# A Topic Modeling Study on Contemporary Latent Topics and Current Trends in Open Access LIS Journals listed in DOAJ

**Abhijit Thakuria** **Parimita Bezbaruah** **Dipen Deka**

**Abstract**

*The study employed Latent Dirichlet Allocation (LDA) topic modeling to analyze bibliographic data (Titles, Abstracts, and Keywords) of 1640 articles published during 2019-2023 from Open Access LIS journals listed in the DOAJ platform and indexed in Scopus database. The study utilized R statistical platform and used coherence method to determine optimal number of topics k=8. This analysis unveiled eight latent topics, with Topic 8 'Library Management' emerging as the most prevalent topic with prevalence score of 23.291. Moreover, Topic 6 'Knowledge Management' and Topic 4 'Bibliometrics' were identified as stable and popular topics, with respective Annual Growth Rates (AAGR) of 15.99% and 12.39% in OA LIS literature.*

**Keywords:** Open Access, Topic Modeling, LDA, Journals

## 1. Introduction

The scholarly communication landscape has shifted towards openness, accessibility and collaboration, propelled by the Open Access movement that aims to eliminate cost barriers and enhance research dissemination. Open Access (OA) literature, as defined by Suber (2012), is freely accessible online, fostering unrestricted access to scholarly content. The emergence of OA journals has revolutionized scholarly communication, offering unrestricted access without financial or technical barriers. The Directory of Open Access Journals (DOAJ) is a popular platform indexing high-quality, peer-reviewed research across diverse disciplines, promoting visibility and accessibility of open access journals globally. In the Library and Information Science (LIS) domain, OA publications are gaining popularity, with major publishers expanding their OA offerings. Barik and Jena (2019) revealed OA LIS journals have shown better visibility and growth than non-OA journals. Moreover, papers per year and total citations of OA LIS journals have higher citation rates compared to non-OA LIS journals (Chen and Du, 2016). The vast array of scholarly literature and interdisciplinary nature of LIS journals poses challenges in identifying trends and topical foci. Traditional methods like literature review and content analysis can be time-consuming and resource-intensive. This has led researchers to turn to computational methods such as text mining and topic modeling, like Latent Dirichlet Allocation (LDA) (Blei et al. 2003), to reveal hidden patterns in scholarly literature. Topic modeling analysis reveals underlying latent semantic structure within large collection of text or collection of documents.



**Corresponding Author**
Abhijit Thakuria
Email: abhijitthakuria97@gmail.com

It automatically clusters frequently co-occurring words across documents to identify group of words, which represent distinct topics (Blei, 2012). This shift towards OA and computational methods reflects a transformative trend in scholarly communication, enhancing accessibility and knowledge dissemination.

## 2. Review of Literature

Majhi and Mukherjee (2024) conducted a study on research trends in three Scopus-indexed Indian LIS journals from 2011 to 2022. The study analyzed 1213 titles and abstracts using LDA topic modeling, revealing common themes like 'Library users' studies' and 'bibliometric indicators'. Additionally, journal-specific topics such as technological innovation, resource utilization, and network analysis were identified. Over time, interests in digital libraries, global output analysis, and online search strategies remained concurrent, while research on academic library resources, electronic resource usage, and open access showed declining interest. Saha and Ghosh (2023) analyzed 736 Open Access LIS articles from Indian and international Scopus indexed journals, revealing that in India, citation analysis, bibliometrics, and scientometrics were prominent, while globally, users' perspectives, information use, data management, open access, library services, and information literacy were emphasized. Cross-country comparisons highlighted socio-cultural differences in LIS research topics, indicating a growing trend of cross-cultural adoption. Kumar and Thakur (2022) studied 1271 doctoral level LIS theses in India, identifying 'ICT and its applications' and 'Information seeking behaviour' as popular research topics, with declining interest in Bibliometrics and webometrics analysis over time.

Miyata et al. (2020) analyzed 1648 research articles from key LIS journals, noting significant topic shifts between 2000-2002 and 2015-2017 due to internet technology advancements. Information retrieval declined while Bibliometrics persisted, with decreasing topic diversity observed over time. Han (2020) analyzed 14,035 documents from 1996 to 2019, observing a decline in library-related research in LIS but stable topics like bibliometrics, particularly citation analysis, and Information Retrieval transitioning to model-based text processing. Information seeking and behaviour remained consistent. Figuerola et al. (2017) used the LISA database to extract bibliographical data (titles and abstracts) of LIS publications from 1978 to 2014. The study identified 19 significant topics grouped into processes, information technology, library studies, and specific information applications.

## 3. Objectives of the study

i. To identify contemporary latent topics in OA LIS journals.

ii. To determine the percentage of research proportion of latent topics in OA LIS journals during the period 2018-2023.

iii. To find the Annual Growth Rate (AGR) and Average Annual Growth Rate (AAGR) of latent topics during the period 2018-2023.

## 4. Methodology

### 4.1 Selection of Journals

The present study selected twelve open access LIS journals indexed in the Scopus database and which were also listed in DOAJ. As LIS is highly interdisciplinary in nature, only core LIS Open Access journals are selected which are not influenced by publications from other disciplines. Furthermore, the prestige and journal quartiles of the selected journals were assessed using the widely recognized Scimago Journal Ranking (SJR), which relies on SCOPUS indicators for its journal rankings. The selected journals are listed in Table-1.

**Table 1: List of Selected OA LIS Journals**

| Sl. no | Journal Titles | Total | Cumulative Publications | Percentage % |
|---|---|---|---|---|
| 1 | Journal Of Data And Information Science | 126 | 126 | 8% |
| 2 | Liber Quarterly | 74 | 200 | 5% |
| 3 | College And Research Libraries | 246 | 446 | 15% |
| 4 | Information Technology And Libraries | 126 | 572 | 8% |
| 5 | Evidence Based Library And Information Practice | 199 | 771 | 12% |
| 6 | Information Research: An International Electronic Journal | 184 | 955 | 11% |
| 7 | Insights: The Uksg Journal | 107 | 1062 | 7% |
| 8 | Journal Of Information Literacy | 97 | 1159 | 6% |
| 9 | Open Information Science | 63 | 1222 | 4% |
| 10 | Annals Of Library And Information Studies | 118 | 1340 | 7% |
| 11 | Journal of Librarianship and Scholarly Communication | 80 | 1420 | 5% |
| 12 | International Journal of Information Science and Management | 220 | 1640 | 13% |
| | **Total** | **1640** | | **100%** |

### 4.2 Data Extraction and Selection Criteria

The data for this study were extracted from the Scopus database for 1640 articles published in 12 Open Access LIS journals indexed in the Scopus database and listed in DOAJ for the period of 5 years (2019-2023). The primary intent of selecting a shorter period is to assess the latent topics of contemporary LIS literature and recent trends in open access avenues. The bibliographic data of 1640 articles containing Titles, Abstracts, keywords and years of publications were extracted in CSV format.

### 4.3 Data Analysis Tools

The study employed R statistical and data analysis tools for Data analysis and, along with MS Excel for preparation of the data. Moreover, the Latent Dirichlet Allocation (LDA) algorithm is used for topic modeling analysis on the R platform, supported by important R packages such as 'topicmodels', 'dplyr', 'LDAVis', 'LDAshiny' and 'ggplot2'.

### 4.4 Data Preprocessing

Data preprocessing is an essential step in preparing raw data for analysis. To remove noise and inconsistencies, the dataset was cleaned and transformed into lowercase text, removing numerals, special symbols, stop words and custom words. The necessary R libraries used for tidying the dataset were 'dplyr', 'tidyr' and 'tibble'. Further, the dataset was stemmed and lemmatized using 'tm' and 'textstem' R libraries, which is a vital procedure for Natural Language Processing (NLP) or text analysis to reduce text to its common base form. This process is important for ensuring accurate and consistent results.

## 5. Analysis and Findings

### 5.1 Selecting the optimal k Value

LDA Topic modeling analysis is widely used in the R statistical and data analysis platform. Title, abstract, and keywords of each article were selected for text analysis. Subsequently, the corpus was pre-processed and a document-term matrix (DTM) was generated. The selection of an optimal number of topics ('k' value) in an LDA model is challenging and typically user-defined; however, the effectiveness of topic classification mostly depends on selecting the optimal number of topics or 'k' value. The popular Coherence method was used and a coherence graph was generated exhibiting optimal values range from k=2 to k=50 (Figure-1) to identify the suitable number of topics for the corpus.

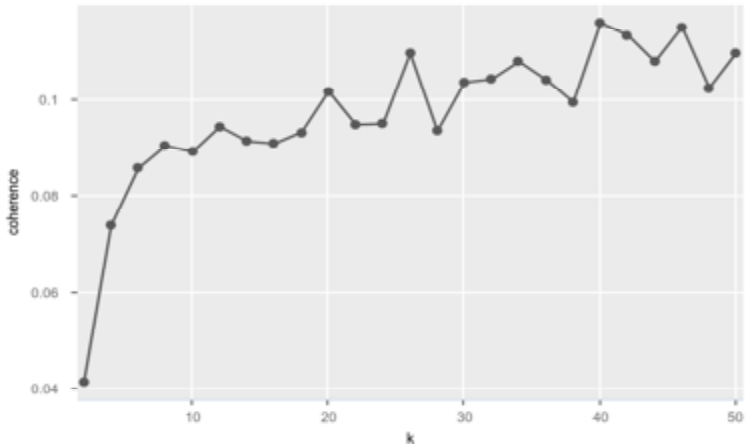


**Figure 1: Coherence graph for optimal values range k=2 to k=50**

The coherence scores for determining an optimal number of topics were computed on the preprocessed dataset in the R environment. The parameters were set at a thousand iterations with a hundred burn-ins and a hyper parameter á of 0.1. Subsequently, high coherence values were observed at 8, 12, 20, 26 and 30 topics. Additionally, the 'LDAvis' library was utilized to visualize topic relationships, with topics represented as circles on a 2D map presented in Figure-2. The Optimal topic values ranging from k=2 to k=30 were explored to minimize overlap and enhance topic separation. After a thorough iterative process of visualization and considering coherence scores, the optimal number of topics was determined to be k=8.

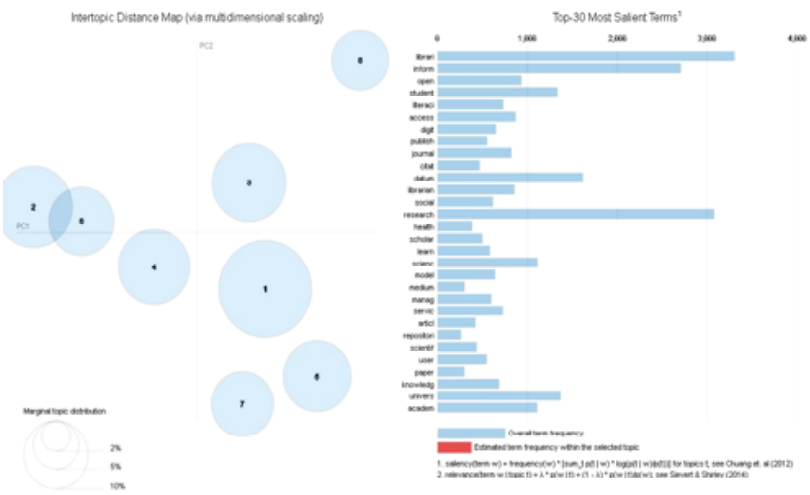


**Figure 2: OA dataset (k=8) 2D topic visualisation and top 30 prevalent terms**

### 5.2 Latent Topics with Top 15 Terms

The LDA Topic Model is trained in R software using popular topic models R library 'topicmodels' setting the optimal number of topics at k=8 and using Gibbs Sampling, with the alpha value set to 0.1 and the seed value set to 1. The LDA model organizes a collection of words into different topics represented as strings. Each word within these strings carries a specific weightage, contributing to the overall association with the topic. LDA facilitates the identification of latent topics based on the association and weightage of these words. The output of the LDA model, comprising the top 15 terms for each topic, is visualized using the 'ggplot2' R package. Furthermore, Figure-3 highlights the 'tidytext' R library beta values for each term.

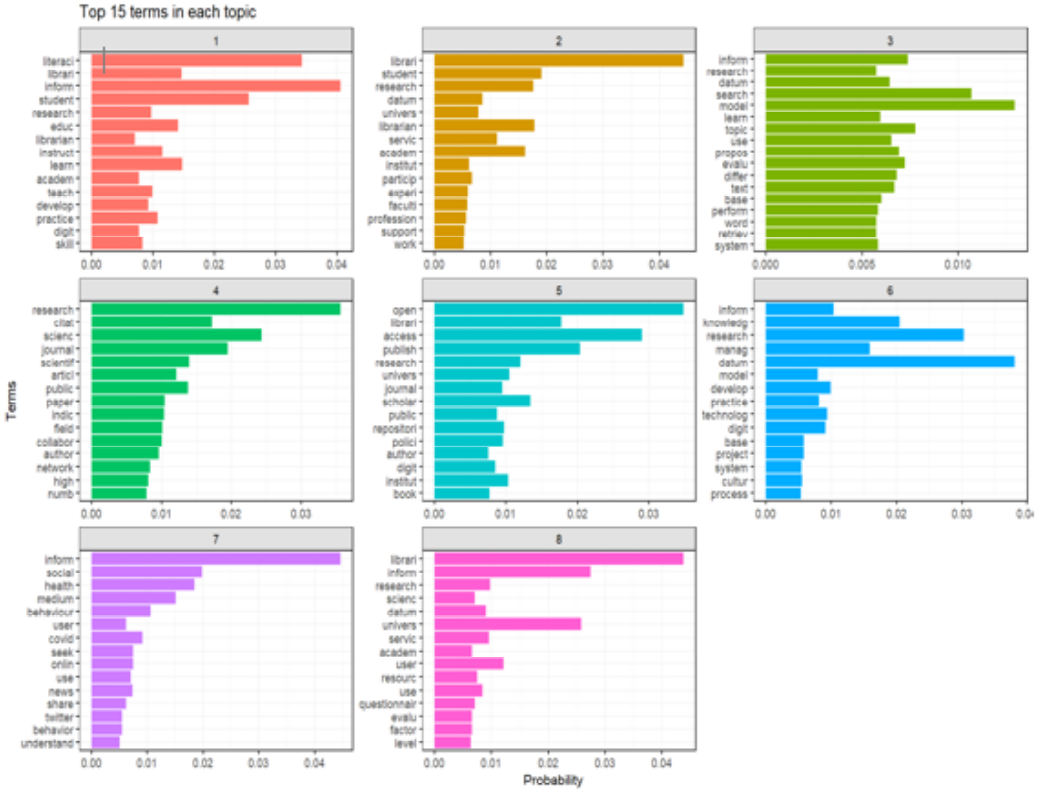


**Figure 3: Identified Latent topics with beta values**

### 5.3 Labeling of Topics

The topics extracted through topic modeling can be challenging to understand, and therefore an expert human interpretation is needed to evaluate and label the topics. To address this, our study used word-intrusion and topic-intrusion methods introduced by Chang et al. (2009) and Newman et al. (2010). These methods involve human evaluation of the top terms of the themes to ensure coherence and meaningful association within the corpus. By enhancing the interpretability of the topics, these methods improve comprehension of the underlying themes in the text data.

The topics were labelled by analysing the top 10 terms of each topic highlighted in Table-2 using word intrusion and topic intrusion methods. The coherence and prevalence scores are presented for each identified topic. Evaluating these scores is important for assessing the quality of the identified topics which helps in accurate labeling. The topic 'Open Access and Scholarly Communication' (T_5) has the highest coherence score of 0.176, followed by 'Information Literacy' (T_1) and 'Academic Library Services' (T_2), indicating that words within these topics are semantically coherent and represent distinct concepts. Conversely, the topic 'Knowledge Management' (T_6) has the lowest coherence score of 0.042, suggesting that the words within this topic are weakly semantically coherent. Whereas the topic 'Library Management' (T_8) has the

highest prevalence score of 23.291, which indicates the topic is discussed frequently across the entire OA corpus, followed by 'User Studies and Perception' (T_2) and 'Bibliometrics' (T_4). Interestingly, despite having the highest prevalence score, 'Library Management' (T_8) has a lower coherence value.

**Table 2: Labeling of identified latent topics**

| Topics | Topic Labels | Coherence | Prevalence | Top 10 terms |
|---|---|---|---|---|
| T_1 | Information Literacy | 0.142 | 10.345 | students, information, literacy, learning, education, skills, instruction, student, librarians, teaching |
| T2 | Academic Library | 0.135 | 18.081 | library, academic, librarians, services, methods, Services university, public, research, participants, students |
| T_3 | Information Retrieval and Machine Learning | 0.047 | 10.18 | search, data, model, information, text, web, models, semantic, topic, retrieval |
| T_4 | Bibliometrics | 0.11 | 12.517 | research, articles, journals, papers, citation, publications, journal, researchers, citations |
| T_5 | Open Access and Scholarly Communication | 0.176 | 8.446 | open, access, publishing, scholarly, policies, journals, publishers, journal, academic, copyright |
| T_6 | Knowledge Management | 0.042 | 8.011 | knowledge, information, university, data, research, universities, management, model, technology, content |
| T_7 | Information Seeking Behaviour | 0.068 | 8.929 | information, social, health, media, behaviour, online, news, users, seeking, sharing |
| T_8 | Library Management | 0.047 | 23.491 | data, digital, library, management, development, practices, services, process, literature, collections |

## 6. Current Trends of Latent Topics

The study employed the R library 'LDAshiny' with a dataset consisting of the year of publication of articles. The analysis was conducted in terms of the proportion of research publications published for each latent topic during the study period 2019-2023. Figure-3 presents the percentage of research proportion for each latent topic for each year. The study revealed that Topic 8, 'Library Management' and Topic 2, 'Academic Library Services' are dominant topics in OA LIS literature. The percentage of research proportion for Topic 8 is between 20-25%, and for Topic 2, it ranges from 15-20% across all the identified latent topics. In contrast, Topic 1 'Information Literacy', Topic 5, 'Open Access and Scholarly Communication' Topic 6, 'Knowledge Management' and Topic 7, 'Information Seeking Behaviour' have lower percentage of research proportion

ranging between 9-10%, 6-9%, 4-10%, and 7-11%, respectively, during the study period. However, Topic 1, Topic 5 and Topic 7 are stable and persistent topics but sometimes exhibit downward trends during the study period. It was observed that Topic 6 is a gradually emerging topic in OA LIS literature. Moreover, Topic 4, 'Bibliometrics' remained a persistent topic and gradually emerged in terms of research output every year. Furthermore, the study observed Topic 3, 'Information Retrieval and Machine Learning' as a stable and newly emerging topic with a percentage of research proportion ranging between 10-13%. Further, the present study revealed that Topic 6, Topic 5, and Topic 7 had positive Average Annual Growth Rate (AAGR) of 15.99%, 12.39%, and 3.49%, respectively, during the study period presented in Table-3. In contrary, all other topics exhibited negative AAGR. It was observed that T4 'Bibliometrics' had a positive AGR each year, indicating it as a popular and stable topic. Moreover, it was observed that Topic 7 had an Annual Growth Rate (AGR) of 50.99% in 2021-22, which was the highest in a single year across all the topics.

**Table 3: AGR and AAGR of latent topics**

| Topic | Topic Labels | Annual Growth Rate of Topics | | | | AAGR |
|---|---|---|---|---|---|---|
| | | 2019-20 | 2020-21 | 2021-22 | 2022-23 | |
| T1 | Information Literacy | -6.84% | -0.19% | -7.17% | 1.65% | -2.51% |
| T2 | Academic Library Services | 1.07% | -9.06% | -23.97% | 13.53% | -3.69% |
| T3 | Information Retrieval and Machine Learning | -16.04% | 0.99% | -1.51% | -7.94% | -4.90% |
| T4 | Bibliometrics | 28.44% | 14.08% | 9.78% | 9.62% | 12.39% |
| T5 | Open Access and Scholarly Communication | -12.53% | 11.81% | -27.96% | 23.85% | -0.96% |
| T6 | Knowledge Management | 27.76% | 33.07% | 25.81% | -6.68% | 15.99% |
| T7 | Information Seeking Behaviour | -9.35% | -1.93% | 50.99% | -22.28% | 3.49% |
| T8 | Library Management | 1.47% | -10.57% | 5.12% | -4.40% | -1.68% |

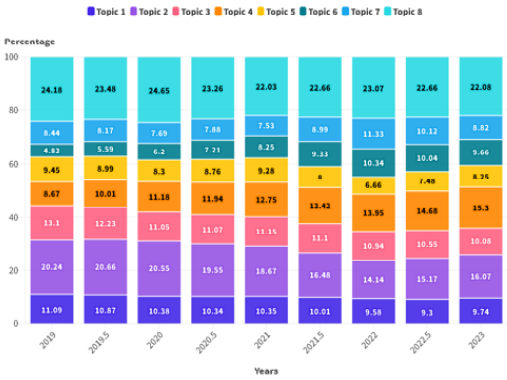


**Figure 3: Column percentage chart of trending OA latent topics**

## 7. Discussion

The present study identified eight latent topics, as shown in Figure-3 and labeled in Table-2. Comparing our findings with prior studies using LDA models by Miyata et al. (2020) and Figuerola et al. (2017), as well as other methods such as content analysis (Jarvelin and Vakkari, 2022), co-citation analysis (Åström, 2007), keyword analysis (Papiã and Buhin, 2019), and bibliometric methods (Chang et al., 2015), the findings of latent topics in LIS align with the present study. Jarvelin and Vakkari (2022) noted the increasing trend of research in Bibliometrics, while Han (2020) found it to be a stable topic in LIS. Our study noticed the emerging trend of T4 'Bibliometrics' which aligns with prior findings. Comparing prior studies by Figuerola et al. (2017) revealed Library Management as one of the vital LIS research clusters, aligning with our findings as Topic 8 'Library Management' has the highest proportion of research in OA LIS journals. Furthermore, Saha and Ghosh (2023) identified Library Services as a prominent LIS topic, consistent with our findings where Topic 2, 'Academic Library Services,' emerged as a dominant research area. Further, Figuerola et al. (2017) mentioned the growth of topic Knowledge Management in LIS due to the adaptation of paperless office and electronic administration, as our study observed Topic 6 'Knowledge Management' as a gradually emerging topic in OA LIS literature recently.

Miyata et al. (2020) asserted Scholarly Communication as a dominant topic, which aligns with our finding of Topic 5 'Open Access and Scholarly Communication' as a consistent and stable topic. Chang et al. (2015) employed bibliometric methods, Han (2020) employed LDA, revealing Information Seeking Behavior and Information Literacy as constant topics in LIS, aligning with the findings of our study as Topic 7, 'Information Seeking Behavior' and Topic 1, 'Information Literacy' as stable and persistent topics in OA LIS avenues. Tuomaala et al. (2014) determined Information Retrieval (IR) as a significant topic in LIS, which resonates with our findings. Our study revealed Topic 3, 'Information Retrieval and Machine Learning' indicating a gradual emergence of machine learning research in open access LIS platforms. However, topics such as governance, big data, tweet analysis, and social media, noted in previous studies (Han, 2020; Papiæ and Buhin, 2019), were absent in OA LIS research, suggesting a lack of diversity in topics.

Open access research literature enhances the accessibility of scholarly knowledge, providing a platform for researchers to present their findings and engage in interdisciplinary collaboration. This study unveils valuable insights into prominent and prevalent topics in OA LIS literature published across various sources. The identified latent topics will help LIS researchers and practitioners understand the research landscape and determine emerging, declining, and historical trends in OA avenues. Researchers can use these topics to identify gaps in the literature and find potential research areas that align with current trends. These insights will aid funding bodies in prioritizing under-researched or emerging areas within LIS. Furthermore, policymakers can make informed decisions based on these trends, aligning their priorities with the innovation and growth of the LIS domain.

## 8. Conclusion

Topic modeling analysis offers valuable insights into the predominant themes and concepts within a subject. Open access research literature plays a crucial role in breaking barriers to scholarly knowledge accessibility,

providing a broad platform for researchers to share findings and engage in interdisciplinary collaboration. Analyzing open access research through topic modeling aids in identifying emerging areas and shifts in interest within the LIS domain. It serves as a vital resource for independent or novice researchers lacking institutional access or funding, facilitating their research journey. Additionally, such analysis assists funding agencies in aligning priorities with the current needs of the domain, fostering innovation and growth. Despite core research areas persisting within OA LIS literature, there is a noticeable lack of depth and diversity in broader topics such as governance, text analysis, the Internet of Things, and artificial intelligence.

**References**

1. Åström, F. (2007). Changes in the LIS research front: Time sliced cocitation analyses of LIS journal articles, 1990–2004. Journal of the American Society for Information Science and Technology, 58(7), 947–957. https://doi.org/10.1002/asi.20567
2. Barik, N., & Jena, P. (2019). Visibility and growth of LIS research publications: A Scopus based analysis of select open access journals during 2001 to 2015. Library Hi Tech News, 36(7), 1–11. https://doi.org/10.1108/LHTN-05-2019-0035
3. Blei, D. M., Ng, Y. A., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. Accessed on 10th May 2013 from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
4. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems, 22. Retrieved on April, 23rd 2024, from https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html
5. Chang, Y.-W., Huang, M.-H., & Lin, C.-W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. Scientometrics, 105(3), 2071–2087. https://doi.org/10.1007/s11192-015-1762-8
6. Chen, M., & Du, Y. (2016). The status of open access library and information science journals in SSCI. The Electronic Library, 34(5), 722–739. https://doi.org/10.1108/EL-05-2015-0070
7. Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. Scientometrics, 112(3), 1507–1535. https://doi.org/10.1007/s11192-017-2432-9
8. Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. Scientometrics, 125(3), 2561–2595. https://doi.org/10.1007/s11192-020-03721-0
9. Järvelin, K., & Vakkari, P. (2022). LIS research across 50 years: Content analysis of journal articles. Journal of Documentation, 78(7), 65–88. https://doi.org/10.1108/JD-03-2021-0062
10. Kumar, V., & Thakur, K. (2022). Using text analysis to study doctoral-level library and information science research trends in India. Annals of Library and Information Studies, 69(3), 191–202. https://doi.org/10.56042/alis.v69i3.58719

11. Majhi, D., & Mukherjee, B. (2024). Analysing Library and Information Science Articles Using Topic Modeling Approaches. DESIDOC Journal of Library & Information Technology, 44(2), 114–123. https://doi.org/10.14429/djlit.44.2.19312

12. Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modeling and visualization. Scientometrics, 125(1), 665–687. https://doi.org/10.1007/s11192-020-03657-5

13. Mukherjee, B. (2009). Scholarly research in LIS open access electronic journals: A bibliometric study. Scientometrics, 80(1), 167–194. https://doi.org/10.1007/s11192-008-2055-2

14. Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. Proceedings of the 10th Annual Joint Conference on Digital Libraries, 215–224. https://doi.org/10.1145/1816123.1816156

15. Papiæ, A., & Buhin, M. (2019, May). Mapping the Hot Topics in Library and Information Science Field in Period 2015-2018 Year. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 508-513). IEEE.

16. Saha, M., & Ghosh, S. (2023). Topic Modelling in Library and Information Science from the Primary Data: Swing in Thrust Areas. International Journal of Information Science and Management (IJISM), 21(3). https://doi.org/10.22034/ijism.2023.1977569.0

17. Suber, P. (2012). Ensuring open access for publicly funded research. BMJ, 345(aug08 1), e5184–e5184. https://doi.org/10.1136/bmj.e5184

18. Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. Journal of the Association for Information Science and Technology, 65(7), 1446–1462. https://doi.org/10.1002/asi.23034

**About Authors**

**Abhijit Thakuria,** Research Scholar, Gauhati University, Guwahati, Assam
Email: abhijitthakuria97@gmail.com
ORCID: https://orcid.org/0000-0002-3852-1982

**Parimita Bezbaruah**, Research Scholar, Gauhati University, Guwahati, Assam
Email: bezparimita@gmail.com
ORCID: https://orcid.org/0009-0008-5768-7495

**Dr. Dipen Deka**, Associate Professor, Gauhati University, Guwahati, Assam
Email: dipendeka@gauhati.ac.in
ORCID: https://orcid.org/0000-0002-2226-8839