Afr. J. Lib, Arch. and Inf. Sc. Vol. 36, No. 1 (January 2026) 174-186

https://doi.org/10.5281/zenodo.21232460

Information Retrieval Challenges in Multilingual Literary Archives

Xiaoliang Wen[1]*

[1]School of Oriental Languages, Mudanjiang Normal University, Mudanjiang, Heilongjiang, 157011, China.

**Corresponding Author: Xiaoliang Wen (xiaoliangwen1@yeah.net)*

Abstract

Multilingual literary archives play a vital role in preserving cultural heritage and helping to conduct comparative research. But there is a significant problem in retrieving information on these archives. This paper discussed the major issues of retrieval in multilingual literature collections, specifically linguistic diversity, semantic ambiguity, metadata inconsistency, and technology constraints. The literature review indicates that morphological variation, polysemy, and culturally embedded meanings all are significant challenges based on a qualitative study of the literature available. They make indexing, querying and cross linguistic retrieval even more difficult. High and low resource languages also cause inequitable performance and limit access to marginalized literary traditions. The paper identifies the destructive nature of poor, culturally blind metadata standards. It further demonstrates that machine translation and key word-based systems tend to overlook the interpretive and contextual layer of the literature. These problems require an interdisciplinary approach. It can be assisted by the combination of computational linguistics, digital humanities, and literary studies. It is also important to develop inclusive metadata structures and development of sophisticated semantic retrieval frameworks. This study contributes to the current debate to transform multilingual libraries into more equitable, findable, and viable. Furthermore, this paper also explains the various application on information retrieval system. This paper recommends that the technical, institutional, and ethical solutions need to be combined to overcome the issues of information retrieval in the multilingual literary archives.

Keywords: Information Retrieval Challenges (IRC), Multilingual Literary (ML), Archives (AA).

Introduction

Information retrieval in multilingual literary archives is designed by a complex relationship of textual ambiguity, linguistic diversity, user-centered constraints, and technological limitations. Literary collections consist of texts whose meanings are culturally embedded, layered, and often unaffected to homogenous symbol (Xia, 2024). When these collections develop multiple languages, the retrieval process becomes significantly more challenging, demanding systems to manage variation not only in grammar and vocabulary but also in interpretive depth and cultural expression. One of the most significant challenges is linguistic heterogeneity. Almost all languages have different morphological richness, syntactic organization, and semantic structures, which can cause complications in fundamental operations of retrieval, including tokenization, query matching, and indexing (Bringmann and Zhukova, 2024). For example, highly modified languages generate multiple word forms from a single root, which make it ineffective to exactly match terms. In multilingual archives, retrieval systems process various kinds of linguistic systems at the same time, which often leads to inconsistent

performance across languages and decreases the accuracy of retrieval for under-resourced or linguistically complex languages (Lo, 2024). Another similar information retrieval challenge is related to semantic ambiguity and polysemy that are especially common in literary texts. Literary words are often multi-faceted and can be determined by context, fiction, or historical context (Kim and Lee, 2024). When these texts are disseminated across languages, it becomes increasingly harder to align semantics. The literal translation usually does not store information that is metaphorical or symbolic, and this leads to the retrieval systems that emphasize surface-level similarity of lexical meanings but not conceptual similarity. This shortcoming confines the accessibility of users to texts using thematic or symbolic specifications. Furthermore, cross-lingual information retrieval is also considered an important challenge (Xia, 2024). Many users may post queries using one language and request documents written in a different language. That is why the system needs to determine semantic equivalence across languages. This process can be facilitated by automated translation tools, which, however, usually fail to deal with idiomatic expressions, archaic language, and stylistic elements in literary works. Therefore, cross-language searches can find irrelevant information or miss essential texts that convey similar concepts with the help of culturally specific words (Gnoli et al., 2024). In addition, the other most significant challenge which information retrieval systems has been facing nowadays is lack of standardized and culturally sensitive metadata. It complicates retrieval in multilingual literary archives (da Silva et al., 2024). Metadata is necessary to describe, categorize, and arrange archival material, but current standards are often based on predominant linguistic or cultural paradigms. Works in non-Western or minority traditions do not always fit into categories and result in an incomplete or incorrect metadata description. This inconsistency restricts discoverability and belief in the possibilities of comparative and cross-cultural literary research. Apart from this, the imbalance between high-resource and low-resource languages also makes retrieval more difficult in multilingual literary archives (Weckmüller et al., 2025). The majority of information retrieval technologies are optimized and created according to the languages that have large digital collections and mathematical equipment. On the other hand, most literary languages do not have annotated datasets, vocabulary resources, and qualified language models. This imbalance leads to unequal retrieval quality, whereas texts written in commonly spoken languages are easier to retrieve as compared to those which are written in marginalized or rare languages, enhancing the inequalities of existing knowledge (Dony et al., 2024). User search intent and behaviour also create difficulties in informational retrieval in multilingual literary Archives. Literary readers and scholars mostly engage in exploratory searches, instead of finding facts accurately. They might find text related to immaterial concept including memory, identity, exile, or resistance concept which are expressed in different ways over different languages and literary traditions (Alon and Krtalić, 2024). Whereas the traditional information retrieval systems, which significantly depend on keyword-based methods, struggle to promote these interpretive queries, particularly in a multilingual system, where theoretical words vary significantly. The historical and contextual dimensions of literary texts add more complexity. The time, place, and sociopolitical conditions under which texts were written determine meanings in literature. Most of the retrieval systems fail to consider the existence of such contextual layers, especially between languages, resulting in decontextualized search results (Priya et al., 2024). This is particularly a deficiency in the case of archival research, whose interpretation and scholarly commentary revolve around the historical distinction. Ethical and institutional factors on retrieval outcomes also influence multilingual literary archives. Whereas systems favour leading languages and literature traditions without active encouragement of linguistic inclusivity (Upadhyay and Viviani, 2025). The challenge of overcoming these challenges requires technical development along with effective policies and ethics. In short, challenges of information retrieval in the multilingual literary archive are caused by linguistic diversity, semantic complexity, poor metadata, gaps in resources, and increasing requirements of users (Musso et al., 2024). These issues need interdisciplinary solutions based on the integration of the fields of computational linguistics, literature studies, and the digital humanities. These are imperative efforts to develop structures of equitable access, quality discovery and sustainable academic participation in multilingual literary heritage (Galuščáková et al., 2024). Different meanings of words and phrases often depend on the context of

culture, history, and literature, and direct translation seldom captures these subtle differences. The problem is intensified by the fact that some users do cross-linguistic searches, where the questions posed in one language are supposed to give relevant texts in another language. Deficiencies in machine translators and cross-linguistic semantic congruency can lead to misrepresentations of meanings and wrong retrieval findings. Retrieval effectiveness is greatly influenced by metadata quality as well as metadata consistency. The multilingual literary archives can also be based on the outdated cataloguing methods, the lack of metadata, or different translation principles, especially in the case of big manuscripts or non-Latin-script manuscripts. Whereas poor information decreases search accuracy and findability. Also, optical character recognition (OCR) software tends to perform poorly with complex scripts, handwritten documents and damaged historical documents, which present additional difficulties with precision in indexing and retrieval. Apart from the technological constraints, multilingual literature collections raise critical ethical and cultural issues. The weakness of having an unfair retrieval system is that there is a risk of maintaining linguistic inequality, especially when the languages that are well-equipped are given more priority than the rest. There is a need to engage with digital humanists, librarians, linguists, and computer scientists in interdisciplinary interaction, in terms of fair representation and accessibility. The need to solve the problem of information search in the multilingual literature collections is, then, not only to the benefit of increasing the efficiency of the system, but also to the development of extensive research, intercultural communication and the subsequent preservation of world literature.

Research Objective

The purpose of this research article is to examine the major information retrieval issues in multilingual literary archives and comprehend the impact of linguistic diversity, cultural diversity, and semantic complexity on the process of organizing, indexing, and retrieving a text in other languages. This paper also evaluates the weaknesses of the existing systems in facilitating cross-language search, precise metadata and appropriate ranking in the context of multilingual archiving systems (Setty, 2024). In addition, metadata standards, controlled vocabularies and new methods of computation, such as cross-lingual retrieval methods and natural language processing, were assessed to enhance effectiveness in the study. Its final aim is to suggest knowledgeable policies and suggestions that can improve retrieval procedures within multilingual literary archives that would promote equal access, cultural conservation, and scholarly involvement.

Literature Review

The increasing development of the process of digitization in literary archives has modernized the worldwide access to cultural, historical, and creative texts. Information retrieval (IR) becomes more complicated with the digitization of manuscripts, poetry, novels and oral traditions of most languages in libraries and cultural institutions (Gagliardi and Artese, 2024). Scholars demonstrate that multilingualism poses linguistic, semantic, and technical problems which old monolingual retrieval systems were unable to address. It was studied that the models of multilingual literary archives require attention towards linguistic diversity, cultural subtlety, and different metadata. Researchers claimed that initial studies of information retrieval were on English or other high-resource languages. These systems assumed the existence of standardized vocabularies, constant grammars and similar scripts. These assumptions cannot be applied to multilingual archives, where the text can be written in several different scripts, dialects, or even history forms. Which in turn led towards a very low retrieval accuracy, particularly in low-resource collections (Builtjes et al., 2025). Furthermore, researchers highlighted that archival studies of languages in Arabic, Devanagari, Latin, and Cyrillic scripts indicate a significant variation in the script that can make the process of indexing, tokenization, and matching of queries more difficult to understand (Gonuguntla et al., 2024). It was claimed that the non-Latin scripts bring difficulty to retrieval systems due to a lack of language tools and variable encoding. This was examined even more problematic in literature archives where antique types of scripts or non-standard spellings were being used in old manuscripts, which can cause further decrease in retrievability (Nguyen et al., 2025). According to the researchers, various words seem similar in different languages, but have different meanings and different things. At the same time, there can also be different words used to denote the same thing in languages. Direct matching of

keywords reduces these semantic connections, resulting in relevant or incomplete answers (Shinde et al., 2024). This problem is enhanced by metaphorical language, symbolism, and metaphorical expressions. It was indicated that modern systems, which are based on surface matching, tend to misunderstand the meaning of literary works (Kumar et al., 2024). Apart from this, researchers found that Cross-lingual information retrieval (CLIR) allows the user to do an inquiry by using a given language and retrieve using another one. Cross-lingual access has been improved by machine translation. However, the errors in translation are detrimental to accuracy, particularly in texts with high culturally-based expression and style. The use of automated tools may put more emphasis on literal meaning compared to literary meaning and lose the contextual meaning. In addition, multilingual archives also tend to contain metadata provided by various institutions, cataloguers over time (Abdullahi and Ekuobase, 2024). It was claimed that the differences in descriptive content, subject headings and the use of language can cause problems in information retrieval. Every language may put a work in its different form, titles or genres, and hence it may not be easy to find all its versions. The absence of standardized multilingual metadata decreases interoperability and restricts extensive search (Phyu et al., 2024). Studied highlighted that the digital literary collections do not represent low-resource languages. The majority of information retrieval research focuses on the most common languages, and regional, indigenous and minority languages have fewer quantities, vocabularies and language models (Zhang et al., 2024). This disproportion results in the lack of equal access while dominant languages can be discovered more than marginalized language traditions. The study of the users indicates that not all of them can design effective cross-language queries. The language storage tower is supported as users are more likely to use familiar languages even in cases where they are irrelevant, but content is available in other languages (Valentini et al., 2025). It was studied that success can be achieved significantly with the help of interface design, language selection tools and multilingual query expansion. Researchers claimed that Linguistic gaps can be reduced with the help of multilingual vocabularies and semantic networks. Studies have shown that concept indexing is a method that connects similar words in different languages, and it is used to improve information retrieval (Ozdemir et al., 2025). However, multilingual knowledge is both time-intensive to create and maintain, and it requires domain knowledge, particularly in the literary field with culturally specific categories. Apart from this, most of the archives use incomplete or monolingual vocabularies, which are ineffective in nature. However, literary texts are not that easy. It involves the complexity of the narratives, the styles and the past language utilization (Girdhar et al., 2024). Researchers found that the models that were made on the basis of modern text might not be accurate when dealing with classical or poetic texts. The measurement tools of evaluation that are used in standard form are unproductive to cultural relevance, which plays a crucial role in literary research. Whereas relevance judgments are more linguistic and cultural relative (Borlund et al., 2024). Therefore, it is not easy to have universal standards. This makes the evaluation difficult and reduces the adoption of new approaches to multilingual literature. According to some researchers, effectiveness is not the only thing that information retrieval systems need to be focused through linguistic diversity and cultural heritage (Alon and Krtalić, 2025). Favouritism models may discriminate against languages or traditions, which supports power relations. Furthermore, researchers claimed that participatory strategies involve inclusion of librarians, scholars and language communities in the design of systems, so that there was an inclusive access to multilingual archives (Tedeschi, 2025). In general, the literature demonstrates that the problem of IR in multilingual literary archives is complicated. Researchers studied the processing of script, inconsistency in metadata, cross-linguistic retrieval, and interaction with the user. The potential of new technologies exists, and the technical advances should be accompanied by cultural sensitivity and interdisciplinary cooperation. The literature-conscious retrieval framework, which is not only multilingual, needs to be discussed in detail in order to provide equitable and significant access to world literary heritage (Ranjgar et al., 2024). The rapid digitalization of literary holdings has transformed the ways in which researchers, academics, and consumers in general access cultural and intellectual resources. Digital literary archives are increasingly being developed by libraries, museums, and other academic institutions throughout the world to preserve literature of different historical times, places, and languages. Though these efforts have

greatly made literature work more accessible, it has also posed significant challenges to information retrieval, especially in multilingual literature archives. Access to the pertinent information in such archives is a complex task that depends on the linguistic differences, cultural background, and technological support differences between languages. Multilingual literary archives hold works written in a wide range of languages, scripts and dialects and often in centuries of literary history. These archives are required to accommodate the shifts in grammar, vocabulary, orthography, and the semantic system in contrast to the monolingual collections. One of the challenges in multilingual retrieval of information is semantic ambiguity. Literary language is figurative and metaphorical, and culturally based, and therefore, to get automated systems to predict meaning in languages reliably is not an easy task.

Methodology

This research applies a qualitative-analytical and systems-oriented approach to explore the major issues in information retrieval in the multilingual literary archives. The design of this study is a combination of a literature review, comparative analysis, and a conceptual framework evaluation that reveal linguistic, technical and cultural barriers to retrieval efficiency in various archival platforms. This approach is suitable in the study of complex, interdisciplinary phenomena where empirical data have to be considered along with theoretical and contextual aspects. It starts with a thorough analysis of the academic publications in the fields of digital humanities, library and information science, computational linguistics and the study of archives.

The recent literature on the topic of multilingual information retrieval systems and their use in literary archives is analyzed with the help of peer-reviewed journal articles and various other sources published within the past 20 years. Special attention is paid to studies that are concerned with cross-lingual information retrieval, metadata standards, semantic indexing, and the processing of low-resource languages. The paper has a theoretical foundation and indicates recurrent problems and research gaps.

The second step is comparative research of selected multilingual literature archives. It includes national digital libraries, literary archives within colleges and universities, and global cultural heritage sites. The appraisal of these archives is conducted by certain criteria, and there is language coverage, the organization of the metadata, search tools, transliteration, and user interface design. The comparison allows the marking of similar boundaries and the identification of best practices in most institutional and technological settings, without any reference to a particular platform.

Figure 1: Conceptual Framework for Multilingual Information Retrieval in Literary Archives.

Applications

Multilingual Literary Archives and Multilingualism

Multilingual literature collections can be applied in a very broad spectrum of scholarly research, education, cultural preservation and art. Problem-solving in the field of retrieval is not merely a technical improvement, but a rudimentary need of influential engagement with the world literary tradition. Improved multilingual search systems can help all types of users, facilitate scholarly research and contribute to equity in the online behavior of languages and cultures.

Literary and Cultural Research

Comparison literature, postcolonial studies, translation studies and world literature should be able to access other languages. Advanced type Multilingual information retrieval systems allow academics to find works that are thematically or artistically related across linguistic boundaries. The cross-lingual search technologies enable individuals to search the archives in one language and find the content that is relevant in another language, enabling comparative inquiry and less dependence on literature that is translated or canonized. This broadens the research field and promotes more accommodative literature research.

Figure 2: User Interface and Search Workflow for Multilingual Literary Archives.

Digital Humanities Projects Multilingual Information Retrieval

The topic modeling as well as sentiment analysis, in addition to the stylistic pattern recognition, are founded on the proper cross-linguistic indexing and semantic alignment. With the development of retrieval systems, it is also possible to build multilingual corpora, analyze the literary trends in various localities, and study the time cycle of narratives, genres, and topics. This program is especially useful in terms of tracking the literary trends throughout the world and cognizing the cultural transfer in the historical contexts. Another use of multilingual retrieval is also found in library and archive management. Librarians and archivists utilize retrieval systems to organize, describe and retrieve collections of different linguistic materials. The use of multilingual metadata standards and improved indexing processes facilitates the ability to manage more complex collections. It can enhance the maintenance of the catalogues, minimize duplication, and give the user delight. Also, multilingual retrieval systems assist archivists in revealing underrepresented languages and authors, which guides their collection development and digitization projections.

Education

Good multilingual literature archives with excellent retrieval opportunities are considered as the precious learning resources. The original texts and translations, critical commentaries, etc., are available to the students who study literature, history, or languages. The search capability of cross-lingual allows the students to learn about literary topics in different cultures, which create a critical thinking and intercultural understanding. The use of real literary books in numerous languages can assist language learners in the acquisition of vocabulary, contextualization in the learning process, and cultural literacy.

Figure 3: Archival Infrastructure for Restoring Marginalized Literary Collections.

Preservation and Renewal of Cultural Heritage

Multilingual literary archives often hold literature in a marginalized or under-resourced language. Better retrieving systems make these works discoverable and available so as not to be marginalized in digital form. Through archives, language differences can be maintained because users may search and locate materials without linguistic challenges. This application highlights the moral aspects of information retrieval as a cultural maintenance tool: translation and publication. Multilingual retrieval systems are essential in search of texts that can be translated and adapted. Cross-lingual search tools are useful to publishers, translators and literary agents to locate less well-known works, authors and culturally relevant writings in a variety of countries. It promotes cultural exchange and literary diversity at international publication markets. The improved retrieval also helps in the study of translation because it enables a comparative examination of the original texts and the translation. Another significant domain of application is the creative industries and entrepreneurship. The influences of foreign literary tradition are becoming significant to writers, artists, filmmakers, and game designers. It promotes innovation and preservation in cultural heritage, especially in cases where the retrieval systems contain contextual metadata and resources of interpretation in addition to the primary texts.

Policy and Institutional Views

Multilingual information retrieval affects decision-making at the national libraries, cultural ministries, and international organizations. Convention data and patterns of retrieval help institutions to determine the way users use multilingual collections, and this helps in making investments in digitalization, language support and development of infrastructure. The program is particularly accessible in universities where the global access and cultural coverage in the local context is balanced.

Figure 4: AI-Driven Semantic Mapping and Machine Translation for Cultural Archives.

Artificial Intelligence and Natural Language Processing Research

Multilingual literature archive serves as a convenient test set to develop and evaluate methods of retrieval. There are semantic ambiguity, figurative language, and historical linguistic variability, which pose challenges to the current NLP models. Better retrieval mechanisms help in the development of multilingual embeddings, machine translation and semantic search, which can be applied to literary archives as well as other multilingual information settings.

Accessibility and Public Involvement

Simple search interfaces, including the support of other languages and scripts, are useful to the general

users, such as independent researchers and lovers of culture. This democratizes literature heritage, not only in the hands of academic elites, but also in enabling more people to participate in cultural knowledge production. Moreover, good information retrieval in multilingual literary archives can be used in various ways that are connected to one another. The retrieval systems may help in enhancing research excellence, educational innovation, cultural preservation, and creative discovery through language, technical and cultural challenges. These applications create awareness of the importance of designing inclusive, accurate and ethically correct information retrieval frameworks in a multilingual literary environment.

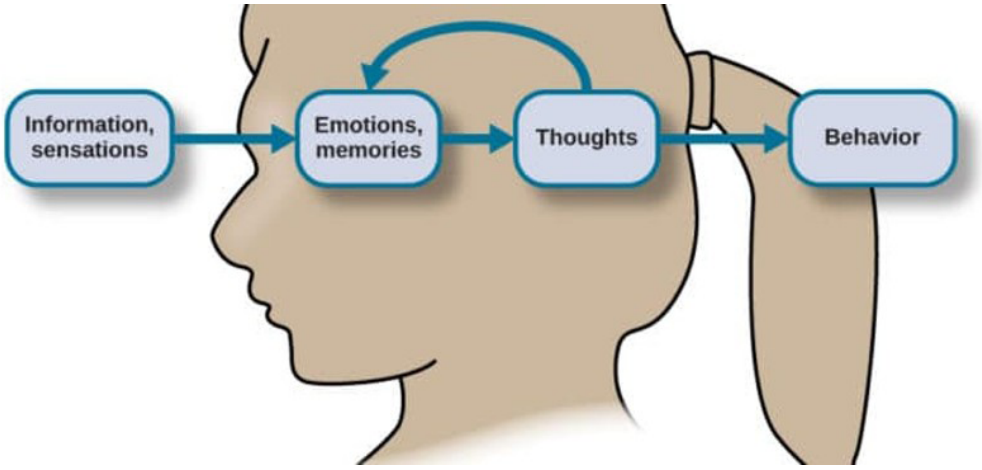

Figure 5: Psychological Framework of Human Information Processing and Behavioral Responses.

Challenges in Multilingual Literary Information Retrieval

Diversity in the Language and in the Scripts: Several languages and scripts can be found in the literary archives (Latin, Arabic, Devanagari, Chinese, etc.), and it is hard to obtain a standard indexing and searching system. In literature, semantic ambiguity is understood as a kind of ambiguity that emerges in the interpretation of a text.

Semantic ambiguity in Literary Texts: The metaphors, symbolism and figures of speech differ among cultures, decreasing the precision of the search systems, which implies the use of keywords and the use of automated systems.

Search Limitations in Cross-Languages: The machine translation and semantic alignment across languages are not consistent, and users may use a different language to search the text with the help of a search engine.

Low-Resource and Endangered Languages: Most languages do not have annotated corpora, lexicons and NLP tools, hence the poor performance of retrieval and poor visibility.

Inconsistency in transliteration and spelling: The problem of several spelling systems of names, places and titles leads to duplication of search results and incomplete search results.

Problem of metadata quality and standardization: Incomplete, inconsistent, or outdated metadata decreases accuracy and recall rate, particularly in non-Western literary collections. The use of legacy cataloguing practices such as subject selection, classification, and data entry is discouraged.

Classification and Entry Data: The archival records of older archives might not conform to the multilingual metadata or contemporary maintenance of the retrieval standards.

Optical Character Recognition (OCR) Errors: OCR does not work well with historical documents, handwritten books and non-Latin text, and this influences the accuracy of indexing.

Cultural and Situational Misinterpretation: Automated systems also tend not to recognize historical, symbolic, and cultural meaning of the literary works.

Prejudice against Dominating Languages: English and other international languages are often given an advantage over minority literary traditions through retrieval systems. Few User Interface Support: Search interfaces might lack support for multilingual input, switching scripts, or culturally diverse searching behaviour.

Evaluation and Relevance Ranking Problems: The process of defining relevance in different languages and cultures is tricky and does not have standardized tests.

Constraints of Digital Preservation and Digitisation: The quality of digitization and the nonexistence of texts restrict the scope and coverage of repossession.

Ethical and inclusion issues: The unequal access to retrieval may lead to the sustenance of linguistic majorities and cultural marginalization.

Discussion

The analysis of the issue of retrieving the information in the literature archives of over one language proves that the interconnections between the linguistic, technological, and cultural aspects that influence the access to the digitalized collections of literary works are associated with complexity. Although digitization has rendered the world significantly more open to literary material, and the technology of information retrieval has made the world a relatively more accessible place, the research findings have indicated that the current systems remain unequal in their ability to accommodate the multilingual and multicultural literary spaces. These thresholds are extremely scholarly, cultural conservation and digital equity. The argument that is raised most significantly in the discussion is the continued prevalence of well-resourced languages in the retrieval systems. The majority of the information retrieval and natural language processing algorithm models are trained and developed using large amounts of information in languages like English, and are more likely to retrieve them with a sidelining effect. This imbalance impacts performance and visibility, where the literature in low-resource / regional languages have lower chances of appearing in the listing. These prejudices might widen the cultural dominance that already exists in online archives, and this is an attempt at promoting an inclusive representation. The other problem that the discussion indicates is the difficulty of the semantic analysis of works of literature. The use of the metaphor, symbolism and culturally embedded meanings becomes very crucial in literary works as opposed to the factual or technical ones. These features are, in most cases, not identified by automated retrieval systems in the right manner, especially when performing an operation on other languages. Even in situations where machine translation is being used, there are small differences in meaning that can confound relevance determination. This limitation implies that the technical solutions cannot do an effective job, and interdisciplinary cooperation is required to ensure the integration of literature theory and cultural knowledge into the retrieval schemes. Metadata quality is another problem, which appears to be a concern in retrieval effectiveness. The inaccuracy in search is caused by inconsistency in cataloguing, insufficient labelling of language, and a mistake in translation. The discussion further implies that the lack of strong computational techniques can be resolved by the fact that the performance of retrieval can be immediately increased with the help of better metadata practices. However, it demands ongoing institutional and professional competence. The dialogue on the side of the user indicates the absence of connections between the system design and user requirements. Occasional cases have been witnessed whereby scholars, students, as well as other general users of multilingual literature archives engage in exploratory as well as interpretive search behaviours as opposed to search queries. However, the majority of retrieval systems focus on the efficiency and perfect matching of retrieval systems, at the cost of discovery and contextual exploration. This incompatibility affects the interaction of users and potentials of archives as dynamic research and learning systems. The deliberation is grounded on moral principles. Such unfair information search is not limited to referring to technical inefficiencies, but also raises the concerns of culture responsibility and digital justice. The challenges are more than enhanced algorithms and need more participants in the process, community involvement, and cultural ownership and representation. Overall, the topic put light on the fact that the issues of multilingual literary information retrieval cannot be solved with the help of autonomous technical interventions. They are rather protestors of all-round initiatives that incorporate technology breakthroughs, metadata advances, user-centricity and moral awareness. Data warehouses should not be used in place of multilingual literary archives since it is necessary to consider them as a cultural ecosystem, which must be accurate, inclusive, and sustainable to set up a retrieval system.

Conclusion

This paper discussed the complicated issues of retrieving information in the literature archive in multilingual literature and, more particularly, in the relationship between linguistic diversity, technological

limitations and cultural context. As digital archives develop and literary materials come into play in various languages and cultures, the efficiency of information search systems is becoming increasingly important in the access, visibility and scholarly use of these collections. The findings demonstrate that multilingual information retrieval is a complicated phenomenon that has important cultural and ethical concerns. Among the key findings, one can include the fact that the current retrieval habits are still biased towards the majority of the popular world languages, which forms an unequal access to the literary content of low-resource and marginalized languages. This imbalance is not only seen in the accuracy of search but also in the diversity of literature traditions in the digital world. Unless there are certain concerted efforts to alleviate the issue of linguistic differences, the multilingual literary archives are in danger of strengthening the hierarchy whereby the literary canons and academic discourse have been historically established. It is emphasized in the conclusion that the semantic and contextual awareness can be applied to the process of literary retrieval. The figurative language, symbolism and allusions of the literary texts can be very full and, in some cases, the automated systems are not able to read them in the proper manner, especially across languages. Despite the openly fascinating opportunities that the advance of the natural language processing and machine translation provide, there are also different levels of their usefulness. This brings out the necessity of having hybrid strategies that would bring together computer strategies and the human experience in literature, linguistics and culture studies. The quality of metadata proves to be a key to augmenting multilingual information search. Consistent and culturally competent metadata can make the discovery of data easier and relevant in the event of the absence of advanced methods of retrieval. The plan to invest in metadata standardization, multilingual cataloguing approaches, or professional training is a workable and effective solution to organizations that deal with multilingual literary archives. Findings of this study emphasize the importance of user-centred design and also ethics responsibility. The retrieval systems must facilitate a great range of search behaviors other than facilitating exploratory and interpretive interaction with the literary material. Simultaneously, the institutions should be aware of their role in the establishment of digital cultural memory and ensure that the retrieval practices are supposed to be inclusive instead of exclusive. Ethical multilingual recovery needs transparency, community input and sensitivity of cultural representation and ownership. In conclusion, it is possible to state that the multilingual literary archive information retrieving issues need a complex and interdisciplinary methodology. Technological innovation, improved metadata practices, inclusive design and ethical awareness are all such aspects that should collaborate to produce retrieval systems that reflect the richness and diversity of the literary history of the world. These kinds of integrated solutions can transform multilingual literary collections into becoming an actualization of easy, fair and culturally competent, digital spaces capable of supporting scholarship, learning and cross-cultural dialogue.

Recommendations and Future Research Directions

- The technical, institutional, and ethical solutions need to be combined to overcome the issues of information retrieval in the multilingual literary archives. These aspects such as the implementation and adoption of multilingual and culturally sensitive metadata standards may be mentioned among others. The archives are meant to highlight homogeneous labelling of language, normal transliteration processes and contextual metadata of both literary and historical and cultural contexts. This also requires training of librarians and archivists to ensure that these standards have been effectively implemented in the various collections.
- The other important concept is the integration of the best cross-lingual information retrieval technology. The models that are set to be used in the future systems are machine translation, multilingual embeddings and semantic search to increase the accuracy of retrieval between languages. It is possible to implement multilingual retrieval solutions in smaller institutions using open-source tools and collaborative platforms without paying large amounts of money. During the development of retrieval systems, a major consideration was the user-centered design. Archives are encouraged to develop search interfaces that are intuitively designed and multilingual and, of course, permit exploratory

searching, script switching, and culturally diversified query behaviours. The addition of user feedback, especially from scholars, students, and language groups who are represented in the archives, can significantly enhance the usability and relevance of the system.

- The research must be carried out in future by working on hybrid forms of retrieval models by integrating computer approaches with human experience in literature, linguistics, and digital humanities. It would be helpful to study the role of cultural context and the interpretation of literature in retrieval systems. Moreover, more empirical studies on the retrieval performance of low-resource and endangered languages, such as the creation of benchmark datasets and evaluation systems, are needed.
- Lastly, it is recommended that future research should be done on the ethical aspect of multilingual information retrieval, i.e. representation, inclusion and digital justice. Research on the community-based methods of archiving and the development of metadata by communities may help make certain that multilingual literary archives evolve into equitable and culturally sensitive knowledge systems.

Acknowledgment

Academic funding project for basic scientific research operating expenses of provincial universities in Heilongjiang Province in 2022 under Grant No.1452MSYYB001.

References

Abdullahi, U. B. and Ekuobase, G. O. (2024). A Lingual Agnostic Information Retrieval System. *The Scientific World Journal*, 2024(1): 6949281. https://doi.org/10.1155/2024/6949281

Alon, L. and Krtalić, M. (2024). "Words Are not just Words; They Carry Experiences Within Them": Navigating Personal Information Management in Multilingual Contexts. In: I. Sserwanga, H. Joho, J. Ma, P. Hansen, D. Wu, M. Koizumi, & A. J. Gilliland (Eds.), *Wisdom, Well-Being, Win-Win*. Springer Nature Switzerland, pp. 333–342. https://doi.org/10.1007/978-3-031-57850-2_25

Alon, L. and Krtalić, M. (2025). "I wish I could use any language as it comes to mind": User experience in digital platforms in the context of multilingual personal information management. *Journal of the Association for Information Science and Technology*, 76(4): 686–702. https://doi.org/10.1002/asi.24964

Borlund, P., Pharo, N. and Liu, Y.-H. (2024). Information searching in cultural heritage archives: a user study. *Journal of Documentation*, 80(4): 978–1002. https://doi.org/10.1108/JD-06-2023-0120

Bringmann, A. and Zhukova, A. (2024). Domain Adaptation of Multilingual Semantic Search--Literature Review. *arXiv preprint arXiv:2402.02932*. https://doi.org/10.48550/arXiv.2402.02932

Builtjes, L., Bosma, J., Prokop, M., van Ginneken, B. and Hering, A. (2025). Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA Open*, 8(5), article no. ooaf109. https://doi.org/10.1093/jamiaopen/ooaf109

da Silva, V. T., Rademaker, A., Lionti, K., Giro, R., Lima, G., Fiorini, S. *et al.* (2024). Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature. *arXiv preprint arXiv:2411.03484*. https://doi.org/10.48550/arXiv.2411.03484

Dony, C., Kuchma, I. and Ševkušić, M. (2024). Dealing with multilingualism and non-English content in open repositories: Challenges and perspectives. *Zenodo*. https://doi.org/10.5281/zenodo.11060284

Gagliardi, I. and Artese, M. T. (2024). Exploring and Visualizing Multilingual Cultural Heritage Data Using Multi-Layer Semantic Graphs and Transformers. *Electronics*, 13(18): 3741. https://doi.org/10.3390/electronics13183741

Galuščáková, P., Oard, D. W. and Nair, S. (2024). Cross-language Retrieval. In: O. Alonso & R. Baeza-Yates (Eds.), *Information Retrieval: Advanced Topics and Techniques*. Association for Computing Machinery, pp. 321–357. https://doi.org/10.1145/3674127.3674136

Girdhar, N., Coustaty, M. and Doucet, A. (2024). Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey. *IEEE Transactions on Computational Social*

Systems, 11(5): 6151–6180. https://doi.org/10.1109/TCSS.2024.3378419

Gnoli, C., Golub, K., Haynes, D., Salaba, A., Shiri, A. and Slavic, A. (2024). Library Catalog's Search Interface: Making the Most of Subject Metadata. *Knowledge Organization*, 51(3): 169–186. https://doi.org/10.5771/0943-7444-2024-3-169

Gonuguntla, H., Meghana, K., Prajwal, T. S., NVSS, K., Sai, K. N. and Jain, C. (2024). Evaluating Modern Information Extraction Techniques for Complex Document Structures. In: *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, pp. 1–5. https://doi.org/10.1109/ICECET61485.2024.10698618

Kim, H. and Lee, S. (2024). POI GPT: Extracting POI Information from Social Media Text Data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 113–118. https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-113-2024

Kumar, U., Shenbaga, R. S., Kasirajan, P. and Sivakamasundari, G. (2024). Smart PDF Inquiry Hub: A Comprehensive Solution for Efficient PDF Document Querying and Information Extraction. In: *2024 International Conference on Expert Clouds and Applications (ICOECA)*. IEEE, pp. 192–198. https://doi.org/10.1109/ICOECA62351.2024.00045

Lo, C.-M. (2024). Multimedia information retrieval using content-based image retrieval and context link for Chinese cultural artifacts. *Library Hi Tech*. https://doi.org/10.1108/LHT-10-2022-0500

Musso, M., Arnold, K., Nanni, F. and Cannelli, B. (2024). What Is in a? Cross-lingual Topic Detection & Information Retrieval in Archives Portal Europe. *ACM Journal on Computing and Cultural Heritage*, 17(2): 1–23. https://doi.org/10.1145/3494572

Nguyen, H. D., Nguyen, T.-H. A. and Nguyen, T. B. (2025). A Proposed Large Language Model-Based Smart Search for Archive System. In: W. Buntine, M. Fjeld, T. Tran, M.-T. Tran, B. Huynh Thi Thanh, & T. Miyoshi (Eds.), *Information and Communication Technology*. Springer Nature Singapore, pp. 210–223. https://doi.org/10.1007/978-981-96-4285-4_18

Ozdemir, A., Odaci, B., Tanatar Baruh, L., Varol, O. and Balcisoy, S. (2025). Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning. *ACM Journal on Computing and Cultural Heritage*, 18(4): 1–25. https://doi.org/10.1145/3746658

Phyu, S. L., Jaman, S., Uchkempirov, M. and Kulkarni, P. (2024). Myanmar Law Cases and Proceedings Retrieval with GraphRAG. In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 2506–2513. https://doi.org/10.1109/BigData62323.2024.10825155

Priya, K., Kamath, A., Chandan, K. M., Praveen, C., Omkar, S. N. and Aaditya, S. J. (2024). Enhancing Q&A Systems with Multilingual Text Conversion and Speech Integration: Harnessing the Power of LangChain and Large Language Models. In: *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, pp. 1–6. https://doi.org/10.1109/CSITSS64042.2024.10816877

Ranjgar, B., Sadeghi-Niaraki, A., Shakeri, M., Rahimi, F. and Choi, S.-M. (2024). Cultural Heritage Information Retrieval: Past, Present, and Future Trends. *IEEE Access*, 12, 42992–43026. https://doi.org/10.1109/ACCESS.2024.3374769

Setty, V. (2024). FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, pp. 2744–2748. https://doi.org/10.1145/3626772.3657663

Shinde, G., Kirstein, T., Ghosh, S. and Franks, P. C. (2024). AI in Archival Science -- A Systematic Review. *arXiv preprint arXiv:2410.09086*. https://doi.org/10.48550/arXiv.2410.09086

Tedeschi, S. (2025). *Towards Comprehensive and Efficient Information Extraction Across Languages* [PhD thesis, Sapienza University of Rome]. https://hdl.handle.net/20.500.14242/189619

Upadhyay, R. and Viviani, M. (2025). Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. *Discover Computing*, 28(1), article no. 27. https://doi.org/10.1007/s10791-025-09505-5

Valentini, F., Kozlowski, D. and Larivière, V. (2025). CLIRudit: Cross-Lingual Information Retrieval of Scientific Documents. *arXiv preprint arXiv:2504.16264*. https://doi.org/10.48550/arXiv.2504.16264

Weckmüller, D., Dunkel, A. and Burghardt, D. (2025). Embedding-Based Multilingual Semantic Search for Geo-Textual Data in Urban Studies. *Journal of Geovisualization and Spatial Analysis*, 9(2), article no. 31. https://doi.org/10.1007/s41651-025-00232-5

Xia, Y. (2024). Advancing Chinese Survey Document Retrieval: Multilingual Models and Semantic Understanding for Structured Information Extraction. *Research Square*. https://doi.org/10.21203/rs.3.rs-5603829/v1

Zhang, Q., Wang, B., Huang, V. S.-J., Zhang, J., Wang, Z., Liang, H. *et al.* (2024). Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. *arXiv preprint arXiv:2410.21169*. https://doi.org/10.48550/arXiv.2410.21169

Xiaoliang Wen was born in Harbin, Heilongjiang Province, China, in 1984. He received his undergraduate and master's degrees from Heilongjiang University, and later earned his Ph.D. from the same institution. Now, he works at Mudanjiang Normal University. His research interests include foreign language education, foreign literature, and Regional and Country Studies.