From 46 Heterogeneous Star Catalogs to One Unified Age Database: An AI-Assisted Integration Workflow

Yi Yang [1]

Abstract

We present a case study of AI-assisted integration of 46 published asteroseismology-related age catalogs into a unified database of 3.85 million stellar records (3.38 million with age estimates). The workflow, built with Claude Code and a few research skills, spans five stages from literature discovery to unified database compilation, handling heterogeneous data sources, inconsistent column naming, and diverse age estimation methods. Verification caught key failure modes including hallucinated column mappings, misclassified derivation methods, and undetected bibcode mismatches. The workflow substantially reduced manual effort in literature discovery, multi-source downloading, and method verification, while catching errors that would have propagated into downstream analyses.

1. Introduction

Asteroseismology—the study of stellar oscillations—provides the most direct observational constraints on stellar interior physics. Global oscillation parameters $\nu_{\max}$ (frequency of maximum power) and $\Delta\nu$ (large frequency separation) measured by the CoRoT, Kepler, K2, and TESS missions encode fundamental stellar properties including mass, radius, and crucially, age (Chaplin & Miglio, 2013; Hekker & Christensen-Dalsgaard, 2017).

However, published asteroseismology-related ages are scattered across dozens of catalogs spanning over a decade of literature, multiple missions, and diverse age estimation methods. These catalogs use inconsistent column naming conventions, mixed units, and are hosted across heterogeneous data repositories. Synthesizing these into a unified database for meta-analysis or training-set construction has traditionally required extensive manual effort.

This paper documents how we used AI assistance to build such a unified database. The focus is on the *research workflow*—how AI tools accelerated literature discovery, data download orchestration, format standardization, method verification, and unified database compilation—rather than the resulting catalog itself.

2. The Cognitive Task

The core task: *identify published asteroseismology-related age catalogs, download them from their respective repositories, standardize columns to a common schema, verify the age derivation methodology, and produce a unified database with provenance tracking.*

This requires navigating four interacting challenges: (1) **literature discovery** across telescope archives (CoRoT, Kepler, K2, TESS), repositories (VizieR, Zenodo, arXiv), and methodological keywords; (2) **multi-source download** with six distinct authentication and format patterns (VizieR/astroquery, Zenodo/REST, arXiv/HTML-parsing, IOP/direct-HTTP, China-VO/FITS, SDSS/proprietary); (3) **column standardization** across $\sim$100 unique identifiers ("age" alone appears as `Age`, `AgeRGB`, `lnAge`, `logAge`, `Final_age`, etc.) with mixed units (Gyr, Myr, log-space); (4) **method classification** requiring original paper reading—ages from grid modeling, ML trained on seismic labels, scaling relations, and isochrone fitting have fundamentally different accuracy properties.

3. AI-Assisted Workflow

3.1. Stage 1: Literature Discovery and Catalog Indexing

AI-assisted approach. A four-layer discovery strategy: (1) Claude Code generated ADQL queries searching VizieR TAP for 192 tables with $\nu_{\max}/\Delta\nu$ columns intersected with 2,257 age-containing tables, yielding 29 candidate tables—one new observational catalog (Fritzewski et al. (2024)) added, three theoretical model grids excluded—and filtering false-positive columns (`stage`, `wage`); (2) `/ads-cli` traversed citation networks and co-author lists of major catalogs (APOKASC-3 (Pinsonneault et al., 2025), K2

[1]Beijing Planetarium, Beijing, China. Correspondence to: Yi Yang <mgcyung@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

GAP DR3 (Zinn et al., 2022), Miglio et al. (2021)), recovering Liagre et al. (2025), Montalbán et al. (2021), and Warfield et al. (2021); (3) literature mining over arXiv, the SDSS Value-Added Catalog (AstroNN DR17; Leung & Bovy (2019)), and VizieR title searches surfaced Vitali et al. (2024) and Lund et al. (2025); (4) iterative gap detection and subagent recommendations added Theodoridis et al. (2026) and Pakštienė et al. (2026).

Result. 46 catalogs identified. Discovery began with an AI-generated seed list of 31 candidates—24 confirmed and 7 rejected (including 3 bibcode hallucinations caught by VizieR/ADS re-resolution)—then added 22 more across systematic VizieR TAP search, citation-network traversal, and literature mining. 13 additional candidates were identified but unavailable (Zenodo pending, data loss, paywalls).

3.2. Stage 2: Multi-Source Download Pipeline

AI-assisted approach. Claude Code generated per-source download scripts with error handling for six repository types. Example VizieR pattern:

```
v = Vizier(row_limit=-1, columns=['**'])
cats = v.get_catalogs('J/ApJS/276/69')
df = cats[0].to_pandas()
```

The AI handled source-specific logic: VizieR table index resolution (some IDs return multiple sub-tables), Zenodo DOI-to-file resolution, arXiv ancillary HTML parsing, IOP MRT direct download, China-VO binary FITS, and SDSS SAS authentication. One authentication failure was caught and corrected (see Section 4).

Result. 46/46 catalogs downloaded (100%), ∼3.85M records (∼3.38M with age estimates from 29 catalogs; ∼0.47M with seismic parameters only from 17).

3.3. Stage 3: Column Standardization

Cognitive challenge. The 46 catalogs used ∼100 unique column names for 12 physical quantities. "Age" alone appeared as `Age`, `AgeRGB`, `lnAge`, `logAge`, `Final_age`, etc., stored in Gyr, Myr, $\log_{10}$(yr), or ln(yr).

AI-assisted approach. For each catalog, Claude Code read the raw column headers and autonomously generated a mapping dictionary aligning them to the 14 standard fields, e.g.:

```
"Map source columns to standard format:
- KIC -> star_id, _RA -> ra, _DE -> dec
- Teff -> teff, Numax -> numax, DNu -> dnu
- AgeRGB -> age, e_AgeRGB -> e_age
Mission: Kepler, Source: apokasc3"
```

Claude generated mapping dictionaries for 46 catalogs from column headers alone—the researcher's role was verification, not specification. Three catalogs stored ages in log/ln space requiring non-trivial conversion: ln(age)→ $e^{\ln(\text{age})}$ for Ness et al. (2016), $\log_{10}$(age,Myr)→ $10^{\log_{10}(\text{age})}/1000$ for Lu et al. (2022). The AI caught these log-space errors during downstream processing; the researcher verified the corrections against original papers (see Section 4).

Table 1. Age estimation methods across 3.38M age records (not deduplicated across catalogs).

Method	N	%
ML (seismic-trained)	3,135,700	92.7
MESA grid	134,709	4.0
ML (mass-to-age)	73,180	2.2
Grid modeling + seismic	26,426	0.8
Pure isochrone fitting	7,993	0.2
Hybrid (scale + isochrone)	3,726	0.1
g-mode pulsator modeling	609	<0.1

Result. 46 catalogs standardized to 14 columns.

3.4. Stage 4: Method Classification Through Paper Reading

Cognitive challenge. Ages from grid modeling with seismic constraints (precision ∼15–30%) are not equivalent to ages from neural networks trained on seismic labels (∼30–50%). Classifying each catalog's method requires reading original papers, not just parsing columns.

AI-assisted approach. For the 29 catalogs with age estimates, Claude Code read arXiv Methods sections, e.g.:

```
"Read Methods of arXiv:2004.14806 (Miglio
    +2021).
How exactly do they derive stellar ages?
    Grid
modeling (PARAM/BASTA/BeSPP) or scaling
    relations?
What model grid and input observables?"
```

This revealed three critical corrections: (1) 12 catalogs classified as "scaling relation" actually use grid modeling (BASTA, PARAM, BeSPP, YREC, AIMS, MESA+GYRE); (2) ML method errors: Wang et al. (2023) uses Random Forest (not XGBoost), Wu et al. (2019) uses KPCA (not CatBoost); (3) several bibcodes did not resolve to the intended paper. The initial over-reliance on column metadata alone produced 12/29 wrong classifications (detailed in Section 4).

Result. 7-category classification (Table 1). 92.7% of ages come from ML trained on seismic labels; only 0.8% from the most reliable method (grid modeling + individual seismic constraints).

3.5. Stage 5: Unified Database and Analysis

Claude Code generated summary compilation scripts aggregating statistics across all mission directories into the unified

database (3.85M records, 3.38M with ages). The method breakdown (Table 1) reveals a dominance of ML-derived ages—critical context for any meta-analysis that aggregates ages without accounting for method-dependent systematics.

4. Failure Modes and Lessons Learned

4.1. Column Mapping Hallucination

Claude generated plausible but incorrect column-to-standard-field associations. In three catalogs, the AI initially misclassified log-space ages as linear, but caught the errors itself during downstream processing and wrote the correct unit conversions. The researcher verified the fixes against original papers—a necessary step, since unit annotations can be inconsistent across sources: e.g., Lu et al. (2022)'s `logAge` column is labeled "Gyr" in the VizieR ReadMe but "Myr" in the arXiv text, and only the Myr interpretation yields physical ages. The remaining 43/46 mappings were correct as AI-generated—column-name semantics alone sufficed, with researcher verification needed only where unit conventions are absent from the headers.

4.2. Bibcode Hallucination

The AI-generated candidate list carried bibcodes that failed to resolve: the "Bellinger+2019" entry had VizieR ID `J/ApJ/878/45` (no such table exists) and ADS bibcode `2019ApJ...878..136Z` (actually Zinn+2019). **Fix:** ADS + VizieR TAP re-resolution flagged 4 non-resolving IDs, excluded from the final 46.

4.3. Method Classification from Surface Features

Initial classification used column metadata alone ("has $\nu_{\max}$" $\rightarrow$ "scaling relation"); 12/29 wrong. **Fix:** Added automated paper-reading verification, reading Methods sections of 16 papers.

4.4. VizieR Authentication Failure

Direct HTTP to CDS VOTable endpoint returned 401; Claude used `requests.get()` which failed silently. **Fix:** Switched to astroquery's VizieR module; lesson: specialized repositories require specialized client libraries.

4.5. Summary of Lessons

1. **Verification precedes execution.** Define physical range checks, bibcode verification, and method verification before writing any download code.

2. **AI excels at semantic column mapping.** AI-generated mappings were correct for 43/46 catalogs from headers alone; researcher intervention was needed only for log-space unit conventions and a few field-level misalignments.

3. **Bibcode integrity matters.** 4 AI-generated candidate bibcodes failed to resolve and were excluded; automated ADS/VizieR verification saved potentially weeks of chasing wrong papers.

4. **Per-catalog unit pattern.** Each catalog treated as a self-contained unit kept the workflow manageable as the count grew from the 31-candidate seed list to 46 confirmed catalogs.

5. Discussion

Several factors made this project a good fit for AI assistance. The task involved high mechanical overhead—46 catalogs, each needing download, parsing, and standardization—but low creative demand per unit. Correctness criteria were well-defined (physical parameter ranges, unit consistency), and the data ecosystem is largely machine-readable (VizieR, ADS, arXiv, Zenodo all provide APIs).

The largest time savings came from literature discovery and download. Discovery required cross-matching 192 VizieR tables containing seismic parameters against 2,257 age-containing tables, then traversing citation networks and mining arXiv/SDSS—a search space that a researcher would navigate manually over weeks of ADS queries, VizieR browsing, and reference-list chasing. Download involved 3.85M records from six repository types (VizieR, Zenodo, arXiv ancillary, IOP, China-VO, SDSS SAS), each with distinct APIs, authentication, and format-parsing requirements; writing and debugging per-source scripts manually is a multi-day effort that AI generated in hours. The most *valuable* saving, however, was in method verification: AI-assisted paper reading caught 12 misclassifications that would have propagated systematic errors into downstream analyses.

The resulting database—46 catalogs, 3.85M records, 3.38M ages in 7 method categories—enables previously infeasible cross-catalog analyses.

Limitations. This is a single-project case study. The original goal targeted ages related to asteroseismology, but the scope expanded during discovery to include catalogs whose age methods are not seismically informed. The 3.38M age records are not deduplicated across catalogs; the same star may appear in multiple catalogs with different method-dependent ages, so the unique-star count is lower. AI aligned column-name semantics reliably, but unit conventions (Gyr vs Myr, linear vs log/ln space) are not encoded in headers and required researcher verification against original papers. Domain expertise was rate-limiting at every stage.

6. Conclusion

We have presented an AI-assisted workflow that integrated 46 asteroseismology-related age catalogs into a unified database (3.85M records, 3.38M ages), substantially reducing manual effort while improving quality: 12 method misclassifications corrected, 4 erroneous bibcodes caught, 3 unit conversion errors identified before propagation.

The key insight: AI assistance transforms catalog integration from an insurmountable manual burden into a feasible project. With 46 catalogs, the traditional one-by-one approach is psychologically overwhelming; researchers either skip it or invest months. AI assistance makes systematic, comprehensive integration the default.

As astronomical surveys expand (Gaia DR4, LSST, Roman), the literature will contain exponentially more specialized catalogs requiring aggregation. The workflow patterns—systematic literature discovery, multi-source orchestration, column standardization with physical sanity checks, method verification via paper reading—are domain-agnostic and transferable.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 32300381, No. 12303035 and No. 12503027), the Young Data Scientist Program of the China National Astronomical Data Center (No. NADC2025YDS-02) and the Financial Program of BJAST (26CA012). This work used data from CoRoT, Kepler, K2, TESS, GALAH, APOGEE, Gaia, CDS VizieR, and ADS. All data sources are publicly available.

References

Chaplin, W. J. and Miglio, A. Asteroseismology of Solar-Type and Red-Giant stars. *Annual Review of Astronomy and Astrophysics*, 51:353–392, 2013.

Fritzewski, D. et al. Age uncertainties of red giants due to cumulative rotational mixing of progenitors calibrated by asteroseismology. *Astronomy & Astrophysics*, 684:A112, 2024.

Hekker, S. and Christensen-Dalsgaard, J. Giant star seismology. *The Astronomy and Astrophysics Review*, 25:1, 2017.

Leung, H. W. and Bovy, J. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, 483:3255–3277, 2019.

Liagre, P. et al. Beyond the Nyquist frequency: Asteroseismic catalog of undersampled Kepler late subgiants and early red giants. *Astronomy & Astrophysics*, 702:A144, 2025.

Lu, Y. et al. Similarities behind the high- and low-α disc: small intrinsic abundance scatter and migrating stars. *Monthly Notices of the Royal Astronomical Society*, 512: 2890–2910, 2022.

Lund, M. N. et al. Luminaries in the sky: The TESS legacy sample of bright stars: I. Asteroseismic detections in naked-eye main-sequence and subgiant solar-like oscillators. *Astronomy & Astrophysics*, 701:A285, 2025.

Miglio, A. et al. Age dissection of the Milky Way discs: Red giants in the Kepler field. *Astronomy & Astrophysics*, 645:A85, 2021.

Montalbán, J. et al. Chronologically dating the early assembly of the Milky Way. *Nature Astronomy*, 5:640–647, 2021.

Ness, M. et al. Spectroscopic determination of masses (and implied ages) for red giants. *The Astrophysical Journal*, 823:114, 2016.

Pakštienė, E. et al. Calibration of the [C/N] and [Y/Mg] chemical clocks with asteroseismic ages from the TESS space mission. *Astronomy & Astrophysics*, 708:A250, 2026.

Pinsonneault, M. et al. APOKASC-3: The third joint spectroscopic and asteroseismic catalog for evolved stars in the Kepler fields. *The Astrophysical Journal Supplement Series*, 276:69, 2025.

Theodoridis, A. et al. Asteroseismically inferred ages of 132,000 red giants with TESS. *The Astrophysical Journal Supplement Series*, 283:4, 2026.

Vitali, S. et al. Exploring the dependence of chemical traits on metallicity. chemical trends for red giant stars with asteroseismic ages. *Astronomy & Astrophysics*, 687:A164, 2024.

Wang, C. et al. Precise masses and ages of 1 million RGB and RC stars observed by LAMOST. *Astronomy & Astrophysics*, 675:A26, 2023.

Warfield, J. T. et al. An intermediate-age alpha-rich galactic population in K2. *The Astronomical Journal*, 161:100, 2021.

Wu, Y. et al. Ages and masses of 0.64 million red giant branch stars from the LAMOST galactic spectroscopic survey. *Monthly Notices of the Royal Astronomical Society*, 484:5315–5329, 2019.

Zinn, J. C. et al. The K2 galactic archaeology program data release 3: Age-abundance patterns in C1-C8 and C10-C18. *The Astrophysical Journal*, 926:191, 2022.