

# Astronomical Alias Lookup Tool

Design & Operations Document — alias\_lookup\_tool.py

*Pipeline Stage 02 — Alias Lookup | July 2026*

## 1. Purpose

The Alias Lookup Tool is stage 02 of the constellation guide pipeline. Astronomical objects accumulate identifiers from many catalogs — a single globular cluster may be known by NGC, Messier, Melotte, GCl, ESO, and survey designations simultaneously. Observation-planning software and star charts differ in which designation they index, so a complete cross-reference list per object makes every target findable in any tool.

The tool takes an observation spreadsheet exported from the cleanup stage (01), merges researched alias identifiers into each object's Name field, and produces a standardized output workbook ("02-Alias Lookup.xlsx"). All merging, normalisation, deduplication, and catalog-order sorting is deterministic and repeatable; the research step (gathering raw aliases from authoritative websites) is performed separately and supplied to the script as input data.

## 2. What the Script Does

Processing proceeds in eight sequential steps for each matched row:

| Step | Action      | Description                                                                                                                                                                                          |
|------|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Read        | Opens the input .xlsx and locates the ID, Type, and Name columns by header name (not position).                                                                                                      |
| 2    | Match       | Finds rows whose ID matches a key in the TARGETS dictionary (case-insensitive, punctuation-insensitive).                                                                                             |
| 3    | Split       | Parses existing Name content on semicolons and separator commas; commas inside brackets or parentheses are protected (e.g. "[66, 67, 68, 70, 72] Oph" stays intact).                                 |
| 4    | Merge       | Combines existing aliases with newly researched aliases from TARGETS.                                                                                                                                |
| 5    | Format      | Inserts a dash after M / NGC / SAO prefixes ("NGC 6284" and "NGC6284" both become "NGC-6284"). All other identifiers are left exactly as found.                                                      |
| 6    | Deduplicate | Removes duplicates using full normalisation (all non-alphanumerics stripped, lowercased) so "GCl 49" = "GCl-49" and "CD-24 13255" = "CD-24°13255". Also removes any alias identical to the row's ID. |
| 7    | Sort        | Orders aliases by the 20-level catalog priority (Section 6).                                                                                                                                         |
| 8    | Write       | Saves the output workbook, preserving all other columns, formatting, and unmatched rows untouched.                                                                                                   |

## 3. Expected Inputs

### 3.1 Input Spreadsheet

An .xlsx workbook whose first row contains column headers. Three columns are required; all others pass through unchanged.

| Column | Role                                                                              | Examples                                                   |
|--------|-----------------------------------------------------------------------------------|------------------------------------------------------------|
| ID     | Object identifier used for row matching                                           | "NGC 6284", "Messier 10", "Barnard 59", "67 Ophiuchi (AB)" |
| Type   | Object classification used for disambiguation during research                     | GC, OC, PN, DN, DS, MS-n, CS, AS, Star                     |
| Name   | Alias field to be updated; may already hold semicolon- or comma-delimited content | "M-10; NGC-6254; GCI-49"                                   |

### 3.2 TARGETS Dictionary

A Python dict at the bottom of the script maps each object ID to its researched alias list. Keys must match the ID column value (comparison is case- and punctuation-insensitive, but component-suffix IDs such as "36 Ophiuchi (AB)" are distinct keys and never match the bare parent ID). An empty list [] instructs the script to normalise and re-sort the row's existing content without adding anything new.

### 3.3 Configuration Constants

INPUT\_FILE (source workbook path), OUTPUT\_FILE (fixed as "02-Alias Lookup.xlsx" by pipeline convention), and COL\_ID / COL\_TYPE / COL\_NAME (header names, editable if a future export renames columns).

## 4. Values Calculated

The script computes no astronomical quantities for the spreadsheet itself; its calculations serve data verification:

- **Alias string** — The merged, normalised, deduplicated, catalog-ordered alias string written to each Name cell.
- **Coordinate separation** — Angular separation between two coordinate pairs, computed on-sky:  $\Delta RA(\text{arcmin}) = |\Delta RA^\circ| \times \cos(\text{Dec}) \times 60$  and  $\Delta \text{Dec}(\text{arcmin}) = |\Delta \text{Dec}^\circ| \times 60$ . The  $\cos(\text{Dec})$  factor converts a raw RA difference into true arc distance.
- **Epoch precession** — Coordinates reported in epochs other than J2000 are precessed to J2000 before comparison, using astropy (FK5 frames) when installed or a built-in simplified IAU-1976 formula (~1 arcsecond accuracy within 100 years of J2000) otherwise. RA/Dec inputs are accepted as decimal degrees or sexagesimal strings ("17 04 28.7").

## 5. Website Resources Utilized

| # | Source                       | Domain                | Where Aliases Are Read                                   |
|---|------------------------------|-----------------------|----------------------------------------------------------|
| 1 | Wikipedia                    | en.wikipedia.org      | Infobox → "Other designations"                           |
| 2 | Stelle Doppie                | stelledoppie.it       | "Other Catalogs and Designations" section (double stars) |
| 3 | NASA/IPAC NED                | ned.ipac.caltech.edu  | Overview tab → "Cross-identifications"                   |
| 4 | SEDs Interactive NGC Catalog | spider.seds.org/ngc   | "Cross Identifications" field                            |
| 5 | VizieR                       | vizie.cfa.harvard.edu | Catalog cross-match queries                              |

| # | Source | Domain              | Where Aliases Are Read               |
|---|--------|---------------------|--------------------------------------|
| 6 | SIMBAD | simbad.u-strasbg.fr | Basic search → "Identifiers" section |

Practical access note: SIMBAD and NED disallow automated fetching (robots.txt). During assisted research sessions, Wikidata is used as a reliable proxy — its identifier lists mirror SIMBAD's — cross-checked against Wikipedia infoboxes, messier.seds.org cross-reference pages, and theskylive.com. Manual browser research against the primary sources remains fully supported.

## 6. Data Integrity Validation

Because catalog identifiers are easily confused between neighbouring objects, aliases from a candidate source are accepted only when the source passes at least ONE of two verification rules against an already-verified source:

### Rule 1 — Two-Alias Match

At least two aliases from the candidate source also appear in a verified source. Comparison strips all non-alphanumeric characters and lowercases, so "NGC-6284", "ngc 6284", and "NGC6284" all match, and punctuation variants such as "CD-24°13255" versus "CD-24 13255" cannot cause a false mismatch.

### Rule 2 — One-Alias Match + Coordinate Agreement

Exactly one alias matches (same normalisation), AND the two sources' positions agree within 5 arcminutes in both RA and Dec. The RA tolerance is evaluated on-sky (cos-Dec corrected), and both positions are precessed to J2000 first when epochs differ — a B1950 position differs from J2000 by roughly 8 arcminutes of RA near the celestial equator, enough to falsely fail the check if uncorrected.

The script exposes `verify_by_alias_match()` and `verify_by_coords()` so these rules can be applied programmatically during research. Additional integrity safeguards built into the merge step:

- **Header lookup** — Columns are located by header name, never by position, so column reordering cannot corrupt output.
- **Non-destructive** — The input file is copied first and modified in the copy; the source workbook is never altered.
- **Unmatched rows** — Rows with no TARGETS entry are left completely untouched.
- **Fill-and-preserve** — Existing cell content is merged, never overwritten — no researched value silently replaces prior data.

## 7. Logic for Using Source Data

- **Bootstrap** — An object's research begins from its own known identifiers (the spreadsheet ID and any existing Name content) as the initial verified set.
- **Chain of trust** — Each website is checked in turn; a source that passes Rule 1 or Rule 2 has ALL of its listed aliases accepted, joining the verified pool for testing subsequent sources.
- **Majority formatting** — When the same catalog number appears with different formatting across sources, the version used by the majority of sources is kept.
- **All-or-nothing** — A source that fails both rules is discarded entirely for that object — no partial acceptance.
- **Deterministic merge** — Aliases are then merged through the deterministic pipeline of Section 2 (format → dedupe → sort), producing one semicolon-delimited string per object.

## Catalog Sort Order (20 levels)

| Priority 1–5 | 6–10         | 11–15         | 16–20                    |
|--------------|--------------|---------------|--------------------------|
| 1. SAO       | 6. HR        | 11. IC        | 16. Sh                   |
| 2. HIP       | 7. ADS       | 12. Barnard   | 17. Melotte              |
| 3. Bayer     | 8. Messier   | 13. LDN       | 18. GCl                  |
| 4. Flamsteed | 9. NGC       | 14. Abell     | 19. Other (alphabetical) |
| 5. HD        | 10. Caldwell | 15. Collinder | 20. Proper names (last)  |

## 8. Outputs

- **Primary output** — "02-Alias Lookup.xlsx" — the updated workbook; every matched row's Name cell holds the final semicolon-delimited alias string; all other cells and rows are byte-identical to the input.
- **Run log** — A console log listing each updated row by ID and a final count, providing an auditable record of exactly which rows were touched.

Reference regression values from the Ophiuchus run (57 rows updated) are documented in the companion guide (alias\_lookup\_reproduction\_guide.md); for example, Messier 10 → "M-10; NGC-6254; C 1654-040; Mel 157; GCl-49; [KPS2012] MWSC 2498".