Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Has Digital Preservation Research Evolved? A Transformer-Based Topic Modeling and Temporal Analysis of Digital Preservation and Curation in Libraries, Archives and Cultural Heritage Institutions

Chandan Jyoti Baishya

chandanj.baishya@cottonuniversity.ac.in

Cotton University

Pompy Deka

University of Delhi

Research Article

Keywords: Digital Preservation, Digital Curation, Web Archiving, Cultural Heritage, BERTopic, Dynamic Topic Modeling

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11731242/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Libraries, archives and cultural heritage organizations now consider digital preservation and curation to be fundamental duties. However, the research literature on these activities is extensive, multidisciplinary and dispersed and its thematic structure and evolution have not been thoroughly mapped. The main research areas in digital preservation and curation are identified in this paper, which also charts their evolution from 2016 to 2025. 4,091 articles, conference papers and data papers were kept after deduplication, screening and eligibility evaluation of the 7,552 records that were obtained from the Web of Science Core Collection and Scopus in accordance with PRISMA 2020 principles. Within the BERTopic framework, titles and abstracts were embedded with a pretrained RoBERTa transformer using mean pooling and whitening, reduced with UMAP, clustered with HDBSCAN and described with class-based TF-IDF; each topic was then tracked over the course of publication years using dynamic topic modeling. Web archiving, born-digital records, institutional repositories, FAIR-oriented standards, DNA data storage, 3D and architectural heritage documentation, digitization of dance, music and crafts, OCR for low-resource scripts and community, queer and feminist archives were among the 27 topics identified by the model. Nine research directions were created from these. The output more than tripled over the course of the decade, according to temporal analysis and the focus of the field has changed. While web and born-digital archiving (− 2.7) and repository and research-data infrastructure (− 2.2) declined in relative share while remaining active, 3D and built heritage documentation (+ 4.7 percentage points of share) and AI and intelligent archival systems (+ 3.9) grew at the fastest rates. The results provide researchers and practitioners with an evidence-based field map and emphasize the necessity to maintain investment in born-digital collections and plan for the long-term preservation of AI-generated and 3D outputs.

Introduction

Libraries, archives, museums and other cultural heritage institutions now hold a large and rapidly growing share of their collections in digital form. Some of this material is digitized from analogue originals; an increasing proportion is born-digital, including e-records transferred from government bodies, research datasets, websites, social media content, email and audiovisual files. Unlike paper, digital objects depend on specific file formats, software, hardware and storage media, all of which become obsolete. Without deliberate and continuous intervention, digital content can become unreadable within years rather than centuries, a risk the field has long described as a potential "digital dark age." Digital preservation, understood as the series of managed activities necessary to ensure continued access to digital materials for as long as necessary and digital curation, which extends that responsibility across the whole lifecycle of creation, appraisal, description, storage, access and reuse (Higgins, 2008), have therefore moved from specialist concerns to core functions of memory institutions.

The period from 2016 to 2025 has been mainly eventful for this field. Reference frameworks such as the Open Archival Information System (OAIS) model (CCSDS, 2012) and preservation metadata standards matured into routine practice, while the FAIR principles (Wilkinson et al., 2016) reframed long-term

preservation as a precondition for findable, accessible, interoperable and reusable research data. At the same time, new technical possibilities and pressures emerged: large-scale web and social media archiving (Brügger, 2018; Costa et al., 2017), 3D documentation of built and intangible heritage, blockchain-based integrity assurance, DNA and other novel storage media and the application of machine learning and artificial intelligence (AI) to archival description, appraisal and access (Colavizza et al., 2021). The COVID-19 pandemic further accelerated digitization programmes and prompted rapid-response collecting of web content.

The resulting literature is large, interdisciplinary and scattered across library and information science, archival science, computer science, digital humanities, heritage science and photogrammetry. Existing reviews typically focus on a single subdomain, such as web archiving or research data repositories, or rely on keyword co-occurrence and citation analysis, which capture structural relations between terms and documents but are less able to represent the semantic content of abstracts. As a result, there is still no fine-grained, data-driven map of what the field as a whole has been researching over the last decade and how its emphases have shifted.

Following the approach of Xu et al. (2025), who combined BERTopic and Dynamic Topic Modeling (DTM) to trace AI research in library information services, this study applies transformer-based topic modeling to the digital preservation and curation literature. BERTopic (Grootendorst, 2022) clusters contextual document embeddings and describes each cluster with class-based TF-IDF, while its dynamic extension tracks how topics change over time. The study is guided by three research questions:

RQ1

What are the key research topics in digital preservation and curation research in libraries, archives and cultural heritage institutions between 2016 and 2025?

RQ2

Into which major research directions can these topics be organized?

RQ3

How have these research directions evolved over time?

By answering these questions, the study provides an evidence-based overview of the field's intellectual structure and trajectory, which can help researchers position new work and help practitioners and funders prioritize investment in preservation capacity.

Related Work

Digital Preservation and Curation in Memory Institutions

Research on digital preservation has long been organized around a small number of reference models. The OAIS reference model, standardized as ISO 14721, defines the functional entities of an archive (ingest, archival storage, data management, administration, preservation planning and access) and the information packages that move between them (CCSDS, 2012). Lavoie (2004) demonstrated that OAIS provided a common terminology for libraries and archives who were developing preservation repositories. Higgins (2008) developed the Digital Curation Centre's lifecycle model which expanded the focus from simply storing content to a series of curation activities throughout its lifetime. Lifecycle thinking has influenced how universities develop research data management services; however, Cox & Tam (2018) caution that lifecycle models may be too simplistic to represent the chaotic and repetitive nature of conducting research.

In the same line with these models, many pieces of work discuss preservation strategies and their evaluation as well. Becker et al. (2009) proposed a systematic preservation planning method that evaluates migration and emulation alternatives against institutional requirements and community tools for format identification, validation and fixity checking have become routine components of repository workflows. Distributed approaches such as LOCKSS ("Lots of Copies Keep Stuff Safe") replicate content across independent institutions to reduce the risk of loss (Rosenthal et al., 2005). Corrado and Sandy (2017) emphasize that technology is only one leg of preservation capacity, alongside organizational commitment, policy and sustainable resourcing.

Web archiving has developed into a distinct strand. National libraries and the Internet Archive collect websites at scale and researchers have examined crawling strategies, quality assurance, access tools and the scholarly use of archived web material (Brügger, 2018; Costa et al., 2017). In parallel, the transfer of born-digital government and personal records has raised questions about appraisal, authenticity and contextual information (Lee, 2011). More recently, AI and machine learning have been proposed for metadata generation, handwritten text recognition, appraisal and sensitivity review, while raising concerns about transparency, bias and the archival record (Colavizza et al., 2021; Jaillant, 2022).

Topic Modeling and BERTopic

Topic modeling is an unsupervised technique for discovering latent themes in large text collections. Classical probabilistic models such as Latent Dirichlet Allocation (LDA) represent documents as mixtures of topics and topics as distributions over words (Blei et al., 2003). Because these models rely on bag-of-words representations, they ignore word order and context and often struggle with short, technical or semantically dense texts such as scientific abstracts. Neural approaches instead build on contextual embeddings from pretrained transformer language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

BERTopic (Grootendorst, 2022) combines four modular steps: (1) contextual document embeddings, (2) dimensionality reduction with UMAP (McInnes et al., 2018), (3) density-based clustering with HDBSCAN (Campello et al., 2013) and (4) topic description with class-based TF-IDF (c-TF-IDF). Comparative studies

report that BERTopic produces more interpretable topics than LDA and NMF on heterogeneous corpora (Egger & Yu, 2022) and it has been used to map research fields, policy documents and patents. Its dynamic topic modeling component recalculates topic representations for each time slice without re-clustering, allowing the semantic evolution of each topic to be followed consistently over time. In library and information science, Xu et al. (2025) used this pipeline to identify 27 topics and ten research directions in AI-enabled library services. To the authors' knowledge, no study has yet applied transformer-based topic modeling to the digital preservation and curation literature as a whole, which is the gap this study addresses.

Research Methods

Data Sources

Following the PRISMA 2020 guidelines (Page et al., 2021), records were retrieved from two complementary bibliographic databases: The Web of Science Core Collection (WoSCC; Science Citation Index Expanded, Social Sciences Citation Index and Arts & Humanities Citation Index) and Scopus. The searches combined terms describing digital preservation activities with terms describing the institutional and collection contexts in which they take place. The exact queries are reported in Table 1. Both searches were run on 4 November 2026 and limited to publications from 2016 to 2025; the Scopus search was additionally limited to English-language documents at the database level.

Table 1
Search queries used in each database

Database	Search query
WoSCC (SCI-E, SSCI, A&HCI)	TS=(("digital preservation" OR "digital curation" OR "web archiv*" OR "born-digital" OR "long-term preservation" OR "digital archiv*") AND (librar* OR archiv* OR repositor* OR "cultural heritage" OR museum*)); Publication Years: 2016–2025
Scopus	TITLE-ABS-KEY(("digital preservation" OR "digital curation" OR "web archiv*" OR "born-digital" OR "long-term preservation" OR "digital archiv*" OR "preservation metadata" OR "digital repositor*") AND (librar* OR archiv* OR repositor* OR "cultural heritage" OR museum* OR "research data")) AND (LIMIT-TO(LANGUAGE, "English")) AND PUBYEAR > 2015 AND PUBYEAR < 2026

The searches returned 1,458 records from WoSCC and 6,094 records from Scopus, a combined total of 7,552. Records were then processed in four stages (Fig. 1). First, 268 records without a usable abstract were removed and 1,222 duplicates were removed by matching normalized DOIs and, for records without a DOI, normalized titles; where a record appeared in both databases the WoSCC version was retained. This left 6,062 unique records for screening. Second, 1,155 records were excluded because they fell outside 2016–2025 (45), were not in English (69), or were not research articles, conference papers or data papers (1,041; mainly reviews, book chapters, books, editorials, errata, notes and letters). Third, the remaining 4,907 records were assessed for eligibility. Because broad terms such as "digital archive" or "digital repository" frequently occur as passing mentions (for example, a note that a dataset was

deposited in a repository), a record was retained only if the core preservation concept appeared in its title or author keywords, appeared at least twice in its abstract, or appeared in the abstract together with an institutional or preservation context term (e.g., library, archives, museum, heritage, collection, curation). This rule excluded 816 records, most of which were clinical, engineering or natural-science papers that mentioned a repository or archive only incidentally. The final dataset comprised **4,091 records** (914 from WoSCC and 3,177 from Scopus).

Research Design

The study used BERTopic for topic identification and its dynamic topic modeling extension for temporal analysis. Each record was represented by its title and abstract. Documents were embedded with a pretrained transformer, the embeddings were projected to a lower-dimensional space with UMAP and clustered with HDBSCAN and each cluster was described with c-TF-IDF keywords. Topics were then grouped into broader research directions and tracked across publication years (Fig. 2).

Preparation and Pre-processing

All analyses were run in Python 3.13 on a Linux workstation (two CPU cores, 8 GB RAM, no GPU). The main libraries and versions are listed in Table 2.

Table 2
Python libraries employed

Libraries with version	
bertopic = = 0.17.4	torch = = 2.15.0
umap-learn = = 0.5.12	scikit-learn = = 1.9.1
hdbscan = = 0.8.44	numpy = = 2.5.3
spacy = = 3.8.18	pandas = = 3.0.5
en_core_web_trf = = 3.8.0	matplotlib = = 3.11.2
spacy-curated-transformers = = 0.3.1	

Pre-processing was kept deliberately light, because transformer models rely on natural word order and context. Two cleaning steps were applied before embedding. First, publisher and licence boilerplate appended to abstracts (copyright statements, "all rights reserved", open-access licence text and conference series notices) was removed with regular expressions, because in a first trial run such text formed its own spurious clusters. Titles were also normalized: titles written entirely in capitals were converted to sentence case and translated titles appended in square brackets were removed, because the case-sensitive transformer otherwise grouped documents by typography rather than content. Second, the headings of structured abstracts used by some publishers ("Purpose",

"Design/methodology/approach", "Findings", "Originality/value" and similar) were stripped so that the remaining sentences read as continuous prose. No stemming or stopword removal was applied to the embedding input. Stopwords were removed only when building topic keyword representations: the vectorizer used the scikit-learn English stopword list extended with generic academic terms (e.g., "paper", "study", "approach", "results") and with frequent function words from the Spanish, Portuguese, German and French text that occurs in some bilingual abstracts. Unigrams and bigrams occurring in at least three documents were retained.

BERTopic Model Construction

Text Embedding

Each document (title followed by abstract) was converted into a dense vector using the transformer component of the spaCy en_core_web_trf pipeline, which is based on RoBERTa-base (Liu et al., 2019), a 12-layer, 768-dimensional transformer pretrained on a large English corpus. The model was run locally. For each document, the hidden states of all word-piece tokens from the final transformer layer were averaged (mean pooling), giving one fixed-length vector per document:

$s = (1/n) \Sigma_{i=1}^{n} h_i$ (1)

where h_i is the final-layer representation of the i-th token and n is the number of tokens. Mean-pooled vectors from a general-purpose transformer are known to be anisotropic, meaning that they occupy a narrow cone in the embedding space so that all documents appear similar (Ethayarajh, 2019). In our first trial, clustering these raw vectors produced only two topics. Following the whitening approach of Su et al. (2021), the vectors were therefore centred, transformed with principal component analysis (PCA) whitening to 256 dimensions and L2-normalized. This step spreads the documents more evenly in the space while preserving their relative semantic positions and it substantially improved cluster separation.

Dimensionality Reduction

The whitened embedding's were reduced to 10 dimensions with UMAP, a non-linear technique that preserves local neighbourhood structure as well as much of the global structure. We set n_neighbors = 15 to balance local and global structure and min_dist = 0.0 to allow tight clusters and used cosine distance as the similarity measure:

$d_{cos}(x_i, x_j) = 1 - (x_i \cdot x_j) / (\|x_i\| \|x_j\|)$ (2)

where x_i and x_j are the embeddings of documents i and j. A fixed random seed (42) was used so that results are reproducible.

Clustering

The reduced vectors were clustered with HDBSCAN (Campello et al., 2013; McInnes et al., 2017), a hierarchical density-based algorithm that does not require the number of clusters to be fixed in advance, can find clusters of arbitrary shape and labels points in low-density regions as noise rather than forcing them into a cluster. HDBSCAN first computes the core distance of each point i as its distance to its k-th nearest neighbour:

core_k(i) = d(i, NN_k(i)) (3)

and then defines the mutual reachability distance between points i and j as:

d_mreach(i, j) = max{core_k(i), core_k(j), d(i, j)} (4)

where d(i, j) is the Euclidean distance in the reduced space. A minimum spanning tree over these distances is condensed into a cluster hierarchy, from which the most stable clusters are selected (excess-of-mass method). In this study, min_cluster_size was set to 25 and min_samples to 10, so that each topic contains at least 25 documents and prediction_data = True was enabled so that documents could later be assigned to existing clusters.

Topic Representation

Each topic was described with c-TF-IDF, which treats all documents in a cluster as one concatenated document and weights terms that are frequent in that cluster but rare across clusters (Grootendorst, 2022):

W(t, c) = tf(t, c) × log(1 + A / f(t)) (5)

where tf(t, c) is the frequency of term t in class c, f(t) is the frequency of term t across all classes and A is the average number of words per class. The ten highest-weighted terms were kept for each topic. The authors then read the keywords and the most representative abstracts of each topic and assigned a short descriptive label.

Dynamic Topic Modeling

To examine temporal change, the static model was first fitted to the whole corpus. Documents were then grouped by publication year (2016–2025) and c-TF-IDF representations were recalculated for each topic in each year without re-clustering:

W(t, c, τ) = tf(t, c, τ) × log(1 + A / f(t)) (6)

where τ denotes the year. Global tuning (averaging each yearly representation with the global one) and evolutionary tuning (averaging with the previous year) were enabled to stabilize the yearly keywords. Because annual output grew markedly over the period, topic frequencies were analysed both as raw counts and as shares of each year's total output, so that growth in a topic could be distinguished from growth of the field as a whole.

Hyperparameter Turning and Outlier Handling

Because the corpus is larger and more heterogeneous than the sample study's, several settings were compared before the final model was chosen. With min_cluster_size values of 15, 20, 25 and 30, the model produced 45, 36, 27 and 22 topics respectively. A minimum cluster size of 25 (with min_samples = 10) was selected because it gave topics that were large enough to be stable across years while still separating substantively different research areas; smaller values split coherent themes into near-duplicates and larger values merged distinct areas such as web archiving and born-digital records. The number of topics was not fixed in advance or reduced afterwards (nr_topics = None), because automatic topic merging collapsed the model to a handful of very broad topics in a pilot run.

HDBSCAN initially labelled 2,291 documents (56.0%) as outliers, which is typical for density-based clustering of heterogeneous abstracts. To retain as much of the corpus as possible without forcing unrelated documents into topics, each outlier was reassigned to its most similar topic only if the cosine similarity between its embedding and the topic centroid was at least 0.15, a value close to the 10th percentile of similarities observed among documents that HDBSCAN had placed in topics (0.14). This reassigned 650 documents; 1,641 documents (40.1%) remained unassigned and were excluded from topic-level counts. Topic keywords were recalculated after reassignment.

Findings

Key Research Topics

The corpus grew from 276 records in 2016 to 896 in 2025, with output roughly stable at 280–390 records per year until 2023 and a sharp increase in 2024 and 2025, which together account for about a third (33.7%) of all included records (Fig. 6a). Conference papers made up about 36% of the corpus, reflecting the strong computer-science component of the field; the most frequent sources included Lecture Notes in Computer Science, the ACM/IEEE Joint Conference on Digital Libraries, CEUR Workshop Proceedings, Library Philosophy and Practice, the Proceedings of the Association for Information Science and Technology and the ISPRS Archives.

The BERTopic model identified **27 topics** (Table 3). Figure 3 shows the five highest-scoring terms of the first 16 topics. The following paragraphs describe the most important topics, grouped by theme.

Table 3
Topics identified, with document counts and representative keywords

No.	Topic label	n	Representative keywords (c-TF-IDF)
0	AI and deep learning for heritage image and text recognition	293	model, accuracy, recognition, cultural, image, digital, data
1	3D digitization, photogrammetry and point clouds	137	heritage, models, cultural, cultural heritage, point, reconstruction, data
2	Web archiving: crawling, replay and archivability	115	web, web archives, archived, pages, page, archives, social
3	Archiving crises, conflict and pandemics	111	pandemic, war, digital, disaster, archives, online, world
4	Physical long-term preservation of specimens and materials	102	longterm, specimens, degradation, preservation, longterm preservation, species, storage
5	Community, queer and feminist digital archives	111	archives, digital, archive, queer, history, archival, digital archives
6	Archaeological and spatial data archiving	103	digital, data, archaeological, preservation, access, archaeology, resources
7	Literary, artistic and personal digital archives	93	archive, digital, digital archive, archives, archival, family, material
8	Architectural heritage documentation (HBIM)	99	heritage, cultural, cultural heritage, architectural, digital, temple, site
9	Institutional repositories and ETD preservation in university libraries	106	repositories, libraries, irs, digital, institutional, policies, university libraries
10	Landscape, urban and site heritage digitization	95	heritage, cultural, digital, preservation, cultural heritage, conservation, river
11	Digitizing dance, music and traditional crafts	90	cultural, traditional, dance, culture, heritage, cultural heritage, preservation
12	Critical perspectives on digital archives and memory	92	digital, archives, archive, data, heritage, cultural, archival
13	Legal frameworks, standards and FAIR principles	81	data, digital, legal, fair, preservation, cultural, heritage
14	Storage media and preservation infrastructures (incl. DNA storage)	92	data, digital, heritage, storage, cultural, information, dna storage
15	Semantic web, linked data and archival search	81	data, web, semantic, documents, archives, information, digital

Note. n = documents assigned after outlier reassignment; 1,641 documents remained unassigned (outliers).

No.	Topic label	n	Representative keywords (c-TF-IDF)
16	Born-digital records: accessioning, processing and transfer	59	borndigital, records, collections, borndigital records, digital, public bodies, archives
17	Intelligent archives management (cloud, blockchain)	62	management, archives, information, technology, data, intelligent, digital archives
18	Language resources and OCR for low-resource scripts	58	language, languages, arabic, english, digital, indigenous, resources
19	Research datasets and digitized collections	56	data, available, digital, project, biographical, collection, repository
20	Preservation practice, skills and risk in institutions	50	digital, data, preservation, archives, information, archival, language
21	Technologies for intangible cultural heritage protection	75	heritage, cultural, cultural heritage, intangible, intangible cultural, preservation, digital
22	Archivists, users, email appraisal and privacy	69	users, archivists, email, digital, privacy, archives, data
23	Regional and non-English-speaking repository initiatives	30	digital, um, preservation, archivo, archaeological, peru, media
24	Research data management, link rot and scholarly repositories	61	digital, data, information, articles, repositories, conclusion, library
25	Digital humanities and written heritage	62	digital, cultural, preservation, heritage, music, knowledge, libraries
26	Library digitization projects and access services	67	digital, library, preservation, archives, university, project, collection

Note. n = documents assigned after outlier reassignment; 1,641 documents remained unassigned (outliers).

The largest topic, **Topic 0 (AI and deep learning for heritage image and text recognition, n = 293)**, with keywords such as "model," "accuracy," "recognition" and "image," brings together work that applies convolutional networks, vision transformers and other deep learning models to cultural heritage materials: handwritten and historical document recognition, binarization of degraded pages, mural and artefact classification and damage detection. Its dynamic keywords show a clear change of character. In 2016 the topic was described by "web," "data" and "query," reflecting information-system and retrieval papers; by 2020 terms such as "open source" dominated; and by 2025 it was defined by "accuracy," "model," "recognition" and "image," the vocabulary of supervised deep learning. Closely related, **Topic 18 (Language resources and OCR for low-resource scripts)** covers OCR and annotation for scripts and languages such as Ge'ez, Telugu, Balinese and classical Arabic, where preservation and machine readability depend on building new language resources.

A second cluster of topics concerns the documentation of physical heritage. **Topic 1 (3D digitization, photogrammetry and point clouds, n = 137)** addresses laser scanning, photogrammetry, point-cloud

processing and 3D model preservation, including ontologies for 3D heritage models and library services for 3D/VR content. **Topic 8 (Architectural heritage documentation)** focuses on monuments, temples and historic buildings, often using Heritage Building Information Modelling (HBIM), while **Topic 10 (Landscape, urban and site heritage digitization)** extends documentation to heritage landscapes, rivers, cities and archaeological sites.

Web and born-digital archiving forms a distinct and long-established strand. **Topic 2 (Web archiving: crawling, replay and archivability, n = 115)** covers crawler design, replay of archived pages, archivability assessment, Memento-based access and the scholarly use of web archives and **Topic 15 (Semantic web, linked data and archival search)** addresses entity linking, temporal search and ranking in archival collections. **Topic 16 (Born-digital records)** deals with the accessioning, processing, description and transfer of born-digital collections, including email, software-dependent literary works and records transferred from public bodies; its 2025 keywords ("accessioning," "workflows," "records") show a shift towards operational workflows. **Topic 22 (Archivists, users, email appraisal and privacy)** examines how archivists and users engage with digital archives, including automated appraisal of email and privacy in access.

Several topics concern institutional infrastructure. **Topic 9 (Institutional repositories and ETD preservation, n = 106)** covers digital preservation in university libraries and institutional repositories, frequently in African and South Asian contexts, with recurring attention to electronic theses and dissertations, OAIS-based information packages and open-source platforms such as DSpace. **Topic 24 (Research data management, link rot and scholarly repositories)** addresses data management services, URL decay in scholarly citations and FAIR-oriented repository design. **Topic 13 (Legal frameworks, standards and FAIR principles)** deals with legal deposit, archival law, metadata standards (e.g., Dublin Core) and FAIR compliance and **Topic 14 (Storage media and preservation infrastructures)** includes preservation storage criteria, infrastructure sustainability and, increasingly, DNA-based data storage.

Intangible heritage forms a fast-growing group. **Topic 11 (Digitizing dance, music and traditional crafts, n = 90)** includes motion capture of folk and classical dance, digital archives of music and opera and the documentation of embroidery, clay figures and other crafts, often combined with recognition models. **Topic 21 (Technologies for intangible cultural heritage protection)** addresses digital platforms, virtual reality and knowledge bases for safeguarding intangible heritage.

A further group of topics takes a critical and social perspective on archives. **Topic 5 (Community, queer and feminist digital archives, n = 111)** examines LGBTQ+, feminist and activist archives, community metadata and questions of representation. **Topic 3 (Archiving crises, conflict and pandemics)** covers rapid-response collecting around COVID-19, war and natural disasters such as the 2011 Great East Japan Earthquake. **Topic 7 (Literary, artistic and personal digital archives)** and **Topic 12 (Critical perspectives on digital archives and memory)** examine digital archives as sites of memory, authorship and knowledge production. Finally, **Topic 4 (Physical long-term preservation of specimens and materials)** is a distinctive topic on the conservation science side of the field, including the long-term

preservation of natural-history specimens, material degradation and microbial deterioration in collections.

Major Research Directions

Figure 4 shows the two-dimensional UMAP projection of all documents, coloured by research direction. Some topics, such as architectural heritage documentation (Topic 8), legal frameworks (Topic 13), language resources (Topic 18), research datasets (Topic 19) and preservation practice (Topic 20), form compact, well-separated clusters. Others, such as critical perspectives on digital archives (Topic 12) and library digitization projects (Topic 26), are more dispersed, indicating that they draw on vocabulary shared with several other areas. The large grey region in the centre consists of documents that the model left unassigned; these are typically general or cross-cutting papers.

To identify broader structure, the similarity between topics was calculated from their c-TF-IDF vectors (Fig. 5a) and the topics were clustered hierarchically (Fig. 5b). Pairwise similarities ranged from 0.16 to 0.72 (median 0.45). The most similar pairs were 3D digitization and architectural documentation (Topics 1 and 8; 0.72), community archives and literary/personal archives (Topics 5 and 7; 0.69), archaeological data archiving and critical perspectives on archives (Topics 6 and 12; 0.68) and institutional repositories and research data management (Topics 9 and 24; 0.63) and the two intangible heritage topics (Topics 11 and 21; 0.63). The least connected topics were the regional, largely non-English repository initiatives (Topic 23; mean similarity 0.25) and web archiving (Topic 2; mean 0.34), whose technical vocabulary (crawling, replay, Memento) is specific to that community.

Drawing on these similarity patterns, the document map and a reading of the representative abstracts, the 27 topics were organized into nine research directions (Table 4).

Table 4
Research directions, their topics and change in share of output

Research direction	Topics	n	Share 2016–20 (%)	Share 2021–25 (%)	Change (pp)
RD1: Web and Born-Digital Archiving	2, 15, 16, 22	324	9.6	6.9	-2.7
RD2: Repositories, Research Data and Scholarly Infrastructure	9, 19, 24, 6	326	9.3	7.1	-2.2
RD3: Preservation Policy, Practice and Storage Infrastructure	13, 14, 20	223	5.0	5.8	+ 0.8
RD4: 3D and Built Heritage Documentation	1, 8, 10	331	5.2	9.9	+ 4.7
RD5: Intangible Heritage and Endangered Languages	11, 21, 18	223	4.4	6.1	+ 1.7
RD6: AI and Intelligent Archival Systems	0, 17	355	6.3	10.2	+ 3.9
RD7: Critical, Community and Memory Archives	3, 5, 7, 12	407	9.1	10.5	+ 1.4
RD8: Digital Humanities and Library Digitization	25, 26, 23	159	2.2	5.0	+ 2.8
RD9: Material and Specimen Preservation	4	102	2.3	2.6	+ 0.3

Note. Shares are the direction's documents as a percentage of all included records published in each period (2016–2020: 1,589; 2021–2025: 2,502). pp = percentage points.

Research Direction 1 (RD1): Web and Born-Digital Archiving (Topics 2, 15, 16 and 22) covers the capture, replay, search and appraisal of web and born-digital content. It addresses crawling and archivability, semantic and temporal access to archived content, accessioning and processing of born-digital records and the archivist and user perspectives that shape these collections.

RD2: Repositories, Research Data and Scholarly Infrastructure (Topics 6, 9, 19 and 24) concerns the repositories that hold scholarly outputs and research data: institutional repositories and ETDs, research data management and link rot, disciplinary archives for archaeological and spatial data and published datasets and digitized collections.

RD3: Preservation Policy, Practice and Storage Infrastructure (Topics 13, 14 and 20) brings together the organizational and technical foundations of preservation: legal deposit and archival law, standards and FAIR principles, storage criteria and new storage media and the skills, risks and practices of preservation in institutions such as university libraries.

RD4: 3D and Built Heritage Documentation (Topics 1, 8 and 10) addresses the use of laser scanning, photogrammetry, HBIM and GIS to document and preserve monuments, buildings, sites and landscapes and the preservation of the resulting 3D data.

RD5: Intangible Heritage and Endangered Languages (Topics 11, 18 and 21) focuses on safeguarding living heritage, including dance, music, crafts, oral traditions and low-resource languages, through digitization, knowledge bases, language resources and recognition technologies.

RD6: AI and Intelligent Archival Systems (Topics 0 and 17) covers the application of machine learning and deep learning to recognition, classification and restoration tasks, together with "intelligent" archives management systems based on cloud computing, knowledge graphs and blockchain.

RD7: Critical, Community and Memory Archives (Topics 3, 5, 7 and 12) treats digital archives as social and cultural institutions, examining community and activist archives, crisis collecting, literary and personal archives and the politics of memory and digitization.

RD8: Digital Humanities and Library Digitization (Topics 23, 25 and 26) covers digitization projects in libraries, digital humanities scholarship on written heritage and regional repository and archive initiatives, many published in languages other than English with English abstracts.

RD9: Material and Specimen Preservation (Topic 4) is a small but distinct direction on the physical conservation of collections, including natural-history specimens and degradation of materials, which shares the "long-term preservation" vocabulary of the digital field but addresses physical objects.

Evolution of Research Directions

Because annual output more than tripled over the decade, absolute counts rose for almost every topic. To separate topic-specific change from overall growth, Table 4 compares each direction's share of total output in 2016–2020 (1,589 records) with its share in 2021–2025 (2,502 records) and Fig. 6 shows annual counts for each topic.

The fastest-growing direction was **RD4 (3D and Built Heritage Documentation)**, whose share almost doubled from 5.2% to 9.9% (+ 4.7 percentage points). All three of its topics rose steeply after 2022: 3D digitization (Topic 1) reached 40 documents in 2025, architectural documentation (Topic 8) 31 and landscape and site digitization (Topic 10) 27. **RD6 (AI and Intelligent Archival Systems)** followed (6.3% to 10.2%; +3.9 points). This growth was driven entirely by Topic 0, which remained at 10–28 documents per year until 2024 and then jumped to 141 documents in 2025, by far the largest single-year count of any topic; Topic 17 (intelligent archives management) peaked earlier, in 2022 (14 documents) and has since declined. **RD8 (Digital Humanities and Library Digitization)** more than doubled its share (2.2% to 5.0%), mainly because of Topic 25 (digital humanities and written heritage), which rose from almost no documents before 2021 to 36 in 2025. **RD5 (Intangible Heritage and Endangered Languages)** grew from 4.4% to 6.1%; within it, Topic 11 (dance, music and crafts) increased from at most six documents per

year up to 2023 to 17 in 2024 and 51 in 2025, often combining heritage documentation with recognition models.

RD7 (Critical, Community and Memory Archives) grew more moderately (9.1% to 10.5%). Community, queer and feminist archives (Topic 5) peaked in 2021 (17 documents), while crisis archiving (Topic 3) climbed after 2020 and peaked in 2024 (21 documents), consistent with the increase attention to collecting following the COVID-19 pandemic and later conflicts. **RD3 (Preservation Policy, Practice and Storage Infrastructure)** gained slightly (5.0% to 5.8%); legal frameworks and FAIR principles (Topic 13) peaked in 2021 and storage infrastructures (Topic 14) rose to 15 documents in 2022 and 19 in 2025, with DNA storage appearing among its yearly keywords from 2020 onwards. **RD9 (Material and Specimen Preservation)** remained small and stady (2.3% to 2.6%).

Two directions declined in relative terms. **RD1 (Web and Born-Digital Archiving)** fell from 9.6% to 6.9% of output (− 2.7 points). Web archiving (Topic 2) was the most active topic in 2016 (18 documents) and has since fluctuated between 8 and 15 documents per year, so that its share of the field fell even though absolute activity did not collapse. Semantic and linked-data access to archives (Topic 15) peaked in 2019. By contrast, born-digital records (Topic 16) and archivist and user studies (Topic 22) held steady or grew, with Topic 16 reaching its highest count in 2025. **RD2 (Repositories, Research Data and Scholarly Infrastructure)** also declined in share (9.3% to 7.1%; −2.2 points). Archaeological and spatial data archiving (Topic 6) peaked in 2017 (24 documents) and fell sharply thereafter, while institutional repositories (Topic 9) have stayed at 10–14 documents per year since 2019 and research data management (Topic 24) rose to a peak of 15 documents in 2024, suggesting that these areas have matured into stable, practice-oriented research rather than disappeared.

Discussion

This study mapped a decade of research on digital preservation and curation in libraries, archives and cultural heritage institutions. Three broad conclusions emerge.

First, the field is thematically broad and only loosely integrated. The 27 topics span conservation science, computer vision, web engineering, information policy, archival theory and digital humanities and the median similarity between topics was moderate. The two least connected topics, web archiving and regional repository initiatives, illustrate different kinds of separation: web archiving has a specialized technical vocabulary and its own venues, while regional initiatives are connected to the rest of the field more by subject than by shared methods. For researchers, this means that advances in one strand, such as automated appraisal or format migration, may not reach adjacent communities without deliberate bridging.

Second, the centre of gravity of the field has shifted from infrastructure-building towards heritage content and AI-based processing. In 2016–2020, web and born-digital archiving and repository-based scholarly infrastructure were the two largest directions. In 2021–2025, they were overtaken by 3D and built heritage documentation, AI and intelligent systems and critical, community and memory archives.

This mirrors a broader movement in the heritage sector from establishing preservation systems towards producing, enriching and using digital surrogates. The very steep rise of Topic 0 in 2025 suggests that deep learning has moved from an experimental technique to a routine component of digitization projects, particularly for recognition, classification and restoration tasks. At the same time, the dynamic keywords of this topic and of intangible heritage research focus mainly on model accuracy, with relatively little attention to how AI outputs, training data and models are themselves preserved, documented and made trustworthy over time. This is a gap where established preservation thinking, such as OAIS and provenance metadata, could make an important contribution (Colavizza et al., 2021; Jaillant, 2022).

Third, declining shares should not be read as declining importance. Web archiving, institutional repositories and research data management remained active in absolute terms and their relative decline largely reflects the rapid growth of heritage documentation and AI research in 2024–2025. The rising attention to born-digital accessioning workflows, user studies and email appraisal suggests that the born-digital strand is moving from proof-of-concept tools towards operational practice. Similarly, the steady output on institutional repositories, frequently from Africa and South Asia, shows that building basic preservation capacity remains a pressing concern in many regions, even as better-resourced institutions explore advanced technologies (Corrado & Sandy, 2017).

The findings suggest three concerns for practitioners. Libraries and archives investing in 3D and AI-based digitization need to consider how complex outputs, such as 3D models, point clouds and machine-generated metadata, should be archived. Born-digital and web collections, which are most vulnerable to loss, need to be invested in, even though they still receive less research attention. And development of capacity within under-resourced regions through skills and a sustainable repository infrastructure, is crucial to ensuring an inclusive digital record.

Several methods have been used to interpret the results. First, the embeddings come from a general-purpose RoBERTa model with mean pooling and whitening, rather than a model tailored for sentence similarity such as all-mpnet-base-v2. Whitening improved separation, but if the model is tuned to sentence similarity, it may produce smaller clusters and less outliers. Second, 40.1% of documents remained unassigned after reassigning outliers, so the topic count describes the core of each theme, rather than the corpus. Third, because the authors assigned the topic labels and research directions, these can be seen as a subjective decision. Fourth, the 2025 increase reflects the increasing coverage of conference proceedings in Scopus, and should be confirmed by future data.

Conclusion

Using BERTopic and dynamic topic modeling on 4,091 records from the Web of Science Core Collection and Scopus, this study identified 27 research topics in digital preservation and curation between 2016 and 2025 and organized them into nine research directions. The results show a field that has grown rapidly, especially since 2024 and whose focus has shifted from web archiving and repository

infrastructure towards 3D documentation of built heritage, AI-based recognition and processing of heritage materials, intangible heritage and critical perspectives on community and memory archives. Established areas such as web archiving, born-digital records and institutional repositories remain active but now account for a smaller share of the literature.

The study has several limitations. Its findings reflect research publications rather than the actual state of preservation practice in institutions. The dataset was restricted to English-language records indexed in two databases, so it underrepresents grey literature, practitioner reports and non-English scholarship. The Web of Science and Scopus queries were not identical (the Scopus query also included "preservation metadata," "digital repositor*" and "research data"), which may have increased the weight of repository and research-data topics in the Scopus portion of the corpus. The eligibility rule was automated rather than based on manual screening of every record. Future research could combine topic modeling with citation and co-authorship analysis, compare embedding models, extend the corpus to practitioner literature and conference programmes and examine in more depth how AI-generated outputs can be preserved responsibly.

Declarations

Author Contribution

Chandan Jyoti Baishya conceptualised the study, developed the research framework and was responsible for data evaluation, interpretation of findings and manuscript formatting. Pompy Deka was responsible for data retrieval, data processing and analysis. Both authors contributed to the development of the manuscript, reviewed the findings and approved the final version for submission.

References

1. Becker C, Kulovits H, Guttenbrunner M, Strodl S, Rauber A, Hofman H (2009) Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. Int J Digit Libr 10(4):133–157
2. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
3. Brügger N (2018) The archived web: Doing history in the digital age. MIT Press
4. Campello RJGB, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013), Lecture Notes in Computer Science, 7819*(pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14
5. CCSDS (Consultative Committee for Space Data Systems) (2012) *Reference model for an Open Archival Information System (OAIS)*(CCSDS 650.0-M-2, Magenta Book). Also published as ISO 14721:2012
6. Colavizza G, Blanke T, Jeurgens C, Noordegraaf J (2021) Archives and AI: An overview of current debates and future perspectives. ACM J Comput Cult Herit 15(1). Article 4.

https://doi.org/10.1145/3479010

7. Corrado EM, Sandy HM (2017) Digital preservation for libraries, archives and museums, 2nd edn. Rowman & Littlefield
8. Costa M, Gomes D, Silva MJ (2017) The evolution of web archiving. Int J Digit Libr 18(3):191–205. https://doi.org/10.1007/s00799-016-0171-9
9. Cox AM, Tam WWT (2018) A critical analysis of lifecycle models of the research process and research data management. Aslib J Inform Manage 70(2):142–157
10. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*(pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423
11. Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec and BERTopic to demystify Twitter posts. Front Sociol 7:886498. https://doi.org/10.3389/fsoc.2022.886498
12. Ethayarajh K (2019) How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP 2019*(pp. 55–65). https://doi.org/10.18653/v1/D19-1006
13. Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://doi.org/10.48550/arXiv.2203.05794
14. Higgins S (2008) The DCC curation lifecycle model. Int J Digit Curation 3(1):134–140. https://doi.org/10.2218/ijdc.v3i1.48
15. Jaillant L (ed) (2022) Archives, access and artificial intelligence: Working with born-digital and digitized archival collections. Bielefeld University
16. Lavoie BF (2004) The Open Archival Information System reference model: Introductory guide. DPC Technology Watch Report. Digital Preservation Coalition
17. Lee CA (2011) A framework for contextual information in digital collections. J Doc 67(1):95–143
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692
19. McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. J Open Source Softw 2(11):205. https://doi.org/10.21105/joss.00205
20. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: Uniform manifold approximation and projection. J Open Source Softw 3(29):861. https://doi.org/10.21105/joss.00861
21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
22. Rosenthal DSH, Robertson T, Lipkis T, Reich V, Morabito S (2005) Requirements for digital preservation systems: A bottom-up approach. D-Lib Magazine 11(11). https://doi.org/10.1045/november2005-rosenthal

23. Su J, Cao J, Liu W, Ou Y (2021) Whitening sentence representations for better semantics and faster retrieval. arXiv. https://doi.org/10.48550/arXiv.2103.15316
24. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18
25. Xu Y-S, Idaya Y, A. M. K., Kassim MSS (2025) How do AI technologies augment intelligent information services in libraries? A deep learning-based topic modeling and temporal analysis. Educ Inform 42(1):59–83. https://doi.org/10.1177/01678329251374501

Figures

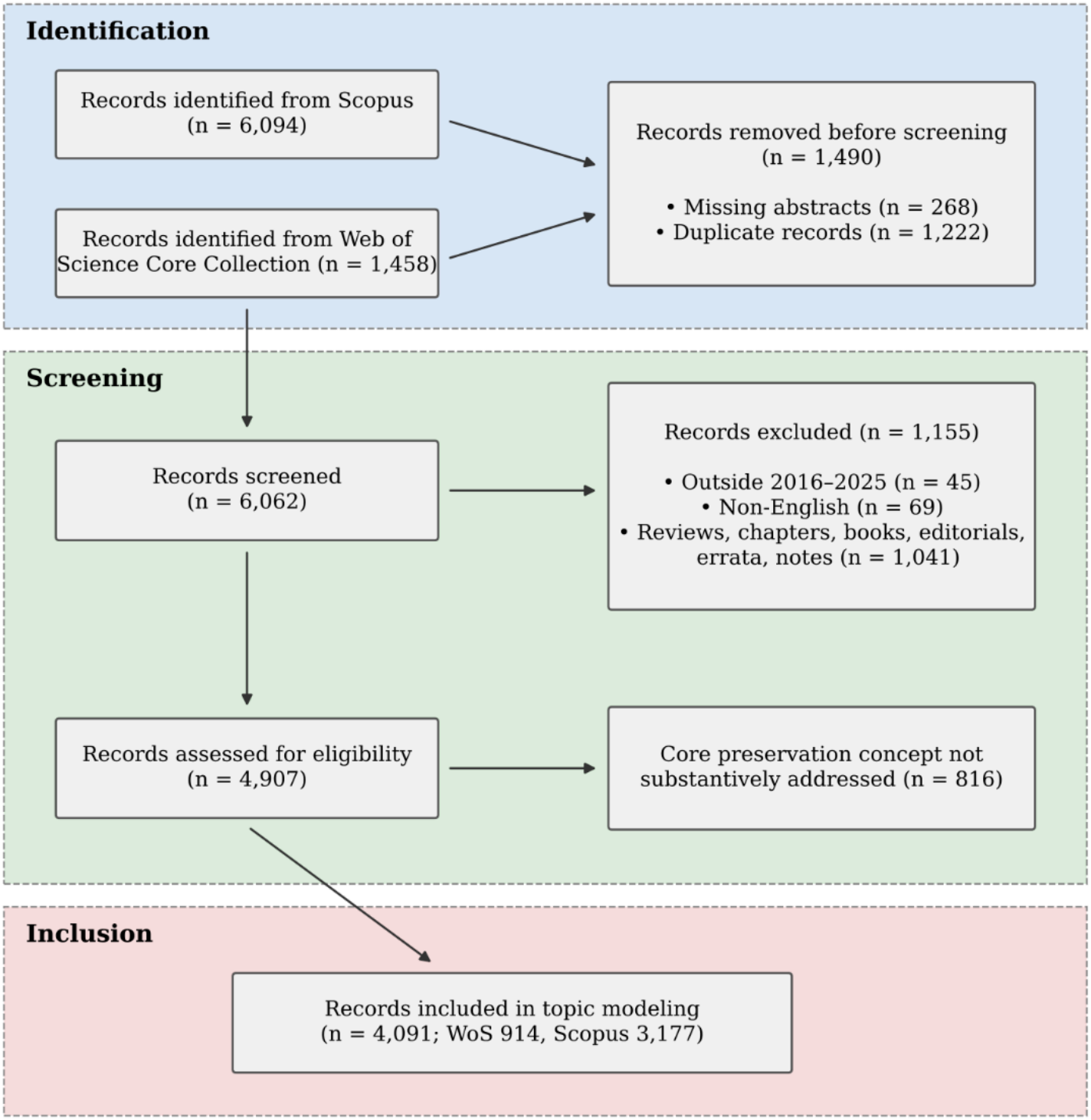

Figure 1

Data collection and screening flowchart (PRISMA 2020)

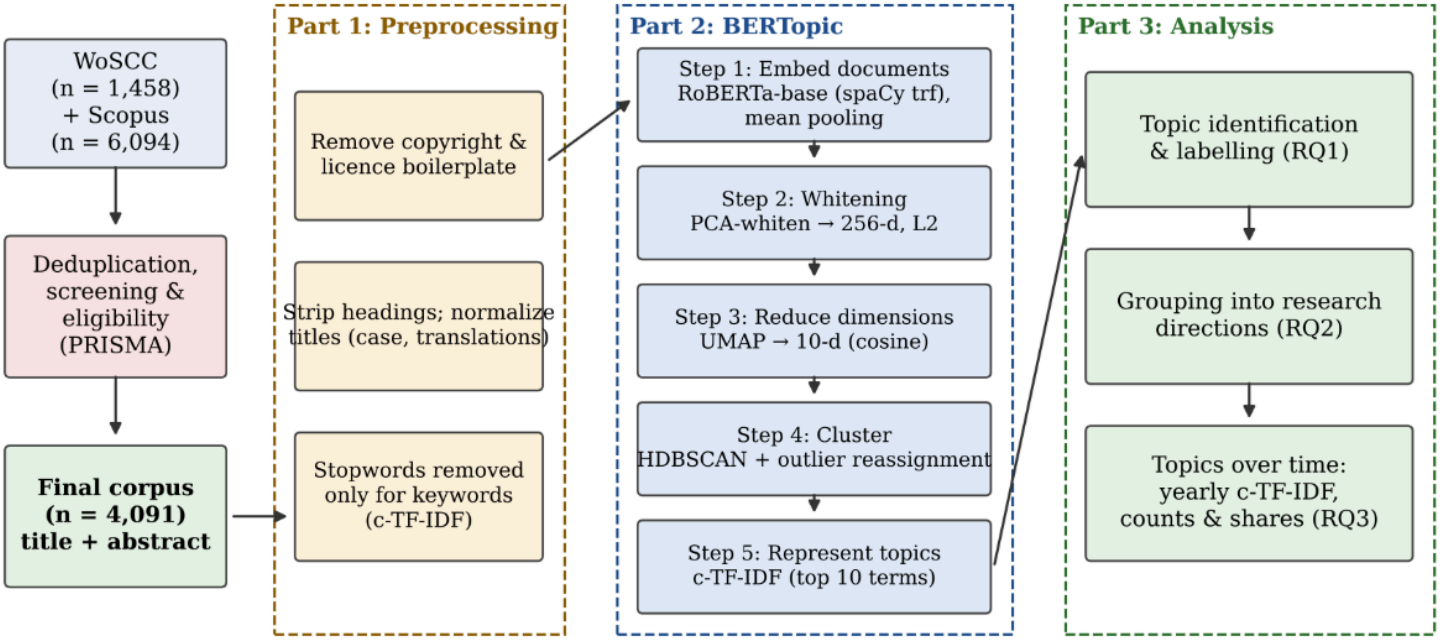

Figure 2

Topic modeling process and methods

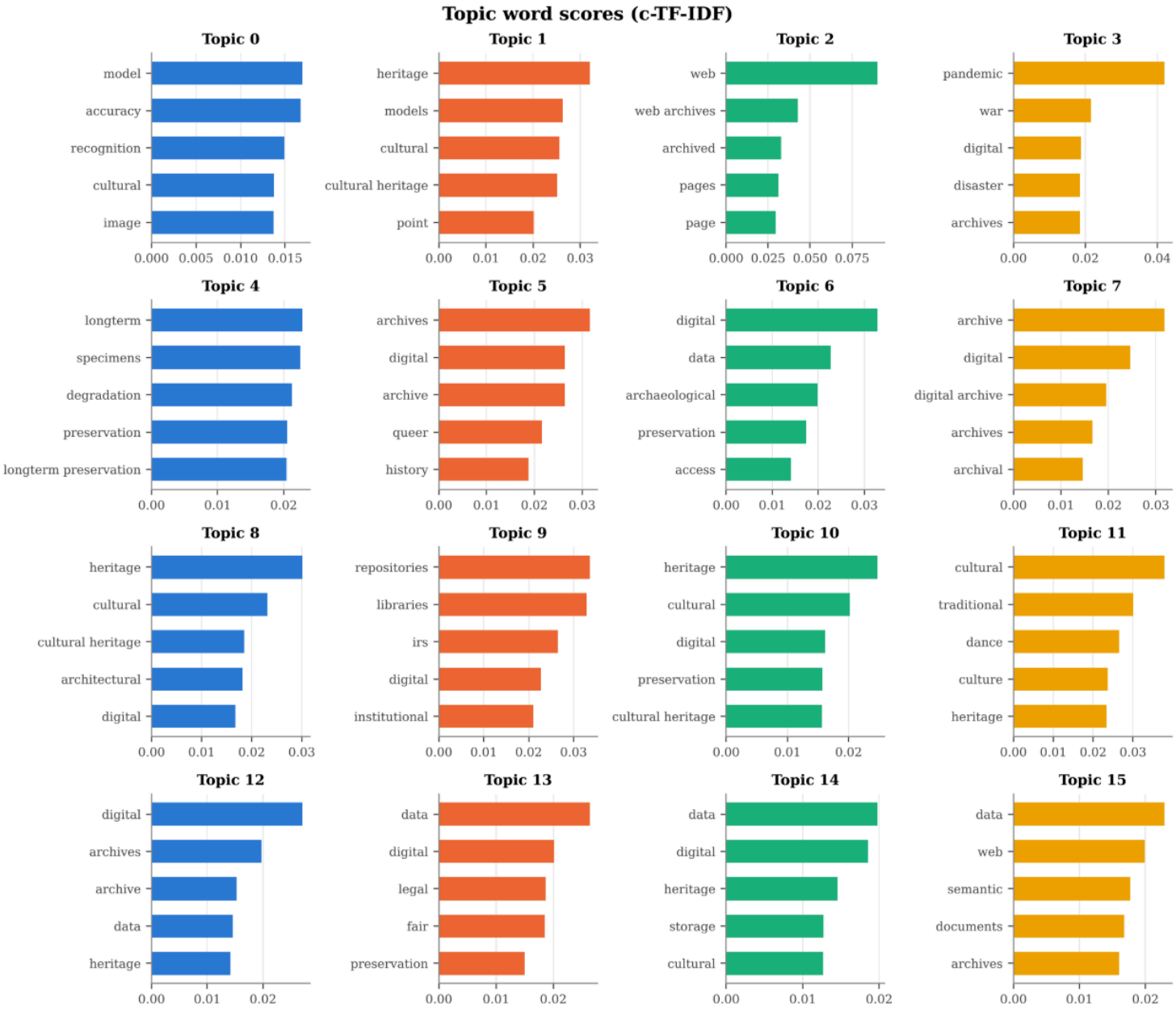

Figure 3

Top five c-TF-IDF terms for the first 16 topics

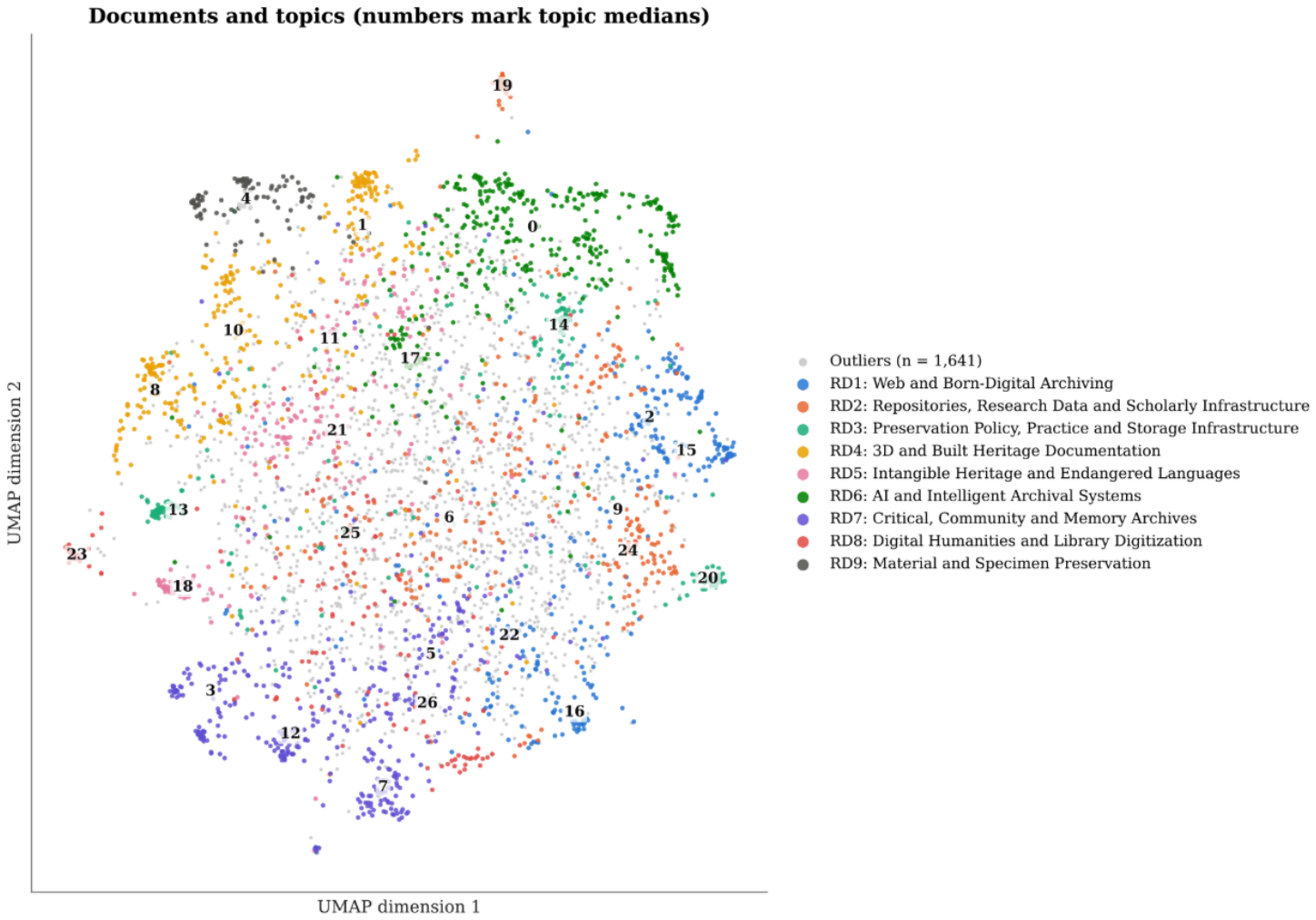

Figure 4

Document–topic distribution (two-dimensional UMAP projection)

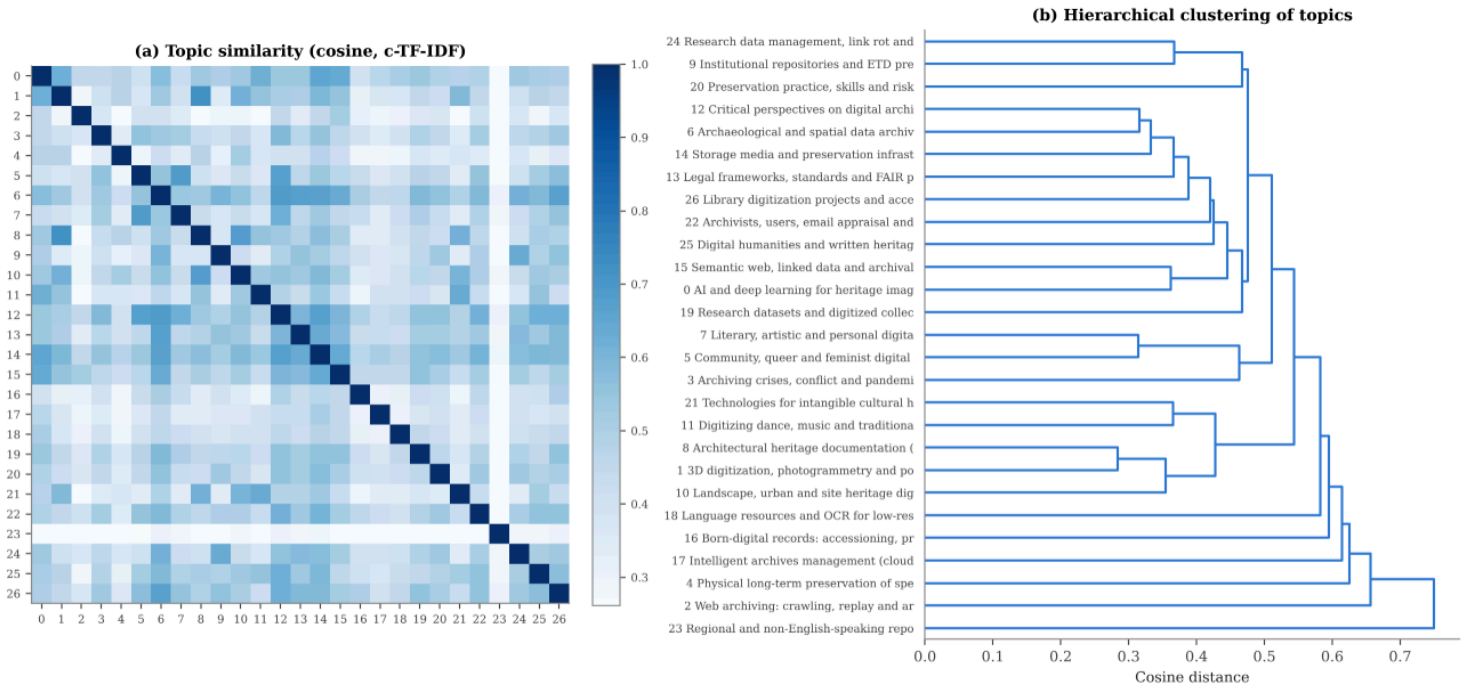

Figure 5

(a) Topic similarity heatmap and (b) hierarchical clustering of topics

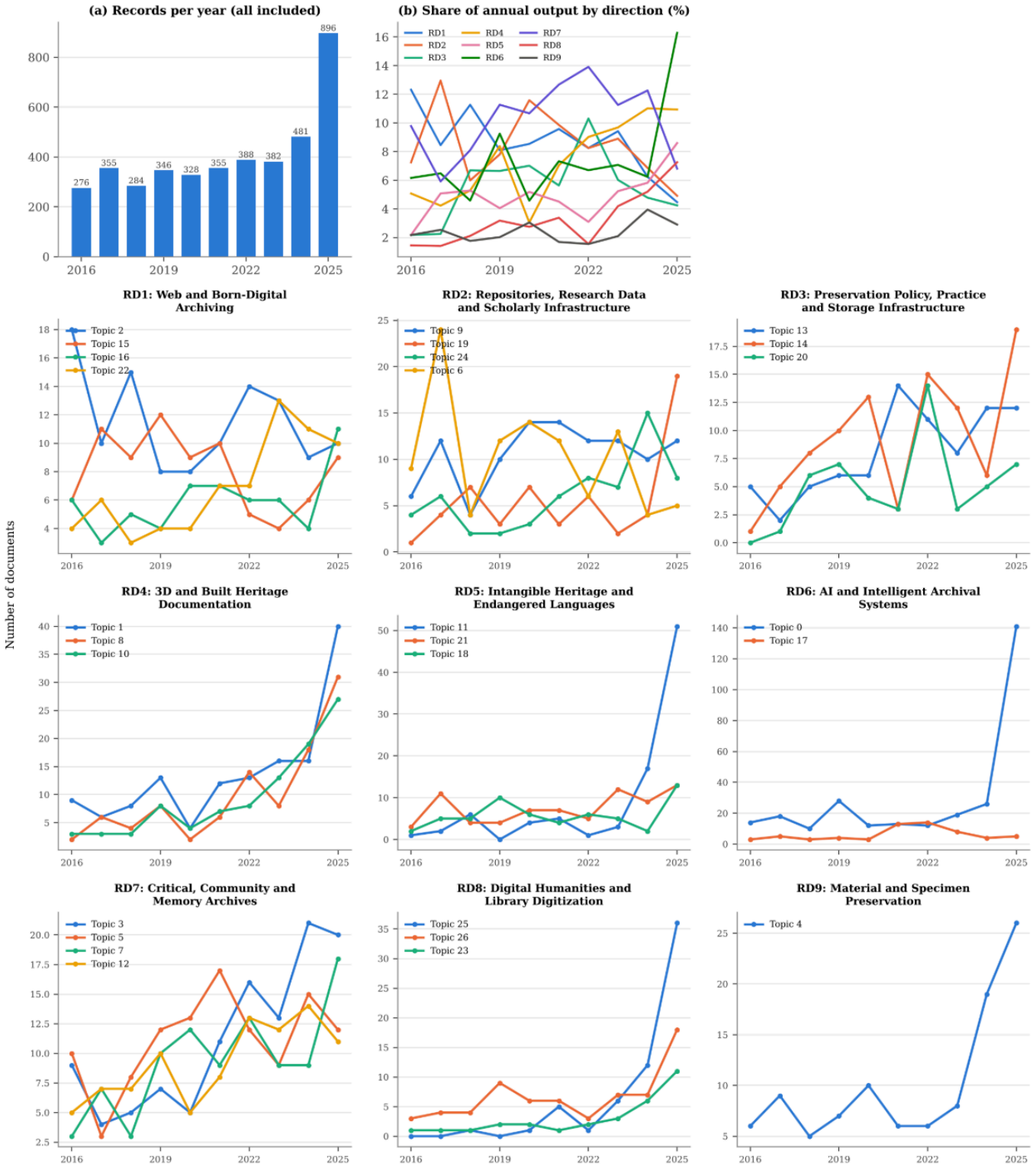

Figure 6

Evolutionary trends of each research direction, 2016–2025