

# Pre-interview survey questions

Intro:

Welcome! You are being asked to complete this survey as part of a research project. This project is being conducted by a graduate student, as a part of a course, to examine dataset search behavior for machine learning datasets.

The survey is 14 questions long and will take anywhere from 5-10 minutes to complete.

You can skip questions that you do not want to answer if the question is not marked mandatory. You must be age 18 or older to participate in the study. There are no known risks or personal benefits to completing this survey. Completing the survey is voluntary and neither completing nor declining to complete the survey will affect standing in relation to ESIP.

Identifiable data will be collected. Participants determined to not be eligible for the survey will have their responses stored for the duration of the study. The survey is not anonymous. Your responses to the survey will be linked to your name and email address. Information provided in this survey can only be kept as secure as any other online communication. Information collected for this study will be used in analyses and produced into a course final report and potentially presentation deliverables for venues such as the 2025 July ESIP conference.

Thank you.

Screeners questions (Yes/No):

- a. Do you have a computer with internet access? **(If no, terminate).**
- b. Are you a machine learning practitioner who uses ML in their regular job? **(If no, terminate).**

- c. Are you willing to have your screen recorded during the interview process? **(If no, terminate)**
- d. “This study will take around 10 minutes for a pre-interview survey and 35 minutes for an interview to describe your experience with ML data searchability and usability (45 minute total maximum). Would this work for you?” **(If no, terminate)**

[if “no” to any of the screening questions]

Based on your responses, you do not qualify for this study. If you would like to be reconsidered for this study, you can write your email below and the research team can discuss any of your concerns.

Identifying questions:

- 1. What is your name (first, last)?
- 2. What is the best email address to contact you?

General questions:

- 3. What is your current job title?
- 4. How many years have you been working professionally with machine learning? [integer type]
- 5. How often are you searching for machine learning training data as part of your job?
  - a. [everyday or every other day]
  - b. [once or twice a week]
  - c. [once or twice a month]
  - d. [once every few months]
  - e. [never]

- i. **Researcher note:** This can create our subgroups of users by frequency of search. More frequent searchers may be asking more specific

questions and have different preferences than someone who doesn't search as much.

6. Are you familiar with Google's Dataset Search? [Yes or No]
7. How frequently do you use Google's Dataset Search? [only serve if "Yes" to previous question]
  - a. [I use Google's Dataset Search everyday or every other day]
  - b. [I Use Google's Dataset Search once or twice a week]
  - c. [I use Google's Dataset Search once or twice a month]
  - d. [I use Google's Dataset Search once every few months]
  - e. [I have never used Google's Dataset Search]
    - i. **Researcher note:** This can create our subgroups of users (novice to experienced). If not, good background information for navigation.
8. Are you familiar with the Croissant metadata schema? [Yes or No]
9. How frequently do you use the Croissant metadata schema? [only serve if "yes" to previous question]
  - a. [I use the Croissant metadata schema everyday or every other day]
  - b. [I Use the Croissant metadata schema once or twice a week]
  - c. [I use the Croissant metadata schema once or twice a month]
  - d. [I use the Croissant metadata schema once every few months]
  - e. [I have never used the Croissant metadata schema]
    - i. **Researcher note:** This can create our subgroups of users (novice to experienced).

Element importance questions:

1. When searching for an ML dataset to use in your work, what metadata elements here would help you determine if the dataset is relevant and/or trustworthy?

- a. Rank the importance of each of these metadata elements in evaluating the usability of a given machine learning dataset.
- i. [Very unimportant]
  - ii. [Unimportant]
  - iii. [Neither important or unimportant]
  - iv. [Important]
  - v. [Very important]

Element	Definition [moderator view only]
name	The name of the dataset.
url	The URL of the dataset. This generally corresponds to the Web page for the dataset.
description	Description of the dataset.
license	The license of the dataset. Croissant recommends using the URL of a known license, e.g., one of the licenses listed at <a href="https://spdx.org/licenses/">https://spdx.org/licenses/</a> .
creator	The creator(s) of the dataset.
datePublished	The date the dataset was published.
keywords	A set of keywords associated with the dataset, either as free text, or a DefinedTerm with a formal definition.
publisher	The publisher of the dataset, which may be distinct from its creator.
version	The version of the dataset following the requirements below.
dateCreated	The date the dataset was initially created.
dateModified	The date the dataset was last modified.
sameAs	The URL of another Web resource that represents the same dataset as this one.

inLanguage	The language(s) of the content of the dataset.
Distribution (file representation/format)	By contrast with schema.org/Dataset, Croissant requires the distribution property to have values of type FileObject or FileSet.
isLiveDataset	Whether the dataset is a live dataset.
citeAs	"A citation to the dataset itself, or a citation for a publication that describes the dataset. Ideally, citations should be expressed using the bibtex format." Note that this is different from schema.org/citation, which is used to make a citation to another publication from this dataset.

- vi. **Researcher note:** We won't have any control over the metadata and datasets they pick so doing this structured task before they've completed the first semi-structured task may help them think about usefulness.
- vii. **Researcher note:** Ranking will provide a general idea of people's determination of relevance and how they evaluate metadata.
- b. Add any fields you don't see that you think would be useful for including in a metadata description of a machine learning dataset.
- 2. When searching for a **spatial** ML dataset to use in your work, what metadata elements help you determine if the dataset is relevant and/or trustworthy?
  - a. Rank the importance of each of these metadata elements in evaluating the usability of a given **spatial** machine learning dataset.
    - i. [Very unimportant]
    - ii. [Unimportant]
    - iii. [Neither important or unimportant]
    - iv. [Important]

v. [Very important]

Element	Definition [moderator view only]
title	A human readable title describing the STAC entity
description	Detailed multi-line description to fully explain the STAC entity
Datetime of acquisition	The searchable date and time of the assets, which must be in UTC
Mode of Acquisition (e.g., Handheld, Spacebourne)	
Sensor (e.g., optical, thermal)	
Sensing Type (e.g., active, passive)	
Wavelength of sensor	
Topology (e.g., image overlap, point density)	
Model category (e.g., classification)	
Scale of processing (e.g., pixel, object)	
Training type (e.g., training, validation)	
File format (is it cloud native?)	
Annotation type (e.g., mask, bbox)	
Bounding box (bbox)	Bounding Box of the asset represented by this Item using either 2D or 3D geometries, formatted according to RFC 7946 (GeoJSON), section 5
geometry/model (e.g., point, grid)	Defines the full footprint of the asset represented by this item, formatted according to RFC 7946 (GeoJSON)
Distribution (file representation/format)	1. An object that contains a URI to data associated with the Item that can be downloaded or streamed, 2. A physical embodiment of the Dataset in a particular format
temporalCoverage (extent)	Typically indicates the relevant time period in

	a precise notation
spatialCoverage (extent)	Indicates areas that the dataset describes

- b. Add any fields you don't see that you think would be useful for including in a metadata description of a **spatial** machine learning dataset.

#### Scheduling:

For the interview portion of the study, please select 3 times you are available to meet online:

[individual selections of dates with hour slots]

Date range:

Potential Hours: 12 pm to 7 pm (specific selection on a day varies based on moderator's schedule)

The study moderator will select one date you provided and email you with a Zoom meeting Google Calendar invitation within 48 hours of survey submission.

If you have any questions or concerns, please email the moderator, Joseph Edgerton at

[either from termination of screening questions or from valid completion of the survey]

We thank you for your time spent taking this survey. Your response has been recorded.