Data Cards: Purposeful and Transparent Documentation for Responsible AI

Mahima Pushkarna
Google Research
mahimap@google.com

Andrew Zaldivar
Google Research
andrewzaldivar@google.com

Abstract

As we move towards large-scale models capable of numerous downstream tasks, the complexity of understanding datasets that give nuance to models rapidly increases. A clear and thorough understanding of a dataset's origins, development, intent, ethical considerations and evolution becomes a necessary step for the responsible and informed deployment of models, especially those in people-facing contexts and high-risk domains. However, the burden of this understanding often falls on the intelligibility, conciseness, and comprehensiveness of its documentation. In this position paper, we propose Data Cards for fostering transparent, purposeful and human-centered documentation of datasets within the practical contexts of industry and research. Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's life cycle for responsible AI development. They provide explanations of processes and rationales that shape the data and consequently the models—such as upstream sources, data collection and annotation methods; training and evaluation methods, intended use, or decisions affecting model performance. Using two case studies, we report on desirable characteristics that support adoption across domains, organizational structures, and audience groups.

1 Introduction

The challenge of transparency in machine learning (ML) models and datasets continues to receive increasing attention from academia and the industry [1, 2]. Often, the goal has been to attain greater visibility into ML models and datasets by exposing source code [5] and contribution trails [6] with diverse oversight [13]. However, attempts to introduce standardized, practical and sustainable mechanisms for transparency that can be used at scale often hit roadblocks. These reflect real world constraints of the diversity of goals, workflows, and backgrounds of individual stakeholders participating in the life cycles of datasets and artificial intelligence (AI) systems [8, 10, 11].

As a step towards these goals, we propose a new framework for transparent and purposeful documentation of datasets, called Data Cards. A Data Card contains a structured collection of summaries gathered over the life cycle of a dataset about observable and unobservable aspects needed for decisions in organizational and practice-oriented contexts. Beyond metadata, Data Cards include explanations, rationales, and instructions pertaining to the provenance, representation, usage, and fairness-informed evaluations of datasets for ML models. These artifacts emphasize information and context that shape the data, but cannot be inferred from the dataset directly.

This short position paper describes the design of Data Cards, and walks through two case studies that describe their creation and use as boundary objects in practice. Creators of Data Cards were able to discover surprising future opportunities to improve their dataset design decisions. Our results suggest that the creation of generative, evaluative, and structured framework around transparency artifacts

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

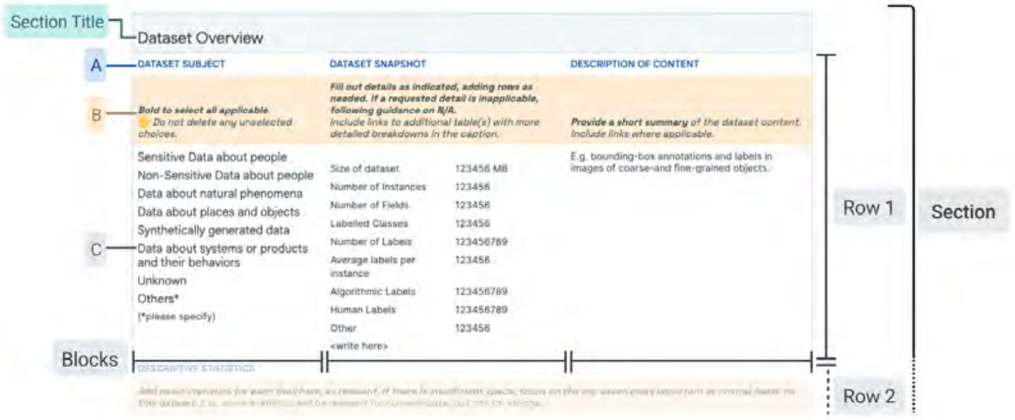

Section Title
Dataset Overview
A
DATASET SUBJECT
DATASET SNAPSHOT
DESCRIPTION OF CONTENT
B
Bold to select all applicable.
Do not delete any unselected choices.
Fill out details as indicated, adding rows as needed. If a requested detail is inapplicable, following guidance on N/A.
Include links to additional table(s) with more detailed breakdowns in the caption.
Provide a short summary of the dataset content. Include links where applicable.
Sensitive Data about people
Non-Sensitive Data about people
Data about natural phenomena
Data about places and objects
Synthetically generated data
C
Data about systems or products and their behaviors
Unknown
Others*
(*please specify)
Size of dataset 123456 MB
Number of Instances 123456
Number of Fields 123456
Labelled Classes 123456
Number of Labels 123456789
Average labels per instance 123456
Algorithmic Labels 123456789
Human Labels 123456789
Other 123456
<write here>
E.g. bounding-box annotations and labels in images of coarse-and fine-grained objects.
Row 1
Section
Blocks
DESCRIPTIVE STATISTICS
Row 2

Figure 1: **A Data Card Template Section:** This section is titled "Dataset Overview", and contains two rows. The first row has three blocks, whereas the second row spans the entire width of the section. Blocks contain (A) A Title, (B) A prompting question, and (C) predetermined choices or suggested answer structures.

is a powerful way of not only adding nuance to the dataset documentation process itself, but also introducing human-centric and responsible practices when using datasets in ML applications.

2 Data Cards

2.1 Methodology & Stakeholder Typology

Over the course of twenty four months, we worked primarily with ML dataset and model owners to produce prototypical transparency artifacts in a real-world setting. We conducted a series of internal and external studies (focus groups, workshops, and surveys) to arrive at our principles, design and structure behind Data Cards. Further, we developed a structured participatory workshop-based approach to engage cross-functional stakeholders when creating transparent metadata schema for dataset documentation [16].

We found that ***Producers*** of datasets and documentation often subscribed to a single, informal notion of "users" of Data Cards – loosely characterized by high data domain expertise, familiarity with similar datasets, and deep technical knowledge. In practice, we found that only a few readers or ***Agents*** actually meet all these requirements. After testing prototypes and proof of concepts with different groups, it became clear that Agents with operational and reviewer needs were distinct categories, and includes stakeholders who may never directly use the dataset, but will engage with a Data Card (e.g., non-technical subject matter experts). Additionally, Agents are distinct from ***Users***, who are individuals and representatives who interact with products that rely on models trained on dataset. Users require a significantly different set of explanations and controls grounded within product experiences. Together, we used this typology of stakeholders in the life cycle of datasets when conceptualizing Data Cards.

2.2 Principles, Design & Structure

Data Cards capture critical information about a dataset across its life cycle. Just as is with every dataset, each Data Card is unique, and no single template satisfactorily captures the nuance of all datasets. While most previous approaches take domain-specific [7] or prescriptive approaches [12, 15] to the creation of transparency artifacts, our novel contributions are the generative design of Data Cards as an underlying framework for transparency reporting, created for readability and scaling in production contexts. As such, Data Cards have been designed along the following principles:

- **Flexible:** Describe a wide range of datasets – live or static, datasets that are actively being curated from single or multiple sources, or those with multiple modalities.

- **Modular:** Organize documentation into meaningful sections that are self-contained repeatable units, able to provide an end-to-end description of a single aspect of the dataset.
- **Extensible:** Components that can be easily reconfigured or extended systematically for novel datasets, analyses, and platforms.
- **Accessible:** Represent content at multiple granularities so readers can efficiently find and effectively navigate detailed descriptions of the dataset.
- **Content-agnostic:** Support diverse media inputs in the form of multiple choice answers, long-form textual answers, visualizations, images, code blocks, tables, and other interactive elements.

For deployment at scale, Data Cards are designed to be extensible, platform- and dataset- agnostic. Content in Data Cards are displayed as *blocks* for easy implementation in interactive browser-based user interfaces and non-interactive platforms. In a Data Card template, these consist of a title, a prompting question with a description, and where relevant, a suggested answer structure to ensure accurate and comparable responses. Blocks are organized into *rows* describing a singular theme, and rows are further stacked into *sections* using meaningful and descriptive titles (Figure 1) that help Agents from different establish mental models for decision making. Fields are further organized into three columns, each describing information with varying levels of detail to support the evolving needs of multiple Agents from different backgrounds. Content increases in both depth and fidelity from left to right such that information within a row is self-contained and easy to find by navigating between rows.

The levels of abstraction in each column are instrumental for setting expectations of answers and scaling adoption. The highest levels describe the obvious and help Agents acquire an overview of the dataset. The intermediate-level abstraction provides technical and commonly observable details of the dataset. Finally, the lowest-level describes the human decisions and explanations in the dataset and why it matters. This provides the necessary context required to interpret quantitaitive characteristics and qualitative claims about the dataset. Together, these layers provide useful details that numerous Agents can understand, including benefits, limitations and corresponding risks, without losing sight of the broader context.

3 Case Studies from a Large Technology Company

3.1 A Computer Vision Dataset for Fairness Research

The first case study involved a research team that created an ML training dataset for computer vision (CV) fairness techniques that described sensitive attributes about people, such as perceived gender and perceived age-range. Using Open Images [14, 17], the dataset included 100,000 bounding boxes over 30,000 images. Each bounding box was manually annotated with perceived gender presentation and perceived age range presentation attributes. Given the risks associated with sensitive labels describing personal attributes weighed against the societal benefit of these labels for fairness analysis and bias mitigation, the team wanted an efficient way to provide an overview of the characteristics, limitations, and communicate acceptable uses of the dataset for internal ethics reviewers and external audiences.

Three groups were involved in the creation of this Data Card [3], which began after the dataset was created. The first group, the dataset authors (Producers), who had deep tacit knowledge of the processes and decisions across the dataset's life cycle, and explicit knowledge from analysis performed for the dataset release—they provided the core of the contents. However, this was distributed across several documents, and the Data Card served as an exercise in organizing knowledge into a "readable format" that could be repeated for multiple datasets.

The next group involved were internal reviewers (Agents) of the dataset and an accompanying paper, conducting an analysis of how the dataset aligns with responsible AI research and development practices. The analysis focused on subgroups in the labels, the trade-offs associated with each subgroup, and clarifying acceptable and unacceptable use cases of the dataset as a whole. The reviewers recommended that the team create a Data Card, which helped reveal differences in perception across experts. Reviewers were unable to ascertain if such discrepancies amongst experts was acceptable, and subsequent conversations raised further questions about the criteria used to label a bounding box with ’unknown’ perceived age-range. As a result, Producers added a custom section about bounding boxes to the Data Card, and created additional supporting visualizations.

The last group were the authors of this paper (also Agents), who provided human-centered design perspectives on the Data Card. Feedback was primarily geared towards uncovering information needed for acceptable conclusions about the accountability, risk, recommendations, uses, consequences, and quality of the dataset. A post-launch retrospective revealed that though the Producers did not have access to dataset consumers (Users), downstream Agents reported finding the Data Card useful, and sought templates for their own use.

3.2 A Geographically Diverse Dataset for Language Translation

The second case study comprised of a team of software engineers and a product manager. They noticed certain models were “picking up” names to define a person’s gender. Upon investigation, they found that previous training datasets did not have sufficient names that were uncommon in English or belonged to a non-American geography. They also found that model creators were making assumptions about these datasets. In response, the team decided to create a geographically diverse dataset from a limited set of publicly curated data from Wikipedia.

However, it became clear that a truly diverse dataset would need to consider race, age, gender, background and profession as well. While countries were acceptable proxies for geographic representation, gender would need to be inferred from the entity descriptions. Without an awareness of the goals of the dataset or the definitions of gender in the data design, the team was concerned that model creators could make assumptions leading to inappropriate dataset use. To communicate these two aspects, the team created a Data Card [4] for readers with and without technical expertise.

Experts responsible for the design, data extraction, cleaning and curation of the dataset worked with a human-centered designer in an iterative process to produce the Data Card after the dataset was created. While the documentation process itself took approximately 20 hours, the Data Card prompted the team to reflect on how data was selected, reviewed and created. In particular, experts stakeholders pointed out that gender is difficult to ascertain in the dataset. These conversations helped the team agree on a definition of perceived gender that relied on gender-indicative terms within the text of the data, using the labels “masculine”, “feminine”, and “neutral” for biographies describing collections of individuals. The team also found that discussions around the Data Card were actually about the dataset, and noted the usefulness of this feedback if received during the design stage.

4 Discussion & Conclusion

While both teams appreciated the transparency added to their respective datasets, creating Data Cards as a final step significantly increased the perception of work required. Rather than a post-implementation task, creating Data Cards as the dataset is created offers several benefits. First, it enables the inclusion of multiple perspectives (engineering, research, user experience, legal, and ethical) to enhance the readability and relevance of documentation, and the dataset quality over time. Then, it forces the aggregation of disparate documentation across the dataset life cycle into a single, ground truth document accessible to stakeholders. Lastly, it facilitates early feedback on responsible AI practices from experts and non-experts that can affect data design and analyses. Of note, teams that developed multiple Data Cards over a period started developing a nuanced vocabulary to express uncertainty that accurately reflected the status of the information.

A limitation of our approach was the use of Google Docs to provide Data Card templates. While this format afforded collaboration across stakeholders and preserved a forensic history of the development of the Data Card through, producers were limited to answering questions in the template with text, tables and images. Iterating on individual fields caused template fragmentation and the loss of the original intent, as observed in our first case study. Additionally, this format did not afford the capacity to automate responses, a much requested feature from dataset producers. Interestingly, we observed that readers had strong opinions about *not* automating certain fields in the Data Card, especially when it contained assumptions and rationales that supported the interpretation of results.

Future work requires a more principled approach for extending and adapting a Data Card template that preserve comparability. Insights from studies call for a participatory approach that can engage diverse, non-traditional stakeholders early into both, the dataset and Data Card development process.

Further, we believe that adopting a co-creative approach that spans the entire dataset and eventually, model life cycle will result in a deliberate approach to automation in documentation. Automated fields should be optimized for accuracy and antifragility of content at any given point in time, preventing the misrepresentation and the subsequent legitimizing of poor quality datasets. We also believe that implicit knowledge cannot be automated, and therefore, demand human-written explanations of methods, assumptions, decisions and baselines within context—explanations generated by approaches that help examine implicit knowledge [9]. Lastly, defining quantitative measures to assess the true value of Data Cards will require adoption at both breadth and depth in the industry. Data Cards can be powerful vehicles that emphasize the ethical considerations of a dataset in ways that can be practically acted upon and support production and research decisions, supporting the data-centric development of large AI models that support multiple, user-facing tasks.

Acknowledgments and Disclosure of Funding

The authors thank Aybuke Turker, Candice Schumman, Reena Jana, Romina Stella, and Susanna Ricco as they contributed to the research that went into the Data Cards template and the case studies. The authors also thank Michael Terry, Meredith Morris, Negar Rostamzadeh, Oddur Kjartansson and Tulsee Doshi, as well as three anonymous reviewers, for their comments on this paper. Mahima Pushkarna and Andrew Zaldivar contributed equally.

References

[1] *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).* https://facctconference.org/

[2] *AI Now Institute.* https://ainowinstitute.org/

[3] *Open Images Extended More Inclusive Annotations for People (MIAP) Data Card.* https://storage.googleapis.com/openimages/open_images_extended_miap/Open%20Images%20Extended%20-%20MIAP%20-%20Data%20Card.pdf

[4] *Translated Wikipedia Biographies Data Card.* https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/Data%20Card.pdf

[5] Antunes, Nuno ; Balby, Leandro ; Figueiredo, Flavio ; Lourenco, Nuno ; Meira, Wagner ; Santos, Walter: Fairness and transparency of machine learning for trustworthy cloud services. In: *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)* IEEE, 2018, S. 188–193

[6] Barclay, Iain ; Taylor, Harrison ; Preece, Alun ; Taylor, Ian ; Verma, Dinesh ; Mel, Geeth de: A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions. In: *Concurrency and Computation: Practice and Experience* (2020), S. e6129

[7] Bender, Emily M. ; Friedman, Batya: Data statements for natural language processing: Toward mitigating system bias and enabling better science. In: *Transactions of the Association for Computational Linguistics* 6 (2018), S. 587–604

[8] Chander, Ajay ; Srinivasan, Ramya ; Chelian, Suhas ; Wang, Jun ; Uchino, Kanji: Working with beliefs: AI transparency in the enterprise. In: *IUI Workshops*, 2018

[9] Denton, Emily ; Hanna, Alex ; Amironesei, Razvan ; Smart, Andrew ; Nicole, Hilary: On the genealogy of machine learning datasets: A critical history of ImageNet. In: *Big Data & Society* 8 (2021), Nr. 2, S. 20539517211035955

[10] Ehsan, Upol ; Liao, Q V. ; Muller, Michael ; Riedl, Mark O. ; Weisz, Justin D.: Expanding explainability: Towards social transparency in ai systems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, S. 1–19

[11] Felzmann, Heike ; Fosch-Villaronga, Eduard ; Lutz, Christoph ; Tamò-Larrieux, Aurelia: Towards transparency by design for artificial intelligence. In: *Science and Engineering Ethics* 26 (2020), Nr. 6, S. 3333–3361

[12] Gebru, Timnit ; Morgenstern, Jamie ; Vecchione, Briana ; Vaughan, Jennifer W. ; Wallach, Hanna ; Daumé III, Hal ; Crawford, Kate: Datasheets for datasets. In: *arXiv preprint arXiv:1803.09010* (2018)

[13] HUTCHINSON, Ben ; SMART, Andrew ; HANNA, Alex ; DENTON, Emily ; GREER, Christina ; KJARTANSSON, Oddur ; BARNES, Parker ; MITCHELL, Margaret: Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, S. 560–575

[14] KUZNETSOVA, Alina ; ROM, Hassan ; ALLDRIN, Neil ; UIJLINGS, Jasper ; KRASIN, Ivan ; PONT-TUSET, Jordi ; KAMALI, Shahab ; POPOV, Stefan ; MALLOCI, Matteo ; KOLESNIKOV, Alexander u. a.: The open images dataset v4. In: *International Journal of Computer Vision* 128 (2020), Nr. 7, S. 1956–1981

[15] MITCHELL, Margaret ; WU, Simone ; ZALDIVAR, Andrew ; BARNES, Parker ; VASSERMAN, Lucy ; HUTCHINSON, Ben ; SPITZER, Elena ; RAJI, Inioluwa D. ; GEBRU, Timnit: Model cards for model reporting. In: *Proceedings of the conference on fairness, accountability, and transparency*, 2019, S. 220–229

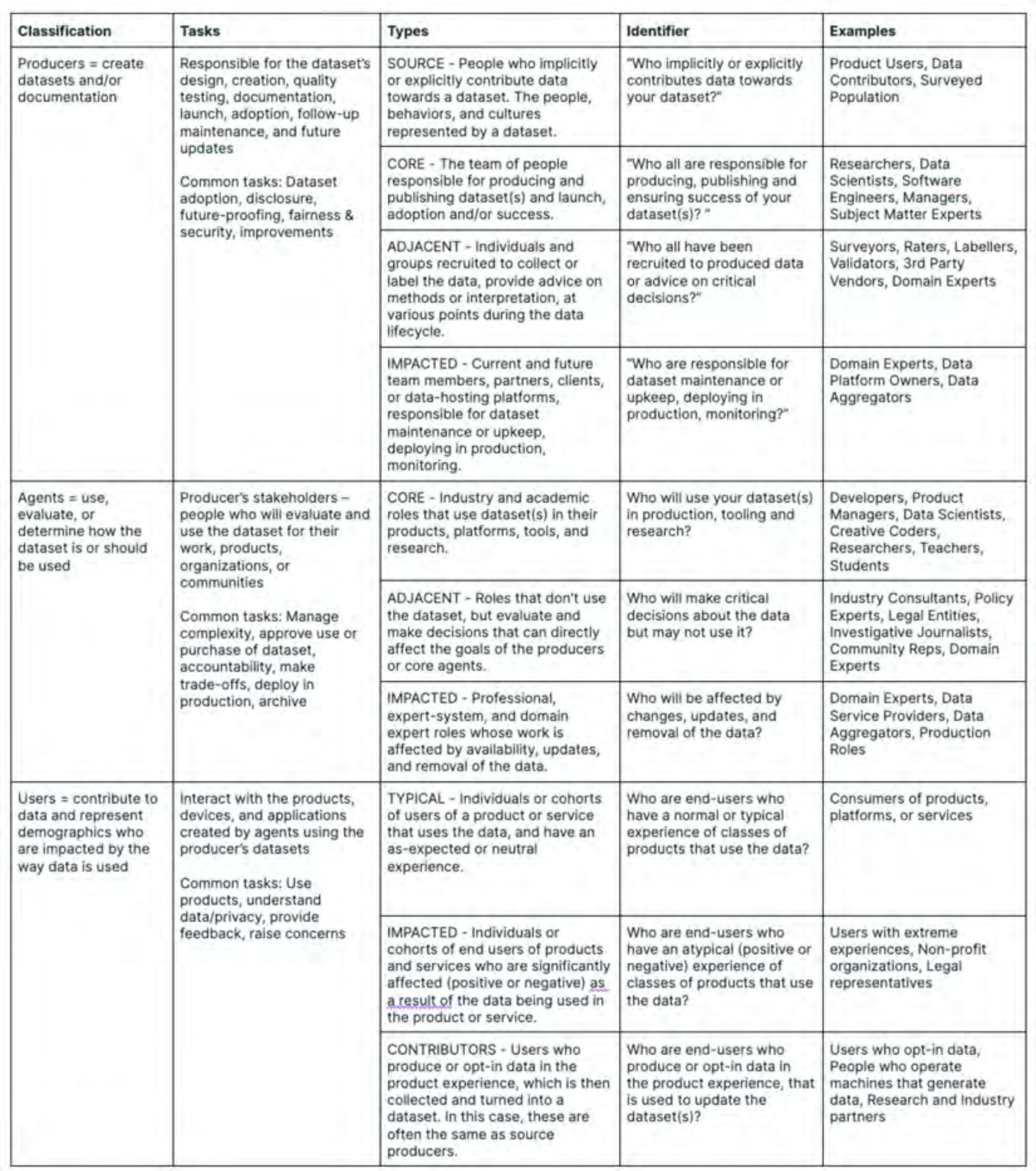
[16] PUSHKARNA, Mahima ; ZALDIVAR, Andrew ; NANAS, Daniel: *Data Cards Playbook: Participatory Activities for Dataset Documentation*. `https://facctconference.org/2021/acceptedcraftsessions.html#data_cards`

[17] SCHUMANN, Candice ; RICCO, Susanna ; PRABHU, Utsav ; FERRARI, Vittorio ; PANTOFARU, Caroline: A Step Toward More Inclusive People Annotations for Fairness. In: *arXiv preprint arXiv:2105.02317* (2021)

A Appendix

In this appendix, we provide a table describing our typology of stakeholders in the dataset's lifecycle, and the Data Cards that describe two datasets, each corresponding to the case studies: More Inclusively Annotated People [17, 3] (figure 3 to figure 9) and the Translated Wikipedia Biographies [4] (figure 10 to figure 12).

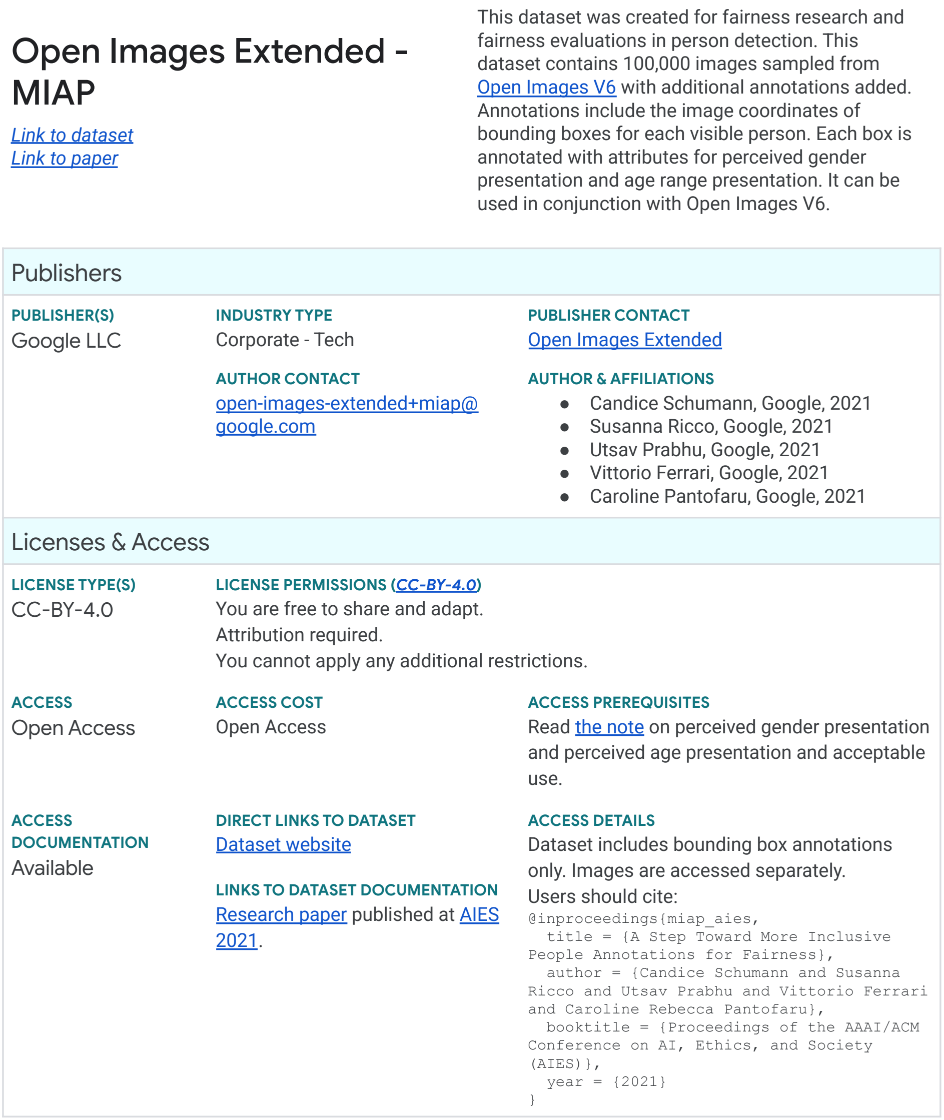
A.1 Disaggregated Typology of Stakeholders



Classification	Tasks	Types	Identifier	Examples
Producers = create datasets and/or documentation	Responsible for the dataset's design, creation, quality testing, documentation, launch, adoption, follow-up maintenance, and future updates Common tasks: Dataset adoption, disclosure, future-proofing, fairness & security, improvements	SOURCE - People who implicitly or explicitly contribute data towards a dataset. The people, behaviors, and cultures represented by a dataset.	"Who implicitly or explicitly contributes data towards your dataset?"	Product Users, Data Contributors, Surveyed Population
		CORE - The team of people responsible for producing and publishing dataset(s) and launch, adoption and/or success.	"Who all are responsible for producing, publishing and ensuring success of your dataset(s)? "	Researchers, Data Scientists, Software Engineers, Managers, Subject Matter Experts
		ADJACENT - Individuals and groups recruited to collect or label the data, provide advice on methods or interpretation, at various points during the data lifecycle.	"Who all have been recruited to produced data or advice on critical decisions?"	Surveyors, Raters, Labellers, Validators, 3rd Party Vendors, Domain Experts
		IMPACTED - Current and future team members, partners, clients, or data-hosting platforms, responsible for dataset maintenance or upkeep, deploying in production, monitoring.	"Who are responsible for dataset maintenance or upkeep, deploying in production, monitoring?"	Domain Experts, Data Platform Owners, Data Aggregators
Agents = use, evaluate, or determine how the dataset is or should be used	Producer's stakeholders – people who will evaluate and use the dataset for their work, products, organizations, or communities Common tasks: Manage complexity, approve use or purchase of dataset, accountability, make trade-offs, deploy in production, archive	CORE - Industry and academic roles that use dataset(s) in their products, platforms, tools, and research.	Who will use your dataset(s) in production, tooling and research?	Developers, Product Managers, Data Scientists, Creative Coders, Researchers, Teachers, Students
		ADJACENT - Roles that don't use the dataset, but evaluate and make decisions that can directly affect the goals of the producers or core agents.	Who will make critical decisions about the data but may not use it?	Industry Consultants, Policy Experts, Legal Entities, Investigative Journalists, Community Reps, Domain Experts
		IMPACTED - Professional, expert-system, and domain expert roles whose work is affected by availability, updates, and removal of the data.	Who will be affected by changes, updates, and removal of the data?	Domain Experts, Data Service Providers, Data Aggregators, Production Roles
Users = contribute to data and represent demographics who are impacted by the way data is used	Interact with the products, devices, and applications created by agents using the producer's datasets Common tasks: Use products, understand data/privacy, provide feedback, raise concerns	TYPICAL - Individuals or cohorts of users of a product or service that uses the data, and have an as-expected or neutral experience.	Who are end-users who have a normal or typical experience of classes of products that use the data?	Consumers of products, platforms, or services
		IMPACTED - Individuals or cohorts of end users of products and services who are significantly affected (positive or negative) as a result of the data being used in the product or service.	Who are end-users who have an atypical (positive or negative) experience of classes of products that use the data?	Users with extreme experiences, Non-profit organizations, Legal representatives
		CONTRIBUTORS - Users who produce or opt-in data in the product experience, which is then collected and turned into a dataset. In this case, these are often the same as source producers.	Who are end-users who produce or opt-in data in the product experience, that is used to update the dataset(s)?	Users who opt-in data, People who operate machines that generate data, Research and Industry partners

Figure 2: **A typology of typical stakeholders in the lifecycle of datasets that we created Data Cards for, broken down by type, identifiers and tasks with example roles.**

A.2 Data Card for Computer Vision Dataset



Open Images Extended - MIAP

Link to dataset
Link to paper

This dataset was created for fairness research and fairness evaluations in person detection. This dataset contains 100,000 images sampled from Open Images V6 with additional annotations added. Annotations include the image coordinates of bounding boxes for each visible person. Each box is annotated with attributes for perceived gender presentation and age range presentation. It can be used in conjunction with Open Images V6.

Publishers

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

AUTHOR CONTACT
open-images-extended+miap@google.com

PUBLISHER CONTACT
Open Images Extended

AUTHOR & AFFILIATIONS

- Candice Schumann, Google, 2021
- Susanna Ricco, Google, 2021
- Utsav Prabhu, Google, 2021
- Vittorio Ferrari, Google, 2021
- Caroline Pantofaru, Google, 2021

Licenses & Access

LICENSE TYPE(S)
CC-BY-4.0

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

ACCESS
Open Access

ACCESS COST
Open Access

ACCESS PREREQUISITES
Read the note on perceived gender presentation and perceived age presentation and acceptable use.

ACCESS DOCUMENTATION
Available

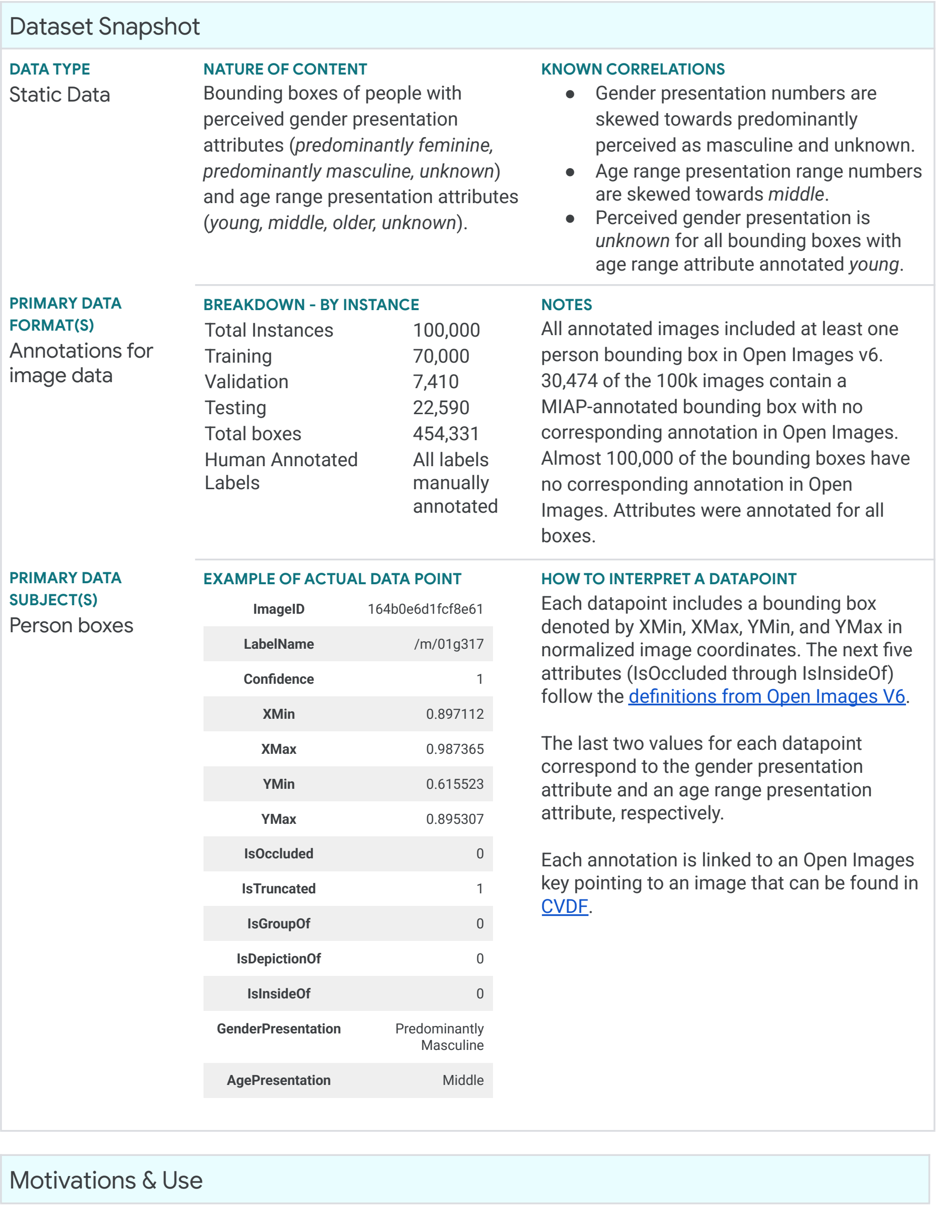
DIRECT LINKS TO DATASET
Dataset website

LINKS TO DATASET DOCUMENTATION
Research paper published at AIES 2021.

ACCESS DETAILS
Dataset includes bounding box annotations only. Images are accessed separately.
Users should cite:

```
@inproceedings{miap_aies,
  title = {A Step Toward More Inclusive
People Annotations for Fairness},
  author = {Candice Schumann and Susanna
Ricco and Utsav Prabhu and Vittorio Ferrari
and Caroline Rebecca Pantofaru},
  booktitle = {Proceedings of the AAAI/ACM
Conference on AI, Ethics, and Society
(AIES)},
  year = {2021}
}
```

Figure 3: **Data Card for Computer Vision Dataset, Page 1 of 7**



Dataset Snapshot

DATA TYPE
Static Data

NATURE OF CONTENT
Bounding boxes of people with perceived gender presentation attributes (*predominantly feminine, predominantly masculine, unknown*) and age range presentation attributes (*young, middle, older, unknown*).

KNOWN CORRELATIONS
- Gender presentation numbers are skewed towards predominantly perceived as masculine and unknown.
- Age range presentation range numbers are skewed towards *middle*.
- Perceived gender presentation is *unknown* for all bounding boxes with age range attribute annotated *young*.

PRIMARY DATA FORMAT(S)
Annotations for image data

BREAKDOWN - BY INSTANCE

Total Instances	100,000
Training	70,000
Validation	7,410
Testing	22,590
Total boxes	454,331
Human Annotated Labels	All labels manually annotated

NOTES
All annotated images included at least one person bounding box in Open Images v6. 30,474 of the 100k images contain a MIAP-annotated bounding box with no corresponding annotation in Open Images. Almost 100,000 of the bounding boxes have no corresponding annotation in Open Images. Attributes were annotated for all boxes.

PRIMARY DATA SUBJECT(S)
Person boxes

EXAMPLE OF ACTUAL DATA POINT

ImageID	164b0e6d1fcf8e61
LabelName	/m/01g317
Confidence	1
XMin	0.897112
XMax	0.987365
YMin	0.615523
YMax	0.895307
IsOccluded	0
IsTruncated	1
IsGroupOf	0
IsDepictionOf	0
IsInsideOf	0
GenderPresentation	Predominantly Masculine
AgePresentation	Middle

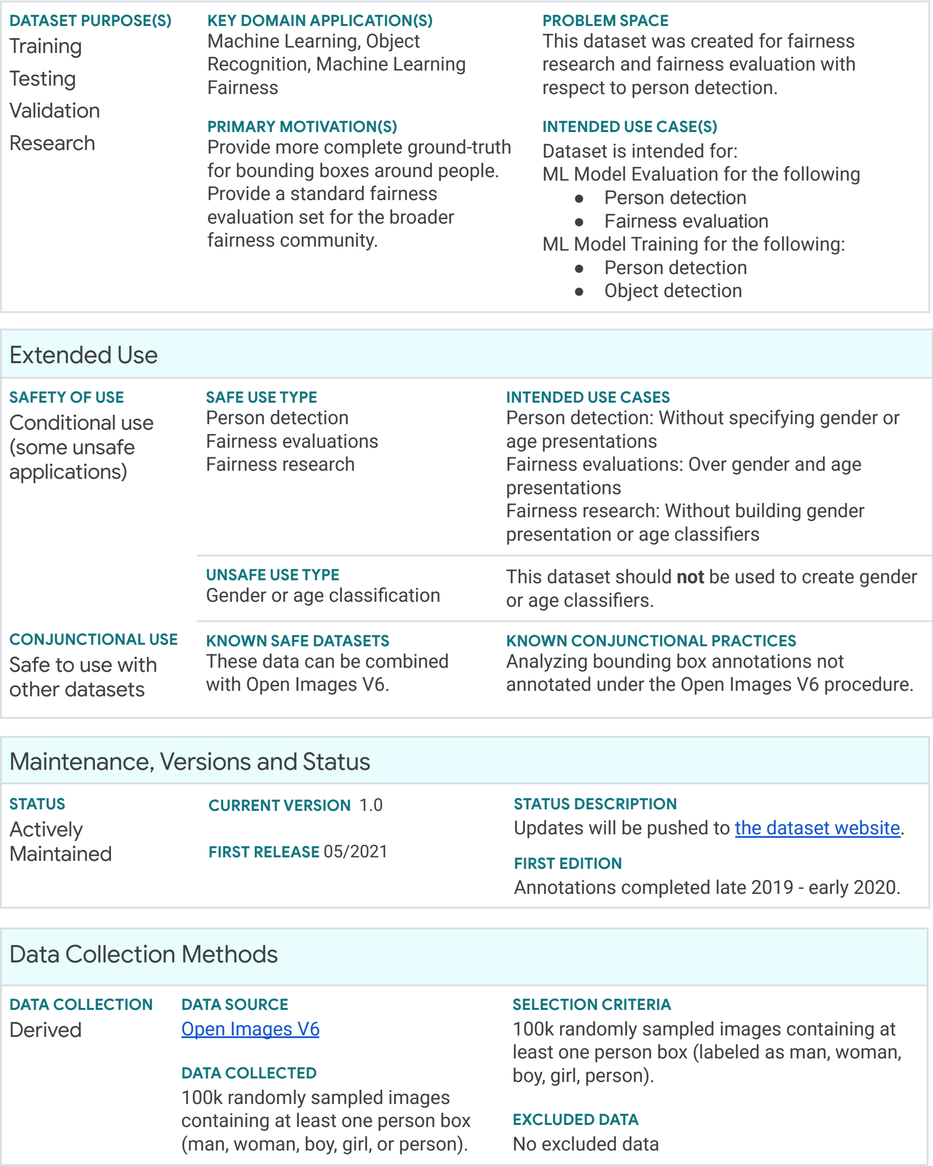
HOW TO INTERPRET A DATAPOINT
Each datapoint includes a bounding box denoted by XMin, XMax, YMin, and YMax in normalized image coordinates. The next five attributes (IsOccluded through IsInsideOf) follow the definitions from Open Images V6.

The last two values for each datapoint correspond to the gender presentation attribute and an age range presentation attribute, respectively.

Each annotation is linked to an Open Images key pointing to an image that can be found in CVDF.

Motivations & Use

Figure 4: **Data Card for Computer Vision Dataset, Page 2 of 7**



DATASET PURPOSE(S)
Training
Testing
Validation
Research

KEY DOMAIN APPLICATION(S)
Machine Learning, Object Recognition, Machine Learning Fairness

PRIMARY MOTIVATION(S)
Provide more complete ground-truth for bounding boxes around people. Provide a standard fairness evaluation set for the broader fairness community.

PROBLEM SPACE
This dataset was created for fairness research and fairness evaluation with respect to person detection.

INTENDED USE CASE(S)
Dataset is intended for:
ML Model Evaluation for the following
- Person detection
- Fairness evaluation

ML Model Training for the following:
- Person detection
- Object detection

Extended Use

SAFETY OF USE
Conditional use (some unsafe applications)

SAFE USE TYPE
Person detection
Fairness evaluations
Fairness research

INTENDED USE CASES
Person detection: Without specifying gender or age presentations
Fairness evaluations: Over gender and age presentations
Fairness research: Without building gender presentation or age classifiers

UNSAFE USE TYPE
Gender or age classification

This dataset should **not** be used to create gender or age classifiers.

CONJUNCTIONAL USE
Safe to use with other datasets

KNOWN SAFE DATASETS
These data can be combined with Open Images V6.

KNOWN CONJUNCTIONAL PRACTICES
Analyzing bounding box annotations not annotated under the Open Images V6 procedure.

Maintenance, Versions and Status

STATUS
Actively Maintained

CURRENT VERSION 1.0

FIRST RELEASE 05/2021

STATUS DESCRIPTION
Updates will be pushed to the dataset website.

FIRST EDITION
Annotations completed late 2019 - early 2020.

Data Collection Methods

DATA COLLECTION
Derived

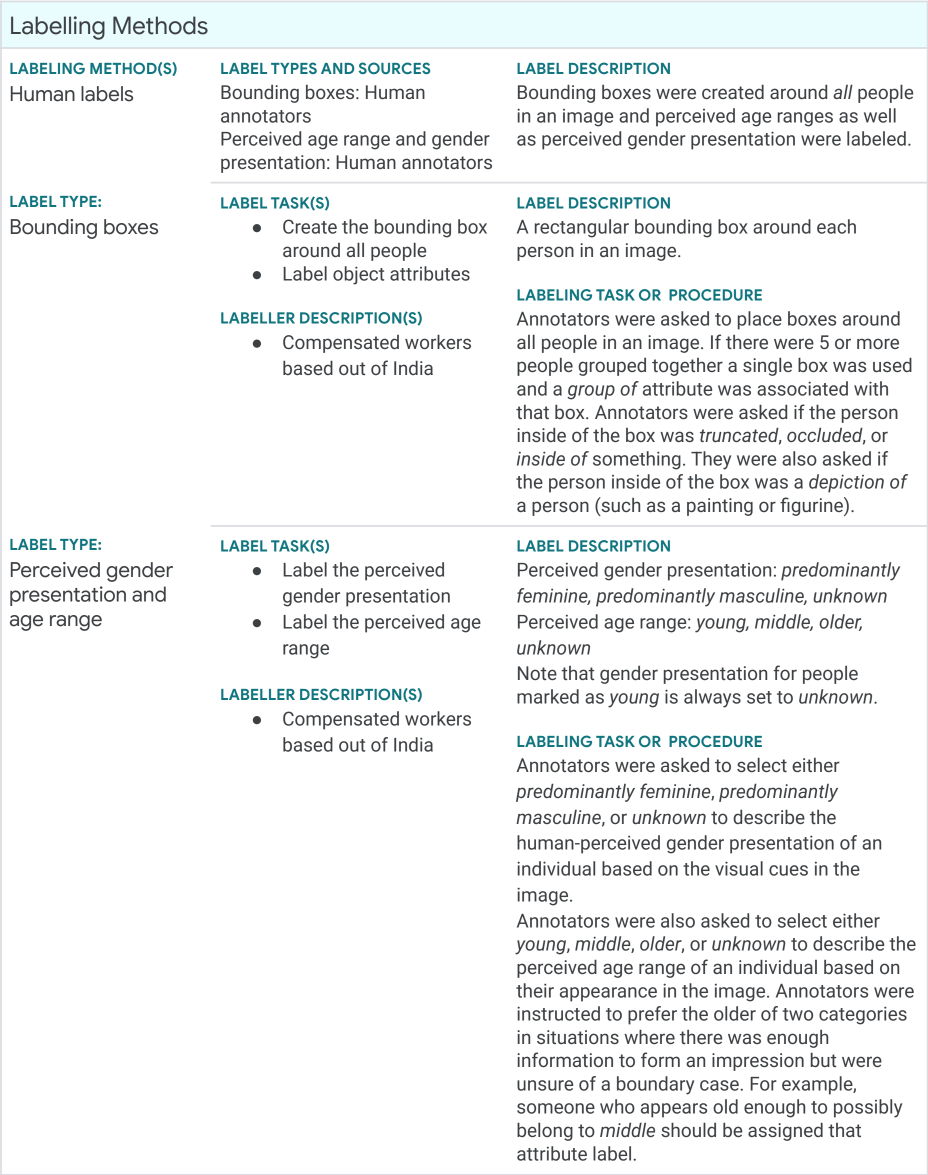
DATA SOURCE
Open Images V6

DATA COLLECTED
100k randomly sampled images containing at least one person box (man, woman, boy, girl, or person).

SELECTION CRITERIA
100k randomly sampled images containing at least one person box (labeled as man, woman, boy, girl, person).

EXCLUDED DATA
No excluded data

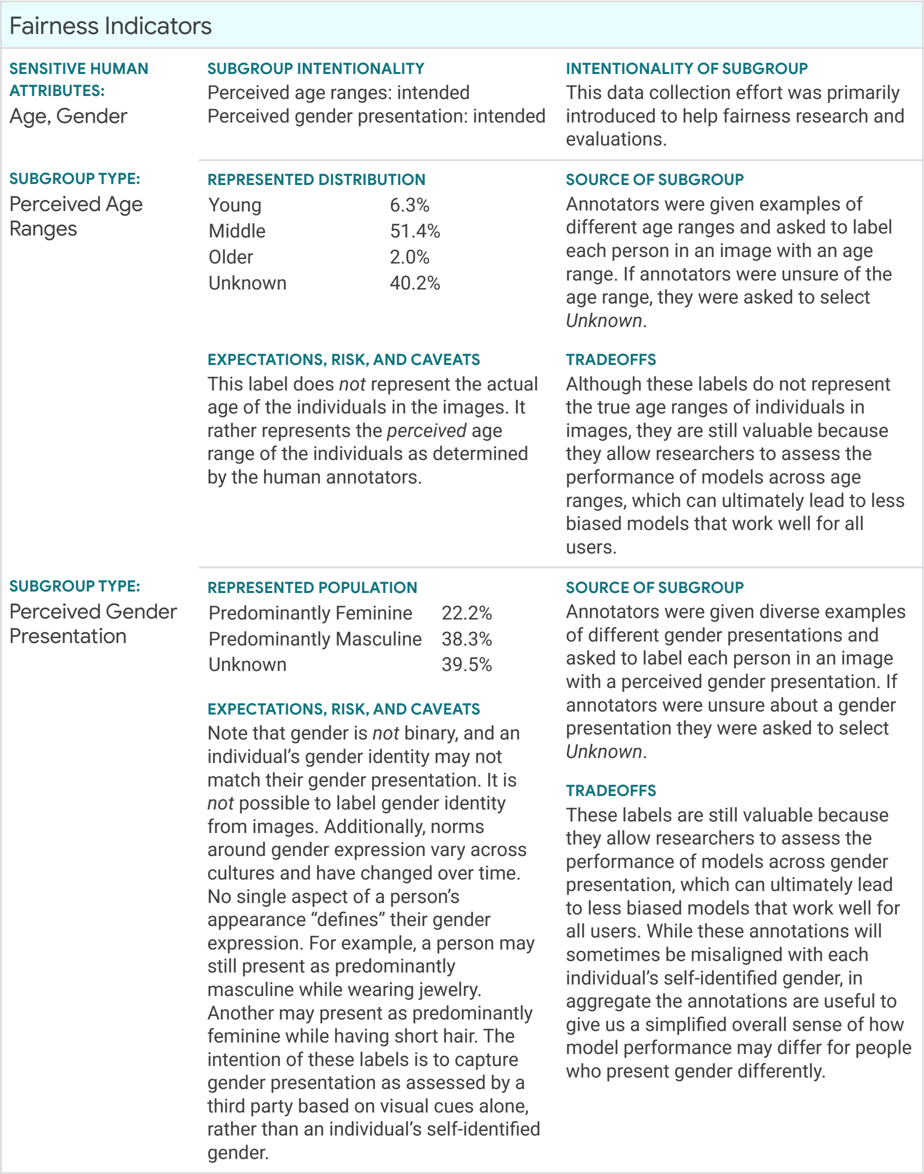
Figure 5: **Data Card for Computer Vision Dataset, Page 3 of 7**



Labelling Methods

LABELING METHOD(S) Human labels	**LABEL TYPES AND SOURCES** Bounding boxes: Human annotators Perceived age range and gender presentation: Human annotators	**LABEL DESCRIPTION** Bounding boxes were created around *all* people in an image and perceived age ranges as well as perceived gender presentation were labeled.
LABEL TYPE: Bounding boxes	**LABEL TASK(S)** ● Create the bounding box around all people ● Label object attributes **LABELLER DESCRIPTION(S)** ● Compensated workers based out of India	**LABEL DESCRIPTION** A rectangular bounding box around each person in an image. **LABELING TASK OR PROCEDURE** Annotators were asked to place boxes around all people in an image. If there were 5 or more people grouped together a single box was used and a *group of* attribute was associated with that box. Annotators were asked if the person inside of the box was *truncated*, *occluded*, or *inside of* something. They were also asked if the person inside of the box was a *depiction of* a person (such as a painting or figurine).
LABEL TYPE: Perceived gender presentation and age range	**LABEL TASK(S)** ● Label the perceived gender presentation ● Label the perceived age range **LABELLER DESCRIPTION(S)** ● Compensated workers based out of India	**LABEL DESCRIPTION** Perceived gender presentation: *predominantly feminine, predominantly masculine, unknown* Perceived age range: *young, middle, older, unknown* Note that gender presentation for people marked as *young* is always set to *unknown*. **LABELING TASK OR PROCEDURE** Annotators were asked to select either *predominantly feminine*, *predominantly masculine*, or *unknown* to describe the human-perceived gender presentation of an individual based on the visual cues in the image. Annotators were also asked to select either *young*, *middle*, *older*, or *unknown* to describe the perceived age range of an individual based on their appearance in the image. Annotators were instructed to prefer the older of two categories in situations where there was enough information to form an impression but were unsure of a boundary case. For example, someone who appears old enough to possibly belong to *middle* should be assigned that attribute label.

Figure 6: **Data Card for Computer Vision Dataset, Page 4 of 7**



Fairness Indicators

SENSITIVE HUMAN ATTRIBUTES:
Age, Gender

SUBGROUP INTENTIONALITY
Perceived age ranges: intended
Perceived gender presentation: intended

INTENTIONALITY OF SUBGROUP
This data collection effort was primarily introduced to help fairness research and evaluations.

SUBGROUP TYPE:
Perceived Age Ranges

REPRESENTED DISTRIBUTION

Young	6.3%
Middle	51.4%
Older	2.0%
Unknown	40.2%

SOURCE OF SUBGROUP
Annotators were given examples of different age ranges and asked to label each person in an image with an age range. If annotators were unsure of the age range, they were asked to select *Unknown*.

EXPECTATIONS, RISK, AND CAVEATS
This label does *not* represent the actual age of the individuals in the images. It rather represents the *perceived* age range of the individuals as determined by the human annotators.

TRADEOFFS
Although these labels do not represent the true age ranges of individuals in images, they are still valuable because they allow researchers to assess the performance of models across age ranges, which can ultimately lead to less biased models that work well for all users.

SUBGROUP TYPE:
Perceived Gender Presentation

REPRESENTED POPULATION

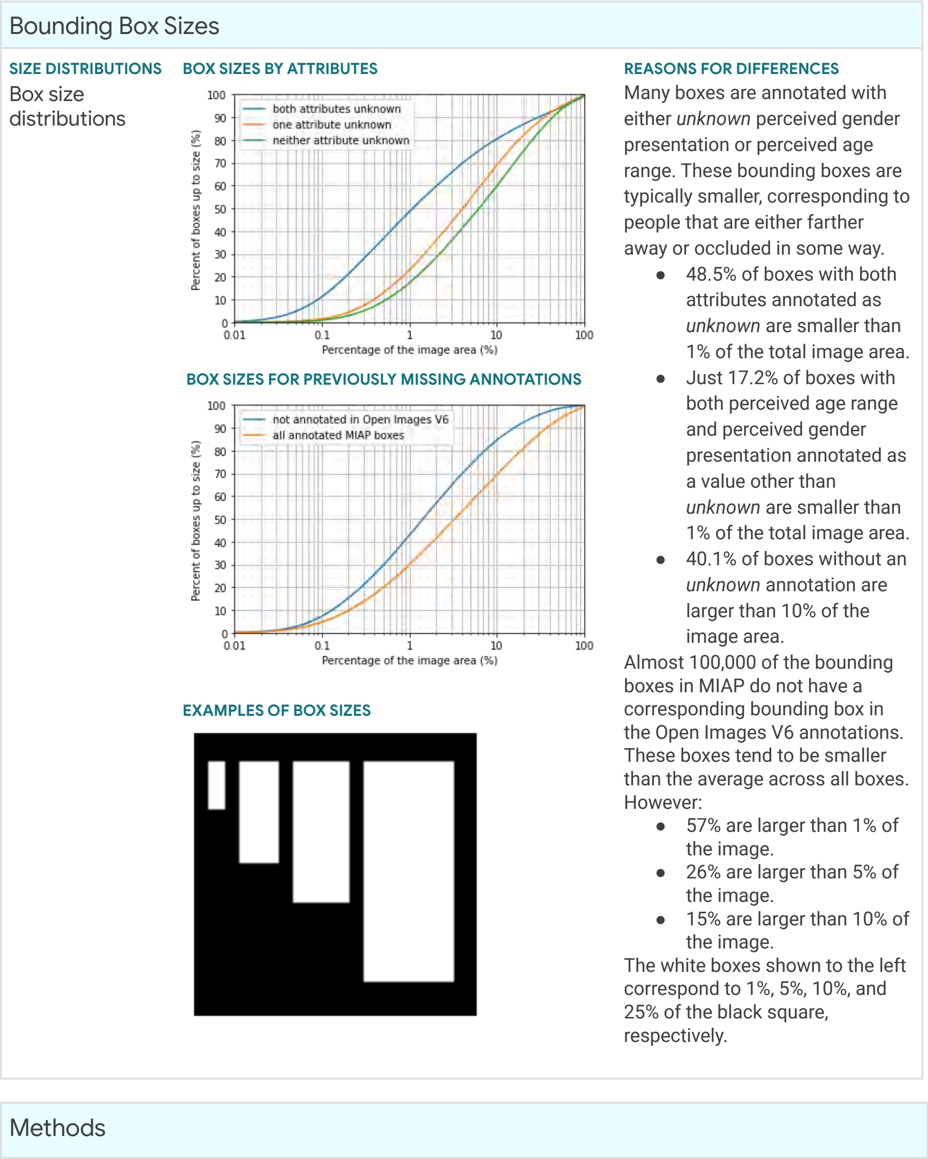
Predominantly Feminine	22.2%
Predominantly Masculine	38.3%
Unknown	39.5%

EXPECTATIONS, RISK, AND CAVEATS
Note that gender is *not* binary, and an individual's gender identity may not match their gender presentation. It is *not* possible to label gender identity from images. Additionally, norms around gender expression vary across cultures and have changed over time. No single aspect of a person's appearance "defines" their gender expression. For example, a person may still present as predominantly masculine while wearing jewelry. Another may present as predominantly feminine while having short hair. The intention of these labels is to capture gender presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender.

SOURCE OF SUBGROUP
Annotators were given diverse examples of different gender presentations and asked to label each person in an image with a perceived gender presentation. If annotators were unsure about a gender presentation they were asked to select *Unknown*.

TRADEOFFS
These labels are still valuable because they allow researchers to assess the performance of models across gender presentation, which can ultimately lead to less biased models that work well for all users. While these annotations will sometimes be misaligned with each individual's self-identified gender, in aggregate the annotations are useful to give us a simplified overall sense of how model performance may differ for people who present gender differently.

Figure 7: **Data Card for Computer Vision Dataset, Page 5 of 7**



Bounding Box Sizes

SIZE DISTRIBUTIONS

Box size distributions

BOX SIZES BY ATTRIBUTES

BOX SIZES FOR PREVIOUSLY MISSING ANNOTATIONS

EXAMPLES OF BOX SIZES

REASONS FOR DIFFERENCES

Many boxes are annotated with either *unknown* perceived gender presentation or perceived age range. These bounding boxes are typically smaller, corresponding to people that are either farther away or occluded in some way.

- 48.5% of boxes with both attributes annotated as *unknown* are smaller than 1% of the total image area.
- Just 17.2% of boxes with both perceived age range and perceived gender presentation annotated as a value other than *unknown* are smaller than 1% of the total image area.
- 40.1% of boxes without an *unknown* annotation are larger than 10% of the image area.

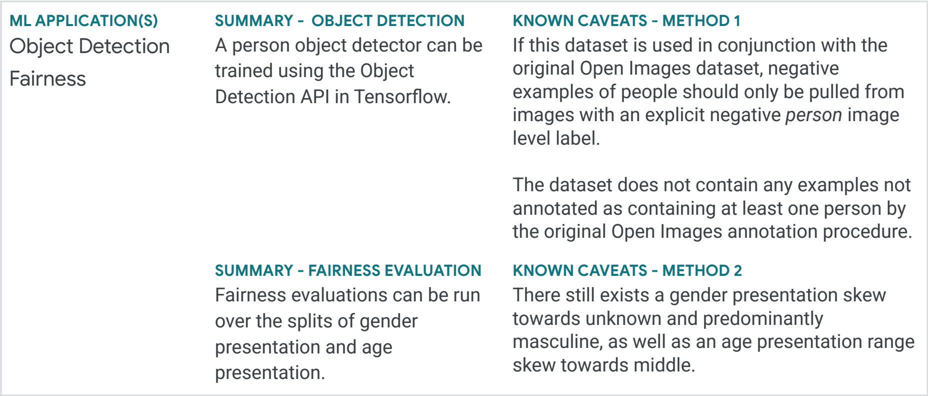
Almost 100,000 of the bounding boxes in MIAP do not have a corresponding bounding box in the Open Images V6 annotations. These boxes tend to be smaller than the average across all boxes. However:

- 57% are larger than 1% of the image.
- 26% are larger than 5% of the image.
- 15% are larger than 10% of the image.

The white boxes shown to the left correspond to 1%, 5%, 10%, and 25% of the black square, respectively.

Methods

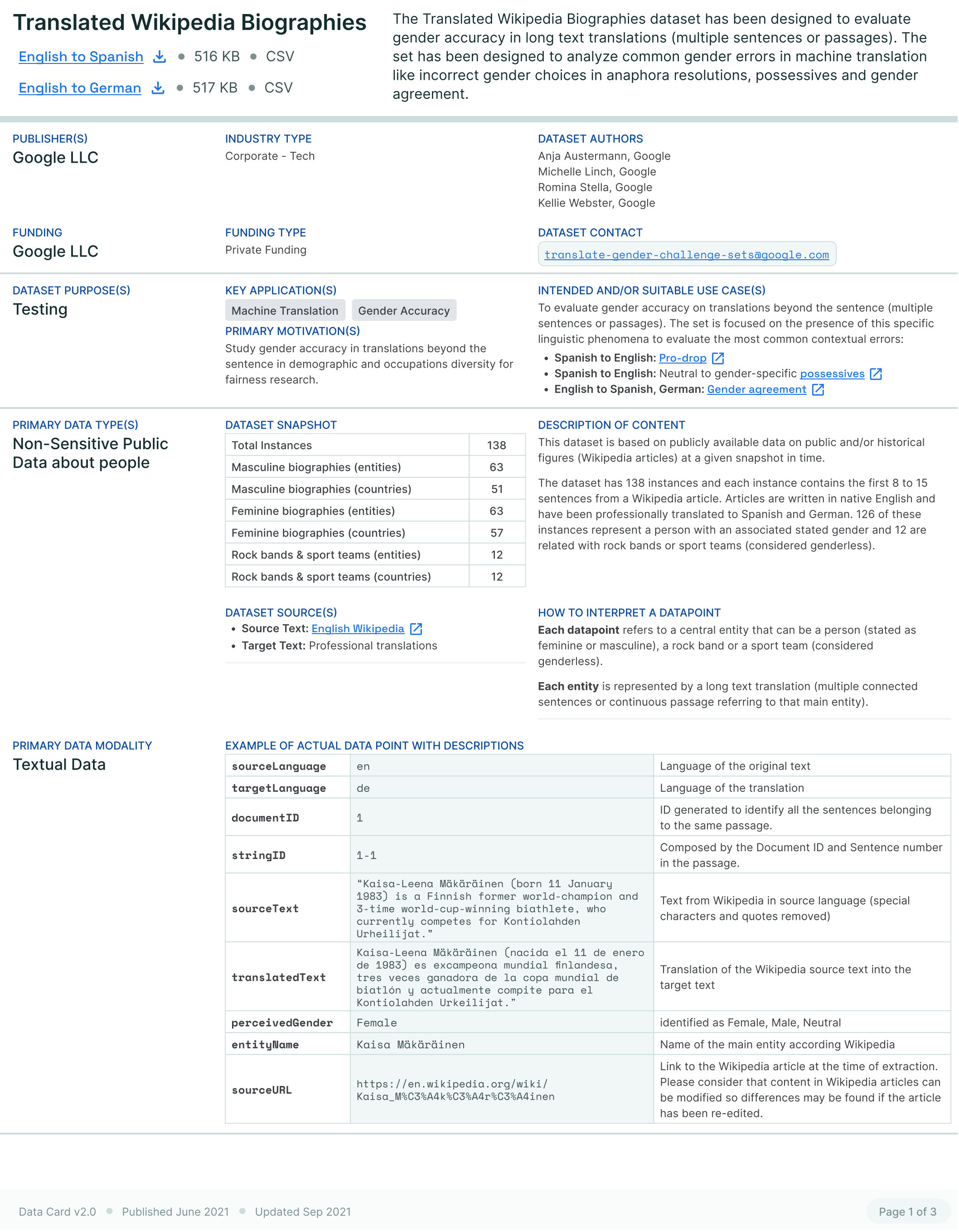
Figure 8: **Data Card for Computer Vision Dataset, Page 6 of 7**



ML APPLICATION(S) Object Detection Fairness	SUMMARY - OBJECT DETECTION A person object detector can be trained using the Object Detection API in Tensorflow.	KNOWN CAVEATS - METHOD 1 If this dataset is used in conjunction with the original Open Images dataset, negative examples of people should only be pulled from images with an explicit negative *person* image level label. The dataset does not contain any examples not annotated as containing at least one person by the original Open Images annotation procedure.
	SUMMARY - FAIRNESS EVALUATION Fairness evaluations can be run over the splits of gender presentation and age presentation.	KNOWN CAVEATS - METHOD 2 There still exists a gender presentation skew towards unknown and predominantly masculine, as well as an age presentation range skew towards middle.

Figure 9: **Data Card for Computer Vision Dataset, Page 7 of 7**

A.3 Data Card for Language Translation Dataset



Translated Wikipedia Biographies

English to Spanish ⤓ • 516 KB • CSV

English to German ⤓ • 517 KB • CSV

The Translated Wikipedia Biographies dataset has been designed to evaluate gender accuracy in long text translations (multiple sentences or passages). The set has been designed to analyze common gender errors in machine translation like incorrect gender choices in anaphora resolutions, possessives and gender agreement.

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

DATASET AUTHORS
Anja Austermann, Google
Michelle Linch, Google
Romina Stella, Google
Kellie Webster, Google

FUNDING
Google LLC

FUNDING TYPE
Private Funding

DATASET CONTACT
translate-gender-challenge-sets@google.com

DATASET PURPOSE(S)
Testing

KEY APPLICATION(S)
Machine Translation   Gender Accuracy

PRIMARY MOTIVATION(S)
Study gender accuracy in translations beyond the sentence in demographic and occupations diversity for fairness research.

INTENDED AND/OR SUITABLE USE CASE(S)
To evaluate gender accuracy on translations beyond the sentence (multiple sentences or passages). The set is focused on the presence of this specific linguistic phenomena to evaluate the most common contextual errors:

- **Spanish to English:** Pro-drop ↗
- **Spanish to English:** Neutral to gender-specific possessives ↗
- **English to Spanish, German:** Gender agreement ↗

PRIMARY DATA TYPE(S)
Non-Sensitive Public Data about people

DATASET SNAPSHOT

Total Instances	138
Masculine biographies (entities)	63
Masculine biographies (countries)	51
Feminine biographies (entities)	63
Feminine biographies (countries)	57
Rock bands & sport teams (entities)	12
Rock bands & sport teams (countries)	12

DESCRIPTION OF CONTENT
This dataset is based on publicly available data on public and/or historical figures (Wikipedia articles) at a given snapshot in time.

The dataset has 138 instances and each instance contains the first 8 to 15 sentences from a Wikipedia article. Articles are written in native English and have been professionally translated to Spanish and German. 126 of these instances represent a person with an associated stated gender and 12 are related with rock bands or sport teams (considered genderless).

DATASET SOURCE(S)

- **Source Text:** English Wikipedia ↗
- **Target Text:** Professional translations

HOW TO INTERPRET A DATAPOINT
Each datapoint refers to a central entity that can be a person (stated as feminine or masculine), a rock band or a sport team (considered genderless).

Each entity is represented by a long text translation (multiple connected sentences or continuous passage referring to that main entity).

PRIMARY DATA MODALITY
Textual Data

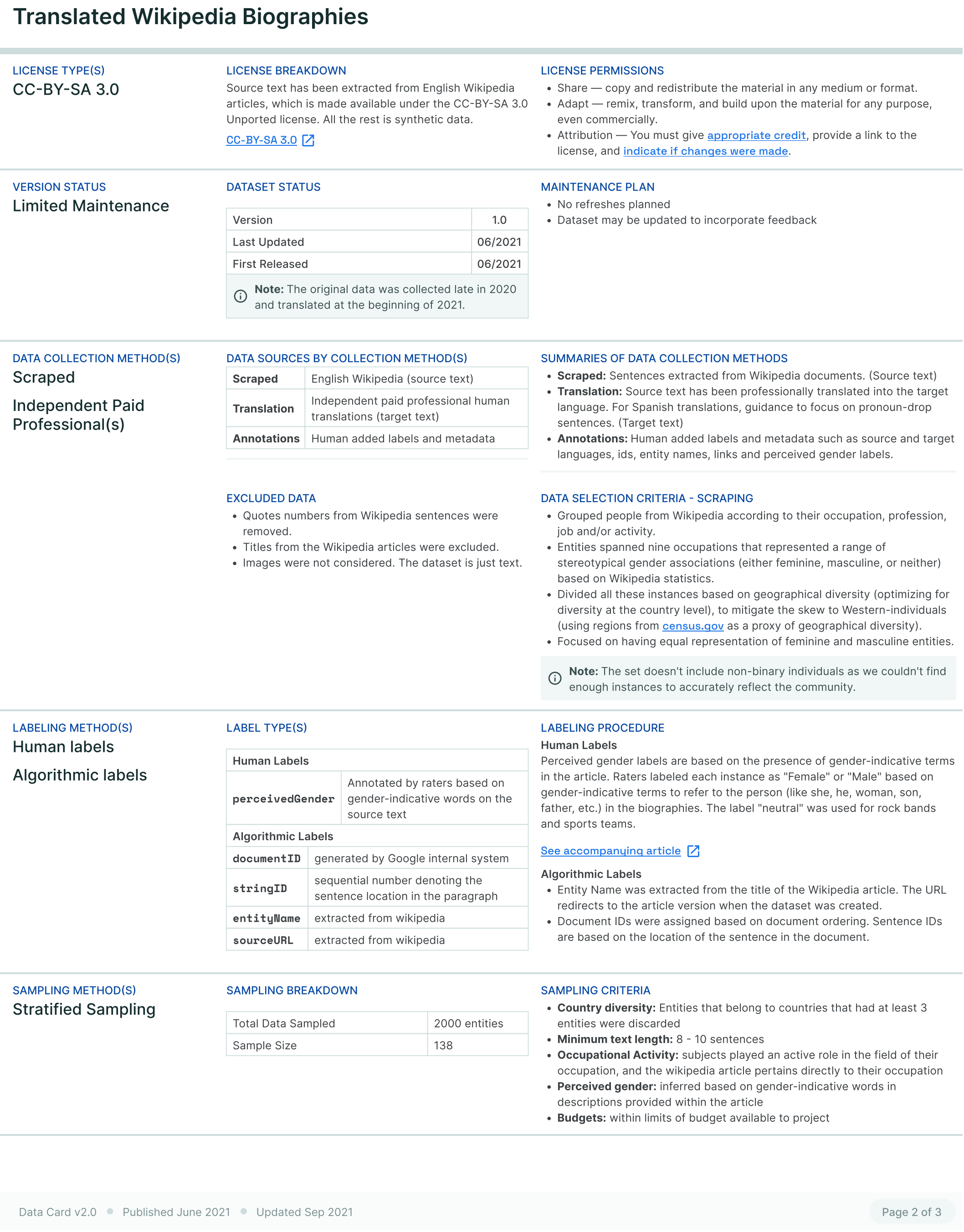
EXAMPLE OF ACTUAL DATA POINT WITH DESCRIPTIONS

sourceLanguage	en	Language of the original text
targetLanguage	de	Language of the translation
documentID	1	ID generated to identify all the sentences belonging to the same passage.
stringID	1-1	Composed by the Document ID and Sentence number in the passage.
sourceText	"Kaisa-Leena Mäkäräinen (born 11 January 1983) is a Finnish former world-champion and 3-time world-cup-winning biathlete, who currently competes for Kontiolahden Urheilijat."	Text from Wikipedia in source language (special characters and quotes removed)
translatedText	Kaisa-Leena Mäkäräinen (nacida el 11 de enero de 1983) es excampeona mundial finlandesa, tres veces ganadora de la copa mundial de biatlón y actualmente compite para el Kontiolahden Urkeilijat."	Translation of the Wikipedia source text into the target text
perceivedGender	Female	identified as Female, Male, Neutral
entityName	Kaisa Mäkäräinen	Name of the main entity according Wikipedia
sourceURL	https://en.wikipedia.org/wiki/Kaisa_M%C3%A4k%C3%A4r%C3%A4inen	Link to the Wikipedia article at the time of extraction. Please consider that content in Wikipedia articles can be modified so differences may be found if the article has been re-edited.

Data Card v2.0 • Published June 2021 • Updated Sep 2021

Page 1 of 3

Figure 10: **Data Card for Language Translation Dataset, Page 1 of 3**



Translated Wikipedia Biographies

LICENSE TYPE(S)

CC-BY-SA 3.0

LICENSE BREAKDOWN

Source text has been extracted from English Wikipedia articles, which is made available under the CC-BY-SA 3.0 Unported license. All the rest is synthetic data.

CC-BY-SA 3.0

LICENSE PERMISSIONS

- Share — copy and redistribute the material in any medium or format.
- Adapt — remix, transform, and build upon the material for any purpose, even commercially.
- Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.

VERSION STATUS

Limited Maintenance

DATASET STATUS

Version	1.0
Last Updated	06/2021
First Released	06/2021

ⓘ **Note:** The original data was collected late in 2020 and translated at the beginning of 2021.

MAINTENANCE PLAN

- No refreshes planned
- Dataset may be updated to incorporate feedback

DATA COLLECTION METHOD(S)

Scraped

Independent Paid Professional(s)

DATA SOURCES BY COLLECTION METHOD(S)

Scraped	English Wikipedia (source text)
Translation	Independent paid professional human translations (target text)
Annotations	Human added labels and metadata

SUMMARIES OF DATA COLLECTION METHODS

- **Scraped:** Sentences extracted from Wikipedia documents. (Source text)
- **Translation:** Source text has been professionally translated into the target language. For Spanish translations, guidance to focus on pronoun-drop sentences. (Target text)
- **Annotations:** Human added labels and metadata such as source and target languages, ids, entity names, links and perceived gender labels.

EXCLUDED DATA

- Quotes numbers from Wikipedia sentences were removed.
- Titles from the Wikipedia articles were excluded.
- Images were not considered. The dataset is just text.

DATA SELECTION CRITERIA - SCRAPING

- Grouped people from Wikipedia according to their occupation, profession, job and/or activity.
- Entities spanned nine occupations that represented a range of stereotypical gender associations (either feminine, masculine, or neither) based on Wikipedia statistics.
- Divided all these instances based on geographical diversity (optimizing for diversity at the country level), to mitigate the skew to Western-individuals (using regions from census.gov as a proxy of geographical diversity).
- Focused on having equal representation of feminine and masculine entities.

ⓘ **Note:** The set doesn't include non-binary individuals as we couldn't find enough instances to accurately reflect the community.

LABELING METHOD(S)

Human labels

Algorithmic labels

LABEL TYPE(S)

Human Labels	
`perceivedGender`	Annotated by raters based on gender-indicative words on the source text
Algorithmic Labels	
`documentID`	generated by Google internal system
`stringID`	sequential number denoting the sentence location in the paragraph
`entityName`	extracted from wikipedia
`sourceURL`	extracted from wikipedia

LABELING PROCEDURE

Human Labels

Perceived gender labels are based on the presence of gender-indicative terms in the article. Raters labeled each instance as "Female" or "Male" based on gender-indicative terms to refer to the person (like she, he, woman, son, father, etc.) in the biographies. The label "neutral" was used for rock bands and sports teams.

See accompanying article

Algorithmic Labels

- Entity Name was extracted from the title of the Wikipedia article. The URL redirects to the article version when the dataset was created.
- Document IDs were assigned based on document ordering. Sentence IDs are based on the location of the sentence in the document.

SAMPLING METHOD(S)

Stratified Sampling

SAMPLING BREAKDOWN

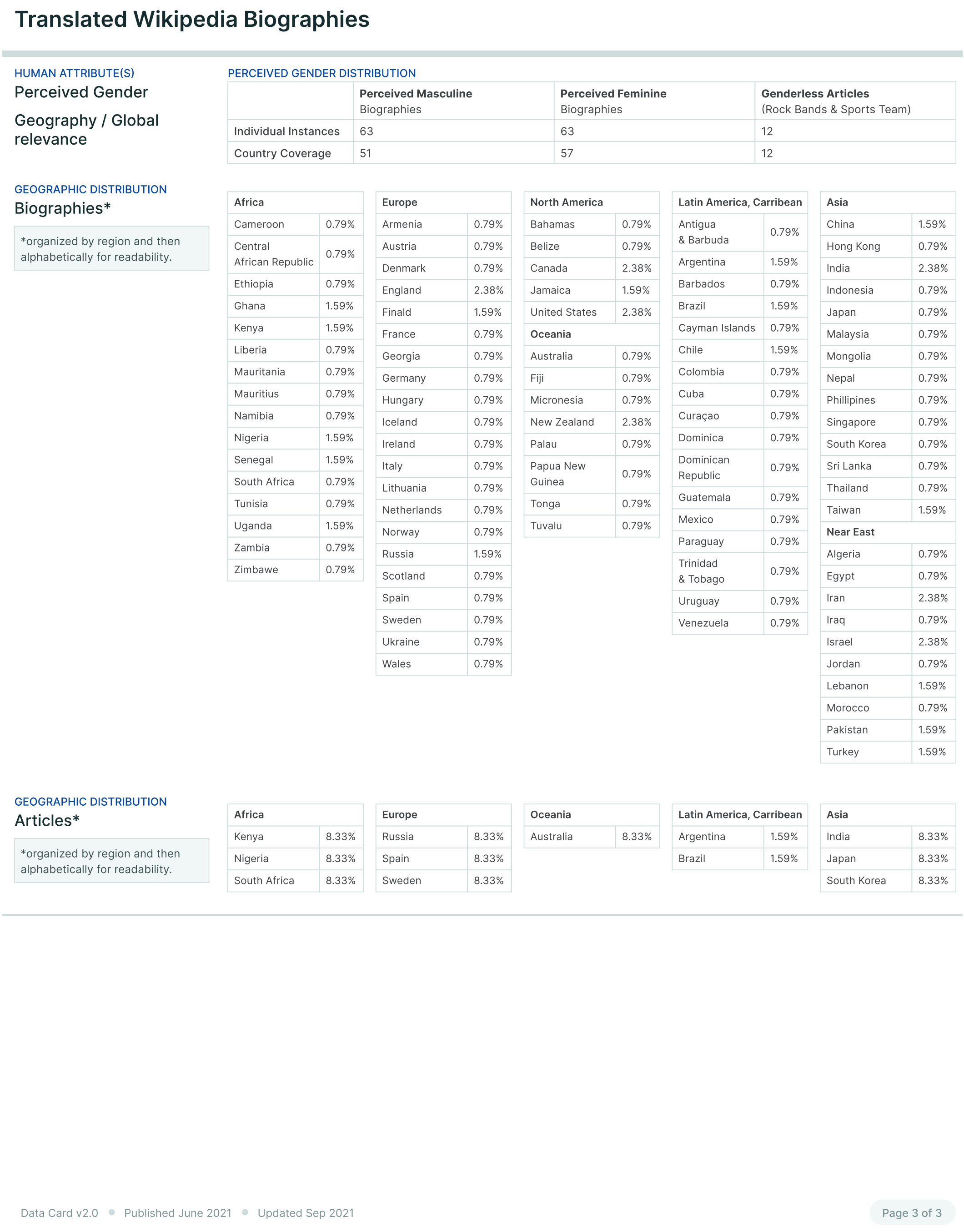
Total Data Sampled	2000 entities
Sample Size	138

SAMPLING CRITERIA

- **Country diversity:** Entities that belong to countries that had at least 3 entities were discarded
- **Minimum text length:** 8 - 10 sentences
- **Occupational Activity:** subjects played an active role in the field of their occupation, and the wikipedia article pertains directly to their occupation
- **Perceived gender:** inferred based on gender-indicative words in descriptions provided within the article
- **Budgets:** within limits of budget available to project

Data Card v2.0 • Published June 2021 • Updated Sep 2021 Page 2 of 3

Figure 11: **Data Card for Language Translation Dataset, Page 2 of 3**



Translated Wikipedia Biographies

HUMAN ATTRIBUTE(S)

Perceived Gender

Geography / Global relevance

PERCEIVED GENDER DISTRIBUTION

	Perceived Masculine Biographies	**Perceived Feminine** Biographies	**Genderless Articles** (Rock Bands & Sports Team)
Individual Instances	63	63	12
Country Coverage	51	57	12

GEOGRAPHIC DISTRIBUTION

Biographies*

*organized by region and then alphabetically for readability.

Africa	
Cameroon	0.79%
Central African Republic	0.79%
Ethiopia	0.79%
Ghana	1.59%
Kenya	1.59%
Liberia	0.79%
Mauritania	0.79%
Mauritius	0.79%
Namibia	0.79%
Nigeria	1.59%
Senegal	1.59%
South Africa	0.79%
Tunisia	0.79%
Uganda	1.59%
Zambia	0.79%
Zimbawe	0.79%

Europe	
Armenia	0.79%
Austria	0.79%
Denmark	0.79%
England	2.38%
Finald	1.59%
France	0.79%
Georgia	0.79%
Germany	0.79%
Hungary	0.79%
Iceland	0.79%
Ireland	0.79%
Italy	0.79%
Lithuania	0.79%
Netherlands	0.79%
Norway	0.79%
Russia	1.59%
Scotland	0.79%
Spain	0.79%
Sweden	0.79%
Ukraine	0.79%
Wales	0.79%

North America	
Bahamas	0.79%
Belize	0.79%
Canada	2.38%
Jamaica	1.59%
United States	2.38%
Oceania	
Australia	0.79%
Fiji	0.79%
Micronesia	0.79%
New Zealand	2.38%
Palau	0.79%
Papua New Guinea	0.79%
Tonga	0.79%
Tuvalu	0.79%

Latin America, Carribean	
Antigua & Barbuda	0.79%
Argentina	1.59%
Barbados	0.79%
Brazil	1.59%
Cayman Islands	0.79%
Chile	1.59%
Colombia	0.79%
Cuba	0.79%
Curaçao	0.79%
Dominica	0.79%
Dominican Republic	0.79%
Guatemala	0.79%
Mexico	0.79%
Paraguay	0.79%
Trinidad & Tobago	0.79%
Uruguay	0.79%
Venezuela	0.79%

Asia	
China	1.59%
Hong Kong	0.79%
India	2.38%
Indonesia	0.79%
Japan	0.79%
Malaysia	0.79%
Mongolia	0.79%
Nepal	0.79%
Phillipines	0.79%
Singapore	0.79%
South Korea	0.79%
Sri Lanka	0.79%
Thailand	0.79%
Taiwan	1.59%
Near East	
Algeria	0.79%
Egypt	0.79%
Iran	2.38%
Iraq	0.79%
Israel	2.38%
Jordan	0.79%
Lebanon	1.59%
Morocco	0.79%
Pakistan	1.59%
Turkey	1.59%

GEOGRAPHIC DISTRIBUTION

Articles*

*organized by region and then alphabetically for readability.

Africa	
Kenya	8.33%
Nigeria	8.33%
South Africa	8.33%

Europe	
Russia	8.33%
Spain	8.33%
Sweden	8.33%

Oceania	
Australia	8.33%

Latin America, Carribean	
Argentina	1.59%
Brazil	1.59%

Asia	
India	8.33%
Japan	8.33%
South Korea	8.33%

Data Card v2.0 • Published June 2021 • Updated Sep 2021 Page 3 of 3

Figure 12: **Data Card for Language Translation Dataset, Page 3 of 3**