

EFRU Data Release V1

Datasheet For Dataset

Based on the [Datasheets for Datasets](#) paper, v7

Motivation

These questions encourage dataset creators to clarify their motivations and funding interests.

For what purpose was the dataset created?

Was there a specific task in mind or a particular gap to fill?

- **Context:** After the January 2025 Eaton Fire in the Los Angeles area, many residents—especially those whose homes remained standing—grew concerned about toxic ash, soot, and other contaminants. Professional environmental testing (e.g., by industrial hygienists) is costly meaning that many residents cannot or could not afford it.
- **Primary Purpose:** Eaton Fire Residents United (EFRU) created this dataset to publicly share pre-remediation contaminant testing data collected and volunteered by community members. It is intended to raise awareness, guide local advocacy efforts, and provide evidence-based policy recommendations. By harmonizing these test results into a map and analysis-ready dataset, EFRU hoped to help individuals who lack the resources for their own testing by providing a broad sense of how contamination appeared in certain neighborhoods.
- **Intended Benefit:** The dataset aims to inform community members and policymakers about exposure risks, encourage further testing, and provide evidence to guide remediation measures.

Who created the dataset and on behalf of which entity?

This dataset was created by EFRU's Data Mapping and Unification subcommittee, a volunteer-led group of local residents and professionals dedicated to ensuring a safe and transparent fire recovery process after the Eaton Fire.

Who funded the creation of the dataset?

The dataset was compiled entirely by community volunteers. No external grants or institutional funding were used.

Any other comments?

None.

Composition

These questions help dataset consumers evaluate the dataset's coverage and representativeness.

What do the instances that comprise the dataset represent?

Each instance (i.e., each row in the CSV) corresponds to a single residence's pre-remediation (i.e., pre-cleaning) test results from a professional testing company (industrial hygienist or certified industrial hygienist) and accredited laboratories. The dataset provides peak (maximum) concentrations for multiple contaminants (e.g., heavy metals, asbestos, ash) found indoors. It does not contain the result of every single test for every single contaminant (e.g., if 5 lead samples were taken, this dataset only reports the highest value).

Important note: No raw or redacted PDFs of the test reports are included in order to protect residents' privacy.

How many instances are there in total?

Currently, there are 201 distinct tests within this dataset, with most of those containing information on multiple different contaminants.

Does the dataset contain all possible instances or is it a sample?

All "valid" submitted test result cases are included—i.e., results from recognized professional testing providers that sampled hard surfaces. About 25 tests were excluded due to either methodological issues (e.g., non-professional test kits, incomplete or compromised samples, tests missing units of measure, or tests of soft materials like couches and clothing), duplicate submissions, tests from other wildfires (e.g., the Palisades Fire), or using methodologies outside those of interest here (e.g., indoor air quality). The dataset is thus comprehensive for this specific scope but does not represent all properties or areas affected by the Eaton Fire as obviously not every affected house was tested, and not all houses that were tested appear in the data set (due to the voluntary nature of participation).

What data does each instance consist of?

Each row contains:

1. Metadata

- Approximate house location (nearest cross street or slightly shifted coordinate)
- Location of test (attic, interior floor, windowsill, etc.)
- User-reported proximity to burned structures
- User-reported property damage level

- Remediation status at the time of sampling
- Date of sample collection

2. Contaminant Information and Measurements

- Wildfire debris (e.g., ash, soot, char)
- Peak lead level (with location notes)
- Asbestos concentration and test method (e.g., TEM vs. PLM)
- Peak metal concentrations such as arsenic, antimony, cadmium, chromium, cobalt, copper, nickel, and zinc

Is there a label or target associated with each instance?

No. The dataset simply reports contaminant concentrations, so no label is provided (in a machine-learning sense).

Is any information missing from individual instances?

- Original and redacted PDFs of the test reports are intentionally withheld.
- Entries will be missing some fields depending on which contaminants were tested (different labs, different heavy metals, whether or not asbestos was tested, etc.).

Are relationships between individual instances made explicit?

Not beyond each row being linked to a specific residence. There is no direct link between residences in the data aside from approximate location references.

Are there recommended data splits?

No. The dataset is not directly intended for machine-learning tasks. However, any future modeling or ML-related work should consider the inherent spatial correlation and large variability in contamination intrusion.

Are there any errors, sources of noise, or redundancies?

- Differences in sampling methods, lab instruments, or lab procedures might introduce noise as the dataset contains test results from many different testing companies.
- Asbestos test methods vary widely in sensitivity, and some labs may rely on less-sensitive methods (like PLM) instead of more-sensitive methods (like TEM).

Is the dataset self-contained?

Yes. No external resources or websites are needed to use the compiled CSV.

Does the dataset contain confidential data?

No. We have anonymized and obfuscated addresses so that specific households cannot be directly identified.

Could the dataset be considered offensive, threatening, or anxiety-inducing?

The data may cause anxiety because it deals with contamination levels in living spaces. This information can be emotionally distressing, especially for those living in fire-affected regions.

Does the dataset relate to people?

Indirectly—these data represent residential properties and contaminants indoors, but they do not contain details on individuals (e.g., name, age, etc.).

Does the dataset identify any subpopulations or sensitive data?

No. We did not collect or include demographic attributes such as race, religion, etc. at any point in the data collection process.

Is it possible to identify individuals?

We have taken measures to prevent re-identification by shifting the reported location to the nearest intersection and removing all personal details. However, individuals who live in very low-density areas should be aware that re-identification is theoretically possible if someone has learned their home's testing status through some other means.

Does the dataset contain sensitive data in any way?

No direct personal or financial information is included. However, contamination levels might be considered sensitive health or environmental data, so we have anonymized addresses to minimize the risk of misuse.

Any other comments?

None.

Collection Process and Preprocessing/Cleaning

How was the data for each instance acquired?

Data was collected via voluntary submissions from community members who had professional indoor environmental testing done at their residences before any remediation. EFRU provided an online form at <https://www.efru.la/share-testing>, and contributors uploaded scanned/PDF copies of their reports. As part of that process, users signed a waiver, which is included in this data release as a separate document.

What mechanisms or procedures were used to collect the data?

1. **Test Exclusions:** Test results were excluded when they were DIY home tests, tests by non-professionals, tests where environmental labs identified problems with sample

integrity/preparation, tests containing measurements that cannot be standardized with other surface tests (e.g., indoor air quality or tests on soft materials like couches or clothing) or tests missing critical information (e.g., the surfaces tested). Such exclusions were relatively rare (~25 tests total).

2. **Redaction:** Each submitted report receives an ID and all personally identifiable information is removed (e.g., address, photos) by one team member with sole access to the original submitted reports.
3. **Geo-anonymization:** To obscure specific households, the latitude and longitude of the testing address are altered to move the location to no closer than the nearest cross street (5 or more homes must be in the general vicinity to be considered available for adjustment, otherwise it is moved further).
4. **Test Type Sorting:** Each test is sorted into the proper categories (e.g., pre- or post-remediation, soil, air, fabric, surfaces).
5. **Extraction and Standardization:** A trained subcommittee member manually extracts key contaminant values (peak lead, asbestos, and other metals) and related metadata to the EFRU dataset. To enable comparisons between standard surface types, names of locations are unified. For example, “Interior Windowsill” is used for all of the following: windowsill, sill, WS, WF, window. All amounts are updated to micrograms per square foot (ug/ft²). The rooms and locations tested were standardized to remove specific information about the number of rooms and other configuration details about a house.
6. **Verification:** For tests flagged for follow up (either by the initial data entry volunteer or through automatic verification) a second team member double-checks the extracted data against the anonymized PDF to fix transcription errors.

All volunteers doing data extraction were required to review our “Standard Operating Procedure” document, which describes best practices for redaction and data handling and also to go through a training process under the supervision of an experienced subcommittee member.

If the dataset is a sample, what was the sampling strategy?

All valid professional test results were included. About five were excluded due to questionable methodology (e.g., consumer-grade test kits) or inappropriate sampling methodologies (like testing soft clothing instead of a hard surface).

Who was involved in the data collection process and how were they compensated?

Test results were collected by testing companies and then the reports were volunteered by the property owners or tenants. Data analysts were volunteers from the local community. No monetary compensation was provided.

Over what timeframe was the data collected?

- The Eaton Fire began on January 7, 2025.
- Testing data in this dataset was collected from late January to late June 2025.
- Additional data may be added in future releases as more residents continue to test or obtain post-remediation testing.

Was the raw data saved in addition to the processed data?

EFRU securely stores the original PDFs (i.e., the “raw” data) on a private, access-restricted cloud drive. These PDFs are not publicly released. Future uses (e.g., follow-up research) would require additional user consent and EFRU’s approval.

Is the software used to preprocess/clean the data available?

Two basic Python scripts were used to import, parse, and cross-check the data before exporting it to a CSV. They are included in the release.

Were any ethical review processes conducted?

EFRU is a grassroots organization and not formally bound by institutional IRB processes. However, the EFRU team worked to follow data protection principles and to minimize harm and a consent process to ensure what the data could and could not be used for was made clear to those that volunteered their test results.

Does the dataset relate to people?

Yes, in that it relates to homeowners/renters, but it focuses on contaminants in the indoor environment. Personal identifiers are removed.

Did you collect data directly from individuals or via third parties?

All data came directly from individuals who chose to share their professional testing reports. We did not purchase or scrape data from external third-party sources.

Were individuals notified about data collection?

Yes, individuals volunteered their data through our website’s submission page and were shown disclaimers as part of that process. They were explicitly told their data would be compiled into an anonymized public dataset.

Did individuals consent to the collection and use of their data?

Yes, they provided explicit consent via the website waiver form, which included a description of how data would be anonymized and used to create a public dataset. A copy of that waiver is included in this data release.

Was a mechanism provided to revoke consent?

The data revocation procedure is provided in the Waiver that each participant must agree to before uploading their test results.

Has an impact assessment been conducted?

No formal privacy impact assessment has been conducted, but we carefully considered potential harms, including the risk of re-identification and misuse (e.g., by insurance companies).

Any other comments?

We welcome further discussion on best practices for data ethics in community-driven environmental testing.

Uses

These questions highlight recommended and discouraged uses of the dataset.

Has the dataset been used for any tasks already?

Yes. Its primary current use is generating a public map on EFRU's website. The map color-codes residences by peak lead concentration and also provides information about other tested contaminants. This map helps neighbors who cannot afford testing understand general contamination patterns and provide evidence-based information to guide local advocacy for cleanup and insurance accountability.

Is there a repository that links to papers or systems that use the dataset?

Not currently. Any relevant academic or policy research that cites the dataset should ideally link back to EFRU's main page. We plan to register the dataset with a DOI on Zenodo; once that is done, relevant references may appear there.

What other tasks could the dataset be used for?

- **Policy Analysis:** Researchers or advocates could correlate contamination data with regional remediation measures to guide legislative or regulatory decisions.
- **Legislative Action:** The map can be leveraged for asking legislators to write new policies supporting our asks for the next legislative cycle.
- **Defining the Zone of Impact:** Garnering support from private, public and academic institutions to support testing of homes and soil will allow expanding the map and defining an "ash zone" and/or other zones of impact.

- **Children’s Health:** The map can be used to leverage testing at public as well as private schools, child care centers, therapeutic and enrichment centers and all other spaces where children gather.
- **Methodological Studies:** Comparing how different laboratories or testing methods measure contaminants.
- **Public Health Research:** Investigating patterns of indoor contamination following wildfires to inform future disaster-response frameworks.

Does the dataset’s composition or collection process impact future uses?

- **Location Shifting:** The approximate location may reduce usefulness for local modeling.
- **Potential Biases:** Houses included may skew toward those whose owners were concerned about environmental hazards (e.g., perhaps families with children or medical conditions) and/or could afford paid professional testing. Residents with fewer concerns and/or lower income may be underrepresented.

Are there tasks for which the dataset should *not* be used?

- **Defining “Safe Zones”:** It is irresponsible to label specific streets as “safe” solely based on these data. The contaminants in wildfire ash dispersal can be extremely sporadic and likely spread well beyond the local region represented by our testing data.
- **Interpolating or Extrapolating Hazard/Risk:** The dataset does not support generating precise maps of contamination that might be used to judge risk at a house-by-house level. It is instead intended to help understand the range of contamination in a broad sense. It also focuses on test results from within and near the burn zone; the lack of results from other (potentially downwind) areas should not be taken to mean that those areas are safe by default.
- **Insurance Underwriting or Rate Setting:** This data is not appropriate for use by insurers or other institutions to inform rate adjustments or any targeted evaluation of potential/current individual policyholders inside or outside the areas covered. We strongly discourage such use.
- **Privacy Violations:** Users should not try to re-identify, dox, or otherwise take action that would harm individual residents.

Any other comments?

This dataset is shared for community benefit; any exploitation that exploits residents or harms their rights or well-being goes against EFRU’s mission.

Distribution

Will the dataset be distributed to third parties?

Yes. The processed CSV (with anonymized data) will be openly accessible. Original/redacted test PDFs will not be distributed.

How will the dataset be distributed?

Results are archived on Zenodo, which provides versioning, assigns a DOI, and makes the data searchable, ensuring easy discoverability and citation.

When will the dataset be distributed?

An initial public release is expected in August 2025. Additional or updated versions may follow every couple of months as new test data becomes available.

Will the dataset be distributed under a copyright or license?

This dataset is released under the Creative Commons Attribution 4.0 International licence (CC BY 4.0) available online (<https://creativecommons.org/licenses/by/4.0/>) and as a text file in the data repository. When using the data, please cite the associated paper and data DOI.

Have any third parties imposed IP-based restrictions?

No. Contributors voluntarily provided their data.

Do any export controls or other regulatory restrictions apply?

No. This is domestic, environmental data with no known export restrictions.

Any other comments?

We attempted to follow the “FAIR Guiding Principles for scientific data management and stewardship” described [online](#) and in a [publication](#). In particular, to make the dataset “Findable,” we have obtained a DOI and uploaded the data on a popular and searchable data hosting service. To make it “Accessible,” we have made the data publicly available on Zenodo indefinitely. To make it “Interoperable,” we compiled the data into a standard CSV file such that it can be loaded into Excel or with standard programming tools (like the open source Pandas package in Python). Finally, we made it “Reusable” by including a Data License and including the standard operating procedures used to generate the dataset from the original industrial hygienist test PDF reports (such that others could extend our dataset using tests they obtained on their own if desired).

EFRU may add disclaimers related to data usage on the Zenodo page.

Maintenance

Who is supporting/hosting/maintaining the dataset?

EFRU maintains the dataset. It will be hosted on Zenodo for long-term stability.

How can the dataset owners be contacted?

Further questions or requests can be directed to EFRU's organizational email:
eaton.fire.residents.united@gmail.com

Is there an erratum?

Each new dataset version published on Zenodo will include a changelog describing any corrections or updates.

Will the dataset be updated?

We anticipate incremental updates every few months as new test results are received or existing data are corrected. Updates will be announced via EFRU's website or email newsletter.

Are there limits on how long data will be retained?

We have not specified a fixed retention period. Personally identifying information is permanently removed before any data release. If a resident requests removal, we will remove their data in the next dataset revision.

Will older versions continue to be supported or hosted?

Zenodo preserves prior dataset versions. While older versions remain accessible, we advise relying on the latest release for the most accurate and complete data.

If others want to build on the dataset, is there a mechanism for them to do so?

EFRU welcomes outside research and community initiatives but cannot provide extensive technical support with integrating data directly into this dataset due to limited volunteer effort. Others wishing to compile data in a similar way should consult the standard operating procedure document.

Any other comments?

We express our profound gratitude to community members who have voluntarily shared their reports and helped us compile this dataset, all in pursuit of rebuilding a healthier, safer community.