The Data Cards Playbook

A toolkit for purposeful and people-centric dataset documentation for transparency in AI systems.

https://pair-code.github.io/datacardsplaybook/

#datacardsplaybook

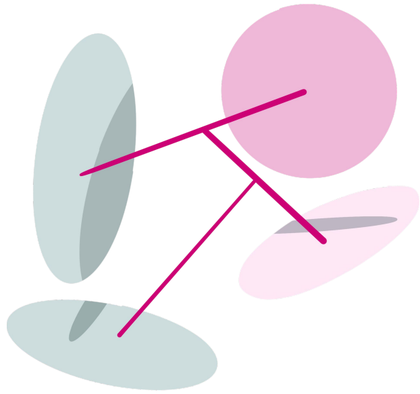
Introduction

01 Ask

02 Inspect

03 Answer

04 Audit



Your Data Card Brief

IN THIS SECTION

Define the scope, values, and vision for your Data Card(s) effort.

INSTRUCTIONS

Through a series of critical reflections and discussions, decide how Data Card(s) might help your datasets' and transparency goals.

OUTCOMES

A brief that describes why you're creating Data Card(s), what success looks like, and what's out of scope.

ACTIVITY LEVEL

Basic

Consider the following types of dataset documentation:

Observable

Shape and size of data, pipelines, access, licenses. Context and information that can be easily acquired from the data.

UTILITARIAN DOCUMENTATION

Explainable

Documentation ++ Processes, rules, rationales that shape the data. Context and information that cannot be learned from the data.

TRANSPARENT DOCUMENTATION

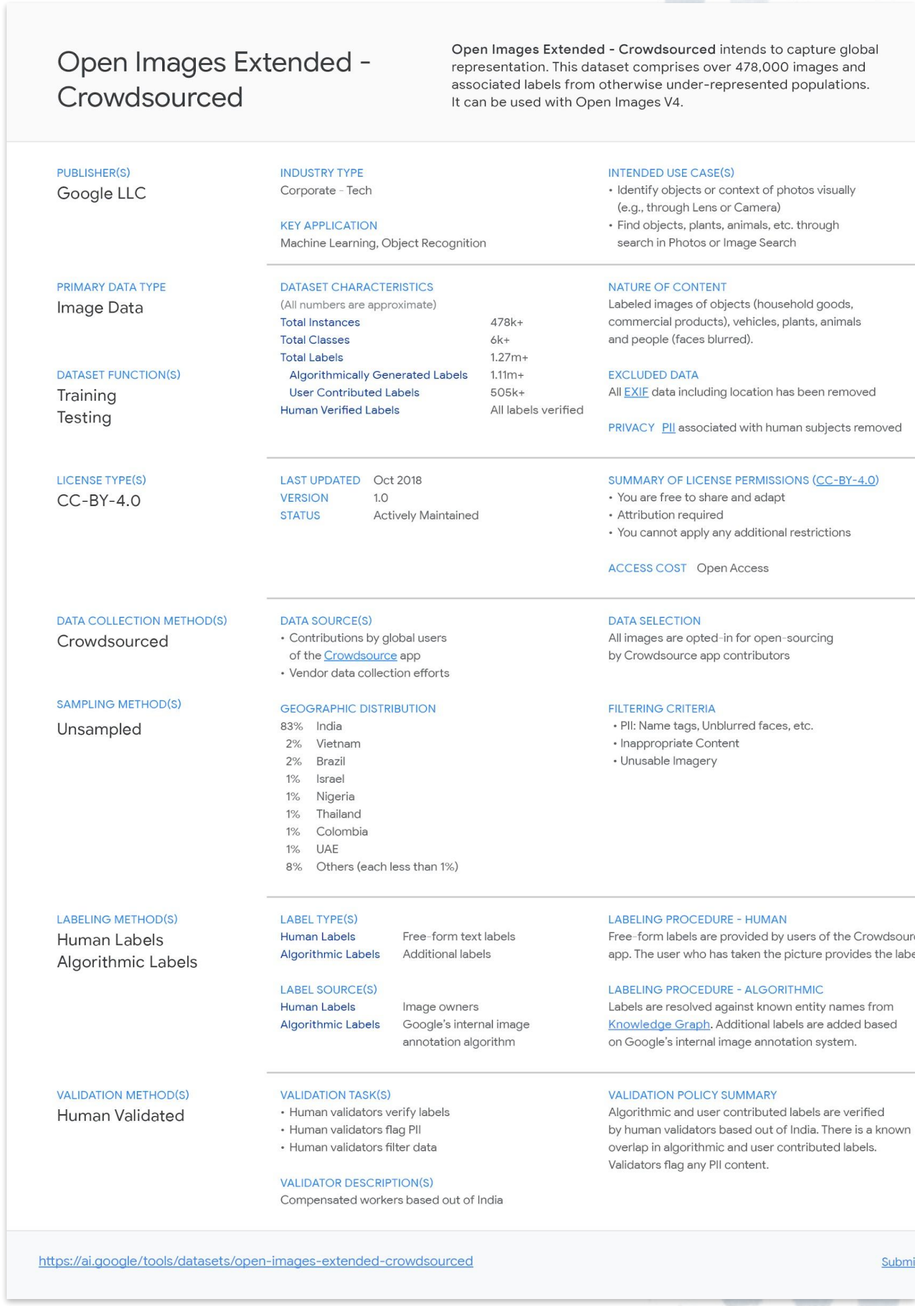
Understandable

Transparent documentation **intentionally written for humans** and required for making responsible decisions about dataset use.

RESPONSIBLE DOCUMENTATION

Data Cards summarize critical information about datasets that help people to make informed decisions about how data is used in ML systems for product, policy, and research.

Translate your definition of transparency into the scope and utility for Data Cards.



Open Images Extended - Crowdsourced

Open Images Extended - Crowdsourced intends to capture global representation. This dataset comprises over 478,000 images and associated labels from otherwise under-represented populations. It can be used with Open Images V4.

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

KEY APPLICATION
Machine Learning, Object Recognition

INTENDED USE CASE(S)
- Identify objects or context of photos visually (e.g., through Lens or Camera)
- Find objects, plants, animals, etc. through search in Photos or Image Search

PRIMARY DATA TYPE
Image Data

DATASET CHARACTERISTICS
(All numbers are approximate)

Total Instances	478k+
Total Classes	6k+
Total Labels	1.27m+
Algorithmically Generated Labels	1.11m+
User Contributed Labels	505k+
Human Verified Labels	All labels verified

NATURE OF CONTENT
Labeled images of objects (household goods, commercial products), vehicles, plants, animals and people (faces blurred).

EXCLUDED DATA
All EXIF data including location has been removed

PRIVACY PII associated with human subjects removed

DATASET FUNCTION(S)
Training
Testing

LICENSE TYPE(S)
CC-BY-4.0

LAST UPDATED Oct 2018
VERSION 1.0
STATUS Actively Maintained

SUMMARY OF LICENSE PERMISSIONS (CC-BY-4.0)
- You are free to share and adapt
- Attribution required
- You cannot apply any additional restrictions

ACCESS COST Open Access

DATA COLLECTION METHOD(S)
Crowdsourced

DATA SOURCE(S)
- Contributions by global users of the Crowdsource app
- Vendor data collection efforts

DATA SELECTION
All images are opted-in for open-sourcing by Crowdsource app contributors

SAMPLING METHOD(S)
Unsampled

GEOGRAPHIC DISTRIBUTION

83%	India
2%	Vietnam
2%	Brazil
1%	Israel
1%	Nigeria
1%	Thailand
1%	Colombia
1%	UAE
8%	Others (each less than 1%)

FILTERING CRITERIA
- PII: Name tags, Unblurred faces, etc.
- Inappropriate Content
- Unusable Imagery

LABELING METHOD(S)
Human Labels
Algorithmic Labels

LABEL TYPE(S)

Human Labels	Free-form text labels
Algorithmic Labels	Additional labels

LABELING PROCEDURE - HUMAN
Free-form labels are provided by users of the Crowdsourc app. The user who has taken the picture provides the labe

LABEL SOURCE(S)

Human Labels	Image owners
Algorithmic Labels	Google's internal image annotation algorithm

LABELING PROCEDURE - ALGORITHMIC
Labels are resolved against known entity names from Knowledge Graph. Additional labels are added based on Google's internal image annotation system.

VALIDATION METHOD(S)
Human Validated

VALIDATION TASK(S)
- Human validators verify labels
- Human validators flag PII
- Human validators filter data

VALIDATION POLICY SUMMARY
Algorithmic and user contributed labels are verified by human validators based out of India. There is a known overlap in algorithmic and user contributed labels. Validators flag any PII content.

VALIDATOR DESCRIPTION(S)
Compensated workers based out of India

https://ai.google/tools/datasets/open-images-extended-crowdsourced

Submi

OPEN IMAGES EXTENDED - CROWDSOURCED DATA CARD ↗

As a group, answer the following questions:

What are the top 3 goals of your dataset(s)?

What are the top 3 goals of your Data Card(s)?

3 circumstances that describe the experience of transparency that your Data Card(s) should deliver?

3 circumstances in which your Data Card(s) are a wasted investment?

5 min

What are the top 3 goals of your dataset(s)?

Consider: What motivates you to create them? What does success of your dataset look like? Is it adoption or purchase? Improvement of production systems? Furthering a research agenda?

1. ✍️

2. ✍️

3. ✍️

5 min

What are the top 3 goals of your Data Card(s)?

Consider: How does the goal of your Data Card support the goals of your dataset(s)? How must a Data Card be used so that it contributes to the success of your dataset(s)?

1. ✍️

2. ✍️

3. ✍️

5 min

3 circumstances that describe success for your Data Card(s)?

Think about your own experience of transparency, the experience of using dataset documentation, and successful outcomes of transparency in data.

1. ✍️

2. ✍️

3. ✍️

5
min

3 circumstances in which your Data Card(s) are a wasted investment?

Consider obvious and mundane tasks, or easily preventable failures of Data Cards. What type of documentation included in a Data Card might make it worthwhile?

1. ✍️

2. ✍️

3. ✍️

15 min

Your Data Card Brief

By using Data Cards, we hope to...

✍️ short statements that describe major use cases, objectives and value of Data Cards

Successful Data Cards look like ...

✍️ explicit statements for measuring the success of implementing Data cards, telling us if they have met our objectives.

Non-goals of Data Cards are...

✍️ statements about what has been purposefully excluded from your Data Card efforts, and things that should not be related.

Checklist

YOU SHOULD NOW HAVE DEFINED

–

- The goals and objectives of your dataset(s)
- The goals and objectives of your Data Card(s)
- What’s in scope and out of scope for your Data Card(s)
- When your Data Card(s) do not work

#datacardsplaybook

The Data Cards Playbook ↗ is an adaptable toolkit of participatory activities, conceptual frameworks, and guidance that support Responsible AI practices for transparency in dataset documentation.

If you've adapted, implemented, or have feedback for this guidance, we'd love to hear from you at https://github.com/pair-code/datacardsplaybook ↗.

Find the complete playbook at
https://pair-code.github.io/datacardsplaybook ↗