Journal of Open Humanities Data

Curating Multimodal Temple Heritage Data for the Study of Public Belonging in Malaysia

DISCUSSION PAPER

GENGYE WU
MUHAMAD AZWAN ABD RAHMAN
TAN RAAN HANN
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Multimodal heritage research combines source narratives, visual materials, and platform comments, but partial records can obscure what each layer supports. This discussion paper develops and demonstrates a documentation protocol using a four-temple corpus from Malaysia. The reported corpus comprises 11 source records, 16 visual coding units, and 1,554 cleaned YouTube comments from 29 comment-bearing videos. The documentation release separates inherited aggregates, retrospective provenance, claim-level evidence, and newly dated source observations. Eleven provenance records and six claim-evidence records are implemented in CSV and SQLite; a runnable script checks aggregate arithmetic and identifier uniqueness. Three source-access observations created on 5 October 2026 include two retrieved pages and one search-only record whose page retrieval timed out. A worked example links a Thean Hou source title with a newly checked institutional architecture narrative while preserving undocumented item-level dates and visual links. The audit reproduces Johor's 94.34% share of reported comments and the six theme counts totaling 908, without treating those counts as an exhaustive partition. The audit does not reconstruct original comments, images, coding assignments, or processing records. The contribution is an inspectable documentation layer and reuse procedure for incomplete heritage corpora, not validation of the original thematic analysis, full empirical replication, or measurement of interpersonal outcomes.

CORRESPONDING AUTHOR:
Gengye Wu
The Institute of Malaysian and International Studies (IKMAS), The National University of Malaysia (UKM), Malaysia
p117845@siswa.ukm.edu.my

KEYWORDS:
open humanities data; multimodal corpus; digital heritage; data provenance; Malaysian Chinese temples; public belonging

TO CITE THIS ARTICLE:
Wu, G., Abd Rahman, M. A., & Raan Hann, T. (2026). Curating Multimodal Temple Heritage Data for the Study of Public Belonging in Malaysia. *Journal of Open Humanities Data*, 12: 146, pp. 1–12. DOI: https://doi.org/10.5334/johd.688

(1) CONTEXT AND MOTIVATION

Chinese temples have long served overseas Chinese communities as religious institutions and as places where social relationships are organized, displayed, and remembered. Lineage associations, dialect-group networks, volunteer groups, worshippers, festival participants, and visitors meet through recurring rituals and shared spaces. In Malaysia's multilingual and multi-religious environment, temples such as Thean Hou Temple in Kuala Lumpur, Kek Lok Si Temple in Penang, Cheng Hoon Teng Temple in Melaka, and Johor Ancient Temple in Johor Bahru also function as heritage landmarks, tourist sites, and public communication centers. Their significance lies not only in what they preserve, but also in how they create visible occasions for affiliation, interaction, and belonging.

These institutions leave varied digital traces. An official architectural description presents an institutional account of a place; a festival image frames a particular event; a viewer comment expresses a situated response to a video. All can inform heritage research, but they serve different evidentiary purposes. Grouping them under a single temple name may obscure who produced each item, what it represents, and which interpretations it can support. A reusable corpus therefore requires more than collecting accessible material: it needs clear distinctions between source objects, analytical annotations, and claims.

The four-temple study considered here illustrates this problem. Its institutional and media layer contains 11 source records, its visual layer contains 16 coding units, and its comment layer contains 1,554 cleaned public YouTube comments. The comments mainly concern Johor Ancient Temple and Chingay. Treating these layers as equal measures of public participation would turn retrieval conditions into apparent differences in belonging. Likewise, treating comments as testimony about durable social ties would go beyond what traces on a public platform establish.

This discussion paper shifts the analytical focus from inferred interpersonal outcomes to documenting and interpreting a multimodal heritage corpus. Public belonging is used as an interpretive concept for represented or expressed affiliation with places, groups, creators, and ritual communities. It is not a measured psychological construct. The paper asks how distinct corpus layers can be related without erasing provenance, what the measured distribution lets researchers compare, and what documentation and access conditions are needed for responsible reuse.

The contribution is a demonstrated documentation protocol for uneven heritage corpora. It separates inherited aggregate claims, retrospective source inventories, new dated source observations, and unverified item-level relationships. Eleven provenance records and six claim-evidence records make this distinction explicit; a script recomputes aggregate consistency, and a SQLite database supports inspection. These operations verify documentation structure and aggregate arithmetic; they do not validate the original collection or thematic analysis. The contribution is a bounded implementation and reuse of a documentation layer, not a new audience study or a validated cultural classification. Other heritage projects can adapt this workflow to link inherited sources, separately dated observations, and bounded claims while preserving gaps in source histories.

(2) RELATED WORK

(2.1) HERITAGE REPRESENTATION AND CULTURAL MEMORY

Cultural memory and social identity theories explain why places, rituals, and group references matter for heritage interpretation (Assmann, 1995; Nora, 1989; Tajfel & Turner, 1979). Here, these concepts guide contextual description rather than psychological measurement. Malaysian Chinese temples link migration, dialect associations, locality, and transregional ties (Dean, 2020, 2022; Yan et al., 2020). Keeping these distinctions prevents generic heritage categories from erasing locally meaningful relations.

Carey's (1989) ritual view of communication and Collins's (2004) interaction-ritual account help interpret shared symbols and collective occasions. Neither supplies missing observations of relationship formation. Their role is to support documentation of context, not to validate outcomes from aggregate platform traces.

(2.2) PLATFORM MEDIATION AND THE STATUS OF DIGITAL TRACES

Mediatization research examines how institutional practices become visible through media forms and logics (Hjarvard, 2008a, 2008b, 2013, 2016). Online and offline practices intersect (Campbell, 2013), but visual social media require attention to their specific forms (Highfield & Leaver, 2016). Hu et al. (2024) provide a computational study of Malaysian Chinese folk-belief communication. In the present protocol, platform and access state are evidence attributes, not proxies for community participation.

Direct creator address resembles parasocial communication (Horton & Wohl, 1956), but one comment cannot establish enduring attachment or belongingness (Baumeister & Leary, 1995). Therefore, the documentation layer preserves discourse labels while withholding interpersonal-outcome claims.

(2.3) MULTIMODAL INTERPRETATION AND REUSABLE ANNOTATION

Representation and visual semiotics distinguish culturally situated meaning from an image's observable composition (Hall, 1997; Kress & van Leeuwen, 2006; Rose, 2016). Architectural research also connects temple forms with regional and cultural histories (Koh & Lim, 2022; Liew et al., 2025). These approaches justify separating source description, observation, and interpretation rather than treating an image as direct evidence of social bonds.

Content-analysis procedures require explicit units, category definitions, and coding decisions (Krippendorff, 2018; Schreier, 2012); term weighting can help navigation but does not replace interpretation (Salton & Buckley, 1988). For documentation, W3C provenance distinguishes objects, activities, and responsible agents (World Wide Web Consortium, 2013). This protocol draws on that distinction without claiming formal PROV compliance. Its contribution is to preserve the boundary between inherited claims and newly observed evidence when original records cannot be recovered.

(3) DATASET DESCRIPTION

Gengye Wu collected the original corpus on 10 July 2026 for this study and retains the original records. These records are shared through individual requests for scholarly exchange rather than deposited openly. The public release contains aggregates and documentation; original records may be made available upon reasonable request.

The working corpus includes Thean Hou Temple in Kuala Lumpur, Kek Lok Si Temple in Penang, Cheng Hoon Teng Temple in Melaka, and Johor Ancient Temple in Johor Bahru. We selected these cases purposively to reflect differences in location, heritage profile, ritual activity, and public visibility. Table 1 records this rationale; it does not define a representative sample of Malaysian temples.

Table 1 Case selection rationale and primary analytical emphasis.

CASE	LOCATION	REASON FOR INCLUSION	PRIMARY ANALYTICAL EMPHASIS
Thean Hou Temple (THT)	Kuala Lumpur	Highly visible urban temple with official web presence, festival imagery, tourism appeal, and large Facebook metadata footprint.	Architectural heritage, Mazu devotion, lantern spectacle, guided cultural display.
Kek Lok Si Temple (KLS)	Penang	Major temple attraction strongly associated with Chinese New Year lighting and tourism circulation.	Festival atmosphere, tourism visibility, visual spectacle.
Cheng Hoon Teng Temple (CHT)	Melaka	Historic temple linked to heritage tourism and UNESCO-oriented representation.	Historical memory, inscriptional heritage, continuity of worship.
Johor Ancient Temple (JAT)	Johor Bahru	Temple strongly associated with Chingay, five dialect groups, and intensive YouTube engagement.	Local identity, procession, dialect-group unity, platformed engagement.

Tables 2 and 3 describe the layers and identify 11 source records: institutional, tourism, news, and public-profile records. The visual layer contains 16 coding units without a complete unit-level inventory. The comment layer contains 1,554 cleaned comments from 29 videos with

retrievable comments among 40 candidates. These counts describe different objects, not a single sample. The deposit preserves reported summaries; corpus completeness and record-level links cannot be independently verified from it.

Table 2 Reported corpus layers, scale, and analytical use.

CORPUS LAYER	MATERIAL	SCALE	ANALYTICAL USE
Official and media texts	Official websites, Facebook page metadata, tourism pages, and news pages.	11 source records	Identifying institutional narratives, heritage framing, and public circulation contexts.
Visual materials	Temple architecture, festival images, processions, inscriptions, crowds, deity images, and lighting displays.	16 visual coding units	Analyzing visual semiotic resources and compositional strategies.
Public comments	Public YouTube comments associated with temple-related videos.	1,554 cleaned comments	Examining platform-visible relational discourse, recurrent themes, case affiliation, and salient terms.

Table 3 Public source inventory retained from the original study.

Note. This inventory retains source identifiers and titles rather than complete preservation records. Original source URLs, access dates, rights information, and links to the 16 visual units remain incompletely documented. Source titles are inventory labels, not verified contemporary guidance.

ID	CASE	SOURCE TYPE	PLATFORM	TITLE/SOURCE
S001_THT	THT	social_profile	Facebook	Thean Hou Temple Malaysia | Kuala Lumpur
S002_THT	THT	official_page	Official website	Thean Hou Temple – HainanNet
S003_THT	THT	official_page	Official website	Thean Hou Temple Architecture – HainanNet
S004_THT	THT	official_page	Official website	Attraction – HainanNet
S005_KLS	KLS	media_page	Online travel media	Chinese New Year Events in Penang 2026: Info & Schedule
S006_CHT	CHT	social_profile	Facebook	The Cheng Hoon Teng Temple | Malacca City
S007_CHT	CHT	heritage_page	Discover Malaysia UNESCO	Cheng Hoon Teng Temple – Discover Malaysia-UNESCO
S008_JAT	JAT	social_profile	Facebook	Johor Ancient Temple | Johor Bahru
S009_JAT	JAT	official_tourism_page	Tourism Johor	Old Chinese Temple
S010_JAT	JAT	news_page	ATV News	Johor Ancient Temples Release SOPs: Parades Permitted, But Followers Prohibited from Accompanying – Asia Television News
S011_THT	THT	news_page	Xinhua	Thean Hou Temple in Kuala Lumpur, Malaysia, hosts a "Celebrating the Lantern Festival with Joy" gala – Xinhua News Agency

(3.1) REPOSITORY AND RELEASE STATUS

Repository location and name: Malaysian Chinese Temple Heritage Public Aggregate Data and Documentation (Version 1.2) is deposited in Zenodo (Wu et al., 2026) at https://doi.org/10.5281/zenodo.23367960 (record: https://zenodo.org/records/23367960). Version 1.0 remains separately citable. CC BY 4.0 covers researcher-authored aggregates, metadata, paraphrases, and documentation; third-party source content is not relicensed, and the repository record does not extend beyond the documented source materials.

Object description: Version 1.2 contains eight aggregate/inventory CSV tables, the missing-metadata register, source_provenance.csv (11 records), claim_evidence_register.csv (six records), and new_web_observations.csv (three observations). It includes heritage_documentation.sqlite, validate_aggregates.py, validation_report.json, README, data dictionary, metadata, changelog, and checksums. The README maps all 20 files to their purposes; README and dictionary align missing-data terms with Section 5.4 while retaining original CSV and SQLite values. SQLite stores fields as text; numerical checks cast counts. Source observations contain researcher-authored paraphrases and access metadata, not redistributed webpages or images. Original comments, identifiers, images, coding assignments, and processing scripts are excluded.

Formats and versions: Version 1.2 provides UTF-8 CSV, JSON, Markdown, SQLite, and a Python standard-library validation script in a ZIP archive. These release formats do not reconstruct the original collection environment.

Creation and publication dates: Versions 1.0, 1.1, and 1.2 were prepared on 4, 5, and 14 October 2026, respectively. Version 1.2 was published on 14 October 2026. New source observations remain dated 5 October 2026; these dates do not reconstruct original empirical collection. Creator roles appear on the title page.

Language and license: documentation is in English under CC BY 4.0. The original corpus is described as multilingual, but language counts and translation procedures are undocumented.

The license covers researcher-authored materials, not third-party images, webpage content, or comments.

(3.2) LINKING LAYERS WITHOUT IMPLYING MATCHED OBSERVATIONS

The case abbreviations THT, KLS, CHT, and JAT connect documentation at the case level. The source provenance table preserves the 11 inherited source IDs and records the absent original URLs, source-access dates, image links, and comment links as explicit documentation gaps. Three entries point to separately identified WEB20261005 observation records. The join denotes a current title-based source match, not recovery of a source snapshot. Original visual and comment units remain unlinked; no artificial item identifiers are assigned to absent observations. These records therefore preserve provenance limits rather than reconstruct full source histories.

A case-level association is weaker than an item-level evidentiary link. For example, an institutional page and comments under a Chingay video may concern Johor Ancient Temple without referring to the same event, period, or audience. Bringing them together supports contextual comparison, not a matched reception study. Researchers should not assume that commenters saw the institutional page or that people pictured in an image authored the comments. Keeping this distinction central to the proposed framework clarifies the limits of provenance.

(4) METHOD

(4.1) COLLECTION BOUNDARIES AND ANALYTICAL UNITS

This study documents aggregate summaries from a purposive qualitative analysis of four Malaysian Chinese temples, supplemented by a separately dated source-retrieval exercise. Table 4 defines the analytical units. The documentation exercise exports aggregate tables, preserves source identifiers, records claim-level uncertainty, and checks which statements are arithmetically reproducible from the public release. It does not independently repeat the original empirical analysis.

Table 4 Analytical units and coding focus.

ANALYTICAL DOMAIN	UNIT	CODING FOCUS
Textual representation	A source page, profile description, or media article	Heritage narrative, institutional identity, ritual reference, tourism framing, multicultural framing.
Visual representation	One image or visually coherent image unit	Symbol, salience, color, composition, spatial framing, crowd/ritual presence, heritage atmosphere.
Comment discourse	One public YouTube comment	Language pattern, theme, case affiliation, recurrent terms, and creator-oriented address.
Cross-case synthesis	Case-level pattern	Narrative emphasis, visual strategy, platform affordance, public engagement pattern.

The platform-access register records YouTube comment retrieval using yt-dlp and limited or unavailable access elsewhere (Table 5). Original video IDs, commands, collection logs, and software versions are not available in the documentation package. Table 5 describes reported collection conditions, not access independently repeated during the documentation exercise.

Table 5 Reported platform-access boundaries.

PLATFORM	STATUS	NOTE
YouTube	retrieved	Public video comments retrieved with yt-dlp from publicly accessible videos; author identifiers are hashed.
Tripadvisor	access_limited	Direct requests returned JS/captcha protection; not bypassed. Search snippets only used for source discovery, not comment corpus.
Google Reviews	not_retrieved	No official API key or authorized export available during collection; scraping dynamic review pages was not attempted.
Facebook	limited	Logged-out/public metadata only in the original corpus; full comment streams require authorization or manual export.
Xiaohongshu/Douyin/ Zhihu/Weibo/Baidu	not_retrieved	Public, stable, non-login comment access was not available within ethical collection limits during collection.

The reported cleaned total does not reveal how many comments were originally retrieved or removed. The available documentation does not specify deduplication, spam handling, treatment of replies, empty comments, emoji, or multilingual segmentation. Accordingly, the word cleaned identifies a reported processing stage, not an independently reproducible procedure. A release should recover these decisions from records where possible and explicitly mark unrecoverable details.

(4.2) CODING AND COMPUTATIONAL ASSISTANCE

The original corpus analysis used theory-informed and inductive coding of ritual, architectural, identity, devotional, tourism, and creator-oriented themes (Krippendorff, 2018; Schreier, 2012). Its analytical categories distinguish person-to-place, person-to-group, creator-viewer, cross-group, intergenerational, and co-presence relations. These are retained analytical descriptions, not a recovered operational codebook. The present documentation does not assign these codes to absent comments or claim validated boundaries.

The reported visual analysis concerns symbols, salience, framing, color, light, crowds, and ritual action across 16 visual units. The images and unit-level assignments are not included in the public package for independent checking. The four modes in Table 6 are interpretive summaries whose evidentiary limits are retained explicitly.

Table 6 Cross-case synthesis of heritage representations and comment evidence.

CASE	DOMINANT NARRATIVE	DOMINANT VISUAL STRATEGY	PLATFORM-VISIBLE RELATIONAL DISCOURSE
THT	Curated Chinese cultural environment and Mazu-centered public attraction.	Architecture, lanterns, deity halls, gardens, festival lights.	Small corpus; exploratory context only.
KLS	Festival spectacle and heritage tourism.	Chinese New Year lighting, temple scale, viewing atmosphere.	Small corpus; exploratory context only.
CHT	Historical continuity and heritage legitimacy.	Historic facade, inscriptions, preserved architectural memory.	Insufficient comment evidence for engagement claims.
JAT	Local identity, Chingay, dialect-group unity, and civic heritage.	Procession, crowd, ritual movement, old temple identity.	Large YouTube comment corpus; local identity, creator mediation, and festival affect are prominent.

The original corpus analysis involved close reading, thematic comparison, recurrent terms, and case-level TF-IDF. Complete tokenization, translation, stopword, weighting settings, and outputs are not available. TF-IDF is therefore reported as part of the original analytical procedure, rather than an independently reproduced component of the documentation exercise. The available records do not support new computational theme or audience inferences.

This study documents no independently coded comparison, agreement statistic, or adjudication record. The reported theme totals therefore remain descriptive coding outputs. Category exclusivity, multiple coding, and treatment of the residual other category require clarification before percentages or inferential comparisons are appropriate.

(4.3) DOCUMENTATION CONSTRUCTION AND SOURCE CHECKING

We copied each source inventory record with its identifier and case code, then augmented it with explicit provenance and missing-field states. A separate claim register distinguishes recomputable sums, counts without record-level verification, and retained interpretations. On 5 October 2026, three title-based searches targeted the HainanNet architecture page, the Discover Malaysia-UNESCO Cheng Hoon Teng page, and the Tourism Johor Old Chinese Temple listing. Two page texts were retrieved; the third timed out after search discovery. Each access observation records its URL, date, locator, method, status, paraphrased observation, interpretation boundary, and relationship to the original inventory entry. This purposive demonstration is not a representative source sample.

CSV tables were mirrored into SQLite with descriptive table names. The validation script uses Python's csv and json modules to recompute count sums, Johor's corpus share, source-ID uniqueness, and access-status totals. SHA-256 hashes support file-integrity checking. These checks were run on the revised release; neither arithmetic agreement nor source rediscovery validates original thematic coding, historical page content, ethical clearance, or license claims.

(4.4) EVIDENCE LEVELS AND PRIVACY

The analytical framework distinguishes three levels of evidence. Representational evidence concerns how texts and images depict affiliation, co-presence, and collective selfhood. Discursive evidence concerns how commenters publicly express recognition, support, place attachment, or creator-oriented engagement. Interpersonal-outcome evidence concerns persistent bonds, relationship quality, or causal effects; the present dataset does not provide that level of evidence. This distinction clarifies the scope of the original corpus analysis and the current documentation.

The comment dataset reportedly uses hashed author identifiers and refrains from reproducing searchable quotations. These precautions reduce direct disclosure but do not establish complete anonymity: original wording, timestamps, video identifiers, and repeat appearances can still permit linkage. The documentation release therefore separates aggregate counts and researcher-authored annotations from potentially traceable comment content. Public accessibility and hashing alone do not establish ethical exemption or redistribution permission.

(5) RESULTS AND DISCUSSION

(5.1) CORPUS IMBALANCE AND THE SCOPE OF COMPARISON

Johor Ancient Temple contributes 1,466 of 1,554 comments (94.34%) and 22 of the 29 comment-bearing videos (Table 7). Thean Hou Temple contributes 44 comments across two videos, Kek Lok Si 40 across four, and Cheng Hoon Teng four from one. Consequently, a pooled comment analysis predominantly describes the Johor subset. A larger retrieved subset does not indicate stronger community belonging: differences may arise from video selection, event visibility, uploader activity, platform access, or retrieval timing.

Table 7 Public YouTube comments and comment-bearing videos by case.

CASE	COMMENTS	VIDEOS	ANALYTICAL STATUS
CHT	4	1	Insufficient for comment interpretation
JAT	1466	22	Primary comment corpus
KLS	40	4	Exploratory context only
THT	44	2	Exploratory context only

The four cases remain analytically useful because their textual and visual materials offer contrasting heritage representations. However, the comparison is asymmetric. Institutional and visual interpretation can inform all four cases, while substantive comment discussion is concentrated on Johor. Small comment subsets supply exploratory context and cannot function as equivalent comparison groups. The full distribution makes this imbalance explicit and prevents the four cases from being treated as equally supported.

(5.2) FOUR MODES OF HERITAGE REPRESENTATION

These four modes summarize the reported interpretations; original visual and comment records are not publicly deposited for independent revalidation. Thean Hou Temple illustrates how religious architecture can be communicated as a setting for shared affiliation. Official materials organize deity worship, gardens, festival events, and visitor attractions into a coherent route. This strategy makes the site legible to worshippers, local visitors, tourists, news audiences, and social-media viewers. It places worship and heritage within a public environment where participants can orient themselves to common symbols.

The architecture material identified in the HainanNet source inventory emphasizes rooflines, lanterns, and layered ornamentation. In the reported visual interpretation, these features make a shared cultural environment recognizable. This reading reflects public representation; it does not show how visitors actually relate to one another. The small Thean Hou comment subset offers insufficient evidence to interpret silence or limited discussion as weak attachment.

Kek Lok Si Temple illustrates festival spectacularization. In the collected materials, Chinese New Year lighting, tourism circulation, and seasonal atmosphere make temple-related heritage visible as an event viewed with others. The relational importance lies in representing coordinated attention and shared festivity rather than proving relationship formation.

Only 40 Kek Lok Si comments were available. They provide limited examples of thankful and tourism-oriented expression, but the subset is too small to establish visitor satisfaction, relationship quality, or representative reception.

Cheng Hoon Teng Temple represents historical and intergenerational continuity. Heritage-oriented materials connect the temple to Melaka's urban history, long-term Chinese presence, and continuing worship. The dominant relational claim is representational: present publics are positioned in relation to predecessors, inscriptions, preserved craft, and an enduring place of worship.

The original Cheng Hoon Teng visual interpretation spotlights a frontal facade, roofline, and inscriptions as resources for communicating historical persistence. These features can anchor a discussion of architectural memory, but viewers' experiences of continuity remain unmeasured. Four retrieved comments cannot substantiate a case-level interpretation of reception. The heritage-oriented sources are therefore the principal evidence for this case.

Johor Ancient Temple provides the strongest evidence of relationship-oriented public discourse. Its materials connect the temple to Johor Bahru, five dialect groups, and the Chingay procession. Heritage appears here as civic ritual: a repeated public event through which local affiliation, organizational unity, co-presence, and participation become visible.

For Johor, the reported visual interpretation emphasizes crowds, ritual objects, cameras, and night lighting as a representation of coordinated civic ritual. This reading should remain linked to an identified source image or frame in a future inventory. The visual inventory requires item-level identifiers and source links before we can evaluate the images independently.

(5.3) JOHOR COMMENT CATEGORIES AS DESCRIPTIVE ANNOTATIONS

Table 8 presents the six reported non-other theme counts for Johor Ancient Temple. These counts are retained as descriptive annotations rather than a verified exhaustive or mutually exclusive classification.

Table 8 Reported non-other theme counts in Johor Ancient Temple comments.

THEME	COMMENTS	RELATIONAL INTERPRETATION
creator-mediated engagement	234	Users interact with video creators and filming practices, showing that heritage circulation is shaped by platform mediation.
local identity	213	Users connect temple ritual with Johor/Johor Bahru belonging and local pride.
festival affect	154	Users respond emotionally to processions, crowds, spectacle, and festive atmosphere.
heritage pride	131	Users frame Chingay and temple culture as valuable Chinese or Malaysian Chinese heritage.
multicultural recognition	109	Users recognize coexistence, cross-community visibility, or Malaysia's plural setting.
devotional blessing	67	Users express blessing, reverence, or ritual respect.

The reported counts are 234 for creator-mediated engagement, 213 for local identity, 154 for festival affect, 131 for heritage pride, 109 for multicultural recognition, and 67 for devotional blessing—the six values sum to 908. Because the complete coding rules, residual categories, and provenance are unavailable, that sum is not treated as the number of uniquely classified comments, and the difference from 1,466 is not assigned to "other". A full codebook and assignment table are needed to distinguish overlap, exclusions, and uncoded records.

Creator-mediated engagement concerns uploaders and discussion of filming; local identity concerns place affiliation; festival affect concerns reactions to event atmosphere; heritage pride concerns cultural valuation; multicultural recognition concerns public recognition of plurality; and devotional blessing concerns ritual respect. These readings retain the analysis's substantive focus. They indicate communicative positioning within the corpus, not nationality, community membership, relationship persistence, or representative attitudes. Critical comments about restrictions or governance should remain visible, although no count is provided for that category.

(5.4) A DOCUMENTATION PROTOCOL AND ITS IMPLEMENTED SCOPE

The corpus suggests four documentation criteria. First, provenance should identify the originating object and its producer, while also marking the limits of what the record can establish. Institutional statements, tourism accounts, news coverage, and viewer comments need distinct source types. Second, analytical units should be explicit: an image crop, a video frame, a complete page, and a comment require different identifiers and contextual fields. Third, annotations should remain separable from source content so alternative readings can be recorded. Fourth, missing metadata and restricted access should have explicit states rather than blank cells that imply absence of cultural activity.

The release implements a documentation subset of this architecture: source provenance, claim evidence, new access observations, aggregate tables, and an SQLite mirror. The source register contains 11 unique IDs; the three newly created observation IDs remain distinct. A full visual-unit register, video register, and comment annotation table would require original records. The implemented subset is inspectable, while those absent components remain outside the current scope.

Missing-data states describe different stages: “not retrieved” denotes unsuccessful or unavailable collection; “not recorded” denotes absent documentation; “not coded” denotes absent analytical assignment; and “not released” denotes exclusion from public dissemination. These states are distinct from restricted access and from a measured zero. Unavailable Facebook discussion, for example, cannot be interpreted as zero engagement.

The validation report confirms 1,554 reported comments, 29 comment-bearing videos, 11 unique source identifiers, and Johor’s 94.34% share. It also confirms the six theme values sum to 908 and that two of three new access observations contain retrieved page text. The report explicitly marks original raw-record availability and theme-partition verification as false. It does not subtract 908 from 1,466 to create an “other” category. That subtraction would silently assume an undocumented, exhaustive, exclusive coding scheme.

(5.5) A WORKED PROVENANCE EXAMPLE

Thean Hou source S003_THT illustrates a three-step sequence: inherited source entry, newly dated observation, and bounded claim. The inherited inventory gives a title and platform but no original visit date or image association. WEB20261005_001 supplies a newly checked URL and identifies the architecture-page introduction as its locator. The association’s narrative describes color contrasts and differing roof-ridge forms (Selangor and Federal Territory Hainan Association, n.d.). This supports the observation that institutions narrate architectural features as cultural resources. It does not prove that the original 16-unit visual corpus included a particular webpage image, or that visitors experienced belonging. A researcher can join the two records by legacy_source_id while preserving these limits.

The Cheng Hoon Teng observation similarly records a heritage-page narrative connecting architecture and continuing community practices (Discover Malaysia-UNESCO, n.d.). The Tourism Johor entry demonstrates the contrary state: search discovery with page-retrieval failure. Its URL is retained for follow-up, but no content code is assigned. Consequently, an accessible title, a retrieved institutional narrative, and a validated historical claim remain distinct evidentiary states.

(5.6) LIMITATIONS AND THE CONTRIBUTION TO HUMANITIES DATA PRACTICE

The main limitation is the gap between the reported corpus and a fully documented reusable release. The first author retains the records collected on 10 July 2026, but original comments, the complete visual inventory, search strings, processing settings, and coding assignments are absent from the public package. File-level checks verify documentation integrity and aggregate arithmetic, but not collection, cleaning, or coding. The documentation protocol makes the consequences of these gaps explicit.

(6) REUSE POTENTIAL

(6.1) REUSE SUPPORTED BY THE CURRENT RELEASE

Aggregate validation requires Python 3 and basic command-line skills; provenance joins require SQLite and basic SQL. Claim-evidence inspection requires a CSV viewer or spreadsheet application and examines evidentiary support rather than recoding comments.

A reproducible reuse exercise starts by running validate_aggregates.py and comparing its JSON output with validation_report.json. A second exercise opens heritage_documentation.sqlite and joins source_provenance to new_web_observations on new_observation_id = observation_id. S003_THT then resolves to a dated architecture-page observation in the documentation release while retaining unknown original dates and item links in the source record. A third exercise selects claim C04 to explain why a reproducible sum of 908 is not evidence of an exhaustive theme partition. These operations demonstrate arithmetic and documentation reuse without reinterpreting absent comments.

The current aggregate and provenance package supports teaching in visual culture, heritage studies, and qualitative digital methods. Students can compare documented observation with interpretation and examine why a large platform subset does not imply representative community reception. Aggregate distributions and researcher-authored annotations support these documentation exercises without redistribution of identifiable comment text or third-party images. Any educational release should state what was omitted, how that affects the exercise, and where provenance cannot be fully recovered.

(6.2) REUSE REQUIRING ADDITIONAL RECORDS

If the retained original records are shared and supplemented with the necessary documentation, they could support comparative research on digital heritage representation. Institutional and media descriptions could be contrasted with visual annotations to examine which architectural, historical, and ritual features receive emphasis. Comment annotations could support close reading of place affiliation, creator address, devotional language, and governance disagreement, with Johor treated as the principal comment case. These uses depend on recoverable source associations and should retain the asymmetry of the collected material, while preserving the limits of provenance and attribution.

Computational reuse requires original records and fuller documentation. Topic or lexical analysis needs documented language handling, tokenization, and preprocessing; cross-case comparison needs sensitivity analysis for unequal sizes. Network analysis requires authorized relational data beyond these counts. Relationship prediction or audience-effect measurement would require a different design, additional consent or authorization where applicable, outcome evidence, and clearer provenance.

(7) CONCLUSION

The four-temple case demonstrates a limited, implemented documentation protocol. It preserves the distinction between 11 source records, 16 reported visual units, and 1,554 reported comments; it makes Johor's 94.34% concentration reproducible; and it separates inherited interpretations from three new dated access observations. Eleven provenance records, six claim-evidence records, a SQLite mirror, and a validation script allow users to inspect what each claim depends on. The contribution is traceable documentation under incomplete evidence, not comprehensive open data, validated thematic categories, or verified interpersonal outcomes. Access to the retained original records and fuller documentation would permit a stronger empirical analysis, but new observations must remain separately dated rather than silently repairing the past.

AI DECLARATION

OpenAI's ChatGPT assisted with language revision, restructuring, source discovery, preparation of dated source-observation metadata and paraphrases, and documentation and validation code. It did not create original comment text, reconstruct missing visual units, assign original empirical codes, or establish ethics approval. Newly created web-access observations are separately dated and distinguished from the original corpus materials. The authors remain responsible for reviewing all AI-assisted text, source matches, code, metadata, interpretations, and declarations.

DATA ACCESSIBILITY STATEMENT

Version 1.2 of the aggregate and provenance package is openly available in Zenodo (Wu et al., 2026): https://doi.org/10.5281/zenodo.23367960. Contents are detailed in Section 3.1. Researcher-authored materials are licensed under CC BY 4.0. Original comments, user identifiers, third-party images, and original coding assignments are not included in the public deposit. The original corpus records are retained by the first author, Gengye Wu, and may be made available upon reasonable request for scholarly exchange.

ETHICS AND CONSENT

This documentation exercise uses aggregate summaries and public institutional pages, without recruiting participants or releasing identifiable comments, searchable quotations, or third-party images. No institutional approval, exemption, or other formal determination is documented in the supplied materials or claimed here. Public accessibility and reported hashing are not treated as ethical exemption or consent to unrestricted redistribution.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the temple representatives, cultural heritage researchers, and community members whose publicly available accounts provided contextual information on Malaysian Chinese temple heritage. The authors also thank the Institute of Malaysian and International Studies (IKMAS), Universiti Kebangsaan Malaysia (UKM), for institutional and academic support.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Gengye Wu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft.

Muhamad Azwan Abd Rahman: Project administration, Supervision, Validation, Writing – review & editing.

Tan Raan Hann: Writing – review & editing, with primary responsibility for language improvement.

AUTHOR AFFILIATIONS

Gengye Wu orcid.org/0009-0004-6390-1030
The Institute of Malaysian and International Studies (IKMAS), The National University of Malaysia (UKM), Malaysia

Muhamad Azwan Abd Rahman orcid.org/0000-0002-9415-0901
The Institute of Malaysian and International Studies (IKMAS), Universiti Kebangsaan Malaysia (UKM), Malaysia

Tan Raan Hann orcid.org/0000-0002-9361-3038
The Institute of Malaysian and International Studies (IKMAS), Universiti Kebangsaan Malaysia (UKM), Malaysia

REFERENCES

Assmann, J. (1995). Collective memory and cultural identity. *New German Critique*, *65*, 125–133. https://doi.org/10.2307/488538

Baumeister, R. F., & **Leary, M. R.** (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Campbell, H. A. (Ed.). (2013). *Digital religion: Understanding religious practice in new media worlds*. Routledge.

Carey, J. W. (1989). *Communication as culture: Essays on media and society*. Routledge.

Collins, R. (2004). *Interaction ritual chains*. Princeton University Press. https://doi.org/10.1515/9781400851744

Dean, K. (2020). Religion and the Chinese diaspora in Southeast Asia. *Review of Religion and Chinese Society*, *7*(2), 213–242. https://doi.org/10.1163/22143955-00702005

Dean, K. (2022). One ocean one temple: Alternative Chinese temple networks in Southeast Asia. *Global Networks*, *22*(4), 699–715. https://doi.org/10.1111/glob.12391

Discover Malaysia-UNESCO. (n.d.). *Cheng Hoon Teng Temple*. Retrieved October 5, 2026, from https://discovermalaysia-unesco.com/portfolio/cheng-hoon-teng-temple/

Hall, S. (Ed.). (1997). *Representation: Cultural representations and signifying practices*. Sage.

Highfield, T., & **Leaver, T.** (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, *2*(1), 47–62. https://doi.org/10.1080/22041451.2016.1155332

Hjarvard, S. (2008a). The mediatization of religion: A theory of the media as agents of religious change. *Northern Lights*, *6*(1), 9–26. https://doi.org/10.1386/nl.6.1.9_1

Hjarvard, S. (2008b). The mediatization of society: A theory of the media as agents of social and cultural change. *Nordicom Review*, *29*(2), 105–134. https://doi.org/10.1515/nor-2017-0181

Hjarvard, S. (2013). *The mediatization of culture and society*. Routledge. https://doi.org/10.4324/9780203155363

Hjarvard, S. (2016). Mediatization and the changing authority of religion. *Media, Culture & Society*, *38*(1), 8–17. https://doi.org/10.1177/0163443715615412

Horton, D., & **Wohl, R. R.** (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, *19*(3), 215–229. https://doi.org/10.1080/00332747.1956.11023049

Hu, N., Ho, K. C., & **Fan, P. S.** (2024). Malaysian Chinese folk beliefs on Facebook based on LDA topic modelling. *Humanities and Social Sciences Communications*, *11*, Article 611. https://doi.org/10.1057/s41599-024-03066-6

Koh, J. H., & **Lim, C. M.** (2022). Architectural language of the Southern Fujianese traditional Chinese temple timber frame structure in West Malaysia: A cultural semiotics analysis. *International Journal of Architectural Research: ArchNet-IJAR*, *17*(1), 42–62. https://doi.org/10.1108/ARCH-12-2021-0351

Kress, G., & **van Leeuwen, T.** (2006). *Reading images: The grammar of visual design* (2nd ed.). Routledge.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Liew, C. W., Tang, H., Liu, X., & **Li, J.** (2025). A comparative analysis of eaves characteristics in Chinese temples in Lingnan, China, and Johor-Malacca, Malaysia. *Journal of Asian Architecture and Building Engineering*. https://doi.org/10.1080/13467581.2025.2514267

Nora, P. (1989). Between memory and history: Les lieux de memoire. *Representations*, *26*, 7–24. https://doi.org/10.2307/2928520

Rose, G. (2016). *Visual methodologies: An introduction to researching with visual materials* (4th ed.). Sage.

Salton, G., & **Buckley, C.** (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage.

Selangor and Federal Territory Hainan Association. (n.d.). *Thean Hou Temple architecture*. Retrieved October 5, 2026, from https://www.hainannet.com.my/thean-hou-temple-architecture/?lang=en

Tajfel, H., & **Turner, J. C.** (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

World Wide Web Consortium. (2013, April 30). *PROV-Overview: An overview of the PROV family of documents*. https://www.w3.org/TR/prov-overview/

Wu, G., Abd Rahman, M. A., & **Tan, R. H.** (2026). *Malaysian Chinese temple heritage public aggregate data and documentation* (Version 1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.23367960

Yan, Y., Dean, K., Feng, C.-C., Hue, G. T., Koh, K.-H., Kong, L., Ong, C. W., & **Tay, A.** (2020). Chinese temple networks in Southeast Asia: A WebGIS digital humanities platform for the collaborative study of the Chinese diaspora in Southeast Asia. *Religions*, *11*(7), 334. https://doi.org/10.3390/rel11070334

TO CITE THIS ARTICLE:
Wu, G., Abd Rahman, M. A., & Raan Hann, T. (2026). Curating Multimodal Temple Heritage Data for the Study of Public Belonging in Malaysia. *Journal of Open Humanities Data*, 12: 146, pp. 1–12. DOI: https://doi.org/10.5334/johd.688

Submitted: 05 October 2026
Accepted: 16 October 2026
Published: 09 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.