

REVIEW PAPER

Defining the Boundaries: Research Objectives and Scope of an AI-Based Bibliometric Framework

Harendra Nath Roy¹ and Jayati Lahiri Dey^{2*}

1Ph.D. Scholar, Department of Computer and Information Science, Raiganj University, Raiganj, West Bengal, India

²Assistant Professor, Department of Computer and Information Science, Raiganj University, Raiganj, West Bengal, India

*Corresponding author: ld.jayatii@gmail.com

Received: 19 Mar., 2026

Revised: 05 May, 2026

Accepted: 25 May., 2026

ABSTRACT

Doctoral theses constitute a significant yet systematically underutilized category of scholarly output. Despite their intellectual depth and methodological rigour, most dissertations remain poorly indexed, rarely cited in commercial databases, and largely invisible to mainstream bibliometric evaluation. This paper presents a structured articulation of the research objectives and analytical scope of an AI-Based Bibliometric Framework designed to address this invisibility—with particular focus on the interdisciplinary domain of Library and Information Science (LIS) and Computer and Information Science (CIS). Six specific research objectives are defined: conceptualizing the AI-driven framework, incorporating citation context analysis using NLP, addressing limitations of traditional bibliometric indicators, designing metadata validation mechanisms, conducting comparative India–Global analysis, and contributing theoretically to AI-enabled scientometrics. The scope encompasses doctoral theses from national repositories (*Shodhganga*), international databases (ProQuest, EThOS, Scopus), and LIS-specific archives, with explicit AI compatibility criteria. A comparative citation analysis of 284 doctoral theses from five Indian universities (2015–2023) empirically grounds the study, revealing that 18–35% of theses carry zero recorded citations—reflecting a discoverability gap rather than low quality. The framework integrates PageRank-based importance scoring, TF-IDF and LDA-based semantic classification, context-aware citation impact metrics, and altmetric aggregation. By bridging LIS bibliometric tradition with CIS computational methodology, the proposed framework advances both the theory and practice of research evaluation in the era of AI-augmented scholarship.

Keywords: Bibliometrics; doctoral theses; artificial intelligence; natural language processing; citation analysis; Library and Information Science; *Shodhganga*; research evaluation; scientometrics; interdisciplinary research

The exponential expansion of global scholarly literature has fundamentally transformed knowledge production, dissemination, and evaluation. In disciplines such as Library and Information Science (LIS)

How to cite this article: Roy, H.N. and Dey, J.L. (2026). Defining the Boundaries: Research Objectives and Scope of an AI-Based Bibliometric Framework. *IJISC*, 13(01): 113-125.

Source of Support: None; **Conflict of Interest:** None



and Computer and Information Science (CIS), publication growth has been particularly pronounced over the past two decades, driven by digitalization, open-access mandates, and interdisciplinary convergence. Traditional bibliometric techniques—rooted in citation counts, co-authorship patterns, and impact indicators—have served as foundational instruments for assessing research performance. However, such metrics primarily capture quantitative citation frequency rather than the qualitative intent or contextual meaning underlying scholarly communication^[1].

A particularly acute dimension of this problem concerns doctoral theses. Despite representing some of the most original and comprehensive academic work produced annually, doctoral dissertations remain under-indexed, poorly cited in major databases, and structurally invisible to standard bibliometric evaluation. A single curator manually processing 50 theses per week would require over 150 years to process the current Shodhganga corpus—illustrating the absolute necessity of intelligent automation^[2]. National repositories such as Shodhganga (India, 400,000+ theses), ProQuest Dissertations and Theses Global (USA, 5 million+), and EThOS (UK, 600,000+) collectively represent an enormous unmined scholarly resource.

Artificial Intelligence (AI)—particularly Natural Language Processing (NLP) and Machine Learning (ML)—offers a transformative paradigm for bibliometric analysis. Through AI-driven automation, bibliographic datasets of any scale can be processed to identify not only who cites whom, but also why and how these citations occur. The integration of deep contextual understanding enables a decisive shift from descriptive to interpretive scientometrics, emphasizing meaning, influence, and knowledge flow rather than mere numerical frequency^[3].

This paper presents a structured articulation of the research objectives and analytical scope of an AI-Based Bibliometric Framework, specifically designed for the interdisciplinary domain of LIS and CIS. The paper proceeds as follows. Section II establishes the scholarly value of doctoral theses and identifies specific gaps. Section III reviews the structural limitations of traditional bibliometric analysis. Section IV presents four core AI techniques with their mathematical foundations and documented applications. Section V presents the comparative citation analysis of Indian doctoral theses. Section VI maps five priority use-case domains. Section VII discusses implementation challenges. Section VIII offers future directions and recommendations. Section IX concludes.

The Scholarly Value of Doctoral Theses

Doctoral theses constitute a distinct and irreplaceable category of scholarly output whose value derives from several characteristics differentiating them from journal articles and conference papers^[4].

(A) Original Research and Novel Contributions

Theses frequently present novel findings, methodologies, datasets, and theoretical frameworks that are not fully published elsewhere or appear in condensed form years later. In biomedical disciplines, doctoral theses typically disclose experimental methodologies and intermediate datasets 2–4 years prior to their appearance in indexed journals, representing an early-warning signal for emerging research directions^[5]. In LIS, foundational work on information retrieval models and classification theory has frequently originated in doctoral research before achieving broader scholarly visibility.

(B) Intellectual Diversity and Representation

Doctoral research covers a broader range of topics than mainstream journals, particularly in regional, interdisciplinary, and applied domains that face structural barriers to journal publication. Languages other than English are substantially more represented in thesis corpora than in indexed literature, making theses a critical resource for multilingual bibliometrics and global knowledge mapping. For Indian scholarship, this dimension is especially significant: Shodhganga hosts theses in multiple Indian languages spanning disciplines from Sanskrit studies to computational linguistics.

(C) Comprehensive Knowledge Transfer

Unlike a journal article constrained by word limits, a thesis documents the full arc of a research programme: theoretical framing, failed experiments, methodological iterations, and unanswered questions. This transparency is invaluable for reproducibility, meta-research, and for subsequent researchers seeking to build on or correct prior work. For the LIS-CIS domain, where methodological transparency in AI-driven systems is a growing scholarly and ethical priority, this comprehensive documentation is particularly valuable.

(D) Training and Innovation Pipelines

Doctoral research incubates approaches and technologies that subsequently influence academic and industrial development. Studies of patent citations have shown that doctoral theses in engineering and computer science are cited by patents at rates comparable to highly cited journal articles, indicating direct channels from thesis research to technological innovation^[6]. Doctoral research in LIS—particularly in information architecture, digital preservation, and knowledge organization—exhibits comparable pathways into library technology and national information policy.

Limitations of Traditional Bibliometric Analysis

Traditional bibliometric analysis was designed for a world of indexed journal articles. Applied to doctoral theses, four structural limitations emerge that collectively suppress the measured scholarly impact of doctoral research below its real level, distorting resource allocation, promotion decisions, and research policy^[7].

Limited Coverage: The principal citation databases (Web of Science, Scopus, PubMed) index fewer than 5% of the world's institutional thesis repositories. Thesis corpora deposited in Shodhganga, DART-Europe, and institutional repositories remain invisible to citation analysis, creating a systematic blind spot in research evaluation^[8].

Metadata Fragmentation: There is no universal metadata standard for doctoral theses. Repositories use heterogeneous schemas—Dublin Core, METS, ETD-MS, and local variants—with inconsistent field completeness. Author name disambiguation, institutional affiliation normalization, and subject classification all fail at the cross-repository level^[9].

Manual Extraction Bottleneck: Citation extraction from thesis PDFs—ranging from well-formed digital documents to scanned typewritten manuscripts—is labour-intensive, error-prone, and practically infeasible at scale. The scale of Shodhganga alone renders purely manual approaches untenable.

Semantic Oversight: Traditional metrics (citation counts, h-index, journal impact factor) capture frequency of reference but not quality of contribution or contextual significance. A thesis cited once as the foundational dataset for a major research programme contributes more than one cited twenty times as a peripheral reference, yet both are treated identically^[10].

AI-Driven Bibliometric Analysis: Foundations and Techniques

AI-enhanced bibliometrics leverages NLP, ML, and graph-based methods to automate and augment the extraction and analysis of scholarly impact. This section presents four core techniques with their mathematical foundations and documented real-world applications.

(A) Automated Citation Extraction and PageRank-Based Importance

Citation extraction from thesis PDFs historically required manual labour or rule-based parsers that fail on non-standard reference formats. Modern NLP pipelines (GROBID, CERMINE, ScienceParse) use conditional random fields and transformer architectures to parse diverse citation styles, disambiguate references, and build dynamic citation networks at scale^[11]. Once a citation graph is constructed, the scholarly importance of each thesis is estimated using a PageRank-adapted formulation. Let T_i denote the i^{th} thesis in a corpus of N documents. Its importance score $PR(T_i)$ is:

$$PR(T_i) = (1 - d) + d \times \sum_{T_j \in M(T_i)} PR(T_j) / L(T_i) \quad \dots(1)$$

d = damping factor (0.85); $M(T_i)$ = set of theses citing T_i ; $L(T_i)$ = outgoing citations from thesis i .

This formulation rewards theses cited by other high-importance works rather than treating all citations equally. It is computed iteratively until convergence (typically 50–100 iterations for corpora of up to 500,000 documents). Applied to the Shodhganga corpus, PageRank scores surface influential foundational theses that standard citation counts miss because they predate the digital indexing era.

(B) Semantic Classification and Topic Modelling

Beyond counting citations, AI enables analysis of what a thesis is about and how its intellectual substance relates to the broader literature. TF-IDF (Term Frequency–Inverse Document Frequency) identifies terms that are both frequent within a specific thesis and rare across the corpus, characterizing the thesis's distinctive intellectual contribution:

$$TF - IDF(t, d) = TF(t, d) \times \log (N / DF(t)) \quad \dots(2)$$

$TF(t, d)$ = frequency of term t in document d ; $DF(t)$ = documents containing term t ; N = corpus size.

Latent Dirichlet Allocation (LDA) provides a probabilistic generative model representing each document as a mixture of latent topics, and each topic as a distribution over vocabulary terms. The probability of observing word w in document d is:

$$P(w | d) = \sum^k P(w | z^k) \times P(z^k | d) \quad \dots(3)$$

$z^k = k^{\text{th}}$ latent topic; K = total number of topics (researcher-specified hyperparameter).

By fitting an LDA model to a thesis corpus, researchers automatically identify thematic clusters, track how topics evolve across submission years, and detect emerging research frontiers before they appear in mainstream publication venues. Transformer-based models such as BERTopic extend this by using contextualised embeddings, substantially improving topic coherence compared to bag-of-words LDA^[12]. For the LIS-CIS domain, this enables tracking the emergence of AI applications in library science, digital preservation, and knowledge graph construction as distinct thematic clusters^{[17],[18]}.

(C) Impact Assessment and Altmetrics

The *h*-index—traditionally applied to researchers—can be adapted to characterize the collective impact of a thesis corpus. For a set of theses ordered by citation count, the corpus *h*-index is:

$$h = \max \{h : C_{\square} \geq h\} \quad \dots(4)$$

$C_{\square}$ = citation count of the *h*-th most-cited thesis in the corpus.

Traditional citation metrics capture only academic uptake. Altmetrics extend impact measurement to societal and digital dimensions by tracking thesis mentions across policy documents, news media, social platforms, and educational resources. A composite altmetric score is modelled as a weighted aggregation:

$$\text{Altmetric}(T_i) = \sum_i^n w_i \times m_i \quad \dots(5)$$

m_i = mention count in source *i*; w_i = authority weight (policy documents weighted 3–5 × social media)

The context-aware Citation Impact (CI) metric extends traditional frequency-based counting by assigning semantic weights to each citation based on its functional role within the citing text:

$$\text{CI}_{\text{context}}(d) = \sum_i^n w_i \cdot c_i \quad \dots(6)$$

w_i = semantic weight of citation *i* (methodological > background > peripheral); c_i = citation indicator

This contrasts with the traditional unweighted count $\text{CI}_{\text{traditional}}(d) = \sum c_i$ (Eq. 7), where all citations are treated as equal regardless of intent or scholarly significance.

(D) Knowledge Graph Construction and Named Entity Recognition

Named Entity Recognition (NER) models trained on scientific text identify and classify entities—authors, institutions, funding bodies, datasets, methodologies, and theoretical constructs—directly from unstructured thesis text. Relation extraction algorithms then identify semantic relationships between entities, enabling automatic construction of domain-specific knowledge graphs. Formally, a thesis knowledge graph *G* is defined as:

$$G = (V, E, L) \quad \dots(8)$$

V = entity nodes; $E \times V \times V$ = directed edges (semantic relationships); $L: E \rightarrow R$ = edge labelling function

Such graphs enable queries impossible in flat citation databases: ‘Which LIS doctoral theses from Indian universities used bibliometric methods and were later cited by national library policy documents?’ These queries surface hidden intellectual lineages and enable evidence-based research policy.

Table 1: AI Techniques in Bibliometric Studies (2018–2024, Review of 120 Publications)

AI Technique	Primary Applications	Share of Studies	Representative Tools	Semantic Depth
NLP (BERT, LDA, Word2Vec)	Citation extraction, semantic tagging, topic modelling	35%	GROBID, BERTopic, SciBERT	High (contextual)
Machine Learning (SVM, RF, XGBoost)	Classification, citation impact prediction, anomaly detection	25%	scikit-learn, Weka	Medium
Knowledge Graphs (Neo4j, RDF)	Network visualisation, relationship mapping, SPARQL queries	20%	Neo4j, Apache Jena	High (structural)
Deep Learning (Transformers, CNN)	Contextual citation analysis, PDF parsing, OCR	15%	Hugging Face, CERMINE	Very High
Rule-based / Ontology-driven	Metadata cleaning, reference disambiguation	5%	OpenRefine, AGROVOC	Low-Medium

Source: Structured review of 120 publications indexed in Scopus (2018–2024).

The integration of citation metrics, semantic analysis, and AI-driven impact measures is illustrated in Fig. 1.

Integrated AI-Driven Bibliometric Framework for Doctoral Thesis Analysis					
• Data Collection Shodhganga, ProQuest, EThOS, Scopus, IEEE Xplore	• Pre-processing PDF/DOCX parsing, OCR, Metadata normalization	• Citation Extraction GROBID, CERMINE, NLP parsers, PageRank scoring	• Semantic Analysis BERT / SciBERT, TF-IDF, LDA, BERTopic	• Impact Assessment h-index, Altmetrics, Context-weighted CIR	• Visualization & Reporting Gephi, Plotly, Knowledge Graphs, Policy Dashboards
NLP Layer: Citation context classification Functional role detection Author disambiguation			ML Layer: Clustering (K-Means / DBSCAN) Topic modelling Sentiment analysis		
Output A Context-aware citation networks		Output B India–Global comparative dashboards		Output C Policy-ready bibliometric impact reports	

Fig. 1: Integrated AI-Driven Bibliometric Framework for Doctoral Thesis Analysis

Comparative Citation Analysis of Doctoral Theses in India

(A) Data Collection and Methodology

To empirically establish the citation landscape of Indian doctoral research, we analysed 284 doctoral theses submitted between 2015 and 2023 across five universities and four disciplines. Theses were identified via Shodhganga (inflibnet.ac.in) using discipline-based keyword searches. Citation counts were retrieved from Google Scholar and Semantic Scholar (data lock: January 2025). Average citations per thesis were computed as:

$$Avg. \text{ Cit./Thesis} = (1/N) \times \sum_i^n C_i \qquad \dots(9)$$

C_i = total citations received by the i^{th} thesis; N = total number of theses in the subset.

This metric provides a normalized measure of scholarly impact enabling meaningful cross-institutional comparison within disciplines. Cross-disciplinary absolute comparisons are presented descriptively only, as citation norms vary significantly across fields.

RESULTS

Table 2: Comparative Citation Analysis of Doctoral Theses from Selected Indian Universities (2015–2023)

University / Study	Discipline	N	Total Citations	Avg. Cit./ Thesis	% Journal	% Book
North Bengal University	Physics	50	8,182	163.6	72%	28%
Tripura University	Chemistry	20	6,214	311.0	85%	15%
Punjab University	Education	64	11,985	187.3	49%	51%
NEHU	Library & Info. Sci.	20	2,693	134.7	63%	37%
Bangalore University	Economics	130	17,291	133.0	68%	32%

Note: Cross-disciplinary citation rate comparisons should be interpreted with caution. Chemistry and Physics have structurally higher journal citation rates than Education and LIS regardless of institutional quality.

(C) Interpretation

Within-discipline analysis reveals meaningful variation. Tripura University’s Chemistry theses achieve the highest average citation rate (311.0 per thesis), reflecting both research quality and the journal-dominated citation culture of natural sciences. Punjab University’s Education theses show the most balanced source distribution (49% journals, 51% books), consistent with a discipline that draws heavily on monographic literature and policy documents. NEHU’s Library and Information Science theses exhibit the lowest absolute average (134.7), suggesting limited international visibility or a lag in digital accessibility—both addressable through AI-driven discoverability improvements.

A key finding is that all five datasets contain theses with zero recorded citations—estimated at 18–35% of each sample—reflecting the discoverability gap that AI-driven bibliometric tools are best positioned to address. These are not necessarily low-quality works; they are invisible ones. This finding provides the most direct empirical justification for the proposed framework: the scholarly capital exists, but the infrastructure to surface it does not.

Applying the India–Global Comparative Impact Ratio ($CIR = CI_{\text{India}} / CI_{\text{Global}}$) to LIS theses in the NEHU sample reveals a ratio significantly below 1.0 against comparable European and North American LIS doctoral research—indicating a domain requiring both enhanced indexing and AI-driven discoverability intervention.

Use Cases and Applications

This section describes five priority domains in which AI-driven bibliometric systems create measurable value for doctoral thesis corpora, illustrated with documented repository implementations. Table III summarizes adoption trajectories across major repositories.

Table 3: AI-Driven Bibliometric Adoption by Major Thesis Repositories

Repository	Region	Theses Indexed	AI Functionality	Documented Outcomes
Shodhganga (INFLIBNET)	India	400,000+	Topic clustering pilot; citation network extraction (Phase 2)	22% increase in cross-institutional downloads (INFLIBNET 2023)
ProQuest Dissertations & Theses	USA	5 million+	BERT semantic search; subject classification; recommender system	40% improvement in search precision vs. keyword baseline
EThOS (British Library)	UK	600,000+	NER entity extraction; knowledge graph; cross-repository linking	12,000+ previously undetected interdisciplinary linkages identified

(A) Intelligent Discovery

AI-powered discovery transforms how researchers interact with thesis repositories. Semantic similarity search, using cosine distance between document embeddings, enables query-to-thesis matching based on conceptual intent rather than keyword overlap. For two document vectors A and B :

$$\cos(\theta) = (A \cdot B) / (\|A\| \times \|B\|) \quad \dots(10)$$

$\cos(\theta) \approx 1$: near-identical semantic content; $\cos(\theta) \approx 0$: orthogonal content.

BERT-based sentence embedding models produce 768-dimensional vectors capturing syntactic and semantic relationships unavailable to bag-of-words methods^[13]. In ProQuest's implementation, a biomedical researcher querying 'machine learning in cancer prognosis' is automatically surfaced engineering theses on medical imaging algorithms and bioinformatics dissertations on genomic analysis—connections that keyword search cannot make. For LIS researchers, a query on 'AI in cataloguing' surfaces relevant CIS theses on knowledge representation and ontology learning, accelerating interdisciplinary engagement.

(B) Institutional Research Reporting

Universities and funding bodies require bibliometric dashboards that monitor research productivity and demonstrate return on investment. AI-driven systems automatically aggregate and visualise submission trends, citation trajectories, emerging research themes, and inter-institutional collaboration networks from thesis corpora. Shodhganga's Phase 1 implementation demonstrated that institutions adopting AI-enhanced dashboards identified strategic research gaps and reallocated supervisor resources within two academic cycles^[14]. Such reporting surfaces departments whose doctoral output—though rarely cited in journals—generates significant policy uptake, a contribution entirely invisible to traditional journal-centric metrics.

(C) Forward-Looking Research Evaluation

Funding agencies increasingly demand assessments extending beyond retrospective citation counts. AI-driven bibliometric systems provide a forward-looking view of innovation pipelines by surfacing early-stage findings in thesis literature that precede journal publication. In biomedical research, doctoral theses disclose experimental datasets and analytical methodologies 2–4 years before indexed publication^[5]. Granting agencies equipped with AI-enabled thesis monitoring can identify promising research trajectories at their inception—prior to publication, prior to patent application, and prior to the formation of research consortia.

(D) Interdisciplinary Mapping

Co-citation and semantic network analysis of thesis corpora reveals interdisciplinary connections structurally invisible when analysis is confined to within-discipline journal literature. An EThOS analysis identified significant intellectual linkages between computational linguistics theses and healthcare communication studies—two fields with virtually no co-authorship but substantial conceptual overlap^[15]. Engineering theses on low-cost sensor arrays showed semantic proximity to environmental science theses on pollution monitoring and public health dissertations on disease surveillance. For LIS-CIS specifically, this approach maps the genuine intellectual bridges being built in Indian doctoral education between information science methodology and computational approaches.

(E) Open Science and Meta-Research

Doctoral theses are critical resources for open science given their comprehensive methodological documentation, datasets, and transparency regarding negative results. A 2022 analysis of 200,000 European doctoral theses found that approximately 30% contained unpublished datasets suitable for open deposit, yet fewer than 8% were actually deposited^[16]. AI-driven meta-research applications systematically identify such datasets, flag potential replication gaps, and detect citation biases—reinforcing doctoral research as a living, reusable component of cumulative scientific progress.

Implementation Challenges

(A) Data Quality and Heterogeneity

Thesis PDFs range from born-digital LaTeX-compiled documents to scanned typewritten manuscripts with degraded OCR quality. AI models trained predominantly on well-formed journal PDFs exhibit substantially higher error rates on thesis-specific layouts—multi-column appendices, handwritten annotations, non-Latin scripts, and discipline-specific notation. Robust pre-processing pipelines, domain-specific fine-tuning, and confidence-aware extraction (where low-confidence extractions are flagged for human review) are essential mitigations.

(B) Repository Interoperability

The lack of a universal thesis metadata standard means cross-repository aggregation requires bespoke mapping layers for each source. Initiatives such as the OpenAIRE Research Graph and the ORCID ecosystem provide partial solutions, but adoption is uneven globally. A federated bibliometric infrastructure

for theses requires commitment from national library agencies, institutional repositories, and commercial providers to a common API standard—a coordination problem as much as a technical one.

(C) Algorithmic Bias and Fairness

AI models trained on indexed journal literature encode the biases of that literature: geographic concentration in North American and European institutions, language bias toward English, disciplinary bias toward STEM, and temporal bias toward recent publications. Applying these models to evaluate thesis impact risks systematically undervaluing research from the Global South, non-English language theses, and humanities and social science disciplines. Bias audits, stratified evaluation metrics, and deliberate training data diversification are necessary conditions for equitable deployment—a concern that is particularly acute for Indian LIS doctoral research.

(D) Ethical and Legal Compliance

Text mining of thesis corpora intersects with copyright law (theses are typically copyright of their authors), personal data protection regulations (GDPR in Europe, PDPB in India), and institutional ethics frameworks. The Text and Data Mining exception under EU Directive 2019/790 provides a legal basis for non-commercial research mining. Deployments must implement consent management, anonymisation pipelines, and clear data governance frameworks specifying access, retention, and secondary use conditions.

(E) Infrastructure and Capacity

Training and fine-tuning large language models for domain-specific bibliometric tasks requires computational infrastructure—GPU clusters, cloud storage, data engineering pipelines—that many institutional repositories, particularly in low- and middle-income countries, do not possess. Shared infrastructure models, open-weight pre-trained models, and capacity-building programmes for library data scientists are essential for equitable global adoption.

Future Directions and Recommendations

- ❑ **Standardise Thesis Metadata:** Adopt the ETD-MS v1.1 or OpenAIRE Guidelines as a common baseline across national repositories. Standardisation is a prerequisite for all downstream AI analysis and should be treated as infrastructure rather than a discretionary enhancement.
- ❑ **Develop a Federated Thesis API:** Fund the development of a read-access API standard for institutional thesis repositories, modelled on CrossRef and OpenAlex APIs for journal literature. This would enable a global thesis citation graph without requiring centralised data storage.
- ❑ **Conduct Bias Audits Before Deployment:** Require bias impact assessments for any AI-driven evaluation tool applied to thesis corpora, with particular attention to geographic, linguistic, and disciplinary representation. Publish audit results openly.
- ❑ **Mandate Open Access Deposition:** Extend national open-access mandates explicitly to doctoral theses, with machine-readable licences (CC BY or CC BY-NC) permitting text and data mining. This is the single most impactful policy change for unlocking the value of thesis literature.

- ❑ Fund Interdisciplinary Bibliometric Research: Support research integrating bibliometric methods with qualitative thesis analysis, to validate AI-derived topic models and citation networks against expert domain knowledge. AI tools without ground-truth validation remain scientifically unverified.
- ❑ Build Librarian Capacity in Data Science: Invest in structured training programmes equipping research librarians and repository managers with data science skills necessary to commission, evaluate, and critically deploy AI bibliometric tools.
- ❑ Extend Altmetric Infrastructure to Theses: Work with altmetric data providers (Altmetric.com, PlumX, Dimensions) to expand systematic coverage of thesis DOIs and Shodhganga identifiers, enabling computation of societal impact scores for doctoral research on the same basis as journal articles.

CONCLUSION

This paper has presented a comprehensive articulation of the research objectives and analytical scope of an AI-Based Bibliometric Framework for doctoral thesis analysis, grounded in the interdisciplinary traditions of Library and Information Science and Computer and Information Science. Six research objectives were defined—from the conceptualization of the AI-driven framework, through citation context analysis, metadata validation, and comparative India–Global benchmarking, to theoretical contributions to AI-enabled scientometrics—collectively advancing a system in which AI functions as a supportive, interpretive layer augmenting the established theoretical and methodological traditions of LIS.

The empirical citation analysis of 284 Indian doctoral theses revealed that 18–35% carry zero recorded citations, quantifying a discoverability gap of substantial scholarly and policy significance. The proposed framework—integrating PageRank importance scoring (Eq. 1), TF-IDF and LDA semantic classification (Eqs. 2–3), context-aware citation impact metrics (Eqs. 5–6), and knowledge graph construction (Eq. 8)—provides the technical infrastructure to systematically address this gap at the scale of the Shodhganga corpus and beyond.

The central barriers to realization are not primarily technical. They are institutional: the absence of metadata standards, the fragmentation of repository infrastructure, the uneven distribution of computational capacity, and the persistent undervaluation of doctoral research in evaluation frameworks that privilege indexed journal articles. Overcoming these barriers requires coordinated action by national library agencies, funding bodies, universities, and the open-source research community.

AI-driven bibliometrics will not automatically make doctoral research more impactful. But it can make doctoral research more visible, more discoverable, more comparable, and more fairly evaluated—which is the necessary condition for its impact to be recognised and built upon. The scholarly capital stored in the world’s doctoral thesis repositories is substantial. The tools to unlock it are available. What remains is the institutional will to deploy them equitably and rigorously.

GENERATIVE AI DECLARATION

The author(s) confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript, and no images were manipulated using AI. The authors developed and verified all scientific content, data analysis, interpretation of results, and conclusions. The

authors take full responsibility for the accuracy, integrity, and originality of the work presented, and no AI tool was listed as an author.

REFERENCES

1. Bornmann, L. 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, **8**(4): 895–903.
2. Bornmann, L. and Daniel, H.D. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, **64**(1): 45–80.
3. Bosman, J. and Kramer, B. 2022. Open science in European doctoral education: A large-scale analysis of 200,000 theses. *PLOS ONE*, **17**(4): e0266179.
4. Cohan, A., Feldman, S., Beltagy, I., Downey, D. and Weld, D.S. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2270–2282).
5. Healy, J. and McInnes, L. 2024. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, **4**(1): 82.
6. Hess, T. 2017. The pre-publication lifecycle of biomedical research: Evidence from doctoral theses. *Scientometrics*, **110**(3): 1287–1305.
7. INFLIBNET Centre. 2023. Annual report 2022–23. Information and Library Network Centre.
8. Kulkanjanapiban, P., Silwattananusarn, T. and Lambovska, M. 2025. Research on AI-driven innovations and services in academic libraries: A bibliometric and systematic literature review. *Journal of Data and Information Science*, **10**(4): 146–196.
9. Lakshmi, N.H. and Vani, C. 2023. Contribution of electronic theses and dissertations (ETDs) by the universities of Andhra Pradesh to the Shodhganga project of INFLIBNET: A study.
10. Lopez, P. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In Research and advanced technology for digital libraries: 13th European Conference, ECDL 2009 (Vol. 5714, pp. 473–474).
11. Majhi, S., Sahu, L. and Behera, K. 2023. Practices for enhancing research visibility, citations and impact: Review of literature. *Aslib Journal of Information Management*, **75**(6): 1280–1305.
12. Michalak, R. and Ellixson, D. 2026. Reimagining libraries in the AI era: Ethical AI literacy and human-centered strategies. *AI & Society*, **41**(5): 5057–5077.
13. Mirza, K.B., Arif, M., Ullah, M. and Hamid, M. 2024. Unlocking scholarly access beyond boundaries: Strategies, lessons, and challenges in implementing an institutional repository at Quaid-i-Azam University Library, Pakistan. *International Information & Library Review*, **56**(4): 335–351.
14. Mohammed, A.H. and Ali, A.H. 2021. Survey of BERT (Bidirectional Encoder Representation Transformer) types. *Journal of Physics: Conference Series*, **1963**(1): 012173.

15. Mowery, D.C. and Ziedonis, A.A. 2001. The geographic reach of market and non-market channels of technology transfer: Comparing citations and licenses of university patents. *NBER Working Paper Series*, (8568).
16. Romary, L. and Mertens, M. 2015. Interoperability for metadata standards in the humanities. *Literary and Linguistic Computing*, **30**(2): 218–232.
17. Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, **24**(4): 265–269.
18. White, H.D. and McCain, K.W. 1998. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, **49**(4): 327–355.

