SEA-HISTIN Technical Documentation

Inventory of Historical Weather Observations of Southeast Asia Version: v0 | Documentation compiled: July 2026

Praveen Teleti[1,2], Fiona Williamson[1]

[1]College of Integrative Studies, Singapore Management University, 10 Canning Rise, Singapore
[2]Dept of Meteorology, University of Reading, United Kingdom

Correspondence to: Praveen Teleti (praveent@smu.edu.sg)

S1. Purpose and scope

SEA-HISTIN is a relational **inventory of source documents**, not a database of digitised observations. Each record describes where a historical weather record survives, what it contains, and how ready it is for digitisation. Downstream users query the inventory to decide *what to rescue next*; the observations themselves live elsewhere, in the archives and repositories the inventory points to.

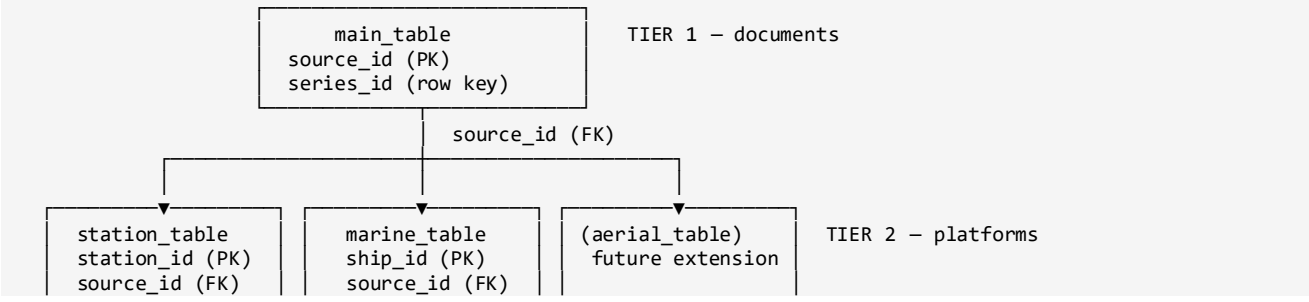
The v0 release comprises four comma-separated files:

File	Tier	Rows	Fields	Grain (one row = …)
`main_table_data_v0.csv`	Primary	183	27	one physical source document
`station_table_data_v0.csv`	Secondary (platform)	1,984	38	one terrestrial station, as reported in one document
`marine_table_data_v0.csv`	Secondary (platform)	10,729	16	one ship voyage / logbook entity, as reported in one document
`platform_instrument_table_v0.csv`	Tertiary (attribute)	4	19	one instrument configuration at one station

All files are UTF-8-compatible plain CSV with a single header row. Note that `marine_table_data_v0.csv` and parts of `station_table_data_v0.csv` contain characters requiring Latin-1 fallback on some readers; UTF-8 normalisation is scheduled for v1.

S2. Relational architecture

The inventory is a three-tier star schema. Documents sit at the centre; observing platforms hang off documents; fine-grained attributes hang off platforms.



S2.1 Why three tiers

A single bulletin may tabulate hundreds of stations; a single station may recur across decades of bulletins. Storing both in one flat table would either duplicate document metadata across every station row, or collapse station detail into an aggregated document record. Separating the tiers resolves the **many-to-many relationship between documents and platforms** without either loss.

Observed in v0:

- **One document → many platforms.** Stations per source range from 1 to **1,156**

(median 1). Ships per source range from 82 to **4,958**.

- **One platform → many documents.** 178 standardised station names appear under more

than one `source_id`; the most-repeated name appears across **6** distinct documents. Each appearance is a separate row with its own `station_id`, preserving provenance while allowing deduplication downstream.

S2.2 Join semantics

Join	Cardinality	Key
main → station	1 : 0..n	`source_id`
main → marine	1 : 0..n	`source_id`
station → instrument	1 : 0..n	`station_id`
main → instrument	1 : 0..n	`source_id` (denormalised convenience key)

`platform_instrument_table` carries `source_id` in addition to `station_id`. This is a deliberate denormalisation: it permits instrument records to be attributed to a *different* document from the one that reported the station (e.g. a station listed in a bulletin, but its gauge specification given in a separate planting manual). Where the instrument metadata derives from the same document, the value simply duplicates the station's `source_id`.

S2.3 Coverage of the linkage in v0

- Referential integrity is **clean**: 0 orphan `source_id` values in either secondary table.
- 33 of 183 sources have station records; 13 have marine records; **0 have both**.
- **137 of 183 sources (75%) have no secondary records at all.** These are documents

catalogued at the document level whose constituent platforms have not yet been enumerated. This is the single largest expansion front for v1 and should not be read as those documents containing no stations.

S3. Identifier scheme

S3.1 Construction

All identifiers are **8 characters**: a one-character type prefix followed by seven characters drawn from base-62 (`0-9`, `A-Z`, `a-z`).

```
I w 4 z 5 n 7 e
```

```
    │  └──────────┘
    │       └─ 7 random base-62 characters  (62^7 ≈ 3.52 × 10^12 combinations)
    └──────── type prefix
```

Prefix	Identifier	Table	Namespace size
A	series_id	main	183 issued
B	source_id	main	183 issued
C	station_id	station	1,984 issued
C	ship_id	marine	10,729 issued
I	instrument_id	instrument	4 issued

Identifiers are **opaque and non-semantic**: they encode no country, date, institution, or sequence. Nothing may be inferred from the character content beyond the type prefix. This is intentional — semantic identifiers become wrong when the underlying facts are revised, whereas opaque identifiers survive recataloguing.

S3.2 Properties and guarantees

- **Persistent.** Once issued, an identifier is never reused, reassigned, or altered, even

if the record it labels is corrected or withdrawn.

- **Unique within namespace.** Verified in v0: no duplicates in any identifier column.
- **Non-sequential.** Random generation avoids implying a chronology or a completeness

that the inventory does not have.

- **Collision probability.** At current scale (~13,000 identifiers), the birthday-collision

probability across a shared 62^7 namespace is on the order of 10^-5. Generation nonetheless checks against issued identifiers and regenerates on collision.

S3.3 `series_id` versus `source_id`

Both are issued one-per-row in `main_table` and both are unique, so they are functionally redundant in v0. Their intended distinction:

- `source_id` — the **stable external identifier** of the document. This is the value

cited, linked, and used as the foreign key in all other tables.

- `series_id` — an **internal row key** for the primary table, reserved for the case where

one physical document is later split into multiple bibliographic series (e.g. a bound volume of annual reports catalogued as one document in v0, resolved into per-year entries in v1). At that point one `series_id` would map to several `source_id` values.

Until that split occurs, `series_id` should be treated as reserved and **not** used as a join key.

S4. Table specifications

S4.1 `main_table` — document tier

Field	Type	Description	
source_id	ID (B+7)	PK. Stable document identifier; FK target for all secondary tables.	

series_id	ID (A+7)	Reserved internal row key (§3.4).	
source_title	text	Title of the document as it appears on the item.	
author_or_agency	text	Responsible person or issuing body.	
publication_year	integer	4-digit year, e.g. 1911.	
archive_or_library	text	Current holding institution.	
call_number_or_url	text	Shelfmark, catalogue reference, or persistent URL.	
citation_recommendation	text	Preferred citation string for the document.	
source_type	vocab	Terrestrial | Marine — determines which secondary table holds its platforms.	
data_format	vocab	Physical presentation of the data (§5.1).	
quantitative	boolean	Whether the document contains numeric observations (inclusion criterion).	
language / script	vocab	Language of text and writing system.	
legibility_score	vocab	Legibility of the surviving copy.	
scan_status	vocab	Quality of available imaging.	
transcription_readiness	vocab	Practical digitisation feasibility (§5.1).	
data_rescue_status	vocab	Imaged | Transcribed — position in the rescue pipeline.	
completeness_category	vocab	Continuity of the record within the document.	
known_gaps	boolean	Whether interruptions are present.	
known_gaps_notes	text	Free-text description of those gaps.	
copyright_status	vocab	Access and reuse constraints.	
related_publications	text	Bibliographically related items.	
related_series_ids	text	Cross-references to other inventory rows.	
remarks	text	Free-text cataloguer notes.	
inventory_version	vocab	Release in which the row was issued (v0).	
created_date, last_updated	date	Record provenance timestamps.	

S4.2 station_table — terrestrial platform tier

Grain: one station **as reported in one document**. A station appearing in three bulletins generates three rows with three station_id values sharing one standardised name.

Group	Fields

Keys	station_id (PK), source_id (FK)
Naming	station_name_original, station_name_standardized, alternative_names, administrative_unit_historical, country_modern
Location	lat_decimal, lon_decimal, latitude_original, longitude_original, coordinate_reference, coordinate_uncertainty_km, elevation_m, distance_to_coast_km
Temporal	start_date, end_date, temporal_resolution, observation_times, time_standard, calendar_type
Variables	variable_rainfall, variable_temperature, variable_pressure, variable_humidity, variable_sunshine, variable_cloud, variable_wind, variable_weather_codes, other_variables
Units	rainfall_units, temperature_units, pressure_units, pressure_reduction, unit_changes_noted
Quality	internal_consistency, location_stability, station_status

The dual naming fields are load-bearing: `station_name_original` preserves the document's own orthography (`Buitenzorg`, `Malacca`), while `station_name_standardized` carries the harmonised form used for matching and deduplication. Harmonisation is performed against the Getty Thesaurus of Geographic Names and GeoNames.

S4.3 `marine_table` — marine platform tier

Grain: one vessel/logbook entity as reported in one document.

Field	Description
ship_id (PK), source_id (FK)	Keys.
Archive No./Folder	Archival reference within the holding repository.
Vessel, Ship Type, Owner of Port of Registry	Platform identity and operator.
Start Date, End Date	Voyage span (§6).
Document Title, Document Description	Item-level description.
Vessel Movements	Free-text itinerary — the marine analogue of a fixed coordinate.
Meteorology, Oceanography, Other data	Free-text statements of recorded content.
Notes, ship_url	Remarks and link to the digitised item.

Naming-convention defect. This table uses Title Case with spaces and one embedded slash (`Archive No./Folder`), whereas every other table uses `snake_case`. This breaks programmatic column access (`df.ship_id` works; `df.Ship Type` does not) and requires quoting in SQL. *Recommendation for v1:* migrate to `snake_case` (`archive_no_folder`, `ship_type`, `owner_or_port_of_registry`, `start_date`, `end_date`, `vessel_movements`, `meteorology`, `oceanography`, `other_data`, `notes`), publishing a column crosswalk.

Content differentiation. Unlike the station table's boolean variable flags, marine content is recorded as free text: 677 distinct `Meteorology` strings, 412 `Oceanography`, 239 `Other data`. This is richer but not machine-queryable. v1 should add parallel boolean flags derived from these strings while retaining the originals as `raw_text`.

S4.4 `platform_instrument_table` — instrument attribute tier

Grain: one instrument configuration at one station, valid over a stated date range.

Field	Description		
instrument_id (PK)	I+7 identifier.		

station_id (FK)	Parent station in station_table.		
source_id (FK)	Document supplying the instrument metadata (§2.2).		
variable	Element measured — one row per element per configuration.		
instrument_type, manufacturer	Equipment identity.		
gauge_diameter_mm, rim_height_m, sensor_height_m, elevation_m	Physical dimensions, all in SI units regardless of the source's original units.		
exposure_notes	Siting, screening, shielding, enclosure.		
units_original	Units as printed in the source (e.g. inches).		
correction_applied	Stated reductions (e.g. reduced to MSL).		
calibration_date	Stated calibration, if any.		
install_date, removal_date	Validity period. Empty removal_date = still in use or unknown.		
standardization_status	Controlled: standardized |	non_standard_described |	unknown.
confidence	Controlled: high |	medium |	low.
raw_text	Verbatim source wording, for provenance and audit.		

The validity period is the point of this table. Because instrument_id — not station_id — is the primary key, one station can hold several dated instrument rows. A station whose 5-inch gauge was replaced by an 8-inch gauge on relocation generates two rows sharing a station_id, distinguished by install_date/removal_date. This preserves the instrument history that downstream homogenisation requires, rather than a single snapshot that silently discards it.

The v0 content (4 rows) derives from Wycherley (1967), *Rainfall Probability Tables for Malaysia*, RRIM Planting Manual No. 12, Appendix D, which specifies the network standard for Malayan rainfall stations: an 8-inch (203.2 mm) manual gauge and a 5-inch (127.0 mm) syphon recorder, each in two permitted mountings (earth-bank deflector, or post with Nipher windshield) that differ in rim height. These four rows therefore describe a **network specification** rather than four physically distinct instruments, and are intended to be cloned onto each station documented as conforming to that standard.

S5. Controlled vocabularies

S5.1 Document tier (main_table)

Field	Permitted values	v0 usage
source_type	Terrestrial, Marine	170 / 13
data_rescue_status	Imaged, Transcribed	77 / 106
known_gaps	True, False	both present
quantitative	True, False	True only

data_format	Printed Table, Manuscript Ledger, Chart, Coded Synoptic Form, Mixed	Printed Table only
completeness_category	Continuous, Intermittent, Fragmentary	Continuous only
legibility_score	High, Medium, Low	High only
scan_status	Good, Fair, Poor, Not scanned	Good only
transcription_readiness	High, Medium, Low	High only
copyright_status	Public, Restricted, Unknown	Public only
language / script	ISO-639 language names / script names	English only
inventory_version	v0, v1, …	v0

S5.2 Platform tier (station_table)

Field	Permitted values	v0 usage
country_modern	15 values: Australia, Brunei, Cambodia, China, Disputed, Federated States of Micronesia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, United States, Vietnam	all in use
coordinate_reference	Exact, Estimated, Inferred	Inferred only
coordinate_uncertainty_km	0, <5, 5-25, >25	0 only
temporal_resolution	Hourly, Synoptic, Daily, Monthly	Monthly only
time_standard	Local, Standard, GMT, Unknown	Local only
calendar_type	Gregorian, Julian, Other	Gregorian only
internal_consistency	Consistent, Unit changes, Ambiguous notation, Mixed conventions	Consistent only
location_stability	Stable, Relocation documented, Relocation suspected, Unknown	Unknown only
observation_times	free text or Unknown	Unknown only
rainfall_units	mm, inches	mm only
temperature_units	degC, degF	degC only
pressure_units	hPa, mb, inHg, mmHg	hPa only
pressure_reduction	Sea-level, Station level, Unknown	Sea-level only
unit_changes_noted	True, False, Unknown	Unknown only
variable_* (8 fields)	True, False	see §9.2

S5.3 Instrument tier

Field	Permitted values
standardization_status	standardized, non_standard_described, unknown
confidence	high, medium, low
variable	rainfall, temperature, pressure, humidity, sunshine, cloud, wind, weather_codes

Case convention is inconsistent between tiers: the document and platform tiers use `Title Case` vocabulary values, the instrument tier uses `lower_snake`. v1 should standardise on `lower_snake` throughout for machine-matching, with display labels handled at presentation time.

S6. Date encoding

Dates in the platform tables are stored as **compact integers**, not ISO strings:

Example	Parsed as	Note
`1041963`	1 Apr 1963	D-MM-YYYY, leading zero dropped
`11101803`	11 Oct 1803	DD-MM-YYYY
`16081843`	16 Aug 1843	DD-MM-YYYY

The encoding is **ambiguous for single-digit days**: a 7-digit value cannot be distinguished algorithmically between `D-MM-YYYY` and `DD-M-YYYY` without external knowledge. Additionally `station_table.start_date` is typed `int64` while `marine_table['Start Date']` is `float64` (carrying `.0` suffixes and NaNs), so the two require different parsing paths.

Only the trailing four digits (the year) should be treated as reliable in v0. Sub-annual analysis must wait for the ISO-8601 migration scheduled for v1 (§9.1).

`main_table.publication_year` is a clean 4-digit integer and is unaffected.

S7. Coordinate encoding

Coordinates appear in two parallel forms:

- `latitude_original` / `longitude_original` — as printed in the source, in whatever notation it used (DMS, degrees-and-minutes, verbal).
- `lat_decimal` / `lon_decimal` — decimal degrees, positive north and east, negative south and west.

Retaining both allows the conversion to be audited and re-derived if a transcription error is found. In v0, 1,980 of 1,984 station rows fall within the Southeast Asian domain (approximately 10°S–27°N, 93°E–145°E); the four outliers are records in adjacent territories (Australia, China, United States, Federated States of Micronesia) and are correct, not errors.

All v0 coordinates carry `coordinate_reference = Inferred`, which should be read as an unset default rather than a determination.

S8. Versioning and citation

The inventory is released as versioned, machine-readable CSV under a persistent DOI. Every row records the release in which it was issued (`inventory_version`) and its provenance timestamps (`created_date`, `last_updated`). Because identifiers are persistent and never reused, a `source_id` cited from v0 remains resolvable in all later releases, whether or not the record's content has since been corrected.

Dataset citation:

> *Teleti, P. and Williamson, F.: SEA-HISTIN: an inventory of historical weather observations of Southeast Asia, Zenodo [data set], https://doi.org/10.5281/zenodo.20663521, 2026.*

Citing an individual source within the inventory: use the document's own `citation_recommendation` field, followed by the inventory citation and the `source_id`.

S9

Table	Keys	Descriptive	Controlled vocab	Free text	Total
main	2	8	13	4	27
station	2	16	15	5	38
marine	2	5	0	9	16
instrument	3	11	3	2	19

S10

Retrieving every terrestrial station catalogued in one document, with its instrument history:

```
import pandas as pd

main = pd.read_csv("main_table_data_v0.csv", encoding="latin-1")
stn  = pd.read_csv("station_table_data_v0.csv", encoding="latin-1")
ins   = pd.read_csv("platform_instrument_table_v0.csv", encoding="latin-1")

doc = main[main.source_id == "B2iRvJNt"]                    # one document
sts = stn[stn.source_id == "B2iRvJNt"]                      # its stations
kit = ins[ins.station_id.isin(sts.station_id)]              # their instruments

# year is the only reliable component of a v0 date (§6)
sts["year"] = sts.start_date.astype(str).str[-4:].astype(int)
```