

Data Sourcing and Intellectual Property Policy

Version 1.0 · Effective 1 November 2026 · Owner: Aleksandr Senkevich, katelabs, Belgrade, Serbia · katelabs.io

1. Purpose and scope

This policy describes where the katelabs dataset comes from, what is stored, what rights katelabs holds in it, and what rights it does not claim. It applies to every record delivered under a sample, a trial, a licence or an API key, and to the research notes published on katelabs.io.

2. What the dataset is

The dataset is a record of English-language cryptocurrency news headlines collected continuously since 4 October 2024. Each record holds the headline, a short publisher-supplied summary where one exists, the canonical URL, the publisher name and domain, the publication time, the time the record was written, the assets the headline refers to, and two values produced by katelabs: a score of expected price effect from -1 to 1 and a type (proactive or reactive). Scores and types are assigned once, at ingestion, by an automated language-model pipeline, and are never revised afterwards. The time of scoring is stored with each record.

3. Sources

- 1. Public material only.** katelabs collects headlines and summaries that publishers make publicly available: RSS and Atom feeds, news sitemaps, public JSON endpoints and licensed news aggregation services. Nothing is collected from behind a login, a paywall, a private group or a messaging channel, and no scraping of article bodies takes place.
- 2. Publisher terms.** Collection follows each publisher's robots directives and feed terms. A publisher that asks to be excluded is removed from collection within five business days; records already delivered to licensees are not recalled, because the dataset is point-in-time, but the publisher is dropped from every subsequent delivery.
- 3. No personal data by design.** Records contain no subscriber, reader or user data. Names appear only where a publisher put them in a headline. katelabs does not enrich records with personal data from any other source.
- 4. Transparency.** Every record carries the publisher's domain and URL, so a licensee can always trace a line back to its public source. The list of domains in the dataset is available to any licensee on request.

4. What katelabs stores and what it does not

Stored: headline text, the summary a publisher attaches to the feed item (typically one to three sentences), URL, publisher, timestamps, asset tags, score, type. Not stored and never redistributed: article bodies, images, author profiles, comments, or any other content beyond the feed item itself. Tracking parameters are stripped from URLs before storage.

5. Intellectual property

- 1. Publisher content.** Headlines and summaries remain the property of their publishers. katelabs claims no copyright in them and reproduces them as short excerpts, with attribution, for the purposes of indexing, classification and quantitative research. A licensee's rights in that material are limited to internal analysis; a licensee may not republish headline text as a news product.
- 2. katelabs derived data.** The scores, types, asset tags, timestamps of ingestion and scoring, the per-asset news index, the dataset structure and selection, the collection pipeline, the prompts, the evaluation methodology and all research notes are the intellectual property of katelabs. They are provided under licence and remain owned by katelabs.
- 3. Licensee outputs.** Models, signals, strategies, reports and other results a licensee builds from the data belong to the licensee. Raw records and derived fields may not be extracted from those results and passed to a third party as data.
- 4. Open slice.** A three-month slice of the dataset is published on Hugging Face under the terms stated on its dataset card; those terms govern that slice only.

6. Point-in-time guarantee

Once written, a record's score, type and timestamps are not changed. Corrections to publisher metadata (for example a fixed URL) are applied as new fields, never by overwriting the original, and are documented in the dataset changelog. Known defects of the pipeline, including gaps in coverage, are disclosed in research note 4 and in the data dictionary that accompanies every delivery.

7. Takedown and questions

A publisher, author or rights holder who believes material is collected or reproduced improperly may write to **aler.senkevich3301@gmail.com** with the URL concerned. katelabs acknowledges within two business days and resolves within ten. Licensees receive notice of any source removal that affects their deliveries.

8. Review

This policy is reviewed at least once a year and whenever the collection method changes materially. The current version is always available at katelabs.io/docs.