

Harnessing AI and big data for intangible cultural heritage: A BERTopic-driven analysis of global trends, emerging themes, and evolutionary dynamics**Guangde Zhu^a, Kanokporn Numtong^{a*} and Limei Wang^a**^a*Faculty of Humanities, Kasetsart University, Bangkok, Thailand***CHRONICLE**

Received August 22, 2025
 Received in revised format
 September 28, 2025
 Accepted November 20 2025
 Available online
 November 20 2025

Keywords:

Intangible Cultural Heritage
Digital Preservation
Heritage Conservation
Artificial Intelligence
Big Data
BERTopic
Topic Modeling
Deep Learning

ABSTRACT

Intangible Cultural Heritage (ICH) is vital to cultural identity but difficult to preserve in a rapidly changing world. Artificial Intelligence (AI) and Big Data offer new opportunities to document and analyze ICH, yet the field lacks a comprehensive overview of research trends. No prior study has combined AI-driven topic modeling with bibliometric analysis to map ICH-AI scholarship. This paper addresses that gap by mining emerging topics and global trends in ICH-AI research over the past decade. We compiled ICH-related publications (2015–2024) from Web of Science and Scopus. Using BERTopic, an AI-based topic modeling technique, we extracted thematic topics from the corpus. We also analyzed publication and citation trends to gauge research growth and impact. Combining these approaches allowed us to track the evolution of research themes and identify emerging topics in ICH–AI research. ICH–AI research has grown steadily, with a surge in publications after 2018. BERTopic uncovered 18 distinct research topics, grouped into four broad thematic directions. Key emerging topics include deep learning for heritage preservation, knowledge graphs for cultural data integration, and immersive technologies (AR/VR) for intangible heritage engagement. These trends reflect a shift toward interdisciplinary, technology-driven approaches to ICH preservation. For researchers, the identified trends highlight new opportunities for collaboration and innovation; for policymakers and cultural institutions, they inform strategic support for AI-driven preservation initiatives. These results underscore the potential of AI and Big Data to enhance ICH safeguarding.

© 2026 by the authors; licensee Growing Science, Canada.

1. Introduction

Intangible cultural heritage (ICH) encompasses the living traditions, expressions, and knowledge that communities transmit from generation to generation, providing a profound sense of identity and continuity (Eichler, 2020). As such, ICH is widely recognized as a pillar of cultural sustainability and a key to preserving collective identity. It binds communities together through shared practices and values, fostering social cohesion and affirming cultural diversity (Shakya & Vagnarelli, 2024). Safeguarding ICH is not only about protecting cultural expressions for their own sake, but also about ensuring that the wealth of knowledge and skills behind those expressions continues to inform and enrich future generations. Indeed, by passing on wisdom, art, and ritual, ICH sustains the “emotional and spiritual genes” of a culture (Zhang et al., 2023), making it an indispensable resource for sustainable development and heritage-driven innovation in society.

Despite its critical importance, ICH today faces numerous challenges in preservation and transmission. UNESCO warns that in the face of rapid globalization, many traditional practices are at risk of “disappear[ing] without help” (UNESCO, 2024). Global cultural homogenization, urbanization, and modernization have increasingly eroded local traditions, leaving some intangible practices struggling to survive in modern society (Zhang et al., 2023). Another major concern is the decline in intergenerational cultural transmission. As social dynamics and lifestyles change, younger generations often have fewer opportunities or less interest to learn ancestral practices, resulting in a breakdown of the chain of custody for cultural knowledge. Recognizing this problem, international agencies have stressed the need to actively involve youth in heritage

* Corresponding author

E-mail address: kankporn.n@ku.th (K. Numtong)

initiatives. For example, UNESCO has prioritized engaging young people in documenting and learning about their community's living heritage so that these practices can be passed on to the next generation (UNESCO, 2024). A related challenge is the lack of systematic documentation methods for many forms of ICH. Traditionally, songs, rituals, oral histories, and other intangible practices were preserved through memory and practice rather than formal records. In a globalizing era, this makes them especially vulnerable to loss. Without comprehensive inventories or digital archives, countless intangible expressions risk being forgotten. Institutions and scholars are therefore seeking new ways to record and safeguard ICH knowledge before it fades away, from community-driven inventories to digital archiving projects (Severo, 2018). The convergence of these challenges, globalization's pressure, weakening cultural transmission, and insufficient documentation, has created an urgent need for innovative, scalable solutions to preserve ICH for the future.

In response to these threats, researchers are increasingly turning to artificial intelligence (AI) and big data techniques to support ICH preservation and analysis. Advances in AI are transforming how cultural heritage is studied and safeguarded across multiple fronts. For instance, machine learning and computer vision have been used to digitize and classify cultural artifacts and performances, enabling large-scale documentation that was previously impractical (Münster et al., 2024). Natural language processing (NLP) techniques allow scholars to mine vast collections of texts, from folklore archives to ethnographic records, extracting patterns and insights that human readers might miss. By analyzing large textual corpora, NLP can reveal latent themes and semantic relationships in cultural materials (Gonzalez-Gomez et al., 2024), supporting deeper understanding of languages, oral histories, and traditional knowledge. Knowledge graph frameworks and ontology-based models have also been applied to ICH, integrating dispersed information into interconnected semantic networks. Such approaches make it possible to map relationships among cultural elements (e.g. practitioners, practices, places, and artifacts) and to uncover complex linkages in intangible heritage domains (Liang et al., 2025). This semantic integration not only aids in knowledge organization but also enables new forms of query and discovery, moving ICH preservation from basic digitization toward "datafication" and intelligent analysis (Liang et al., 2025). Moreover, AI-driven analysis is powering trend detection in the digital humanities, including the cultural heritage field. Data-driven techniques like bibliometric analysis combined with machine learning are now used to identify emerging topics and research trends in large bodies of literature (Münster et al., 2021). Such methods can automatically detect shifts in scholarly focus or the rise of new interdisciplinary themes, offering valuable foresight for policymakers and researchers. Notably, UNESCO and other organizations have highlighted AI's potential in language preservation: for example, automated translation systems can accurately translate and revitalize endangered proverbs and oral literature, greatly improving their accessibility to global audiences (UNESCO, 2024). In sum, AI and big data are ushering in a new era for ICH research – one in which digital humanities tools help bridge the gap between rich cultural traditions and modern analytical capabilities. These technologies enable the cultural sector to tackle scale and complexity, from parsing millions of data points to visualizing knowledge networks, ultimately contributing to more effective safeguarding and understanding of living heritage. Among these emerging technologies, advanced topic modeling techniques have become particularly valuable for extracting insights from large textual datasets in the humanities. In cultural heritage research, topic modeling has been used to discover thematic structures in collections of documents, helping scholars identify what subjects or concerns dominate the discourse. Traditionally, a popular method for this task has been Latent Dirichlet Allocation (LDA), which statistically infers topics based on word co-occurrence. However, LDA and similar early models often struggle with short texts or capturing context, and their results can require extensive tuning. Recent developments in NLP have led to more sophisticated models that leverage contextual embeddings from transformers.

BERTopic is one such state-of-the-art topic modeling approach that has gained traction in the past few years. Introduced by Grootendorst (2022), BERTopic uses pre-trained language models (like BERT) to generate embeddings of documents or sentences, then applies clustering algorithms and a class-based TF-IDF procedure to group semantically similar texts into topics (Samsir et al., 2023). This hybrid approach means that BERTopic can capture nuanced semantic relationships far better than purely probabilistic models. Studies have found that BERTopic often produces more coherent and human-interpretable topics compared to classical models like LDA (Egger & Yu, 2022). For example, in a recent comparative study on student feedback data, BERTopic outperformed LDA in semantic relevance and topic coherence, even though LDA excelled at keeping topics distinct (Wang & Ma, 2024). The BERTopic model identified meaningful themes (such as health-related challenges) that aligned well with expert interpretation, highlighting its ability to extract salient topics from real-world text corpora (Gabarron et al., 2023). Likewise, researchers have successfully applied BERTopic in diverse domains – from analyzing social media discourse to mapping research trends in scientific literature – underscoring its flexibility and effectiveness (Alhaj et al., 2022; Wang et al., 2023). The appeal of BERTopic for digital heritage and ICH studies lies in its capacity to handle heterogeneous and complex textual data (e.g. scholarly articles, field reports, interviews) and distill them into coherent thematic clusters. By moving beyond the "bag-of-words" assumptions of LDA and embracing contextual language understanding, BERTopic opens up new possibilities for uncovering latent topics that reflect the rich subtleties of cultural discourse. Despite the proven utility of AI and topic modeling in related fields, there remains a notable research gap in applying these tools to ICH scholarship at a holistic level. To date, few studies have combined AI-based topic modeling with bibliometric analysis to systematically map the intellectual landscape of ICH research. Most existing reviews of ICH literature rely on conventional bibliometric methods (e.g. citation analysis, co-authorship networks) or manual content analysis of themes (Liu & Pan, 2023). While these approaches have provided valuable snapshots of the field, they may not capture emergent research directions or the full complexity of interdisciplinary trends. In particular, no prior work has employed an advanced model like BERTopic in conjunction with bibliometric techniques to mine emerging topics and global

trends in ICH research. This represents a critical gap, given the rapid growth of publications at the intersection of ICH and digital technology in recent years. To address this shortcoming, our study harnesses BERTopic to perform topic modeling on a large dataset of ICH-related publications (2015–2024), integrated with bibliometric analyses. Through this approach, we aim to reveal the evolving thematic structure of the field and pinpoint burgeoning areas of inquiry that merit attention. Specifically, we seek to answer the following research questions:

1. What major trends can be observed in the scholarly publications on ICH and AI from 2015 to 2024?
2. Based on BERTopic modeling, how many distinct themes can be identified in the ICH–AI research corpus?
3. Upon closer examination of these themes, into which overarching research directions can they be grouped?
4. Which emerging research themes have surfaced in the last decade, and in what ways have they evolved over time?

By answering these questions, the study aims to provide a data-driven overview of the evolving landscape of intangible cultural heritage research in the era of AI and big data. In doing so, we illustrate how modern computational tools can illuminate scholarly trends in the humanities, and we discuss future prospects for integrating AI techniques in the ongoing effort to preserve and understand the world’s living heritage.

2. Research Methods

2.1. Data Sources

To ensure comprehensive coverage of relevant literature, we drew data from two major scholarly databases: Web of Science and Scopus. These databases were chosen because of their broad, multidisciplinary coverage and inclusion of peer-reviewed literature across the sciences, social sciences, and humanities. By using both Web of Science and Scopus, we aimed to capture a wide range of studies on intangible cultural heritage (ICH) in the context of artificial intelligence (AI) and big data, minimizing the risk of missing pertinent publications. All searches and data collection were confined to documents published between 2015 and 2024, aligning with our research focus on the last decade of developments in this domain.

2.1.1. Search Queries

We developed a rigorous search strategy using a combination of keywords related to intangible cultural heritage and AI/big data to identify relevant studies. The search strings were tailored to each database’s query format and designed to be inclusive of various synonyms and related terms. The exact search queries used were as follows:

- **Web of Science:** TS=((“Intangible Cultural Heritage” OR “Cultural Heritage Preservation” OR “Traditional Knowledge” OR “Folklore” OR “Oral Traditions” OR “Traditional Craftsmanship” OR “Heritage Conservation”) AND (“Artificial Intelligence” OR “AI” OR “Machine Learning” OR “Deep Learning” OR “Natural Language Processing” OR “NLP” OR “Big Data” OR “Data Analytics” OR “Data Mining” OR “Knowledge Graph” OR “Computer Vision” OR “Digital Humanities”)).
- **Scopus:** TITLE-ABS-KEY((“Intangible Cultural Heritage” OR “Cultural Heritage Preservation” OR “Traditional Knowledge” OR “Folklore” OR “Oral Traditions” OR “Traditional Craftsmanship” OR “Heritage Conservation”) AND (“Artificial Intelligence” OR “AI” OR “Machine Learning” OR “Deep Learning” OR “Natural Language Processing” OR “NLP” OR “Big Data” OR “Data Analytics” OR “Data Mining” OR “Knowledge Graph” OR “Computer Vision” OR “Digital Humanities”)).

These queries combined terms for intangible cultural heritage (e.g., cultural heritage preservation, folklore, oral traditions) with terms for AI and data-intensive technologies (e.g., machine learning, natural language processing, knowledge graph). Using the logical operator AND between the two groups ensured that retrieved records explicitly mentioned both an ICH-related term and an AI/big-data term. We searched within titles, abstracts, and keywords (Topic field in Web of Science; Title/Abstract/Keywords in Scopus) to capture any study where these concepts intersect. This search strategy is explicitly documented to facilitate replication in future research.

2.1.2. Data Retrieval and Deduplication

Using the above queries, we conducted the searches in each database and retrieved the initial set of records. The Web of Science query yielded 599 records, and the Scopus query returned 1,392 records, for a total of 1,991 raw results. All records including titles, authors, abstracts, and other bibliographic information were exported from both databases. We then merged the results and removed duplicate entries that appeared in both Web of Science and Scopus. Duplicate detection was performed by comparing unique identifiers (like DOIs) and bibliographic details (title, authors, year). In total, 554 duplicate records were identified and removed. This deduplication process resulted in a unique dataset of approximately 1,437 records (i.e., the combined set of publications after duplicates were eliminated). These unique records formed the basis for the subsequent

screening steps. By removing overlaps between the databases, we ensured that each study would only appear once in our analysis.

2.1.3. Screening Criteria

We applied a set of inclusion and exclusion criteria to screen the 1,437 unique records for relevance and suitability. This screening was done in two stages – first by filtering based on document characteristics and then by assessing content relevance. In the initial screening stage, we excluded any records that did not meet our predefined criteria regarding publication year, document type, or language. Specifically, 363 records were removed at this stage for the following reasons:

- **Publication Date Outside 2015–2024:** Publications that were not within the 2015–2024 timeframe were excluded to focus the analysis on the last decade.
- **Document Type:** We excluded items that were not original research contributions, such as review articles, editorials, book chapters, book reviews, corrections, and letters. Our aim was to include only primary research articles and conference papers/proceedings that present original findings.
- **Language:** Publications in languages other than English were excluded, as our analysis was restricted to English-language literature for consistency and because English is the dominant language in the indexed databases used.

After applying these criteria, only records that were published in 2015–2024, written in English, and classified as research articles or conference proceedings remained. This initial filtering reduced the dataset to 1,072 records that met the basic inclusion criteria. We carefully documented these criteria so that future researchers can apply the same filters for a comparable dataset.

2.1.4. Eligibility Assessment

The remaining 1,072 records were then subject to a more in-depth eligibility assessment to ensure each publication was truly relevant to intangible cultural heritage in the context of AI and big data. In this stage, we reviewed titles and abstracts – and in many cases the full text – of each publication to evaluate its topical relevance. We looked for a clear focus on intangible or living cultural heritage (such as traditions, folklore, traditional knowledge, etc.) combined with the application or discussion of AI, machine learning, data analytics, or related big data techniques. Studies that only mentioned these terms in passing or did not substantively connect AI/big data with intangible cultural heritage were considered outside the scope of our research. As a result of this content-based screening, 115 records were removed due to irrelevance. Common reasons for exclusion at this stage included papers that turned out to focus on tangible cultural heritage (physical artifacts) rather than intangible heritage, studies centered on digital heritage or humanities without a clear link to AI techniques, or papers on AI in cultural contexts that did not address heritage preservation or transmission. By the end of the eligibility assessment, we ensured that all remaining studies directly pertained to the intersection of AI (and data-driven methods) with intangible cultural heritage.

2.1.5. Final Dataset

After the rigorous screening and eligibility checks, a total of 957 unique studies were confirmed for inclusion in our analysis. This final dataset of 957 publications represents the body of literature, published from 2015 through 2024, that squarely deals with intangible cultural heritage in the era of AI and big data, as shown in Fig. 1.

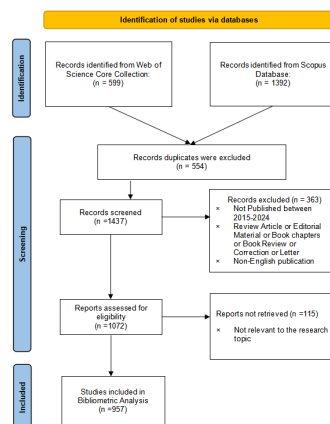


Fig. 1. Workflow of data filtering

Each of these studies meets all the inclusion criteria and collectively they form the basis for our analysis of emerging topics, global trends, and research prospects in the field. We proceeded to compile bibliographic information (such as titles, authors, source titles, publication year, keywords, and abstracts) from these records for further analysis. The data collection and preprocessing steps described above were carried out systematically and are documented in detail to ensure that this study is replicable. Future researchers can reproduce our dataset by using the same databases, running the provided search queries, and applying the stated deduplication and screening criteria, thereby achieving a comparable collection of literature for verification or extended analysis.

2.2. BERTopic Modeling Approach

To extract and analyze key research themes from the corpus, this study employed BERTopic, a state-of-the-art topic modeling technique that leverages transformer-based embeddings and clustering algorithms to generate coherent topic representations (Grootendorst, 2022). Unlike traditional topic modeling approaches such as Latent Dirichlet Allocation (LDA), BERTopic utilizes contextual word embeddings, which improve topic coherence by capturing semantic relationships in text (Samsir et al., 2023). This method is particularly effective for analyzing bibliometric data, as it allows for a nuanced representation of research trends (Banerjee & Pan, 2024; Meitei et al., 2024). The BERTopic pipeline in this study consisted of five key stages: text preprocessing, embedding generation, dimensionality reduction, clustering, and topic representation and interpretation, as shown in Fig. 2.

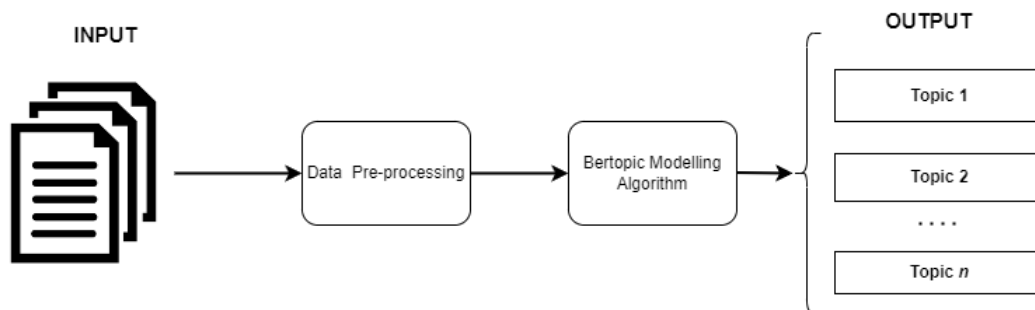


Fig. 2. Workflow of topic modelling

The first step involved text preprocessing, which was essential for enhancing the coherence and quality of topic extraction. The textual data, consisting of titles, abstracts, and author keywords, underwent several preprocessing operations. All text was first converted to lowercase to ensure uniformity, followed by the removal of common English stopwords using the NLTK library. To further improve text representation, lemmatization was applied using the SpaCy package, reducing words to their root forms (e.g., studying to study). Punctuation, numbers, and special characters were eliminated to retain only meaningful textual content. Additionally, bigram and trigram detection was conducted using Gensim's Phrases Model, allowing multi-word phrases such as cultural heritage to be recognized as single entities. Finally, named entity recognition (NER) filtering was performed to exclude proper nouns unrelated to topic formation, such as institution names or locations, ensuring that only conceptually relevant terms contributed to the topic modeling process.

Following text preprocessing, the next step involved embedding generation, where textual data was transformed into numerical vectors. This study employed Sentence-BERT, specifically the "all-MiniLM-L6-v2" model, to encode each document into a 768-dimensional vector representation. The Sentence-BERT embeddings allowed for a high-dimensional representation of semantic similarities between research topics, ensuring that related studies were positioned closely together in the vector space. Each document, composed of its title, abstract, and keywords, was thus converted into an embedding, forming the basis for the subsequent clustering process.

Since raw text embeddings exist in a high-dimensional space, dimensionality reduction was necessary to improve clustering performance while preserving key structural relationships. This was achieved using Uniform Manifold Approximation and Projection (UMAP), a widely used non-linear dimensionality reduction technique. The parameters were set such that $n_neighbors=15$ ensured a balance between local and global relationships, while $n_components=5$ reduced the embedding dimensions to five to facilitate computational efficiency. Additionally, the cosine similarity metric was used to measure distances between points, ensuring that semantic similarities were effectively captured despite the dimensionality reduction.

Once the embeddings were reduced to a manageable feature space, clustering was performed using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This clustering algorithm was chosen due to its ability to

identify arbitrary-shaped clusters while filtering out noise. The HDBSCAN model was configured with `min_cluster_size=10`, ensuring that each topic contained at least ten documents, while `min_samples=5` controlled the cluster granularity to avoid excessive fragmentation. The clustering process grouped similar research articles into distinct thematic categories, allowing for a structured representation of research trends. Documents that did not fit well into any topic were classified as outliers and assigned to a general category labeled -1, which was subsequently excluded from the final topic analysis.

Following clustering, topic representation and interpretation were conducted using Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF), a refined variation of the traditional TF-IDF algorithm that adjusts term importance within specific topic clusters. The c-TF-IDF approach computed standard TF-IDF scores across all terms and then weighted these scores based on their distribution within each identified topic. The most representative terms—the top 10 words per topic—were extracted as defining keywords, forming the basis for topic interpretation. To facilitate visualization and further analysis, topic distributions were represented through bar charts, highlighting the most frequent keywords within each topic, as well as inter-topic distance maps generated via t-distributed Stochastic Neighbor Embedding (t-SNE) to illustrate relationships between topics.

2.3. R-Based Bibliometric Analysis

To analyze the publication trends, emerging topics, and thematic evolution of research on intangible cultural heritage in the context of artificial intelligence and big data, this study employed R-based bibliometric analysis using a set of specialized R packages. The analysis was conducted using *bibliometrix*, a comprehensive bibliometric analysis package (Aria & Cuccurullo, 2017), along with additional visualization tools such as *ggplot2*, *igraph*, and *wordcloud* to present the results in an interpretable and visually engaging manner. First, descriptive statistics were computed using *bibliometrix* to quantify research output and citation impact. Next, a word cloud analysis was performed using *wordcloud* to visualize dominant research themes. Finally, a trend topics and thematic evolution analysis was carried out using *ggplot2* and *bibliometrix* to track the progression of research themes over time, identifying key emerging topics and their historical development.

3. Results and Data analysis

3.1. Descriptive Statistics of the Dataset

Table 1 provides an overview of statistical information of the dataset used in this study, covering key aspects such as timespan, sources, document characteristics, author contributions, collaboration patterns, and document types. The dataset spans from 2015 to 2024, encompassing a total of 957 documents retrieved from 625 different sources, including journals, books, and conference proceedings. The annual growth rate of publications in this field is recorded at 40.79%, reflecting a significant and increasing research interest in the intersection of intangible cultural heritage and artificial intelligence. The average document age is approximately 3.2 years, indicating that the research is relatively recent and evolving. Furthermore, each document, on average, has received 7.595 citations, suggesting a moderate level of academic impact. Notably, the dataset contains no references, which may be a characteristic of the bibliometric extraction process.

Table 1
Descriptive Statistics of the Dataset

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2015:2024
Sources (Journals, Books, etc.)	625
Documents	957
Annual Growth Rate %	40.79
Document Average Age	3.2
Average citations per doc	7.595
DOCUMENT CONTENTS	
Keywords Plus (ID)	4538
Author's Keywords (DE)	2813
AUTHORS	
Authors	2446
Authors of single-authored docs	140
AUTHORS COLLABORATION	
Single-authored docs	144
Co-Authors per Doc	3.67
International co-authorships %	6.897
DOCUMENT TYPES	
article	445
article; early access	5
article; proceedings paper	4
conference paper	392
proceedings paper	111

Regarding the content of the documents, a total of 4538 Keywords Plus (ID) were identified, along with 2813 unique author keywords (DE), illustrating a diverse and expansive thematic landscape within the research domain. The involvement of authors in this field is also substantial, with 2446 individual researchers contributing to the publications. Among them, 140 are authors of single-authored documents, while the total number of single-authored documents is slightly higher at 144, indicating that some authors have contributed more than one single-authored study. Collaboration patterns among authors reveal that each document has, on average, 3.67 co-authors, demonstrating a strong tendency towards collaborative research. However, the percentage of international co-authorship is relatively low, standing at 6.897%, which suggests that while collaboration within national or institutional networks is prevalent, cross-border academic cooperation in this field remains somewhat limited. In terms of document types, the dataset is predominantly composed of journal articles, with 445 standard research articles and an additional five categorized as early access articles. Additionally, four documents are classified as both articles and proceedings papers. Conference-related publications also play a significant role in this research domain, with 392 conference papers and 111 proceedings papers included in the dataset. This distribution suggests that while journal articles remain the primary medium for disseminating research findings, conferences serve as an important platform for presenting emerging studies and fostering academic discussions.

3.2. Publication Trends and Citation Status

Table 2 presents a longitudinal analysis of research output, citation impact, and citable years within the domain of intangible cultural heritage and artificial intelligence from 2015 to 2024. The number of publications (N) has shown a steady increase over the years, reflecting the growing academic interest in this interdisciplinary field. In 2015, only 15 articles were published, whereas by 2024, the number had surged to 326, marking a substantial expansion of scholarly contributions over the past decade. The mean total citations per article (MeanTCperArt) exhibit considerable variation across the years, with the highest recorded in 2019 at 33.98, followed by 2016 at 23.31. These figures suggest that articles published during these years have had a significant academic impact, potentially due to pioneering research contributions, early adoption of AI-driven methods in cultural heritage studies, or the presence of highly cited seminal works. In contrast, more recent years, particularly 2023 and 2024, show lower citation counts per article, with 3.6 and 0.9, respectively. This trend is expected, as newer publications typically require time to accumulate citations. A similar pattern is observed in the mean total citations per year (MeanTCperYear), which accounts for the varying number of citable years. The peak citation impact per year occurred in 2019, reaching 4.85, followed by 2020 at 2.61 and 2016 at 2.33. These peaks indicate periods when key studies gained significant recognition and influence in the academic community.

Table 2
Publication Trends and Citation Status

Year	N	MeanTCperArt	MeanTCperYear
2015	15	4.53	0.41
2016	29	23.31	2.33
2017	38	14.53	1.61
2018	38	14.21	1.78
2019	48	33.98	4.85
2020	70	15.69	2.61
2021	102	10.87	2.17
2022	128	5.59	1.4
2023	163	3.6	1.2
2024	326	0.9	0.45

3.3. Topic Modeling Results

Fig. 3 illustrates the topic modelling results, with a total of 18 topics mined, illustrating the distribution of key terms across different topic clusters within the ICH and AI research domains. Each topic is represented by a bar chart, with the horizontal axis representing the word scores, reflecting the weight or importance of each word in the given topic. The vertical axis shows the most representative words that define the theme, giving insight into the thematic structure of the dataset.

The topic distributions reveal diverse areas of research focus. Several topics, such as Topic 0 and Topic 13, prominently feature terms like “cultural heritage” and “intangible cultural heritage”, indicating that these clusters are fundamentally concerned with heritage studies. In contrast, other topics, such as Topic 1 and Topic 9, highlight technical concepts including “knowledge graph”, “temporal information”, and “distributed collision”, suggesting a strong computational and artificial intelligence dimension within the research field. Additionally, Topic 2 and Topic 3 emphasize elements of traditional culture, with terms like “traditional music”, “music appreciation”, and “Chinese opera”, signifying scholarly engagement with intangible cultural expressions.

Certain topics demonstrate an interdisciplinary connection between artificial intelligence and cultural heritage preservation. For instance, Topic 5 includes references to “big data” and “data environment”, indicating the integration of large-scale computational analysis in heritage studies. Similarly, Topic 6 contains terms such as “knowledge extraction” and “extraction

technology”, reflecting the application of advanced AI methodologies to the processing and analysis of cultural data. Moreover, Topic 7's inclusion of “medical knowledge” and “colorectal cancer” suggests an unexpected intersection between heritage-related knowledge and biomedical applications, potentially exploring traditional medicine and its modern computational study.

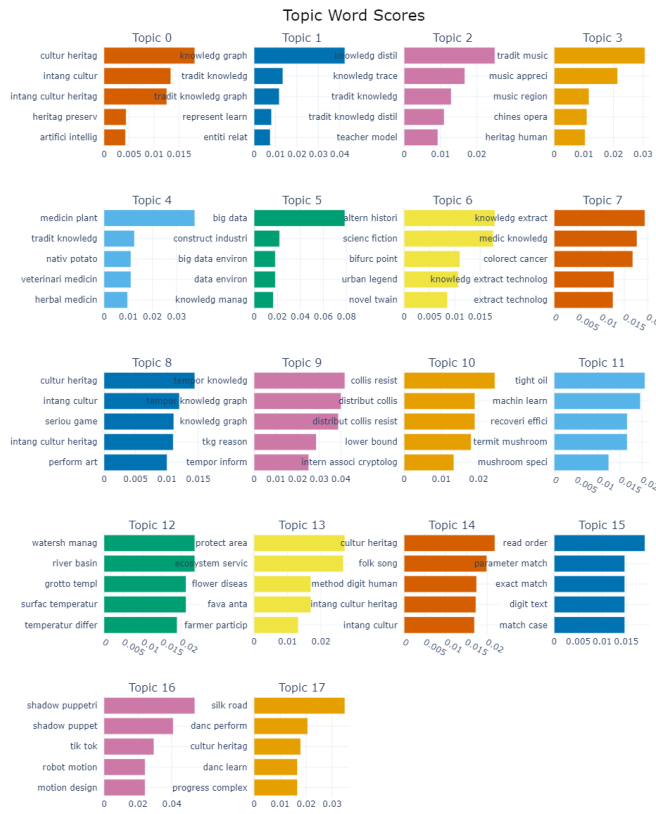


Fig. 3. Topic Modeling Barchart

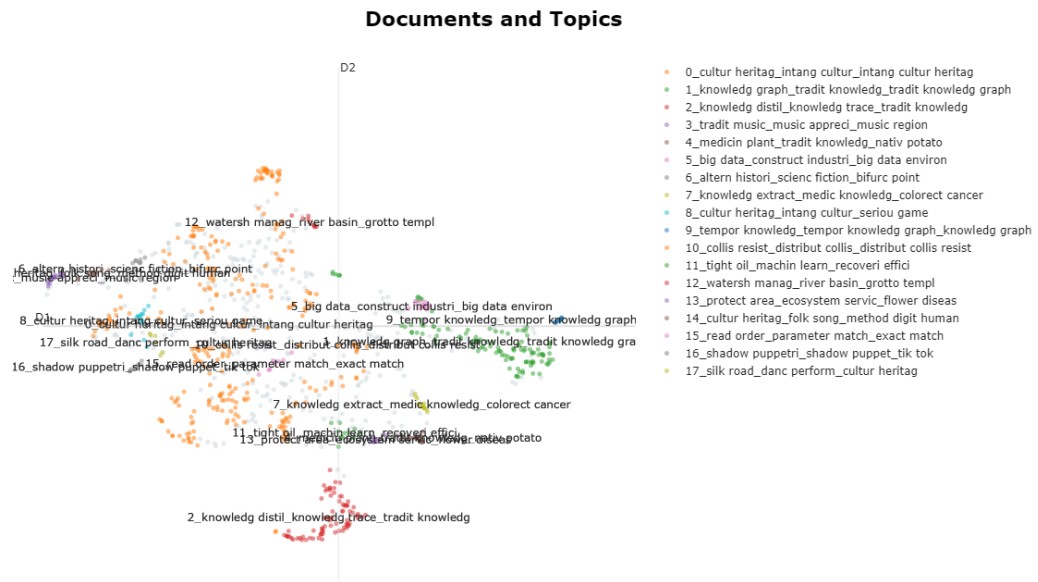


Fig. 4. Documents - Topics Clusters

Several topics also highlight the role of technological advancements in heritage research and preservation. Topic 10 references “tight oil”, “lower bound”, and “inter-associative cryptology”, which may indicate the use of cryptographic or optimization techniques in digital heritage management. Similarly, Topic 11, which includes “machine learning” and “recovery efficiency”, points towards AI-driven techniques for analyzing and restoring cultural artifacts or historical data. Additionally, Topic 16,

which features terms like “shadow puppetry”, “TikTok”, and “robot motion”, suggests an exploration of digital media and automation in heritage performance and storytelling. The emergence of topics related to natural environments, such as Topic 12 with terms like “watershed management”, “river basin”, and “ecosystem services”, suggests that cultural heritage research increasingly engages with ecological and environmental concerns. This reflects a broader shift towards understanding cultural heritage in relation to sustainable development and environmental conservation.

Based on the topics identification and clustering presented in Figure 3 and Figure 4, the identified topics can be categorized into four overarching research directions. Table 3 reveals a structured classification of research directions within the domain of intangible cultural heritage (ICH) and artificial intelligence (AI). Each topic cluster corresponds to a specific thematic orientation, demonstrating how AI-driven methodologies are shaping cultural heritage research across diverse disciplines. These thematic clusters illustrate the breadth of scholarship in this domain, encompassing cultural heritage preservation, technological innovations for cultural heritage, theoretical perspectives on cultural heritage, and sustainability concerns.

Table 3
Classification of Research Topics

Direction	Topic	Representative Keywords
1	0	Cultural heritage, intangible cultural heritage, heritage preservation.
	3	Traditional music, music appreciation, Chinese opera.
	13	Folk songs, digital human modeling for heritage preservation.
	16	Shadow puppetry, TikTok, digital performance.
	17	Silk Road, dance performance, intangible cultural heritage.
2	1	Knowledge graph, traditional knowledge distillation, entity relationships.
	2	Knowledge tracing, teacher models, AI-driven cultural studies.
	5	Big data, data environments, knowledge management.
	6	Knowledge extraction, extraction technology, AI-driven cultural data processing.
	7	Medical knowledge, AI in knowledge retrieval, intersection with health sciences.
3	11	Machine learning, recovery efficiency, AI-based restoration.
	15	Text recognition, parameter matching, AI-driven document analysis.
	9	Distributed collision resistance, cryptology, AI security applications.
	10	Tight oil modeling, optimization, AI applications beyond cultural domains.
4	12	Watershed management, ecosystem services, environmental monitoring in heritage sites.
	14	Text processing, parameter matching, digital text preservation.
	4	Medicinal plants, traditional medicine, native heritage.
	8	Serious games, temporal knowledge graphs, gamification in heritage studies.

The first major research direction focuses on core studies of preservation and digitization for intangible cultural heritage, encompassing topics that primarily engage with traditional cultural expressions, heritage conservation, and digital documentation. Topic 0, which includes terms such as cultural heritage, intangible cultural heritage, and heritage preservation, represents a foundational discourse on safeguarding heritage. Similarly, Topic 3, characterized by traditional music, music appreciation, and Chinese opera, highlights studies that examine the role of AI in preserving and analyzing traditional performing arts. Further expanding this domain, Topic 13 explores folk songs and digital human modeling, reflecting efforts to digitize and reconstruct traditional cultural expressions through AI technologies. The presence of Topic 16, with terms such as shadow puppetry, TikTok, and digital performance, suggests that digital media and social platforms are being increasingly integrated into the preservation and dissemination of traditional art forms. Additionally, Topic 17, which incorporates Silk Road, dance performance, and intangible cultural heritage, illustrates the intersection of historical trade routes and performing arts within AI-driven heritage studies.

A second major research direction emerges in the AI-Driven knowledge management, where various computational techniques are employed to analyze and manage heritage-related knowledge. Topic 1, which features knowledge graphs, traditional knowledge distillation, and entity relationships, indicates a strong emphasis on AI-driven knowledge representation for cultural heritage studies. The presence of Topic 2, with terms such as knowledge tracing and teacher models, suggests that AI methodologies are being leveraged to enhance the accessibility and transfer of cultural knowledge. Additionally, Topic 5, which includes big data, data environments, and knowledge management, highlights the role of large-scale computational analysis in heritage studies. Topic 6 further underscores this trend with its focus on knowledge extraction and AI-driven cultural data processing, illustrating the increasing use of machine learning in heritage digitization. The inclusion of Topic 7, which references medical knowledge and AI-driven knowledge retrieval, suggests an interdisciplinary intersection where AI is used to analyze traditional medicine within the broader context of intangible heritage. The presence of Topic 11, featuring machine learning and AI-based restoration, points toward the application of advanced computational techniques in the recovery and conservation of cultural artifacts. Finally, Topic 15, which highlights text recognition and parameter matching, illustrates how AI is being employed for automated document analysis and digital text preservation, further reinforcing the role of AI in archival and heritage research.

A third research direction is centered on technological and Cross-domain Innovations for cultural heritage applications, encompassing topics that incorporate cryptography, optimization techniques, and digital security for heritage preservation. Topic 9, which features distributed collision resistance and cryptology, suggests that AI-driven security applications are being integrated into heritage data protection. The presence of Topic 10, which includes tight oil modeling and optimization,

indicates that AI methodologies developed in other domains may also be applied to heritage research. Additionally, Topic 12, which references watershed management and environmental monitoring, reflects a growing concern with the environmental impact of heritage sites and the role of AI in sustainable conservation. Topic 14, which focuses on text processing and digital text preservation, underscores the use of computational linguistics and AI-driven methods in the digitization and maintenance of historical texts.

The final research direction is rooted in historical, literary, and theoretical perspectives on cultural heritage, which reflects an engagement with traditional knowledge systems, historical narratives, and interactive digital humanities. Topic 4, which features medicinal plants, traditional medicine, and native heritage, suggests an intersection between historical knowledge and AI-based analysis, particularly in understanding indigenous and historical medicinal practices. Topic 8, which includes serious games and temporal knowledge graphs, suggests that gamification and immersive technologies are becoming increasingly relevant in heritage studies, allowing for interactive and dynamic representations of cultural history.

By categorizing the topics into these overarching research directions, it becomes evident that the field of intangible cultural heritage and artificial intelligence is characterized by a rich interplay between traditional cultural studies, advanced computational techniques, and interdisciplinary applications. While some topics emphasize heritage conservation and digital preservation, others explore AI-driven knowledge extraction, cryptographic security, and environmental sustainability. Furthermore, the presence of historical and literary topics suggests that AI is not only being used as a tool for analysis but is also reshaping the way cultural narratives are reconstructed and understood. This classification underscores the evolving nature of AI-driven heritage research, highlighting opportunities for cross-disciplinary collaborations and technological advancements in cultural heritage preservation and analysis.

Figure 5 presents a similarity matrix, illustrating the degree of similarity among different research topics identified in the study. The matrix is structured as a heatmap, where the intensity of the color corresponds to the similarity score between topic pairs, with darker shades indicating higher similarity and lighter shades reflecting lower similarity. This visualization provides an important perspective on how different thematic clusters within the research on intangible cultural heritage and artificial intelligence relate to each other.

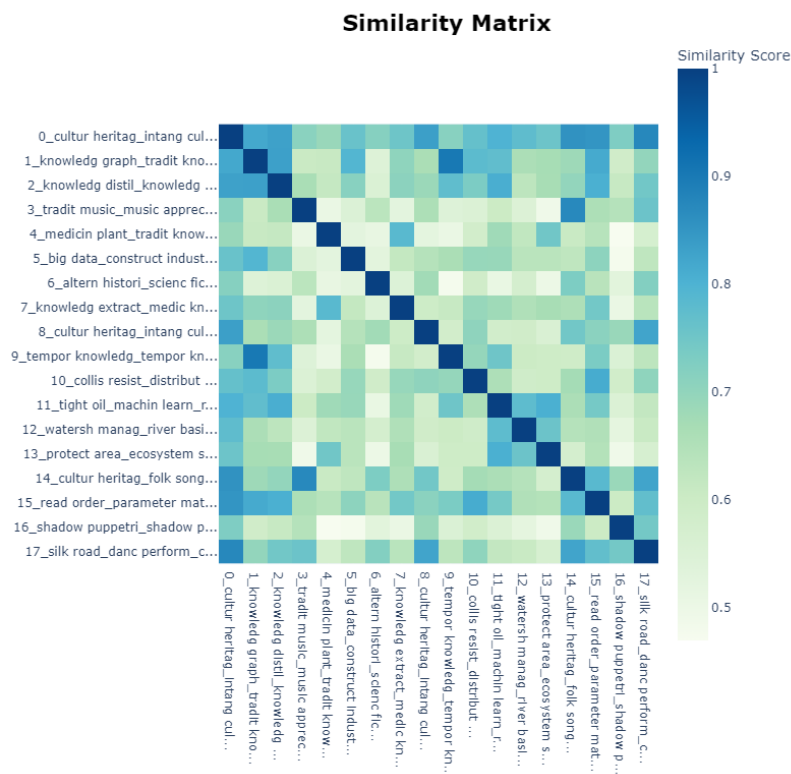


Fig. 5. Similarity Matrix of Research Topics

From the matrix, it is evident that some topics exhibit strong thematic associations, as indicated by the darker blue areas. For instance, Topic 0 (cultural heritage, intangible cultural heritage, heritage preservation) shows high similarity with Topic 8 (cultural heritage, serious games, temporal knowledge graphs) and Topic 13 (cultural heritage, folk songs, digital humanities). This suggests that research on heritage preservation frequently intersects with studies on digital engagement, serious gaming applications, and computational methods for preserving cultural narratives. The presence of strong similarity between these topics indicates that digital tools and AI-driven simulations are increasingly employed to enhance the accessibility and preservation of intangible cultural heritage.

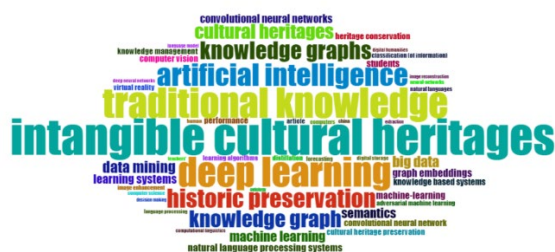
Similarly, topics related to artificial intelligence and data-driven methodologies demonstrate significant interconnections. Topic 1 (knowledge graph, traditional knowledge distillation, entity relationships) exhibits strong similarity with Topic 6 (knowledge extraction, medical knowledge, AI applications in heritage studies) and Topic 5 (big data, data environments, knowledge management). This relationship suggests that knowledge representation and AI-driven extraction techniques are fundamental components in the analysis and preservation of intangible heritage, highlighting how computational techniques are being leveraged for structuring and processing heritage-related data.

Another notable pattern in the similarity matrix is the clustering of topics related to environmental and sustainability aspects of heritage research. Topic 12 (watershed management, river basins, environmental monitoring) and Topic 13 (protected areas, ecosystem services, conservation strategies) exhibit strong interconnectivity, reinforcing the emerging trend of integrating environmental conservation frameworks into heritage preservation. The correlation between these topics suggests that climate change, sustainable conservation, and ecological factors are increasingly being recognized as critical elements in the safeguarding of heritage sites. Moreover, topics associated with technological and computational innovations also display meaningful similarities. Topic 9 (distributed collision resistance, cryptology, AI security applications) and Topic 10 (tight oil modeling, optimization techniques, AI applications) demonstrate considerable thematic overlap, indicating that secure data management, AI-based optimization, and cryptographic methodologies are becoming relevant concerns in heritage studies (Chen et al., 2024). The presence of similarity among these topics highlights the necessity of ensuring the integrity and protection of digital heritage data, particularly in the context of AI-driven documentation and archiving.

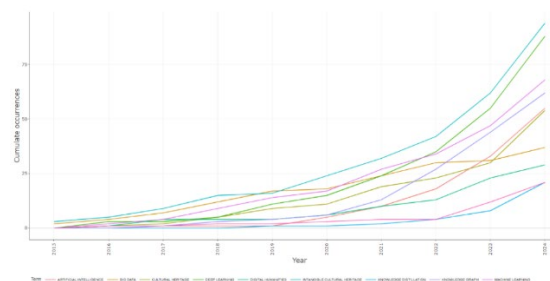
Despite these strong thematic connections, certain topics appear more thematically distinct, as indicated by the lighter-colored areas in the matrix. For example, Topic 3 (traditional music, music appreciation, Chinese opera) exhibits lower similarity with topics related to big data and AI-driven knowledge extraction, suggesting that studies on traditional music preservation remain somewhat independent from computational heritage methodologies. However, its modest similarity with Topic 16 (shadow puppetry, TikTok, digital performance) suggests that performing arts and digital media technologies share a degree of thematic convergence, particularly in discussions related to the digitization of traditional performing arts.

3.4. Key Terms analysis

Fig. 6(a) presents a word cloud that visually represents the most frequently occurring terms in the research domain of intangible cultural heritage (ICH) and artificial intelligence (AI). The size of each word corresponds to its frequency in the dataset, with larger words indicating a higher occurrence and, by extension, greater significance within the field. The most prominent terms include “intangible cultural heritages”, “traditional knowledge”, “deep learning”, “artificial intelligence”, “historic preservation”, and “knowledge graphs”, suggesting that research in this area is heavily centered on the digital preservation of cultural heritage through AI-driven methodologies. The strong presence of “deep learning”, “machine learning”, “data mining”, and “natural language processing” indicates that computational techniques play a central role in contemporary heritage research. The term “knowledge graph” also appears frequently, reflecting the increasing use of structured knowledge representation methods for organizing and retrieving information related to cultural heritage. Additionally, the inclusion of “virtual reality” and “computer vision” suggests that immersive technologies and automated image analysis are being employed to digitize and reconstruct traditional cultural expressions. The intersection of cultural heritage and AI is further highlighted by terms such as “heritage conservation”, “historic preservation”, and “cultural heritage preservation”, which reinforce the notion that digital tools are being leveraged to protect and sustain traditional knowledge. The presence of terms related to learning systems, semantics, and classification suggests that AI-driven methodologies are not only being used to preserve heritage but also to analyze, categorize, and enhance accessibility to heritage-related data.



6(a). Word Cloud



6(b). Cumulative Occurrence of Key Terms Over Time

Fig. 6. Word Cloud and Cumulative Occurrence of Key Terms Over Time

Fig. 6(b) complements this analysis by illustrating the cumulative occurrences of key terms over time, providing insight into the evolution of research trends within the field. The graph demonstrates a notable increase in scholarly interest in intangible cultural heritage and AI-related topics from 2015 to 2024, with an accelerated growth trajectory observed after 2019. This upward trend suggests that the integration of AI and big data in cultural heritage studies has gained significant momentum in recent years.

Among the tracked terms, “artificial intelligence” and “deep learning” exhibit the most pronounced growth, reflecting the increasing application of advanced machine learning techniques in heritage research. Similarly, “big data” and “knowledge graphs” show a continuous rise, indicating the expanding role of large-scale data management and semantic representation in cultural heritage preservation. The presence of “cultural heritage” and “intangible cultural heritage” as steadily growing terms reinforces the fact that digital humanities research remains deeply engaged with heritage-related inquiries, while the term “digital humanities” follows a comparable trajectory, suggesting that computational methods are increasingly being adopted within humanities-based research. The emergence of “knowledge distillation” and “natural language processing” as rising trends underscores the increasing emphasis on AI-driven language models and knowledge extraction techniques, which may be used for digitizing historical texts, analyzing oral traditions, and enhancing accessibility to archival data. The steep growth in these terms after 2020 suggests a rapid expansion in the use of AI to process and structure heritage-related knowledge.

3.5. Trend analysis of emerging topics

Fig. 7 presents a trend analysis of emerging topics within the intersection of intangible cultural heritage and artificial intelligence, mapping their temporal evolution from 2017 to 2024. The horizontal axis represents time, while the vertical axis lists key research terms that have gained prominence over the years. The size of each data point reflects the relative impact or frequency of a given term within the research landscape, offering insights into the chronological progression of scholarly focus in this field.

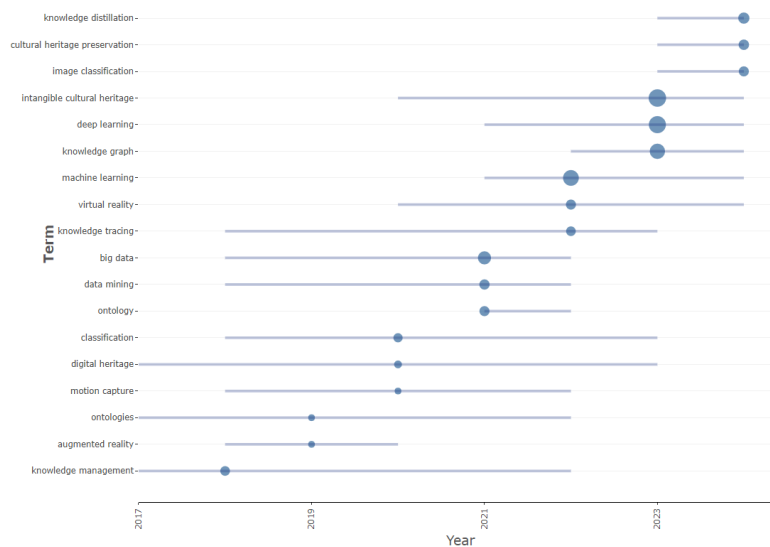


Fig. 7. Emerging Research Trends

The visualization indicates that early AI applications in cultural heritage research primarily focused on knowledge management, augmented reality, and ontologies, which appeared in the pre-2020 period. These early topics reflect foundational efforts in structuring and organizing cultural heritage data, as well as initial explorations into virtual and augmented reality applications for digital preservation. The emergence of motion capture and digital heritage around 2019 suggests an increasing interest in the digitization of cultural expressions and performative traditions, possibly leveraging computational models for recording and analyzing traditional knowledge. A noticeable shift in focus occurs around 2020, where topics such as classification, ontology, data mining, and big data become more prominent. This shift signals the increasing adoption of machine learning and data-driven methodologies in processing and analyzing intangible cultural heritage. The presence of knowledge tracing as an emerging term around the same period suggests that researchers began to explore mechanisms for structuring and exchanging heritage-related knowledge, possibly through AI-enhanced digital platforms and knowledge graph applications.

Post-2021, the research landscape sees a rapid expansion in topics that are directly tied to advanced AI methodologies, including machine learning, deep learning, knowledge graphs, and virtual reality. The increasing prominence of deep learning and knowledge graphs, particularly after 2022, indicates a growing reliance on semantic representation, automated learning models, and neural networks for managing and analyzing intangible cultural heritage data. The rise of cultural heritage preservation and image classification further suggests that researchers are applying computer vision and AI-based recognition techniques to digitally document, classify, and restore cultural artifacts and traditions. One of the most notable trends in the 2023–2024 period is the rise of knowledge distillation, which appears as one of the most recent and impactful research topics. This suggests that AI applications in cultural heritage are shifting toward optimization and compression techniques, enabling efficient knowledge transfer and the distillation of complex cultural data into more accessible forms. The concurrent increase in virtual reality applications further highlights a growing emphasis on immersive heritage experiences, where AI-driven

simulations and VR technologies are being leveraged to recreate historical environments and intangible cultural traditions.

Fig. 8 illustrates the evolution of research topics in the intersection of intangible cultural heritage and artificial intelligence across three distinct time periods: 2015–2018, 2019–2022, and 2023–2024. The Sankey diagram visualizes how research themes have transitioned over time, with connections between different timeframes representing the continuity and transformation of scholarly focus.

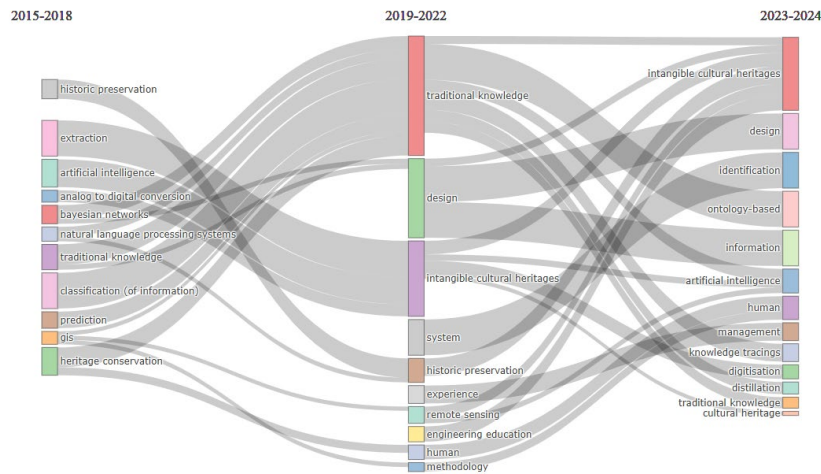


Fig. 8. Evolution of Research Topics across Time Periods (2015–2024)

During the 2015–2018 period, research primarily centered on historic preservation, artificial intelligence, extraction, classification, traditional knowledge, and heritage conservation. These early themes indicate that the initial application of AI in heritage studies focused on digitization, data extraction, and the foundational use of AI-based classification techniques. The presence of natural language processing systems and Bayesian networks suggests an early interest in computational approaches to heritage knowledge representation and semantic structuring. Additionally, topics such as analog-to-digital conversion and prediction imply that AI methodologies were being explored for digitizing historical records and forecasting cultural trends.

Transitioning into the 2019–2022 period, research themes became more refined and interdisciplinary, with a strong emphasis on traditional knowledge, design, intangible cultural heritage, systems, historic preservation, and remote sensing. The persistence of historic preservation and intangible cultural heritage as central themes indicates continuity in efforts to safeguard heritage assets, while the emergence of design and system-based approaches suggests an increasing reliance on structured frameworks and AI-driven methodologies for heritage preservation. The inclusion of engineering education and human experience in this phase reflects an expansion of AI applications beyond heritage documentation, incorporating interactive and educational aspects into digital heritage initiatives.

By the 2023–2024 period, research themes exhibit a significant shift towards advanced AI methodologies, knowledge management, and immersive heritage technologies. The prominence of intangible cultural heritage, artificial intelligence, ontology-based approaches, and identification signals a growing emphasis on semantic modeling, AI-driven knowledge organization, and automated classification techniques. The emergence of digitization, distillation, and knowledge tracing suggests that current research is increasingly concerned with optimizing AI models for cultural data processing, ensuring efficient knowledge transfer, and improving accessibility to heritage-related information. Additionally, the presence of human-centric research themes, such as management, human experience, and information systems, indicates a stronger alignment with user engagement and community-driven heritage conservation initiatives.

The evolution of research themes from digitization and classification (2015–2018) to system design and knowledge structuring (2019–2022), and finally to AI-driven automation, optimization, and management (2023–2024) suggests a progressive refinement of AI applications in cultural heritage research. Early studies focused on foundational digital transformation, while later phases introduced systematic frameworks and interactive technologies, eventually leading to highly automated, knowledge-driven approaches to heritage preservation. Overall, Figure 8 underscores the growing complexity and sophistication of AI applications in cultural heritage studies, highlighting how the field has evolved from basic digital preservation to advanced AI-powered knowledge extraction, interactive heritage experiences, and structured management of cultural information. This trajectory suggests that future research will likely continue to integrate AI, knowledge engineering, and human-centered design principles, further enhancing the role of artificial intelligence in preserving, interpreting, and disseminating intangible cultural heritage.

4. Discussion

4.1. Major Publication Trends

The bibliometric analysis reveals a dramatic rise in scholarly publications at the intersection of ICH and AI over the last decade. Annual output grew from just 15 publications in 2015 to 326 in 2024, reflecting an approximate annual growth rate of 40%. This surge underscores a rapidly expanding academic interest in applying AI and big data techniques to intangible heritage. Citation patterns show that mid-decade works achieved especially high impact: for instance, studies around 2016–2019 garnered the most citations per article (peaking at ~33.98 in 2019), likely due to pioneering contributions that became foundational in this nascent field. By contrast, recent publications (2023–2024) have lower citation averages (e.g. 3.6 in 2023) simply because they have had less time to accumulate citations. These trends suggest an early exploratory phase (pre-2020) that produced influential works, followed by an accelerating phase (2020 onward) with a flood of new research. Over time, the focus of publications has also gradually shifted: early studies often emphasized basic digitization and knowledge organization (e.g. using ontologies or augmented reality for heritage), whereas more recent works increasingly explore advanced AI applications (like deep learning for cultural data and immersive virtual heritage experiences). This shift in academic focus aligns with the maturation of AI technology itself, moving from initial experiments to more sophisticated, data-driven heritage analysis in the late 2010s and early 2020s. In sum, the past decade witnessed both exponential growth in research volume and an evolution in the nature of ICH–AI scholarship – from foundational digital preservation efforts to cutting-edge AI-driven methodologies.

4.2. Distinct Themes from BERTopic Modeling

Using BERTopic, we identified 18 distinct research themes (topic clusters) within the ICH–AI literature. Each theme represents a set of semantically related publications, and together they paint a rich portrait of the field’s diversity. Some themes are anchored in core heritage topics – for example, one prominent cluster (Topic 0) centers on “cultural heritage”, “intangible cultural heritage”, and “heritage preservation”, reflecting foundational concerns of safeguarding and documenting heritage. Similarly, another theme (Topic 3) highlights “traditional music”, “music appreciation”, and “Chinese opera”, indicating a focus on preserving performing arts through digital means. In contrast, several themes emphasize technical AI methodologies in the heritage context. For instance, Topic 1 is characterized by terms like “knowledge graph” and “entity relationships”, signifying work on semantic networks and knowledge representation for ICH. Another theme (Topic 5) features “big data” and “data environments”, underlining large-scale data analytics and infrastructure for cultural heritage. Notably, a few themes illustrate unexpected interdisciplinary crossovers – for example, Topic 7 includes “medical knowledge” and even “colorectal cancer”, suggesting research at the intersection of traditional medicinal heritage and modern biomedical data analysis. Likewise, Topic 12’s keywords (e.g. “watershed management”, “ecosystem services”) show a linkage between heritage and environmental science, pointing to studies that connect cultural practices with natural ecosystems. Overall, the 18 themes range from traditional cultural expressions (e.g. folklore, music, dance) to AI techniques and applications (e.g. machine learning, knowledge extraction, data security) and cross-domain topics (e.g. heritage in healthcare or environmental conservation). This confirms that the research corpus is highly interdisciplinary, blending humanities and arts topics with computer science and other domains.

4.3. Broader Research Directions

Upon examination, these 18 themes can be grouped into four overarching research directions that characterize the ICH–AI field. First is a Heritage Preservation and Digitization direction, encompassing topics focused on documenting and safeguarding intangible culture. This includes themes like traditional music, dance, folk art, and the use of digital tools to preserve them (e.g. Topics 0, 3, 13, 16, 17). Research in this category is rooted in heritage studies and digital humanities, exploring how AI can record, reconstruct, or disseminate traditional knowledge and performances. For example, studies of shadow puppetry and dance preservation fall here, often leveraging technologies like motion capture or social media platforms (as seen with “TikTok” in Topic 16) to engage new audiences in heritage (Herrow & Azraai, 2021; Ami-Williams et al., 2024). The second direction can be described as AI-Driven Knowledge management for ICH, covering themes where advanced computational techniques are applied to heritage data. This group (e.g. Topics 1, 2, 5, 6, 7, 11, 15) includes work on knowledge graphs and ontology-based representations of ICH, machine learning for heritage content analysis, natural language processing of oral histories, and AI-assisted restoration of cultural artifacts. The unifying feature is an emphasis on managing and extracting knowledge from large heritage datasets – for instance, using big data to discover patterns in folkloric texts, or applying text recognition and NLP to historical documents (Topic 15). The third direction highlights Technological and Cross-Domain Innovations in ICH research. Here, scholars borrow and adapt methods from other fields (or address new challenges) in the service of heritage. Included are topics involving cryptography and cybersecurity for heritage data (Topic 9), optimization algorithms and even oil-industry models repurposed for cultural datasets (Topic 10), and the integration of environmental monitoring and GIS (Topic 12) to protect heritage. Also in this category are studies ensuring the integrity and security of digital heritage archives (e.g. blockchain or distributed ledgers suggested by “distributed collision resistance” in Topic 9). The fourth direction revolves around Historical and theoretical integrations on ICH. This includes more theoretical or human-centered approaches (e.g. Topics 4 and 8), such as exploring historical knowledge (like traditional medicine and

indigenous practices in Topic 4) with AI tools, or employing gamification and interactive media in heritage education (Topic 8 on “serious games” and “temporal knowledge graphs”). Work in this vein often intersects with public history, museum studies, and pedagogy – for example, developing serious games to engage people with historical traditions, or using AI to simulate cultural experiences for learning (Kara, 2022). The significance of these four broad directions is that they demonstrate how deeply interdisciplinary the ICH–AI domain is: researchers are not only preserving cultural heritage with digital means, but also pushing technological frontiers (by adapting AI techniques to new contexts) and reimagining how people interact with heritage (through games, virtual reality, and other immersive experiences). This broad categorization underscores that the field operates at the nexus of AI, digital humanities, and heritage studies, drawing methods and perspectives from all three. It also highlights that heritage research in the AI era is multi-faceted – simultaneously concerned with saving the past, innovating for the future, and connecting with global issues like sustainability and education.

4.4. Emerging Research Themes and Evolution

Over the last decade, several research themes have clearly emerged and evolved in prominence, as evidenced by our temporal analysis. In the early years (2015–2018), the dominant topics were foundational: scholars concentrated on historic preservation, digital archiving, and introducing AI to heritage through basic tools like classification algorithms and NLP for knowledge extraction. For example, techniques such as Bayesian networks and early ontology models appear in this period, indicating initial attempts to formally represent cultural knowledge (Ranjgar et al., 2022; Chen et al., 2023). Researchers were also concerned with digitization of analog materials (e.g. converting oral recordings or artifacts into digital form) and even exploratory efforts at predictive modeling of cultural trends. This suggests that the community was laying groundwork by creating digital datasets and trying out how AI might forecast or classify heritage information. Moving into the 2019–2022 period, the field’s focus became more refined and interdisciplinary. Topics like “intangible cultural heritage” and “heritage conservation” remained central, but new emphases on design, systems, and remote sensing emerged. This reflects a shift toward building structured frameworks (e.g. ontology-driven systems or knowledge management platforms) for heritage, as well as leveraging technologies like GIS and satellite sensing to document heritage sites and landscapes (Yao et al., 2023). There is also evidence of broader engagement with human-centered design and education: terms such as “engineering education” and “human experience” start to appear. This indicates that AI in heritage was not just a technical endeavor but began integrating user experience, perhaps through interactive exhibits or educational tools that involve AI (for instance, AR/VR apps for museum education). By the late stage (2023 – 2024), the research themes reflect a marked turn toward advanced AI and knowledge-centric approaches. Terms like “ontology-based approaches”, “knowledge management”, and “identification” (automated identification of heritage elements) become especially prominent. Notably, “knowledge distillation” emerges as a cutting-edge topic in this period, signaling interest in optimizing AI models to handle cultural data more efficiently, for example, simplifying complex deep learning models for practical use in heritage archives (Siountri & Anagnostopoulos, 2023). The increasing visibility of deep learning and virtual reality in the last few years also shows that the field has embraced state-of-the-art AI: researchers are training neural networks for tasks like image classification of artifacts and using VR to create immersive reconstructions of intangible heritage practices (Barbara, 2022). An important emerging theme is the integration of immersive technologies – by 2024, virtual reality and related tools are frequently mentioned, indicating a push toward experiential preservation (allowing people to “experience” traditions digitally) as a complement to merely recording them. We also see continuity and convergence among themes: for instance, the enduring focus on “intangible cultural heritage” itself remains at the core of many topics throughout all time periods, but it gradually becomes intertwined with new techniques (ontologies, AI-driven identification) as the field matures. Some early themes have merged into broader ones – digital preservation and heritage conservation are now often discussed alongside AI methods, rather than in isolation, showing a convergence of heritage expertise with technical innovation (Yan & Li, 2023). Conversely, a few niche topics have diverged or spun off into distinct subfields; for example, the intersection of ICH with environmental sustainability (e.g. climate impact on heritage, noted by terms like “ecosystem services” in Topic 12) was not a major focus in 2015 but gained momentum in the 2020s, evolving into a standalone concern at the crossroads of heritage and ecology. In summary, the last decade’s emerging themes illustrate a trajectory from digitization and basic AI adoption (mid-2010s) to interdisciplinary expansion (around 2020) and finally to cutting-edge AI techniques and new domains (early 2020s). This evolution demonstrates how the field has grown in complexity and scope, continuously incorporating the latest technological advancements (from semantic web to deep learning) and branching into globally relevant areas (like sustainability and health) as it moves forward.

5. Conclusion

This study provides a comprehensive analysis of ICH research in the AI and big data era, revealing significant trends and emerging topics from 2015 to 2024. Scholarly output has grown exponentially, with BERTopic modeling identifying 18 distinct themes grouped into four major research directions: heritage preservation and digitization, AI-driven knowledge management, technological and Cross-domain innovations, and historical and theoretical integrations. Early research focused on digitization and ontology-based knowledge systems, whereas recent studies have embraced deep learning, knowledge graphs, and immersive technologies such as virtual reality for heritage experiences. Notably, emerging topics like knowledge distillation, NLP for folklore analysis, and AI-driven heritage storytelling have gained prominence, indicating a shift toward more advanced and interactive applications.

These findings have practical implications for researchers, policymakers, and cultural institutions. For scholars, this study offers a structured map of ICH–AI research, helping to identify trends and gaps. For policymakers, understanding these emerging themes can guide funding strategies and AI-driven heritage preservation policies. For cultural institutions, leveraging AI methodologies can enhance documentation, public engagement, and accessibility of intangible heritage. Additionally, interdisciplinary collaboration between AI experts, heritage scholars, and practitioners is crucial to ensuring culturally sensitive and sustainable applications of these technologies.

Despite its contributions, this study has limitations. The dataset, limited to indexed academic literature, may not capture non-English publications or community-driven heritage documentation. While BERTopic modeling provides valuable thematic insights, some topics may be oversimplified or require further refinement. Future research should expand datasets, integrate multimodal AI analysis (e.g., combining text with image and audio analysis), and strengthen global collaboration in ICH–AI studies. Addressing these gaps will enhance the role of AI in safeguarding intangible cultural heritage, ensuring its continued preservation and transmission in the digital age.

References

- Alhaj, F., Al-Haj, A., Sharieh, A., & Jabri, R. (2022). Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic. *International Journal of Advanced Computer Science and Applications*, 13(1). <https://doi.org/10.14569/ijacsa.2022.0130199>
- Ami-Williams, T., Serghides, C., & Aristidou, A. (2024). Digitizing traditional dances under extreme clothing: The case study of Eyo. *Journal of Cultural Heritage*, 67, 145–157. <https://doi.org/10.1016/j.culher.2024.02.011>
- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Banerjee, S., & Pan, A. (2024). From Colonial Legacies to Linguistic Inclusion: A BERTopic Enhanced Bibliometric Insight into Global South Higher Education. *IEEE Access*, 12, 117418–117435. <https://doi.org/10.1109/access.2024.3447894>
- Barbara, J. (2022). Re-Live History: An immersive virtual reality learning experience of prehistoric intangible cultural heritage. *Frontiers in Education*, 7. <https://doi.org/10.3389/educ.2022.1032108>
- Chen, D., Sun, N., Lee, J., Zou, C., & Jeon, W. (2024). Digital Technology in Cultural Heritage: Construction and Evaluation Methods of AI-Based Ethnic Music Dataset. *Applied Sciences*, 14(23), 10811. <https://doi.org/10.3390/app142310811>
- Chen, J., Ding, L., Ji, J., & Zhu, J. (2023). A Combined Method to Build Bayesian Network for Fire Risk Assessment of Historical Buildings. *Fire Technology*, 59(6), 3525–3563. <https://doi.org/10.1007/s10694-023-01475-8>
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2VEC, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7. <https://doi.org/10.3389/fsoc.2022.886498>
- Eichler, J. (2020). Intangible cultural heritage, inequalities and participation: Who decides on heritage? *The International Journal of Human Rights*, 25(5), 793–814. <https://doi.org/10.1080/13642987.2020.1822821>
- Gabarron, E., Dorrnzoro, E., Reichenpfader, D., & Denecke, K. (2023). What do autistic people discuss on Twitter? An approach using BERTopic modelling. *Studies in Health Technology and Informatics*. <https://doi.org/10.3233/shiti230161>
- Gonzalez-Gomez, L. J., Hernandez-Munoz, S. M., Borja, A., Azoifeifa, J. D., Noguez, J., & Caratozzolo, P. (2024). Analyzing Natural Language Processing Techniques to Extract Meaningful Information on Skills Acquisition from Textual Content. *IEEE Access*, 12, 139742–139757. <https://doi.org/10.1109/access.2024.3465409>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2203.05794>
- Herrow, M. F. M., & Azraai, N. Z. (2021). DIGITAL PRESERVATION OF INTANGIBLE CULTURAL HERITAGE OF JOGET DANCE MOVEMENT USING MOTION CAPTURE TECHNOLOGY. *International Journal of Heritage Art and Multimedia*, 4(15), 01–13. <https://doi.org/10.35631/ijham.415001>
- Kara, N. (2022). A Mixed-Methods Study of Cultural Heritage Learning through Playing a Serious Game. *International Journal of Human-Computer Interaction*, 40(6), 1397–1408. <https://doi.org/10.1080/10447318.2022.2125627>
- Liang, Y., Xie, B., Tan, W., & Zhang, Q. (2025). Ontology-based construction of embroidery intangible cultural heritage knowledge graph: A case study of Qingyang sachets. *PLoS ONE*, 20(1), e0317447. <https://doi.org/10.1371/journal.pone.0317447>
- Liu, S., & Pan, Y. (2023). Exploring Trends in Intangible Cultural Heritage Design: A Bibliometric and Content analysis. *Sustainability*, 15(13), 10049. <https://doi.org/10.3390/su151310049>
- Meitei, C. M., Dabas, S., & Kumar, A. (2024). Image Compression for Communication: Topic modelling of Scopus Abstracts using BERTopic. *2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 1–6. <https://doi.org/10.1109/wpmc63271.2024.10863726>
- Münster, S., Maiwald, F., Di Lenardo, I., Henriksson, J., Isaac, A., Graf, M. M., Beck, C., & Oomen, J. (2024). Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe. *Heritage*, 7(2), 794–816. <https://doi.org/10.3390/heritage7020038>
- Münster, S., Utescher, R., & Aydogan, S. U. (2021). Digital topics on cultural heritage investigated: how can data-driven and data-guided methods support to identify current topics and trends in digital heritage? *Built Heritage*, 5(1). <https://doi.org/10.1186/s43238-021-00045-7>

- Ranjgar, B., Sadeghi-Niaraki, A., Shakeri, M., & Choi, S. (2022). An ontological data model for points of interest (POI) in a cultural heritage site. *Heritage Science*, 10(1). <https://doi.org/10.1186/s40494-021-00635-9>
- Samsir, S., Saragih, R. S., Subagio, S., Aditiya, R., & Watrianthos, R. (2023). BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(3), 1514. <https://doi.org/10.30865/mib.v7i3.6426>
- Severo, M. (2018). Safeguarding without a record? The digital inventories of intangible cultural heritage. In *Springer eBooks* (pp. 165–182). https://doi.org/10.1007/978-3-319-75759-9_9
- Shakya, M., & Vagnarelli, G. (2024). Creating value from intangible cultural heritage—the role of innovation for sustainable tourism and regional rural development. *European Journal of Cultural Management and Policy*, 14. <https://doi.org/10.3389/ejcmp.2024.12057>
- Siountri, K., & Anagnostopoulos, C. (2023). The Classification of Cultural Heritage Buildings in Athens Using Deep Learning Techniques. *Heritage*, 6(4), 3673–3705. <https://doi.org/10.3390/heritage6040195>
- UNESCO: *Exploring the impact of artificial intelligence and intangible cultural heritage*. (2024). <https://ich.unesco.org/en/news/exploring-the-impact-of-artificial-intelligence-and-intangible-cultural-heritage-13536#:~:text=The%20benefits%20of%20AI%20in,proverbs%2C%20leading%20to%20improved%20global>
- Wang, Q., & Ma, B. (2024). Enhancing BERTopic with Pre-Clustered Knowledge: Reducing Feature Sparsity in Short Text Topic Modeling. *Journal of Data Analysis and Information Processing*, 12(04), 597–611. <https://doi.org/10.4236/jdaip.2024.124032>
- Wang, Z., Chen, J., Chen, J., & Chen, H. (2023). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*. <https://doi.org/10.1007/s11192-023-04776-5>
- Yan, W., & Li, K. (2023). Sustainable Cultural Innovation Practice: Heritage Education in Universities and Creative Inheritance of Intangible Cultural Heritage Craft. *Sustainability*, 15(2), 1194. <https://doi.org/10.3390/su15021194>
- Yao, Y., Wang, X., Luo, L., Wan, H., & Ren, H. (2023). An Overview of GIS-RS Applications for Archaeological and Cultural Heritage under the DBAR-Heritage Mission. *Remote Sensing*, 15(24), 5766. <https://doi.org/10.3390/rs15245766>
- Zhang, B., Cheng, P., Deng, L., Romainoor, N. H., Han, J., Luo, G., & Gao, T. (2023). Can AI-generated art stimulate the sustainability of intangible cultural heritage? A quantitative research on cultural and creative products of New Year Prints generated by AI. *Heliyon*, 9(10), e20477. <https://doi.org/10.1016/j.heliyon.2023.e20477>



© 2026 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).