# Introduction to the dataset: 2024 Open Source Developer Survey

This dataset contains three PDF files (Introduction to the dataset, survey instrument, and survey frequencies) and one Full Survey Reference (FSR) spreadsheet with four tabs. Each of the four files is described below. This introduction is recommended reading as it explains what each of the four files contains. The creative commons license that governs use of this content can be found here: <https://cdla.dev/permissive-2-0/>.

## Introduction to the dataset (PDF)

This document describes the three PDFs and the spreadsheet that have become our preferred deliverables for each dataset. This introduction provides information useful to understand the context of each of the four files. While each of the PDF files is generally self-explanatory, the spreadsheet does benefit from an explanation of the nomenclature and conventions that have been used to convey data and information.

## Survey instrument (PDF)

The survey instrument is a PDF export from the production data collection survey. It identifies each question in the survey and its possible responses. Questions are sequentially numbered.

## Survey frequencies (PDF)

Frequencies for each question asked in the survey can be found in this file. The frequencies are based on the production data that we analyzed from the survey. These frequencies have been generated by our data analysis tool and in some cases have received additional post-processing to reformat the results. All data included in these frequencies include only those records that meet the following conditions:

* **Data quality.** Raw data is subject to a variety of tests to determine its quality. Records that fail one or more of these data quality checks are deleted from the raw dataset.
* **Screening and disqualification.** All surveys include screening criteria which are used to disqualify respondents who do not meet our target sample specifications.

The frequencies in this file provide a distribution of responses for all of the respondents eligible to answer each question. Because we use complete responses as well as partial responses, the sample size tends to decrease some as you progress through the survey due to voluntary respondent survey terminations. In other cases, branching logic in the survey determines which questions are shown to respondents and will also causes the stated sample size to decrease for some questions.

## Full survey reference (spreadsheet)

### Survey reference model (spreadsheet tab)

The survey reference model (SRM) tab contains the survey questionnaire used to program the survey and the branching logic for disqualifying, filtering, or determining which questions a respondent sees. Our data collection tool is programmed from the SRM. Column Headings in SRM tab are defined as follows.

* **P# (Page number).** The Page within our data collection tool that contains the referenced content (question, text, or page title) defined in the SRM.
* **Q# (Question number).** A sequential integer value starting at "1" and increments by “1” that uniquely identifies the survey question number. SRM questions numbers correspond to the question numbers shown in survey instrument.
* **R# (Response Number).** Values (generally numeric integers) that conforms to how our data collection tool encodes survey response values.
* **Questions and Responses.** Each question that is asked to respondents and allowable responses to each question. Also can contain contextual information such as the introduction to the survey or instructions to the respondent.
* **Show/Hide Rules and Logic.** Branching rules for screening or disqualifying respondents and navigation through survey questions based on prior responses.
* **Survey review/Coding notes/Changes made.** Commentary on how to addressing survey programming.

### Filtering Criteria (spreadsheet tab)

The questions and responses that were used by Linux Foundation Research to define what responses to keep, disqualify (DQ), or filter out prior to our analysis and reporting. Records that fail to meet basic data quality standards are deleted from the dataset. The purpose of these filter criteria are so that the user is able to identify the exact data set that Linux Foundation Research used for analysis in case you want to use our analysis as a baseline.

Raw data expanded text (spreadsheet tab)

All personal or private information has been removed from this dataset. This tab contains a column for each survey response with the textual responses made by respondents (rows).

* **Row 1.** Columns entries contain the first occurrence of the Q# of the survey question.
* **Row 2.** Columns entries the description of the survey question asked.
* **Row 3.** Column entries include each response description to the question that was asked. The number of columns per question depend on the number of responses possible for the question.
* **Row 4 🡪 Row N.** Each row includes the response text selected by the respondent to each question.

Raw data expanded numeric (spreadsheet tab)

All personal or private information has been removed from this dataset. This tab contains a column for each survey response with the numerically encoded responses made by respondents (rows).

* **Row 1.** Columns entries contain the first occurrence of the Q# of the survey question.
* **Row 2.** Columns entries the description of the survey question asked.
* **Row 3.** Column entries include each response description to the question that was asked. The number of columns per question depend on the number of responses possible for the question.
* **Row 4 🡪 Row N.** Each row includes the numeric encoding selected by the respondent to each question.

If you have any questions regarding use of this dataset, please do not hesitate to contact Linux Foundation Research using the following link: https://www.linuxfoundation.org/about/contact