Ariadne

A Navigable Etymological Thread, and an Honest Account of Its Difficulties

Alejandro Toledo Martínez

Abstract

Ariadne is an open-source tool that answers a deceptively simple question: given two words, what is the thread that connects them through languages and time? It builds a single, distributable graph of etymological relations — inheritance, derivation, borrowing, calque, cognacy — from Wiktionary and a small set of scholarly resources, and it lets a user pull the *thread* between any two forms: `pico` (Spanish) up to `bekkos` (Gaulish), `ojo` (Spanish) across to `eye` (English) at the `Proto-Indo-European root h ek -`. This paper describes what Ariadne is, what it is for, and — most importantly — the difficulties we met, because those difficulties are the real content. Etymology at scale is not a clean database problem: proto- language reconstructions are neither unique nor agreed upon, dictionaries encode ancestry inconsistently, homographs collapse distinct histories, and coverage is deeply uneven. We argue that a useful etymological tool must therefore be *honest*: it must distinguish what it knows from what it infers, label the seams rather than hide them, and treat the boundary between genuine structure and coincidence as a first-class engineering concern. We report two structural-coverage metrics — C12 (lemmas with no captured ancestry) and C13 (fraction of a language that reaches a Proto-Indo-European root) — that make these gaps measurable rather than assumed. Evaluated against an *independent* cognacy standard (IE-CoR), Ariadne recovers a thread for 81% of expert-attested cognate pairs (a coverage figure) while connecting 0% of unrelated pairs across every random seed — the inferred bridge/variant layer is specific, not apophenic. We release the tool (686k nodes, 803k typed provenance-bearing edges over 1,720 languages), the reproducible build, and the registry — openly at github.com/toledoal/ariadne — and we invite others to build more of this kind. The wider claim is not about one program: large-scale etymological navigation is feasible only when uncertainty is represented explicitly rather than hidden — Ariadne is the experiment that makes that lesson concrete.

1. Motivation: the thread, not the tree

Historical linguistics has spent two centuries building *trees*: family trees, sound laws, reconstructed proto-forms. The results are extraordinary. But a learner, a comparativist, or a curious reader rarely wants the whole tree; they want one thread: *how is this word connected to that one?* Spanish *pico* "beak" turns out to end in a Gaulish `bekkos`, `borrowed into Latin, cousin to English` beak. *That single narrative — word to word, across inheritance and borrowing, up an ancestry and back down another — is what most people mean by "etymology," and it is exactly what standard resources make hard to see. Dictionaries give you one word's ancestry as prose; family trees give you the abstraction; neither lets you* trace a path between two concrete words*.

Ariadne is named for the thread of Ariadne — the string that lets you traverse a labyrinth and find your way back. Its goal is to make the etymological thread navigable: type a start word and an end word, and see the route, as a chain, a graph, a map, or statistics.

2. What Ariadne is

Ariadne is three things in one artifact:

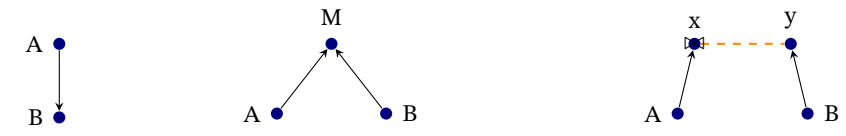
1. A distributable graph. A single SQLite file (`ariadne.sqlite`) is the canonical object: 686k nodes (word forms and reconstructed etyma) and 803k typed, provenance-bearing edges. One file, versionable, copyable, queryable with nothing but SQLite.
2. A pathfinding engine. A Python/NetworkX layer that resolves the *thread* between two words, plus graph statistics (borrowing profiles, hub etyma, depth distributions) as a first-class output rather than a build check.
3. A self-contained viewer. A D3.js web viewer with no backend: a *chain* view (a vertical timeline of the thread), a *graph* view (force-directed neighborhood), a *map* view (arcs over a vendored world map, no external tiles), a *search* view over the whole lexicon, and a *stats* view. A 2,500-node demo runs by double-clicking `index.html`; a full viewer loads the entire database in the browser via sql.js.

Edges are typed, and the type determines how they behave:

- `inherited`, `derived`, `borrowed`, `calque` — asymmetric ancestry. These, and only these, form threads.
- `cognate`, `mention` — lateral relations (siblings, references). They are shown but never routed through as if they were descent (see §5.4).
- `variant` — an inferred bridge between ablaut/allomorph variants of one root (see §4 and §5.5).

The engine finds a thread in three escalating ways, mirroring how an etymologist actually reasons:

1. Lineage — B is a direct ancestor of A: walk A's ancestry.
2. MRCA (most-recent common ancestor) — A and B climb to a shared node (e.g. `padre` and `father` meet at PIE `*ph tḗr`).
3. Bridge — A and B do not share a captured ancestor, but their ancestries can be joined by crossing one lateral link (an attested cognate, or an inferred root variant) at the top: `rojo` → *h rewd-* ⧓ h `rowd ós` → `rot`. This is the *up-over-down* path, and it is what makes Ariadne answer "what is the connection?" rather than merely "is one descended from the other?".



(a) lineage (b) MRCA (c) bridge

Figure 1: The three ways a thread is found. (a) *Lineage*: B is a direct ancestor of A. (b) *MRCA*: A and B climb to a shared ancestor M. (c) *Bridge*: the ancestries are joined by crossing one lateral link (an attested cognate or an inferred root variant, ⋈) at the top — the *up-over-down* path.

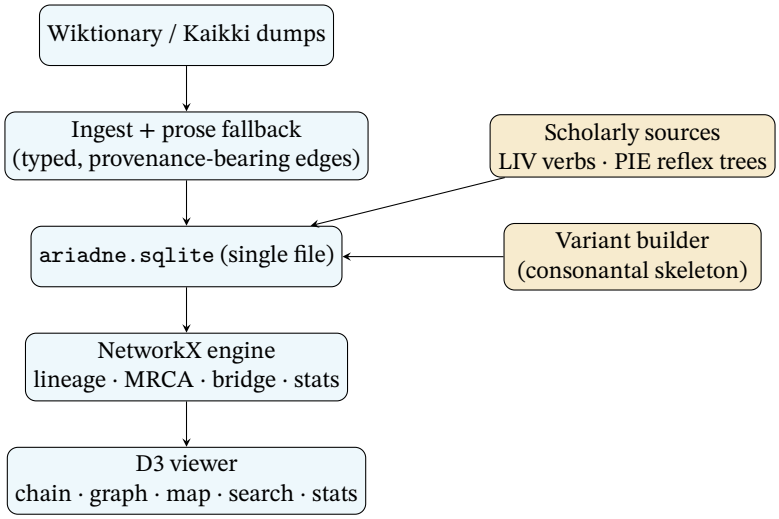


Figure 2: Pipeline. A single SQLite file is the canonical artifact; scholarly sources and the inferred variant layer feed it with their own provenance; engine and viewer are read-only consumers. Every edge keeps a `source_id`, so SOURCE, MACHINE PROPOSAL and REVIEWED never blur.

3. Data and provenance

The base source is Wiktionary, via the wiktextract/Kaikki extraction (Ylönen 2022), whose structured etymology templates give typed edges with standard language codes. On top of it we merge a small set of scholarly resources, each with a dedicated source id so every edge keeps its origin:

- LIV (*Lexikon der indogermanischen Verben*, Rix et al. 2001), via the LiLa project's Linked-Open-Data edition (CC-BY-SA 4.0): PIE verbal roots → Latin reflexes.
- Proto-Indo-European reflex trees from Wiktionary's reconstruction dictionary, which encode the reflex inventories of Pokorny, LIV, NIL, de Vaan, Kroonen, Derksen and others (each cited in the source). This is how a Latin *noun* like *oculus* finally reaches `*h ek -`.
- Proto-Germanic / Proto-Slavic / Proto-Italic reconstruction dictionaries — the branch tops, which in practice deliver Kroonen and Derksen (Wiktionary's proto entries cite them verbatim).

A recurring lesson deserves emphasis: much scholarly apparatus is already encoded inside Wiktionary. We did not need a bespoke Kroonen parser; Wiktionary's Proto-Germanic entry for *raudaz* literally carries inherited from h rowd ós with the citation *"Kroonen 2013, Etymological Dictionary of Proto-Germanic."* Ingesting the right dictionary often beats writing a new adapter.

Every edge records `source_id`, `confidence`, and a `hedge` (prose uncertainty markers such as *possibly*, *perhaps*, *substrate* lower confidence and are surfaced in the viewer). We adopt a three-state discipline borrowed from the wider research program: SOURCE (attested/asserted by a resource) → MACHINE PROPOSAL (inferred by us, e.g. a variant bridge) → REVIEWED (human-checked). Machine proposals are never silently promoted to facts; they are stored, labeled, and rendered differently.

Scale and positioning

The resulting artifact is a single 180 MB SQLite file. Its shape (all figures from the released build):

Property	Value
Nodes (forms + reconstructed etyma)	686,206
Edges (typed, provenance-bearing)	802,589
Languages	1,720
Edge types	inherited 224k · derived 183k · cognate 182k · borrowed 149k · variant (inferred) 64k · calque 983
Edges by source	Wiktionary 704k · inferred-variant 64k · PIE reflex trees 35k · LIV 374
Mean degree	2.29
Weakly-connected components	116,864 (largest = 45% of nodes)
Build	reproducible from public dumps via `make all` (order: ingest → sources → variants → analysis)

The many small components are expected and honest: most of the lexicon is *not* connected into one PIE tree, and the graph does not pretend otherwise. Positioned against comparable resources:

Tool	Word→word pathfinding	Per-edge provenance	In-browser viewer	Marks inferred vs attested
Etymological Wordnet (de Melo 2014)	no (relation table)	source-uniform	no	no
EtymDB (Sagot & Fourrier 2020)	no (relation table)	confidence scores	no	partially
etytree / Etymology Explorer	tree/graph browse	Wiktionary only	yes	no
Ariadne	yes (lineage/MRCA/bridge)	yes (`source_id` per edge)	yes (self-contained)	yes (labeled bridges, homographs, hedges)

The distinguishing commitments are *pathfinding as the primary operation* and *honesty as a first-class feature*: every thread is annotated with how it was found, and every edge with where it came from.

4. Inference done honestly: the variant layer

The hardest structural problem in etymology-at-scale (§5.1) is that the *same* proto-root is written many different ways. *h rewd* - (e-grade) and h rowd ós (o-grade derivative) are the same root of "red," but Wiktionary stores them as separate nodes and often does not link them. So `rojo` climbs to one and `rot` to the other, and the thread breaks — even though the words are obviously related.

We refuse two easy but dishonest answers: (a) declaring "no relation," which betrays the tool's purpose; and (b) fusing the nodes, which fabricates a fact. Instead we add an inferred `variant` layer grounded in the project's own theory: the consonantal skeleton is the invariant, the vowel is the ablaut operator. Two proto-forms of the same language whose consonantal skeleton ($\geq$3 consonants) is identical, or one a prefix of the other, are joined by a `variant` edge that is symmetric, low-confidence, and

explicitly attributed to inference (a distinct source, not Wiktionary). The pathfinder may cross one such bridge; the viewer draws it differently and states plainly: *"crosses an inferred bridge by shared consonantal root — not Wiktionary ancestry."* The user sees the connection *and* sees exactly what kind of connection it is.

Why the consonantal skeleton, and why three? The choice is not arbitrary. Ablaut in Indo-European operates precisely on the *vowel* while the consonantal frame is preserved (*h rewd* - / h rowd ós / *h rud* - share h -r-w-d); the skeleton is therefore the natural invariant for exactly the fragmentation we face, and this is also the operative claim of the wider research program's two-layer model of the sign. We deliberately do not *use a general string metric (Levenshtein) or a full phonological-feature distance: those would score* any *similar forms as related, re-introducing the false positives we are trying to avoid, and they have no principled tie to the ablaut mechanism. Vowels are excluded because they are the* variable *the operator moves. The* $\geq$*3-consonant floor is a precision guard, validated against the data: two-consonant skeletons (e.g. ker)* collapse unrelated roots (*ker-* "horn" with kreu- "blood"), whereas three or more consonants are specific enough that identity-or-prefix rarely over-merges. On the released build the skeleton key reconciled 94% of imported LIV roots against existing nodes. Two honest caveats: (i) in languages where vowel length or quality is phonemic and root-distinguishing the heuristic can over-connect — hence it is *inferred*, *bounded to one hop*, and *labeled*, never a hard merge; (ii) it is a heuristic, not a theory of ablaut. A regular vowel *operator* — treated as a first-class object — is the proper solution, and is deliberately left to a sibling project rather than smuggled in here.

5. The difficulties (the real content)

We report these in detail because they are what anyone building such a tool will hit, and because glossing over them is how tools come to overclaim.

5.1 Reconstruction is not canonical

Proto-Indo-European is not a single agreed object. Different specialists write the same root with different laryngeals, ablaut grades, and notation; the same reflex is hung on different reconstructed stems. A graph keyed on surface strings therefore fragments one root into many nodes. This is not a data-cleaning bug; it is the epistemic status of reconstruction leaking into the data model. Our partial answer is the skeleton-based variant layer (§4); the general answer is a project of its own (a regular *operator* over ablaut grades), which we deliberately did not try to solve inside Ariadne.

5.2 Wiktionary hangs all ancestors on the modern word

Wiktionary typically lists *every* ancestor directly under the descendant, and the intermediate forms do not re-link to each other. ojo's entry names Old Spanish *ojo*, Vulgar Latin *oclus*, and Classical Latin *oculus* as siblings under ojo, rather than as a chain. A naive lineage that "climbs one edge at a time" dead-ends. We reconstruct the chain by taking a word's whole ancestral closure and ordering it by language stage, which recovers the full ojo $\rightarrow$ oclus $\rightarrow$ oculus $\rightarrow$ *h ek - line.

5.3 Etymology hidden in prose (a systematic ancestry gap)

Wiktionary's newer "Etymology tree" format stores only *cognates* in the structured templates and puts the actual ancestry in prose. As a result ojo — and 42,245 Spanish entries — had *zero* structured

ancestry. We added a prose fallback that parses *"from LANGUAGE FORM"* when the structured templates carry no ancestry. A tool that reads only the "clean" structured field silently loses half the ancestry in a language.

5.4 Cognate is not a path

Early versions produced absurd threads by walking chains of *cognates* (a colour word reaching a house word through a sequence of siblings). Cognacy is a lateral relation; treating it as descent connects almost everything. We excluded `cognate/mention` from pathfinding entirely, admitting them only as a single, labeled lateral hop at the top of a bridge (§2).

5.5 Homographs collapse distinct histories

Node identity is `(language, normalized form, part of speech)`. English *red* (colour, from Old English *rēad*) and *reed* (plant, from Old English *hrēod*) both pass through a Middle English spelling *red* — different words, same spelling — which our key merges into one node, producing a spurious "common ancestor." We keep these cases (they are informative) but detect and label them: if the meeting node is *attested* (not a reconstruction) and *bare* (has no etymology of its own), it is almost certainly a homograph collision, and the viewer says so: *"similar spelling, different origin."* A genuine confluence happens at a reconstructed root or at a node that has its own ancestry.

5.6 Silent normalization traps

Latin is transcribed with vowel-length macrons inconsistently: *fīlius* and *filius* became two nodes, so *hijo* reached one and a query for the other found nothing — 11,399 such Latin pairs. We strip only the length macron (not phonemic accents) in the normalization used for *both* ingestion and lookup; the two must match exactly or the search silently fails.

5.7 Coverage is deeply uneven

A structural-completeness metric we call C12 (real dictionary lemmas with *zero* ancestry) shows that 72% of Latin lemmas have no captured etymology. Some of that is genuine (technical, derived, or unknown-origin words), but much of it is the *ojo*-class hole. A companion metric C13 (fraction of a language reaching a PIE root) showed that the *tops* of whole branches were disconnected from PIE — Proto-Germanic, Proto-Slavic and Proto-Italic each reached PIE 0% of the time — because those nodes existed only as stubs cited by descendants, never ingested as entries. Filling them (§3) raised Proto-Italic to 68%. Coverage is not a number you can assume; it is something you must measure per language and per layer.

5.8 Multi-source reconciliation

When you merge LIV, Pokorny, and Wiktionary, the same root appears as `reudh-`, *`h rewd -`*, and `h rewd( )-`. Merging wrongly fabricates identity; not merging leaves the sources as disconnected islands that fail to fill the hole. We reuse the consonantal skeleton (§4) as the cross-source alignment key and reconciled 94% of LIV's roots against existing nodes — but the residual 6%, and the general problem, remain a genuine open challenge.

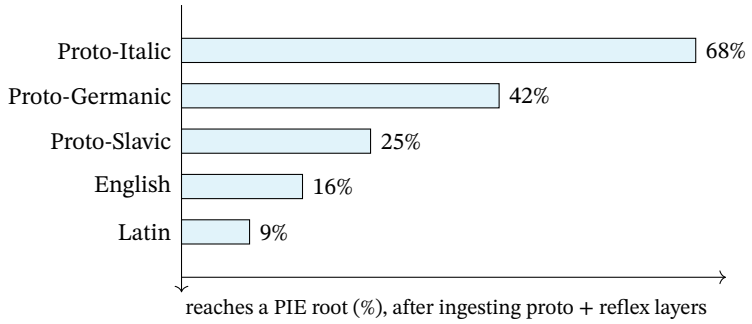


Figure 3: C13 (fraction of a language reaching a Proto-Indo-European root) after the source-merge phases. Before merging, the *tops* of these branches (Proto-Italic, Proto-Germanic, Proto-Slavic) reached PIE 0% of the time — they existed only as stubs cited by descendants. The metric turns "coverage" from an assumption into a measurement.

5.9 Delivering a 172 MB database to a browser

The full graph does not fit the constraints of free static hosting (GitHub's 100 MB per-file limit; Git LFS is not served usefully by Pages). The self-contained demo deploys anywhere; the full viewer needs either lazy range-request access to a chunked database or an external, CORS-enabled host. Distribution of large linguistic artifacts is an unglamorous but real obstacle to reproducibility.

6. Guarding against apophenia

Because the tool *finds connections*, it is structurally prone to finding connections that are not there. We treat this as a primary concern, not an afterthought:

- A living case registry. A JSON file of word pairs, each with an expected outcome (`path` / `bridge` / `none`), is evaluated on every build. It includes anti-apophenia guards — pairs that must *not* connect (`ojo`?`guerra`, `sol`?`luna`). If a guard starts connecting, the inference has become too loose. This same registry caught a real regression: enlarging one data source silently widened the variant buckets and dropped a real bridge, flipping `rojo`↔`rot` to "no relation" until we fixed the threshold.
- Honest labels over hidden heuristics. Threads are annotated with *how* they connect: a genuine lineage, a most-recent common ancestor, an inferred variant bridge, an attested cognate hop, or a homograph collision. The user is never shown a bare line that conflates these.
- Provenance everywhere. Every edge carries its source; inferred edges are a separate, visibly-marked layer.

This is the tool-scale version of a discipline the wider program applies to the science itself: form and evidence as two prostheses against pattern-hunger.

7. Evaluation against an independent cognacy standard

To measure Ariadne without inventing ground truth, we evaluate it against IE-CoR (the Indo-European Cognate Relationship database; expert cognacy judgements over 160 Indo-European languages) — a resource Ariadne does *not* consume. The design has two sides:

- Positives — pairs of forms that IE-CoR places in the *same cognate set* but in *different* languages. These are expert-attested cognates, so Ariadne *should* connect them (by lineage, MRCA, or bridge). This measures recall / coverage.
- Negatives — pairs from *different* cognate sets *and* different concepts, in different languages. These are (overwhelmingly) unrelated, so Ariadne *should not* connect them. This measures the false-positive rate — the critical test of whether the inferred bridge/variant layer over-connects.

We restrict to Latin-script languages, because IE-CoR renders non-Latin scripts in transliteration that would not match Ariadne's orthographic nodes; forms are matched to nodes by (language code, normalized form). The fraction of IE-CoR forms that resolve to a node is reported as resolution coverage — it reflects matching and data gaps, not algorithmic behaviour. We draw 100 pairs (60 positive, 40 negative) and repeat over five random seeds. Fully reproducible: `python3 src/eval_iecor.py`.

Measure	Result
Resolution coverage (IE-CoR Latin-script forms → node)	66.5%
Recall on positives (should connect)	mean 81%, range 65–92% across seeds
— of which, by mode (representative seed)	pure ancestry path 78% · inferred bridge 13% · missed 9%
False positives on negatives (should not connect)	0.0% in every one of the five seeds

Two readings matter. First, recall is a coverage measure, not an accuracy score: the 9–19% of true cognate pairs Ariadne fails to connect are cases where the underlying data lacks the ancestry (e.g. a thin Icelandic or Faroese line), not cases where the algorithm errs; its variance across seeds tracks which branches are sampled. Second, and more importantly, the false-positive rate is 0% under every seed. The layers that most risk fabricating structure — the attested-cognate hop and the inferred consonantal-skeleton variant bridge — did not connect a single unrelated pair. This is the quantitative counterpart of the design stance in §6: inference is admitted only where it is bounded, labeled, and — as measured here — specific.

The evaluation is deliberately modest (100 pairs, Latin-script Indo-European, negatives that are an *upper bound* since a few random pairs could be genuine deep cognates). A larger study across scripts and families, and against other tools and human raters, is the obvious next step (§10).

8. What it is for

- Accelerated language learning and reading. The thread makes the historical/comparative knowledge that already exists *self-discoverable*: a learner sees *why* two words feel related.
- Comparative and areal exploration. Borrowing profiles (donors vs. borrowers), hub etyma, and contact pairs are quantitative outputs, not just visuals.
- A substrate for other tools. The graph is a clean, provenance-bearing base that downstream projects can consume rather than re-extract.

9. An invitation

Ariadne is open source and lives at https://github.com/toledoal/ariadne (code MIT; data CC-BY-SA 4.0, honoring Wiktionary's ShareAlike and attribution). The repository ships *code, not data*: the 180

MB database is rebuilt from public dumps with a single `make all`, and the test registry and the evaluation script (`src/eval_iecor.py`) are public. We built it in the open, with the seams showing, precisely so that others can do better.

We think the useful move is not one monolithic "correct" etymology engine but a *family* of small, honest, composable tools, each answering one question and declaring its limits. Ariadne answers *"where did this go, historically?"* A sibling project, Isogloss, deliberately suspends history to ask *"what formal relations emerge between languages when we assume no genealogy at all?"* — the same data, the opposite premise. Others could tackle the ablaut-operator problem (§5.1) as a first-class object, or the semantic layer, or better multi-source reconciliation (§5.8). The graph, the metrics (C12/C13), and the case registry are offered as a shared starting point.

10. Limitations and future work

Ariadne inherits every limitation of its sources: gaps, errors, and the non-canonical status of reconstruction. Its inferred layer is a heuristic, not a theory of sound change. Coverage of non-inherited, borrowed, and technical vocabulary is thin by construction. Immediate next steps: merge broader modern-language sources (Etymological Wordnet, EtymDB) with cross-validation; improve proto-form reconciliation; and split the database for fully in-browser access. The deeper open problem — unifying ablaut variants under a regular operator — is, appropriately, someone's next tool.

References (to be completed for submission)

- Ylönen, T. (2022). *wiktextract* / Kaikki.org. — structured Wiktionary extraction.
- Rix, H. et al. (2001). *Lexikon der indogermanischen Verben* (LIV); LiLa LOD edition (CIRCSE).
- Pokorny, J. (1959). *Indogermanisches etymologisches Wörterbuch*; digitizations (Starostin/Lubotsky; Köbler).
- Kroonen, G. (2013). *Etymological Dictionary of Proto-Germanic*. Brill.
- Derksen, R. (2008). *Etymological Dictionary of the Slavic Inherited Lexicon*. Brill.
- de Melo, G. (2014). *Etymological Wordnet*. — CC-BY-SA.
- Sagot, B. & Fourrier, C. (2020). *EtymDB 2.0*, LREC. — CC-BY-SA.
- List, J.-M. et al. — CLTS / CLDF / Lexibank (comparative infrastructure).
- Related tools: etytree; Etymology Explorer; Wiktionary "Descendants" trees.