"DIGITIZATION AND PRESERVATION OF INDIAN MANUSCRIPTS AND CULTURAL HERITAGE: LEVERAGING AI AND BIG DATA FOR ENHANCING ACCESS TO TRADITIONAL KNOWLEDGE SYSTEMS"

Researcher
Swati P. Dongre
Research Scholar
RTMNU Nagpur
Mob. No. 9518915081
Email id – swatikadu1931987@gmail.com

Guide
Dr. Aparna Chaudhari
Librarian

Abstract :

India's rich collection of manuscripts and cultural heritage materials represents a vast reservoir of indigenous knowledge across disciplines such as philosophy, science, medicine, and the arts. However, their fragile condition, linguistic diversity, and limited accessibility pose major challenges for preservation and scholarly use. With the advent of the National Education Policy (NEP) 2020, there is a renewed emphasis on integrating Indian Knowledge Systems (IKS) into education and research. This paper explores how emerging technologies—particularly Artificial Intelligence (AI) and Big Data—can transform the digitization and preservation of manuscripts and cultural heritage resources. It highlights the role of AI in optical character recognition (OCR), natural language processing, and multilingual translation for manuscripts, while Big Data analytics can support large-scale organization, retrieval, and dissemination of traditional knowledge. The study also examines existing digital repositories and initiatives in India, identifying gaps and opportunities in leveraging technology for inclusive access. Ethical considerations, policy implications, and skill development for library professionals are discussed within the NEP 2020 framework. The paper proposes a strategic model for creating intelligent, multilingual, and sustainable digital platforms that not only preserve but also revitalize traditional knowledge systems for future generations.

Keywords : Indian Knowledge Systems (IKS), Artificial Intelligence, Big Data, Digitization, NEP 2020, Manuscript Preservation, Library and Information Science (LIS).

Introduction :

India has always been known as a cradle of civilization, rich in intellectual and cultural heritage. Its manuscripts, written in Sanskrit, Pali, Prakrit, Persian, and various regional languages, carry centuries of knowledge in areas like philosophy, medicine, astronomy, mathematics, literature, and the arts. These documents form an important part of the Indian Knowledge System (IKS), which continues to influence learning and practice today. However,

many manuscripts are fragile and face threats from environmental conditions, poor preservation facilities, and a decline in experts who can read ancient scripts. If urgent action is not taken, a large portion of this knowledge could be lost forever. The National Education Policy (NEP) 2020 encourages a multidisciplinary approach to reviving IKS, offering both a challenge and an opportunity for Library and Information Science (LIS) professionals to preserve heritage while using modern technologies to make it more accessible.

Digitization is a key step in protecting manuscripts, allowing libraries and archives to create digital copies, reduce physical handling, and provide wider access for research and education. However, digitization alone is not enough. Emerging technologies like Artificial Intelligence (AI) and Big Data can transform preservation by recognizing complex scripts, improving image quality, translating texts, creating smart search tools, and analyzing large collections to discover new insights. By adopting these tools, LIS professionals can develop new skills, create innovative services, and bring India's cultural heritage to a global audience. At the same time, challenges such as infrastructure, ethical issues, copyright concerns, and community involvement must be addressed. This study explores how AI and Big Data can help preserve and share Indian manuscripts, examining both the opportunities and limitations of these technologies, and suggesting ways to integrate them into LIS practice under NEP 2020.

Problem Statement :

Indian manuscripts and cultural heritage materials hold immense historical, philosophical, and scientific knowledge. However, due to challenges like fragile condition, language barriers, lack of standard metadata, and limited accessibility, much of this knowledge remains underutilized. Traditional digitization initiatives often fail to integrate advanced technologies that can improve discoverability, multilingual access, and long-term preservation. This creates a pressing need to explore how **AI and Big Data** can modernize the preservation and dissemination of Indian Knowledge Systems (IKS) within the framework of NEP 2020.

Objectives of the Study :

1. To examine the current status of digitization and preservation of Indian manuscripts and cultural heritage materials.
2. To explore the potential role of **AI tools (OCR, NLP, machine learning)** in improving accessibility and retrieval of manuscripts in multiple Indian languages.
3. To analyze how **Big Data analytics** can support the management, curation, and dissemination of large-scale traditional knowledge resources.
4. To evaluate challenges such as copyright, ethics, data standards, and resource sustainability in applying emerging technologies.
5. To recommend strategies for integrating digitized Indian Knowledge Systems into **Library and Information Science (LIS) education and practice** in alignment with NEP 2020.

Research Methodology :

This study adopts a **qualitative and analytical approach** supported by case studies and document analysis.

- **Data Sources :** Secondary sources including published research papers, government reports (NEP 2020, NMM documents), digital repository portals (e.g., Digital Library of India, IGNCA), and conference proceedings.
- **Case Studies :** Examination of selected digital initiatives such as the National Mission for Manuscripts, Europeana, and British Library projects to draw comparative insights.
- **Analytical Tools :** Content analysis to evaluate the scope and limitations of current digitization practices; technology mapping to understand AI and Big Data applications in LIS.
- **Scope :** Focus is restricted to manuscripts and cultural heritage collections in India, with references to global best practices for contextual understanding.

Limitations : The study relies on secondary data due to the constraints of time and resources; primary surveys of LIS professionals may be considered for future research.

Literature Review :

Global Digitization Efforts: Around the world, digitization has become key to preserving cultural heritage. UNESCO's *Memory of the World* programme focuses on protecting important documents through digital technologies (UNESCO, 2017). In Europe, projects like Europeana collect cultural resources from multiple institutions, making them accessible across borders (Koltay, 2016). Libraries such as the British Library and the Library of Congress have also digitized large collections of manuscripts and rare books, offering open access to researchers worldwide (Smith, 2018). These initiatives reduce the handling of fragile originals and encourage collaborative research.

Indian Initiatives: In India, the National Mission for Manuscripts (NMM) and the Indira Gandhi National Centre for the Arts (IGNCA) have led efforts to document, conserve, and digitize manuscripts, paintings, and archival materials (NMM Report, 2019). Platforms like the Digital Library of India (DLI) and Bharatiya Digital Sanskriti (BDSL) aim to provide multilingual access to these resources. While these initiatives have made significant progress, challenges remain, including inconsistent metadata, limited interoperability, and user-friendly access issues (Chowdhury & Kumar, 2020).

Role of AI in Preservation: Artificial Intelligence (AI) has brought transformative changes to manuscript preservation. OCR technology can now read ancient handwritten scripts, making them searchable (Terras, 2019). NLP enables automated translation and cross-lingual information retrieval, which is vital for India's diverse languages (Patel, 2021). Machine learning further helps improve image quality, detect damaged text, and automate cataloguing

(Liang et al., 2020). However, faded manuscripts, unique calligraphy, and obsolete scripts still pose challenges.

Big Data for Knowledge Access: Big Data analytics improves management of large digital collections. Metadata-driven analysis enhances discoverability, helps researchers find patterns across manuscripts, and supports long-term preservation through cloud-based infrastructure (Singh & Prasad, 2022; Zhou & Chen, 2019). Despite its potential, issues of ethics, standardization, and privacy need attention.

NEP 2020 and LIS Education: The National Education Policy (NEP) 2020 highlights the importance of reviving Indian Knowledge Systems (IKS) and promoting multidisciplinary research. LIS education must now integrate emerging technologies with traditional knowledge management, focus on digital humanities skills, and support AI-driven preservation research (Hirwade, 2021; Nikose & Naidu, 2022). Libraries are increasingly seen as central hubs for heritage preservation, innovation, and lifelong learning.

Challenges and Summary: Indian digitization projects still face funding, infrastructure, and personnel limitations, as well as ethical concerns over intellectual property and community ownership (Pande, 2020; Ramesh, 2019). Overall, literature shows that digitization, AI, and Big Data together can significantly enhance heritage preservation. Lessons from global projects highlight open access and interoperability, while Indian initiatives emphasize aligning technology with cultural needs. NEP 2020 provides a timely framework for LIS professionals to move from being custodians of knowledge to innovators in digital preservation.

Discussion and Analysis :

The digitization and preservation of Indian manuscripts and cultural heritage have emerged as critical areas of focus in the knowledge economy of the 21st century. The vast repositories of handwritten manuscripts, palm-leaf documents, temple records, and rare cultural artifacts represent centuries of intellectual, spiritual, and social evolution. However, the challenges of accessibility, preservation, and interpretation demand innovative approaches that go beyond conventional archival methods. Artificial Intelligence (AI) and Big Data technologies are increasingly positioned as transformative tools to address these issues.

Preservation and Digitization: A Foundational Step :

Digitization of manuscripts ensures that fragile materials—often centuries old—are preserved in digital formats for long-term access. This reduces the risks associated with physical handling and environmental degradation. However, digitization alone does not guarantee meaningful access. For example, manuscripts written in Sanskrit, Pali, Prakrit, or regional scripts require advanced optical character recognition (OCR) systems capable of handling complex scripts and diverse writing styles. AI-driven OCR and natural language processing (NLP) tools can help overcome these limitations by enabling script recognition, translation, and semantic interpretation, thereby bridging the gap between physical

preservation and intellectual access.

Role of AI in Interpretation and Access :

AI systems can analyze massive amounts of manuscript data to identify patterns, classify texts, and create metadata for improved cataloging. Machine learning algorithms enhance searchability by recognizing keywords, themes, and cross-references across thousands of documents. This significantly improves accessibility for researchers, students, and cultural enthusiasts. For instance, AI-based tools can provide automated summaries or even voice-assisted retrieval of manuscript knowledge, making traditional wisdom more approachable for a wider audience.

Big Data as a Catalyst for Knowledge Integration :

The integration of Big Data analytics allows for cross-comparison of digitized manuscripts across regions, languages, and historical periods. This enables scholars to uncover hidden linkages between philosophical texts, scientific treatises, and cultural practices. Big Data platforms can also integrate manuscript archives with other datasets—such as archaeological findings, oral histories, and modern scientific research—creating a holistic understanding of India's intellectual traditions. The ability to analyze patterns at scale makes Big Data an invaluable resource in connecting traditional knowledge with contemporary research.

Challenges in Application :

Despite the potential, several challenges persist. First, the diversity of languages and scripts presents a major obstacle in building universal AI models. Second, digitization efforts are often fragmented, with different institutions using varied standards, making interoperability difficult. Third, ethical concerns regarding data ownership, cultural sensitivity, and intellectual property rights require careful handling to avoid misrepresentation or exploitation of traditional knowledge. Finally, the costs of infrastructure, training, and sustained maintenance pose practical difficulties for large-scale implementation.

Socio-Cultural Impact :

Beyond technological dimensions, digitization and AI-driven access have profound socio-cultural implications. Democratizing access to manuscripts allows younger generations, diaspora communities, and international scholars to engage with India's cultural heritage. This, in turn, fosters cultural pride, global academic collaboration, and innovation in fields such as Ayurveda, astronomy, metallurgy, and performing arts—domains deeply rooted in traditional manuscripts. However, care must be taken to preserve the authenticity of knowledge systems while adapting them for modern contexts.

Towards a Hybrid Future :

The future lies in creating hybrid models that blend AI-powered digital access with traditional custodianship. Community participation, collaboration between technologists and historians, and government support are vital to ensure that technological interventions respect cultural integrity. Public–private partnerships can help overcome funding constraints, while open-access digital libraries can democratize knowledge. Thus, AI and Big Data should not be seen as replacements but as enablers that extend the lifespan and reach of India's heritage.

Conclusion and Recommendations :

The digitization and preservation of Indian manuscripts and cultural heritage represent a vital step in safeguarding the intellectual wealth of India. Manuscripts, often centuries old, contain vast knowledge in medicine, astronomy, philosophy, literature, and other domains. While traditional preservation techniques are important, they are not sufficient in the digital age. The integration of Artificial Intelligence (AI) and Big Data provides transformative opportunities to preserve, interpret, and share this heritage with the world.

AI enables the recognition of complex scripts, translation of ancient languages, and creation of searchable metadata, making manuscripts more accessible to researchers and the general public. Big Data allows for large-scale analysis, enabling cross-disciplinary research and the discovery of new connections within India's intellectual traditions. Together, these technologies have the potential to democratize knowledge and enhance cultural pride, while also linking traditional wisdom to contemporary science and innovation.

However, challenges remain. Linguistic diversity, lack of standardization in digitization, ethical concerns related to intellectual property and cultural sensitivity, and high costs of infrastructure continue to hinder progress. Without proper frameworks, there is a risk of reducing manuscripts to mere "data," stripping them of their cultural and contextual value.

Key Recommendations :

1. **Creation of a Unified National Repository** – Establish a centralized digital platform integrating manuscript collections from universities, libraries, and cultural institutions.
2. **Development of AI for Indian Languages** – Invest in research to improve OCR, NLP, and translation tools tailored to diverse Indian scripts and dialects.
3. **Standardization of Practices** – Adopt common digitization standards across institutions to ensure interoperability and long-term sustainability.
4. **Ethical and Legal Safeguards** – Frame policies for intellectual property rights, cultural sensitivity, and responsible use of traditional knowledge.
5. **Public–Private Partnerships** – Encourage collaboration between government, academia, and technology companies to share resources and expertise.
6. **Capacity Building and Training** – Train librarians, archivists, and cultural

custodians in digital tools, AI applications, and data management.

7. **Community Participation** – Involve local scholars, traditional custodians, and communities to maintain authenticity and cultural integrity.

In conclusion, digitization supported by AI and Big Data should not be viewed merely as a technological project but as a cultural mission. By blending advanced technologies with ethical, inclusive, and sustainable practices, India can preserve its invaluable manuscripts while making them accessible for global scholarship and future generations.

References :

- **Fredriksson, M. (2022).** India's Traditional Knowledge Digital Library and the politics of documentation. *Third World Quarterly, 43*(3), 602–619. https://doi.org/10.1080/01436597.2021.2012397
  Examines TKDL's role in protecting Indian heritage from biopiracy; useful for understanding socio-political challenges.
- **WIPO (2018).** *A Guide to Intellectual Property Issues in Access and Benefit-Sharing.* World Intellectual Property Organization. https://www.wipo.int/publications/en/details.jsp?id=4383
  Outlines IPR frameworks related to traditional knowledge and biodiversity.
- **Secretariat of the Convention on Biological Diversity. (2011).** *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits.* CBD. https://www.cbd.int/abs
  Provides the global legal framework for protecting and sharing TK.
- **UNESCO. (2015).** *Recommendation concerning the Preservation of, and Access to, Documentary Heritage, including in Digital Form.* Paris: UNESCO.
  Key policy guideline on digital preservation and access.
- **National Mission for Manuscripts (NMM). (2019).** *Annual Report.* Ministry of Culture, Government of India. http://www.namami.gov.in
  Official report on Indian manuscript digitization and conservation efforts.
- **Garg, A., Tiwari, L., Juj, T., Indu, S., & Jayanthi, N. (2022).** Language and era prediction of digitized Indian manuscripts using CNNs. In *Advances in Intelligent Systems and Computing* (pp. 333–344). Springer. https://doi.org/10.1007/978-981-16-4615-3_29
  Shows how AI can classify manuscripts by era and language.
- **Maheshwari, A., Singh, N., Krishna, A., & Ramakrishnan, G. (2022).** A benchmark and dataset for post-OCR text correction in Sanskrit. *arXiv preprint arXiv:2202.02549.* https://arxiv.org/abs/2202.02549
  Contributes dataset + methods to improve Sanskrit OCR accuracy.
- **Sharma, R. (2020).** A survey on offline recognition of handwritten Indic scripts. *Pattern Recognition, 93*, 1–15. https://doi.org/10.1016/j.patcog.2019.07.017
  Comprehensive survey of OCR challenges in Indic scripts.

- **Kulkarni, I. (2022).** Proposed design to recognize ancient Sanskrit manuscripts using machine learning. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4216721
  Early-stage ML framework for manuscript recognition.
- **Kudale, D., et al. (2024).** TEXTRON: Weakly supervised multilingual text detection through data programming. *arXiv preprint arXiv:2403.01054.* https://arxiv.org/abs/2403.01054
  Improves multilingual manuscript digitization by combining weak supervision with OCR.
- **Oxford Centre for Hindu Studies. (2025).** *Creating AI models for handwriting and text recognition in South Asian manuscripts.* OCHS Project Page. https://ochs.org.uk
  Ongoing project using AI to process palm-leaf manuscripts.
- **Colace, F., et al. (2025).** New AI challenges for cultural heritage protection. *Journal of Cultural Heritage, 67*, 221–234. https://doi.org/10.1016/j.culher.2024.07.011
  Discusses AI-driven threats/opportunities in cultural heritage digitization.
- **Harisanty, D. (2024).** Cultural heritage preservation in the digital age: harnessing AI. *Library Hi Tech, 42*(1), 53–68. https://doi.org/10.1108/LHT-03-2022-0103
  Explores AI applications in global cultural heritage preservation.
- **Kumar, A., & Sanyal, S. (2021).** Big Data analytics in cultural heritage: Opportunities for India. *International Journal of Digital Humanities, 2*(1), 45–63. https://doi.org/10.1007/s42803-021-00030-5
  Focuses on Big Data frameworks for Indian heritage datasets.
- **IFLA / UNESCO PERSIST. (2021).** *Guidelines for the Selection of Digital Heritage for Long-Term Preservation.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377330
  Offers international standards for digital preservation and access.
- **Kataria, B., & Jethva, H. B. (2021).** OCR of Indian language manuscripts using CNNs. *Design Engineering, 12*(3), 345–360.
  Technical approach to Indian language manuscript recognition.
- **Sharma, S. (2017).** Traditional Knowledge Digital Library: "A silver bullet" in the war against biopiracy? *The John Marshall Review of Intellectual Property Law, 16*(3), 285–320.
  Critically reviews TKDL as a model for TK preservation.
- **Tuli, S., & Gill, S. S. (2020).** Blockchain for preserving indigenous knowledge systems. *Future Generation Computer Systems, 110*, 175–189. https://doi.org/10.1016/j.future.2020.04.019
  Introduces blockchain as a complementary tool for safeguarding TK.
- **PTI. (2025).** Parliamentary panel seeks AI-based platform for manuscript-to-text conversion. *The Week.* https://www.theweek.in/news/india/2025/02/20/parliamentary-panel-ai-manuscripts.html
  Latest government initiative for AI-based digitization in India.