fully autonomous cataloguing solutions.

In conclusion, AI has the capacity to reshape metadata generation and multilingual access in libraries, but its successful implementation depends on robust validation workflows, continuous staff training, and adherence to ethical and professional standards. Future research should focus on improving multilingual robustness, reducing

bias, and developing hybrid human–AI models that ensure accuracy, inclusivity, and long-term sustainability in library cataloguing practices.

References

Brzustowicz, R. (2023). From ChatGPT to CatGPT. *Information Technology and Libraries*. https://doi.org/10.5860/ital.v42i3.16295

Chisaba Pereira, C. A., Herrera-Calero, R., Niño-Neira, S.-A., & Hurtado-Ortiz, B.-A. (2025). Datalogación: evaluación de herramientas de inteligencia artificial basadas en el Modelo Extenso de Lenguaje (Large Language Model) para la automatización de la descripción de libros. *Infonomy*, *3*(4). https://doi.org/10.3145/infonomy.25.023

Eger, B.-L., Ullmann, F., Dinter, B., & Gluchowski, P. (2025). Mit Open-Source-LLMs zu aussagekräftigen Metadaten: Sprachmodelle als Schlüssel für nutzerfreundliche Open-Data-Portale. *HMD. Praxis Der Wirtschaftsinformatik*. https://doi.org/10.1365/s40702-025-01197-1

Michos, S. E., Stamatatos, E., & Fakotakis, N. (1999). Supporting multilinguality in library automation systems using ai tools. *Applied Artificial Intelligence*, *13*(7), 679–703. https://doi.org/10.1080/088395199117243

Stepanov, V. K., Madzhumder, M. S., & Begunova, D. D. (2024). Application of the big language model – ChatGPT in the librarianship and bibliographical work. *Naučnye i Tehničeskie Biblioteki*. https://doi.org/10.33186/1027-3689-2024-4-86-108

Volker, T., Pfister, J., Koopmann, T., & Hotho, A. (2024). *BibSonomy Meets ChatLLMs for Publication Management: From Chat to Publication Management: Organizing your related work using BibSonomy&LLMs*. https://doi.org/10.1145/3627508.3638298

Xu, C. (2024). *Construction of library subject information intelligent perception system integrating large language model*. https://doi.org/10.1109/imcec59810.2024.10575069

Bevara, R. V. K., Lund, B., Mannuru, N. R., Karedla, S. P., Mohammed, Y., Kolapudi, S. T., & Mannuru, A. (2025). Prospects of Retrieval Augmented Generation (RAG) for Academic Library Search and Retrieval. *Information Technology and Libraries*, *44*(2). https://doi.org/10.5860/ital.v44i2.17361

Kusumaningtiyas, T., Nugroho, P. A., & Noor Azizi, N. A. (2024). Seamless M4T for librarians to communicate and provide multilingual collection services. *Library Hi Tech News*. https://doi.org/10.1108/lhtn-11-2023-0205

Holstrom, C. (2025). Large Language Models (LLMs) and Cataloging: Exploring How ChatGPT and Copilot Assign Subject Headings and Call Numbers. *Proceedings from North American Symposium on Knowledge Organization*, 78–95. https://doi.org/10.7152/nasko.v7i1.95648

Patil, S. S., Kamble, L. Y., & Bagewadi, P. (2025). Transformative role of artificial intelligence in library management: A review. *Gyankosh-The Journal of Library and Information Management*, *16*(1), 27–40. https://doi.org/10.5958/2249-3182.2025.00003.0

Suominen, O., Inkinen, J., & Lehtinen, M. (2025). Annif at the GermEval-2025 LLMs4Subjects Task: Traditional XMTC Augmented by Efficient LLMs. *arXiv.Org*, *abs/2508.15877*. https://doi.org/10.48550/arxiv.2508.15877

Sokół, M., & Andrukhiv, A. (2024). Перспективи / можливості застосування chat gpt для створення метаданих бібліотечних ресурсів. *Vìsnik Hmel'nic'kogo Nacìonal'nogo Unìversitetu*, *341*(5), 417–422. https://doi.org/10.31891/2307-5732-2024-341-5-60

Gamage, R., & Wanigasooriya, P. (2024). Using Generative AI for Bibliographic Description: A Study with ChatGPT 4. *Journal of The University Librarians Association of Sri Lanka*, *27*(2). https://doi.org/10.4038/jula.v27i2.8083

Feng, Z., Feng, X., Zhao, D., Yang, M. Q., & Qin, B. (2024). *Retrieval-Generation Synergy Augmented Large Language Models*. https://doi.org/10.1109/icassp48485.2024.10448015

Larson, R. R., Gey, F. C., & Chen, A. (2002). Harvesting translingual vocabulary mappings for multilingual digital libraries. *ACM/IEEE Joint Conference on Digital Libraries*, 185–190. https://doi.org/10.1145/544220.544259

Mala, J. M. (2024). *From Dewey to Deep Learning: Exploring the Intellectual Renaissance of Libraries through Artificial Intelligence*. https://doi.org/10.17821/srels/2024/v61i1/171001

Mirović, D. (2025). *Sanjaju li androidi električne bibliotekare/ke? Primjena AI tehnologija u nacionalnim bibliotekama Evrope*. *11*(11), 185–213. https://doi.org/10.71271/issn.2303-520x.2024.11