How Probative builds its open datasets: methods report

Version 1.0. Published by Probative Growth, a marketing agency for US law firms. CC BY 4.0. Not legal advice.

What this report covers

This report describes the methods shared by the open datasets Probative Growth publishes on US law-firm marketing: where the data comes from, how it is read and checked, how it is versioned and corrected, and how to cite it. Each dataset's own page and README give its sources, columns, coverage and limits. The current list, with every version and DOI, is at probativegrowth.com/data/ and in machine-readable form at probativegrowth.com/data/catalog.json.

Where the data comes from

We start from the body that publishes or enforces the fact: federal and state agencies, courts, state bars and legislatures, or the platform's own published policy. Every record names its source, links to it where a link exists, and carries the date we last read it.

Each dataset's README explains the status codes it uses, so a reader can tell a value read from a primary source from one that is not.

How it is read

Where a source publishes a file or an official API, a script reads it on the source's own release schedule, at a polite rate and with an honest user agent. Where a source cannot be read by script, a person reads it in a browser, and the read is recorded as a dated check. We do not get past logins, CAPTCHAs or other access controls.

How it is checked

Figures on our pages are generated from the dataset files or registered against them, and the build fails when a page's figure no longer matches its registered source.

A reader can repeat the check: copy a value and its record id, open the source link on that record, then compare the value, the unit and the read date. A newer edition of a source can differ from our record without either being wrong, so the read date matters.

Versions, DOIs and mirrors

Each dataset has version numbers and a DOI on Zenodo: one for each version and one concept DOI that always points to the newest. Changes are published in each dataset's change log and in the changes feed. The same files are published on GitHub, Hugging Face and Kaggle, and a check compares those copies with the live catalog.

Keeping it current

Each dataset declares how often it is re-read. A dataset, page or download that passes its review date fails our build until it is checked again. The data calendar lists the next update of each dataset.

Corrections

A correction is a value we published wrong: a misread table, a wrong join or a typo. We check each report against the official source. If the data is wrong, we fix it, publish a new patch version and log the old value, the new value and the cause at probativegrowth.com/data/corrections/. When an agency revises its own figures, that is an update, and it goes in the change log.

Probative benchmarks

The tables labelled “Probative benchmarks, 2025–2026” are aggregated, rounded, and small groups suppressed. No firm is named in them.

How to read the data

The datasets count and describe; they do not rate a firm, give legal advice or say that an ad is allowed. The lawyer responsible, the state bar or counsel decides that. Search results differ by place, so a figure read from a search result holds for the location it was read from. Two figures that look alike may measure different things: check the place, the period, the unit and the definition first.

How to cite and reuse

The data is licensed CC BY 4.0: you can reuse it, including commercially, with credit and a link. Standard form: Probative Growth, <dataset title>, <year>, https://doi.org/<DOI>. The citation kit at probativegrowth.com/data/citation-kit/ has the exact line for each dataset.

To cite this report: Probative Growth, How Probative builds its open datasets: methods report, 2026, https://probativegrowth.com/reports/probative-data-methods.pdf.