The Variable Zoo Project (variablezoo.org)

The Variable Zoo Project: AI-Driven Variable Star Astronomy

John Rachlin · Khoury College of Computer Sciences, Northeastern University
j.rachlin@northeastern.edu · john@variablezoo.org · The Variable Zoo Project (variablezoo.org)

A vision and status white paper · 2026-07-24

Abstract. The Variable Zoo Project is an effort to build not merely a photometric reduction pipeline but a complete, AI-assisted *research* pipeline for variable-star astronomy — a closed loop from choosing what to observe, through acquiring and measuring images on a small telescope, to analysis and reporting. Beneath that loop sits the project's real bulk of engineering: an **observatory data-management platform** of some 24,819 lines of code spanning a 10.3-million-object reference catalog (derived from the **AAVSO VSX cross-matched with Gaia DR3**), an observations database of 528,753 of our own measurements, and the tooling around them — image/calibration cataloging, identification and classification, a management console, catalog search, the Variable Zoo Explorer, and a public education site. Throughout, artificial intelligence is a **tool and research assistant, not an author**: a human sets the questions, judges the science, and is accountable for anything shared. We are at the earliest stages of an R&D programme; the emphasis is on provenance, reproducibility, and credibility over volume.

1 Aims

Variable stars are among the oldest sustained observational programmes in stellar astrophysics, and one of the few in which modest instruments still make real contributions [1]. Even in the era of all-sky surveys such as ASAS-SN, ZTF and TESS [2–4], a dedicated small telescope can still supply the targeted, sustained cadence those surveys do not. Yet the path from a small telescope to a result worth sharing is long, fragmented, and mostly manual. The Variable Zoo Project asks a single question: **how far can that entire path be made systematic, reproducible, and AI-assisted — without sacrificing scientific rigour?**

Our aim is a **research pipeline, not just a measurement pipeline**. Measuring the brightness of a variable star from calibrated images is a solved-in-principle problem and only the middle of the story. The project treats the whole arc — deciding *which* stars are worth observing, planning the night, acquiring and cataloguing the data, measuring it, analysing the accumulated record, and reporting what is found — as one connected system, resting on a shared data-management foundation. Most data come from a single small telescope today; the architecture anticipates a future in which several independent observers contribute to a common, provenance-tracked record.

This work builds squarely on the **American Association of Variable Star Observers (AAVSO)**. For more than a century the AAVSO and its global network of observers have assembled the definitive record of variable-star behaviour, and its **International Variable Star Index (VSX)** remains the community's authoritative catalog of variable stars [5, 6]. The Variable Zoo's own reference catalog is derived directly from **VSX cross-matched with ESA's Gaia DR3** astrometry and photometry [7, 8]. We see the project as extending that venerable tradition with new tools and methods — not replacing it — and we intend any findings as contributions back to the community the AAVSO has long served.

2 The seven-stage research pipeline

We organise the work as seven stages forming a closed loop: the output of a research cycle — new light curves, refined ephemerides, flagged anomalies — feeds the next round of candidate selection (Fig. 1).

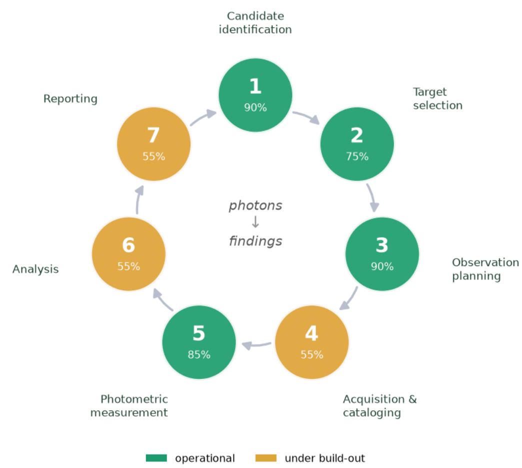

Figure 1. The Variable Zoo research loop. Photons enter at candidate identification and leave as reported findings; results close the loop by informing the next round of targets. Node colour marks how far each stage is built out today (see Fig. 3).

1. **Candidate identification.** Filter the sky by variability type, magnitude, position and season to find objects worth observing — from a reference catalog cross-matched to Gaia.
2. **Target selection.** Choose a specific object, usually a single variable placed at the centre of the field, weighing scientific interest against what the available instrument can reach.
3. **Observation planning.** Turn a target into an executable night: finder charts, exposures, cadence, camera angle, mosaic tiling, and observatory-ready import files.
4. **Data acquisition & image cataloging.** Acquire the frames at the telescope and index every image — lights and calibration — in place, with a persistent identity and provenance.
5. **Photometric measurement.** Apply the current reduction pipeline to the raw images and write calibrated, star-resolved measurements to the database.
6. **Analysis.** Fold, characterise, and derive — periods, amplitudes, ephemerides, statistics, and flagged anomalies — from the accumulated measurements.
7. **Reporting.** Communicate what is found: as PDF reports and interactive console dashboards, and, where a finding genuinely warrants it, as a human-authored, human-owned contribution to the community.

3 Observatory data management: the infrastructure layer

The seven stages are only as good as the data plumbing beneath them, and that is where most of the project actually lives. Of the project's **24,819 lines of Python** (≈30,246 counting the web front-ends), much is not photometry at all but **observatory data management**: the systems that give every object, image, and measurement a durable identity and provenance, and that make the whole record queryable, curatable, and publishable. This platform (Fig. 2) is what turns a growing pile of FITS files into a research programme, and it is the real engine driving the loop.

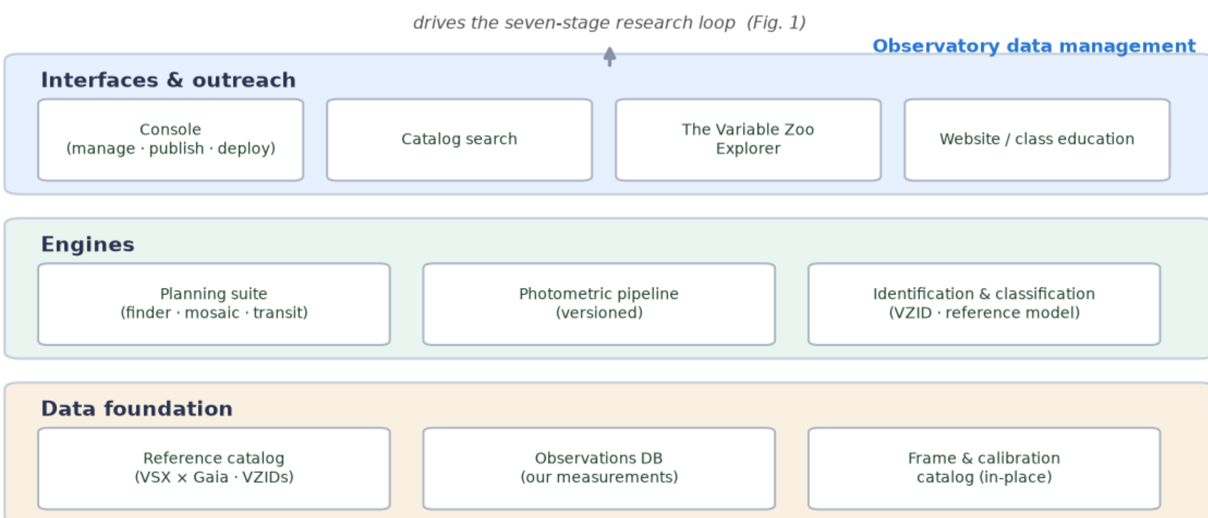

Figure 2. The Variable Zoo platform. A data foundation (catalog, measurements, and in-place frame/calibration cataloging) is acted on by engines (planning, the versioned photometric pipeline, and identification/classification) and surfaced through interfaces and outreach (management console, catalog search, sky explorer, and the public education site). Together they are the observatory data-management infrastructure that powers every stage of the research loop.

Concretely, the platform already includes: a rebuilt **reference catalog** of the AAVSO VSX cross-matched to Gaia DR3, with a permanent Variable Zoo identifier minted for every object; a separate **observations database** of our own measurements, deliberately kept apart from the rebuildable catalog; **in-place image and calibration-frame cataloging** with a managed library of calibration masters; our own initiatives in variable-star **identification and classification** (see below); a local **management console** that is the only tool allowed to curate and publish; a **catalog-search** application; the **Variable Zoo Explorer** (an interactive sky map); and a public **website** that doubles as an educational-outreach platform with per-class variable-star portals.

One early result deserves particular emphasis. Beyond cataloguing existing classifications, the project has developed **its own model for identifying and classifying variable stars**: a supplemental identifier scheme paired with a mechanism-based logical type system that separates a star's discrete physical *type* from its continuous observational *attributes*, collapsing a sprawl of historical labels into a compact, consistent hierarchy. It is deliberately **supplemental to — never authoritative over** — the established AAVSO VSX standards, and is set out in full in a companion technical report [9]. We regard it as already a substantive contribution to how the variable sky can be organised.

4 Current status

The project is young but already has a working spine. Two databases anchor it: the **reference catalog** of 10,304,607 variable and suspected-variable objects, and the **observations database** holding our own 528,753 photometric measurements. The reduction pipeline — built on the open scientific-Python ecosystem [10] — is at **version 1.3.0**, and every run is version-stamped for reproducibility.

Reference catalog (VSX × Gaia DR3)	**10,304,607**	objects, all with a permanent VZID
Our photometric measurements	**528,753**	across 9,508 stars, 99.2% Gaia-resolved
Observation datasets / targets	**37 / 39**	reduced sequences and planned/observed fields
Frames indexed in place	**5,482**	lights + calibration, with provenance
Catalogued documents	**133**	reports, plans, figures and notes
Observers / instruments	**2 / 3**	multi-observer, multi-telescope ready

Maturity is uneven by design (Fig. 3). The pipeline is strongest where it has been exercised most — **candidate identification**, **observation planning**, and **photometric measurement** are operational — while the back half of the loop (**analysis** and **reporting**) and unattended **data acquisition** in the field are the parts still under active build-out.

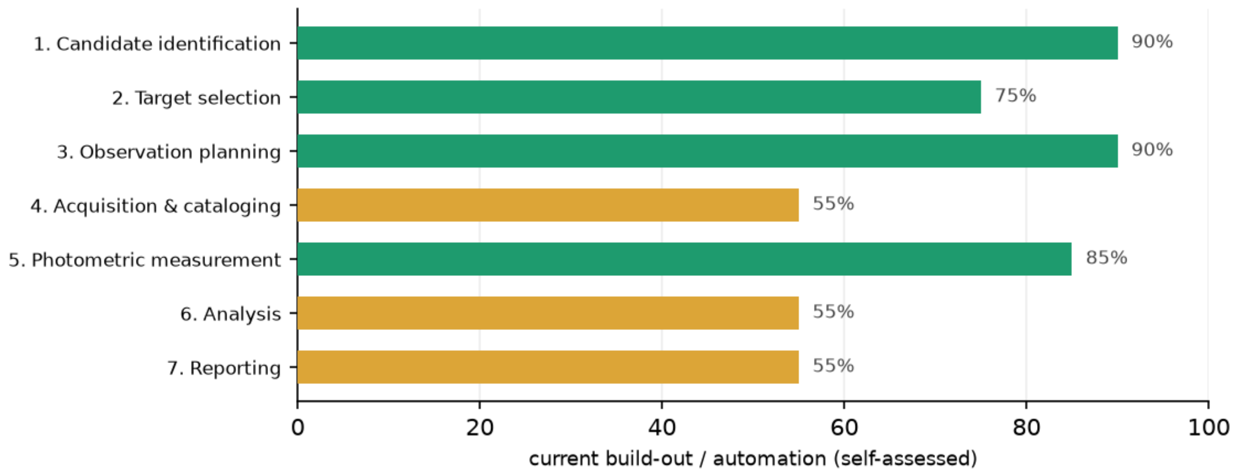

Figure 3. Self-assessed build-out of each pipeline stage. Teal marks stages in routine use; amber marks stages still under active development.

Measurements have accrued from roughly Aug 2025 to Jul 2026 (Fig. 4), almost entirely from a single 60 mm refractor — a demonstration that a disciplined, well-instrumented small telescope produces a research-grade data stream. Around this spine sit the working tools already surveyed in Section 3.

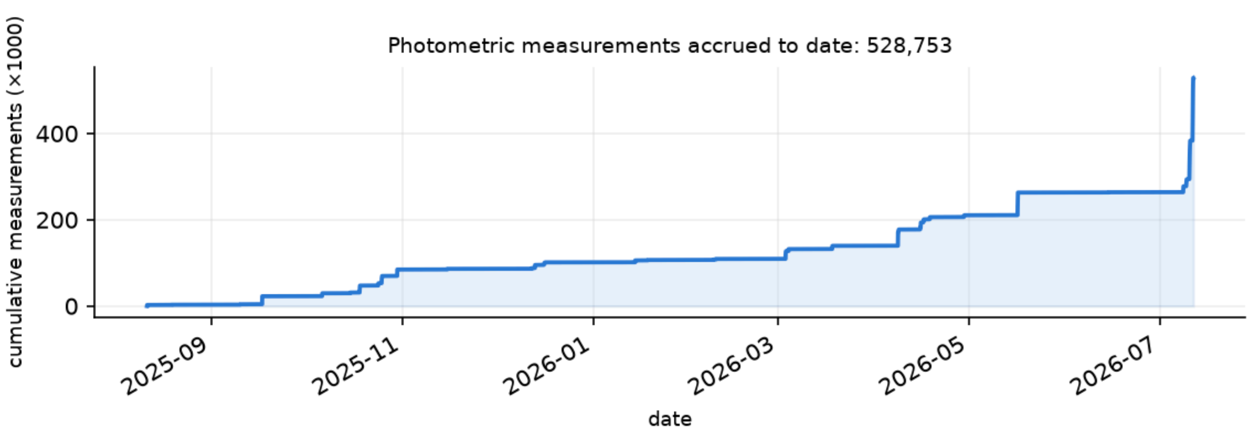

Figure 4. Cumulative photometric measurements to date, almost all from a single small telescope. The record is the raw material of every downstream research stage.

5 The role of AI — and of the human

Machine learning and AI are increasingly part of the astronomer's toolkit [11]. In this project, artificial intelligence is not a bolt-on analysis step; it is the connective tissue of the whole loop. Much of the Variable Zoo — its tools, its data plumbing, and documents like this one — is built and assembled by an AI agent working against the project's own databases. But the right word for that agent is **research assistant**, not author. The division of labour is deliberate and, we think, the whole point:

- **What the AI does.** The tireless, reproducible work that humans find tedious and error-prone — cross-matching catalogs, running and version-stamping the reduction, tracking provenance, drafting figures, tables and summaries, and keeping thousands of records consistent.
- **What the human does.** The judgement that must stay human — deciding which questions are worth asking, whether a signal is real, what a result *means*, how it relates to the community's existing knowledge, and, above all, taking personal responsibility for anything published or claimed.

A word on the final stage, which we have deliberately named **Reporting** rather than "paper generation." Automating a research pipeline's last step raises a fair concern: that it will add to the flood of machine-written, low-quality manuscripts now straining editors and feeding "paper mills." **That is emphatically not the intent.** The system's job is to make careful work *easier*, not to mass-produce submissions. Volume is explicitly a non-goal. No result is submitted anywhere on the strength of automation alone; a human scientist authors, vets, and owns any external contribution, and the project would rather produce one fully provenanced, defensible result than a hundred unvetted ones. Credibility is treated as the scarce resource.

Two engineering commitments keep that stance honest: **provenance and reproducibility** — permanent object identifiers, version-stamped pipeline runs, in-place frame cataloging, and a strict separation between the rebuildable reference catalog and our own measurements; and **transparency** — generated documents that carry measurements or scientific claims are marked as AI-generated and not peer-reviewed, and the underlying data and code remain inspectable.

6 Long-term direction

The direction of travel is to **close and then accelerate the loop**:

- **From measurement to discovery.** Push the two weak ends of the pipeline — turning the catalog and the accumulating light curves into a ranked, science-driven stream of candidates, and turning results into publishable findings (novel classifications, refined periods and ephemerides, anomaly and outburst alerts).
- **Science-grade calibration.** Carry the photometry from differential light curves to standardised, comparable magnitudes so that results are interoperable with the wider community's databases.
- **A multi-observer network.** Extend the provenance model — already multi-observer and multi-instrument at the schema level — so that several independent telescopes can contribute to one coordinated record, sharing targets and cross-checking each other's photometry.
- **Increasing autonomy with human judgement.** Let the AI agent take on more of each stage end-to-end, always with a scientist in the loop for the questions that matter — what is worth observing, what a result means, and what is fit to publish.
- **Open, communicated science.** Produce concrete outputs within defined scopes and share them respectfully with the established community, above all the AAVSO, whose long record this project builds upon rather than supplants.

> **The one-sentence vision — *photons to findings*.** A single scientist, working with an AI collaborator and a small telescope — and eventually a network of them — should be able to run the full arc of variable-star research, from choosing a target to a carefully vetted, communicated result, as one reproducible, provenance-tracked system.

7 Outlook

None of this replaces the astronomer or the century of careful visual and photometric work the field rests on. The wager of the Variable Zoo Project is narrower and, we think, timely: that the *research process* itself — not just the measurement — can be made systematic and AI-driven, and that doing so lets a small observatory contribute more, and more reliably, than its aperture alone would suggest. The stages ahead are the hard ones. This paper is the map we intend to work from.

References

1. Percy, J. R. 2007, *Understanding Variable Stars* (Cambridge Univ. Press).
2. Kochanek, C. S. et al. 2017, *PASP*, 129, 104502 — the All-Sky Automated Survey for Supernovae (ASAS-SN).
3. Bellm, E. C. et al. 2019, *PASP*, 131, 018002 — the Zwicky Transient Facility (ZTF).
4. Ricker, G. R. et al. 2015, *JATIS*, 1, 014003 — the Transiting Exoplanet Survey Satellite (TESS).
5. Watson, C. L., Henden, A. A. & Price, A. 2006, *Society for Astronomical Sciences 25th Annual Symposium*, 25, 47 — the International Variable Star Index (VSX).
6. Kafka, S. 2021, *Observations from the AAVSO International Database*, AAVSO (aavso.org) — the century-scale record of variable-star photometry.
7. Gaia Collaboration (Vallenari, A. et al.) 2023, *A&A*, 674, A1 — Gaia Data Release 3.
8. Gaia Collaboration (Eyer, L. et al.) 2023, *A&A*, 674, A13 — Gaia DR3: variability processing and catalogues.
9. Rachlin, J. 2026, *An Identification and Classification Model for Variable Stars* (The Variable Zoo Reference Model), The Variable Zoo Project (Unpublished).
10. Astropy Collaboration 2022, *ApJ*, 935, 167 — the Astropy Project (v5.0).
11. Baron, D. 2019, *arXiv*:1904.07248 — Machine Learning in Astronomy: a practical overview.

Prepared by The Variable Zoo Project · variablezoo.org · database and line-of-code counts read live at build time (2026-07-24). Figures 1 and 3 are editorial self-assessments; Figure 2 is a schematic of the platform; Figure 4 uses observation-database records. This white paper was written in collaboration with Claude Code (Anthropic).