

Supplementary Materials

Knowledge Structure and Evolution of Industrial Heritage Digitalization: A Comparative Bibliometric Analysis Using CiteSpace and AI-Assisted Methods

S1. Scope of the supplementary package

This package supports the methods and results reported in the main manuscript. It provides the analytical sample, the database-specific literature selection record, the principal parameter settings, a matched WoSCC Pathfinder sensitivity analysis, transparency information for the AI-assisted workflow, finalized cluster labels, and complete term-level outputs. The accompanying workbook (Supplementary-Analysis-Outputs.xlsx) contains the machine-readable tables referenced below.

Source-language keyword strings are retained for traceability and accompanied by English translations in the workbook. Candidate categories, correspondence decisions, and explanatory notes are provided in English. Translation does not change the underlying data or analytical results.

The package distinguishes two temporal outputs: CiteSpace burst terms are generated with CiteSpace/Kleinberg burst detection; AI-assisted candidate growth terms are derived from annual normalized-keyword changes. They are reported together for comparison, but are not treated as equivalent measures.

S2. Analytical sample and temporal-record rule

Dataset	Retained bibliographic records	Valid-year records used for temporal analysis	Note
CNKI	260	259	One CNKI record lacked a valid year. It remains in the retained corpus and cluster-associated-publication counts, but not in annual node selection, slice-based networks, or temporal calculations.
WoSCC	171	171	All retained WoSCC records contained a valid publication year.
Total	431	430	Database-specific analyses were conducted independently before cross-database comparison.

S2.1. Literature retrieval and selection

The database searches returned 918 records: 433 from CNKI and 485 from WoSCC. Following export, two reviewers independently screened records against the predefined criteria for study object, substantive digital content, scholarly character, and minimum bibliographic completeness. CNKI and WoSCC records were screened separately. No automated or LLM-assisted tool was used to make inclusion or exclusion decisions.

Preliminary screening removed duplicate records, records clearly outside the study scope, and records failing basic eligibility. Title, abstract, and keyword screening then assessed thematic relevance. At the final eligibility-verification stage, the reviewers re-examined the bibliographic record and full text where available to confirm substantive relevance and the availability of information required for bibliometric analysis. When initial decisions differed, the two reviewers re-examined the record against the predefined criteria and resolved the decision through discussion and consensus; the exclusion reason was recorded at the relevant stage.

Table S3. Database-specific literature retrieval and selection process.

Stage	Main reason(s) for exclusion	CNKI excluded / retained	WoSCC excluded / retained	Total retained
Initial retrieval	No document-type restriction at search stage	0 / 433	0 / 485	918

Stage	Main reason(s) for exclusion	CNKI excluded / retained	WoSCC excluded / retained	Total retained
Preliminary screening	Duplicates, clearly out-of-scope records, or failure to meet basic eligibility	145 / 288	209 / 276	564
Title, abstract, and keyword screening	Study object or research content did not meet the thematic eligibility criteria	14 / 274	87 / 189	463
Eligibility verification	Insufficient substantive relevance, no clear scholarly focus, or incomplete essential bibliographic information	14 / 260	18 / 171	431

Note. Stage-level counts reconcile to the final 431-record analytical sample. Reasons within preliminary screening were applied in one combined pass and were therefore not further disaggregated. The reporting structure adapts the selection-process transparency items in PRISMA 2020; it does not claim a registered systematic-review protocol.

Reporting guidance: Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021, 372, n71. <https://doi.org/10.1136/bmj.n71>; Page, M.J.; Moher, D.; Bossuyt, P.M.; et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021, 372, n160. <https://doi.org/10.1136/bmj.n160>.

S3. Analytical settings and workflow

Component	Setting or rule	Purpose
CiteSpace	Version 6.4.R1 Advanced; time span 2008–2026; one-year slices; title, abstract, and author keywords; Keywords Plus added for WoSCC.	Construct comparable bibliometric networks from the two database-specific samples.
CiteSpace selection	g-index (k = 25). Pathfinder pruning: CNKI yes; WoSCC primary analysis no. A matched WoSCC Pathfinder rerun is reported in Table S2 and Figure S1.	Retain the original primary analysis and assess the effect of pruning.
AI keyword processing	Database-specific field extraction, rule-based normalization, and LLM-assisted suggestions. LLM outputs were reviewed by the researcher before use.	Reduce superficial variants without replacing researcher judgement.
AI node selection	Approximate g-index $\times 3$ was used as the primary selection rule; Top N = 25 was retained as a supplementary comparison.	Select a stable set of keyword nodes for the AI-assisted network.
AI network construction	Edges represent document-level keyword co-occurrence; edge weights equal co-occurrence frequency. Network simplification was applied to both datasets and is not equivalent to CiteSpace Pathfinder pruning.	Obtain comparable AI-assisted network summaries.
AI clustering	Spectral clustering; candidate K values were compared using eigengap, modularity, network-distance silhouette, cluster size, and interpretability. Final K: CNKI = 5; WoSCC = 9.	Produce descriptive thematic groups for cross-method comparison.
Temporal comparison	CiteSpace burst detection and annual-keyword growth screening were retained as separate outputs.	Avoid conflating different temporal detection mechanisms.

S3.1. Matched WoSCC Pathfinder sensitivity analysis

The primary WoSCC CiteSpace visualization was retained without Pathfinder pruning. To assess the effect of this processing choice, the WoSCC dataset was rerun with Pathfinder pruning while retaining the same corpus, time span, term sources, slice length, and g-index selection rule. The matched run retained the same node count and largest connected component, while reducing the number of links. Major thematic domains remained recognizable, whereas the

boundaries and labels of smaller clusters were more sensitive to pruning. This rerun is used as a robustness check and does not replace the primary results.

Table S2. Matched WoSCC CiteSpace Pathfinder sensitivity comparison.

Item	Primary WoSCC network	Matched Pathfinder sensitivity network
Literature sample	171 WoSCC records	171 WoSCC records
Time span and slice	2008–2026; 1 year	2008–2026; 1 year
Term sources	Title, abstract, author keywords, and Keywords Plus	Same as primary network
Node selection	g-index (k = 25)	g-index (k = 25)
Pruning	No Pathfinder pruning	Pathfinder on merged network; slice pruning and MST disabled
Nodes	347	347
Edges	906	782
Largest connected component	228 nodes (65%)	228 nodes (65%)
Interpretive role	Primary WoSCC CiteSpace results	Robustness check: major domains retained; smaller-cluster boundaries and labels may vary

Note. Edge counts were calculated from the archived GraphML exports. CiteSpace display-level link accounting may differ slightly from GraphML totals.

Figure S1 shows the matched Pathfinder-pruned WoSCC cluster solution.

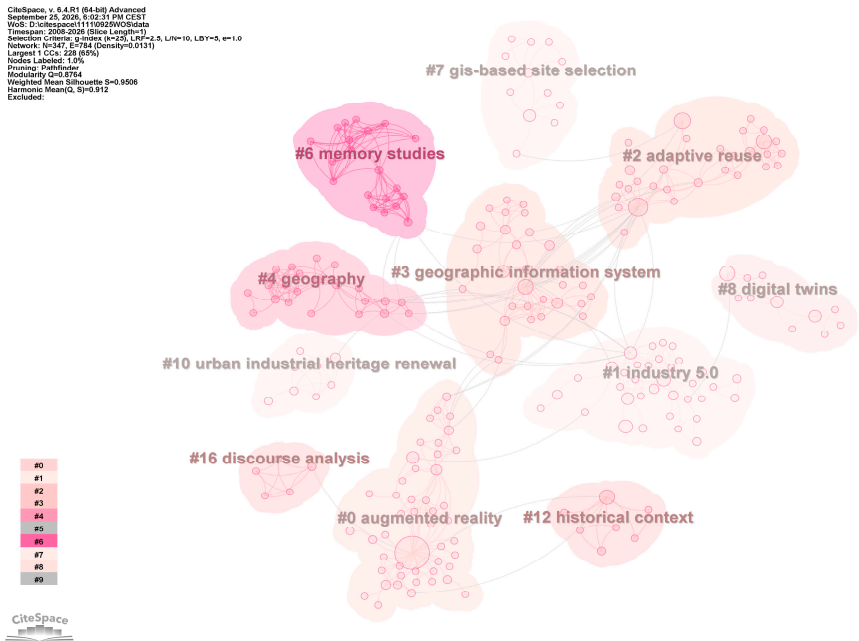


Figure S1. WoSCC CiteSpace keyword-cluster map generated with Pathfinder pruning under matched sensitivity-analysis settings. Node size represents keyword frequency, and colors identify clusters. Labels retain the original, non-consecutive CiteSpace cluster IDs. The original export displays only IDs 0–9 in its color key; the labels identify additional displayed clusters. The displayed network contains 347 nodes and 784 links; Table S2 reports 782 edges counted from the archived GraphML export.

S3.2. Reproducible AI-assisted network computation

The code archive includes a portable implementation of steps 5–9, derived normalized-keyword inputs, the archived computational functions, and a verification record. Running “node core/run_core.mjs --verify” from the extracted code-archive root reproduces node selection, co-occurrence networks, pruning, centrality, spectral candidates, and quality measures without an LLM call or additional Node.js packages. Input schemas, file roles, and execution instructions are provided in core/README.md.

For each database and one-year slice from 2008 to 2026, normalized terms are counted once per document and ranked by decreasing frequency. The approximate g value is the largest rank for which cumulative frequency is at least the squared rank. The primary rule retains the first min(number of terms, ceil(3g)) terms. Co-occurrence is counted only when both terms were selected in the document’s year; each distinct pair contributes once per document. This frequency-based rule is not algorithmically equivalent to CiteSpace’s g-index setting.

The custom Pathfinder-inspired simplification assigns distance 1/w to an edge with co-occurrence weight w. Each edge is tested against the original graph with that edge omitted and removed only if an alternative path is shorter by more than 1e–10. Equal-length alternatives are retained. No MST is applied, and only nodes incident to retained edges enter the final active network. Degree and weighted degree are calculated directly; normalized betweenness uses unweighted shortest paths across all active nodes.

Spectral clustering uses the normalized weighted Laplacian and the three largest eligible eigengaps. The smallest-eigenvalue eigenvectors are row-normalized before deterministic k-means++ clustering. The archived seed is 20260804 + 100 × database index + K, with indices 0 for CNKI and 1 for WoSCC. Candidate K values are 2, 4, and 5 for CNKI and 6, 8, and 9 for WoSCC. The researcher-confirmed choices remain K = 5 and K = 9; they are not automatic maximum-silhouette choices. All implementation details, including numerical tolerances and stopping rules, are supplied in the code.

The reproducibility check matched all selected-node identities and years, document frequencies, unpruned and pruned edges and weights, centrality values, and memberships of all six candidate cluster solutions against the archived results. It also reproduced candidate modularity and silhouette values within a tolerance of 1e–10. Final networks contain 122 nodes and 290 edges for CNKI and 78 nodes and 142 edges for WoSCC. Distance-based silhouette uses 1/w shortest paths within the largest connected component, covering 122 CNKI nodes and 66 WoSCC nodes. These computational checks reproduce the recorded analysis; they do not independently validate its thematic interpretation.

S4. Transparency information for AI-assisted processing

Codex assisted script development, while numerical results were produced by the computational scripts rather than generated directly by the LLM. GPT-5.5 was accessed through Codex and GPT interfaces. LLM-assisted semantic tasks comprised terminology-normalization suggestions, provisional cluster labels, and evidence-bounded interpretation. The LLM did not independently select publications, assign final network or cluster membership, or determine final labels. Semantic suggestions and final analytical choices were reviewed by the researchers.

Item	Disclosure
Tool family	CiteSpace 6.4.R1 Advanced; Python and JavaScript scripts; GPT-5.5 accessed through Codex and GPT interfaces. Codex assisted script development.
Model identifier and version	GPT-5.5, accessed through Codex and GPT interfaces.
Operational prompt template: normalization	“Review the candidate normalization of the following keyword forms. Propose only synonym, abbreviation, capitalization, or morphological merges. Do not introduce a new concept or alter the source record.”
Operational prompt template: labeling	“Using the supplied cluster keywords, representative nodes, and representative publications, propose a concise descriptive cluster label. Do not alter membership; flag uncertainty and possible overlap.”
Operational prompt template: interpretation	“Summarize the supplied output only. Separate direct evidence from interpretation and flag duplicate, overly broad, or low-stability themes for researcher review.”
Human review	Suggestions were accepted, revised, merged, renamed, or rejected through manual checking. The final cluster labels and decisions are recorded in the accompanying workbook.

Item	Disclosure
User-controlled parameters	No user-specified LLM sampling parameters were used. Temperature, top-p, seed, and other system-side generation settings were not exposed or controlled in the interfaces. Computational seeds and numerical tolerances are separately documented in Section S3.2 and the code archive.
Input and output boundary	Codex-assisted code development produced scripts, not analytical measurements. Executed scripts determined numerical metrics and computational membership. Semantic inputs comprised keyword lists, representative nodes, publications, and intermediate tables; outputs comprised suggestions and explanatory text subject to researcher review. The LLM did not make final inclusion, membership, or labeling decisions.
Researcher review	All suggestions were checked against the supplied evidence. Final terminology, labels, and interpretations were retained, revised, or rejected by the researchers.

S5. Finalized AI-assisted thematic clusters

Dataset	Cluster ID	Nodes	Associated publications	Final researcher-reviewed label	Decision
CNKI	CNKI_K05_C01	47	214	Adaptive Reuse and Revitalization of Industrial Heritage	AI-suggested label retained
CNKI	CNKI_K05_C02	36	91	Spatial Pattern Analysis of Industrial Heritage	AI-suggested label retained
CNKI	CNKI_K05_C03	29	137	Digital Conservation of Industrial Heritage	AI-suggested label retained
CNKI	CNKI_K05_C04	7	8	Evolution of Urban Landscape Patterns	AI-suggested label retained
CNKI	CNKI_K05_C05	3	2	Evolution of Industrial Historical Patterns	Manually revised
WoSCC	WoS_K09_C01	26	102	Adaptive Reuse of Industrial Heritage	AI-suggested label retained
WoSCC	WoS_K09_C02	14	39	Digital Twin	AI-suggested label retained
WoSCC	WoS_K09_C03	12	21	Post-Industrial Landscape	AI-suggested label retained
WoSCC	WoS_K09_C04	11	70	Spatial Analysis of Industrial Heritage	AI-suggested label retained
WoSCC	WoS_K09_C05	4	2	Conversion	AI-suggested label retained
WoSCC	WoS_K09_C06	3	1	Coal-Mining Heritage and Cultural Memory	Manually revised
WoSCC	WoS_K09_C07	3	2	Cultural Geography	AI-suggested label retained
WoSCC	WoS_K09_C08	3	1	Mining Regions and Regeneration Cooperation	Manually revised
WoSCC	WoS_K09_C09	2	1	Advance Education	AI-suggested label retained

Associated publications refer to distinct records matched to at least one cluster keyword after stop-term filtering. Records matching only excluded generic terms, such as “保护” (“protection”), were not counted. This counting rule did not alter keyword-cluster membership. A publication may be associated with more than one cluster; counts therefore should not be summed across clusters.

S5.1. Scope of cross-method thematic correspondence

The Cross_Method_Mapping worksheet uses the same major-domain criterion as manuscript Table 4 for its principal themes. “Broad” denotes shared substantive coverage, not identical cluster membership or boundaries. Digital Twin is classified as broad at the digital-technology-domain level: digital twin is directly represented in both procedures, while the AI cluster also contains artificial intelligence, Industry 4.0/5.0, and sustainability. Its finer-grained cluster organization is only partially corresponding. This qualification is retained in the worksheet, and peripheral mappings remain exploratory. Correspondence categories are evidence-informed researcher judgments rather than numerical overlap thresholds.

S6. CiteSpace burst terms and AI-assisted candidate growth terms

Table S1 lists the complete CiteSpace burst-term output. Figure 8 in the main manuscript displays the selected Top 10 terms for each dataset. The workbook additionally records the full candidate-growth-term table and the manual term-to-term correspondence check.

Table S1. Complete CiteSpace keyword burst-term results for the CNKI and WoSCC datasets.

Dataset	Burst term	Strength	Active period
CNKI	city image	0.9767	2024–2026
CNKI	industrial pattern	1.5182	2017–2021
CNKI	cultural tourism	1.0508	2021–2022
CNKI	sense of place	0.9767	2024–2026
CNKI	exhibition design	1.2729	2022–2023
CNKI	digitalization	1.1958	2022–2023
CNKI	digital technology	1.9136	2022–2023
CNKI	model	1.1946	2013–2016
CNKI	Daqing City	0.9767	2024–2026
CNKI	urbanization	1.8164	2017–2020
CNKI	architectural heritage	1.1161	2015–2018
CNKI	Yiwu	1.9483	2009–2011
CNKI	landscape index	2.1498	2010–2017
CNKI	point cloud	1.0303	2012–2020
CNKI	landscape pattern	3.8333	2009–2012
CNKI	urban renewal	1.2636	2024–2026
CNKI	industrial heritage	1.4525	2020–2023
CNKI	information collection	1.6139	2014–2018

Dataset	Burst term	Strength	Active period
WoSCC	Industry 4.0	2.2642	2024–2026
WoSCC	digital twins	1.1410	2024–2026
WoSCC	conservation	1.2772	2021–2024
WoSCC	laser scanning	1.1136	2013–2016
WoSCC	industrial heritage	2.4132	2023–2024
WoSCC	management	1.8800	2020–2024
WoSCC	adaptive reuse	2.4278	2024–2026
WoSCC	brownfield	1.1536	2021–2022
WoSCC	cultural landscape	1.1346	2016–2023
WoSCC	3d modeling	1.2081	2015–2016
WoSCC	post-industrial landscape	1.2153	2016–2021
WoSCC	augmented reality	1.5529	2016–2018
WoSCC	circular economy	1.5500	2024–2026
WoSCC	artificial intelligence	1.1606	2024–2026
WoSCC	digital transformation	1.1410	2024–2026
WoSCC	city	1.4919	2011–2014
WoSCC	geography	1.1238	2019–2020

S7. Data and code availability

The original CNKI and WoSCC bibliographic exports are not redistributed because access is governed by database licences. The supplementary workbook contains derived analytical outputs. Supplementary-Code.zip provides a self-contained core implementation for steps 5–9, with derived document-identifier/year/normalized-term records, archived reference outputs, and a verification report. No titles, abstracts, authors, affiliations, or full texts are included in these core inputs. Historical downstream scripts for labeling, temporal analysis, and comparison tables are retained separately within the archive and require the intermediate files and dependencies documented in its README. Reproducing the numerical core does not reproduce database retrieval, manual screening, or historical LLM interactions.

Supplementary files: Supplementary-Materials.docx; Supplementary-Analysis-Outputs.xlsx; Supplementary-Code.zip; Figure_S1_WoSCC_Pathfinder_Cluster_Map.png; README-Supplementary-Package.md.