Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI

Mahima Pushkarna, Andrew Zaldivar, Oddur Kjartansson
mahimap@google.com,andrewzaldivar@google.com,oddur@google.com
Google Research
Canada, United States, United Kingdom

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 Urheilijat."	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 1: A page from a Data Card summarizing the lifecycle of a text translation dataset. Data Cards organize a variety of content thematically in a row-and-column structure for easy indexing and finding. Blocks increase in detail from left to right, and authors have introduced links to elegantly expose readers to additional documentation using context offered in the Data Card.

ABSTRACT

As research and industry moves towards large-scale models capable of numerous downstream tasks, the complexity of understanding multi-modal 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 the documentation. It requires consistency and comparability across the documentation of all datasets involved, and as such documentation must be treated as a user-centric product in and of itself. In this 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 lifecycle for responsible AI development. These summaries 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. We also present frameworks that ground Data Cards in real-world utility and human-centricity. Using two case studies, we report on desirable characteristics that support adoption across domains, organizational structures, and audience groups. Finally, we present lessons learned from deploying over 20 Data Cards.x

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **General and reference** → *Evaluation*; • **Software and its engineering** → **Software creation and management**; • **Human-centered computing**;

KEYWORDS

data cards, dataset documentation, transparency, responsible AI, datasheets, model cards

ACM Reference Format:
Mahima Pushkarna, Andrew Zaldivar, Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 51 pages. https://doi.org/10.1145/3531146.3533231

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533231

1 INTRODUCTION

The challenge of transparency in machine learning (ML) models and datasets continues to receive increasing attention from academia and industry [1, 2]. Often, the goal has been to attain greater visibility into ML models and datasets by exposing source code [4], contribution trails [8], introducing ML-drive data analysis methods [19], and introducing diverse oversight [18]. Transparency and explainability of model outcomes through the lens of datasets has become a huge concern in regulation from government bodies internationally. However, attempts to introduce standardized, practical and sustainable mechanisms for transparency that create value at scale meet limited success in research and production contexts. This reflects 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 [11, 13, 14].

As a step towards creating value that connects dataset success to research and production experiences, we propose a new framework for transparent and purposeful documentation of datasets, called Data Cards [26]. A Data Card contains a structured collection of summaries gathered over the life cycle of a dataset about observable (e.g., dataset attributes) and unobservable (e.g., intended use cases) 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.

Data Cards emphasize information and context that shape the data, but cannot be inferred from the dataset directly. These are designed as boundary objects [28] that should be easily available in accessible formats at important steps of a user journey for a diverse set of readers. Data Cards encourage informed decision making about data usage when building and evaluating ML models for products, policy and research. Data Cards complement other longer-form and domain-specific documentation frameworks for ethical reporting (See Appendix A) , such as Model Cards [23], Data Statements [9], Datasheets for Datasets [15], and [6] FactSheets.

Data Cards are accompanied by frameworks to adapt them to a variety of datasets and organizational contexts. These frameworks are pivotal to establishing common ground across stakeholders and enable diverse input into decisions. Our case studies demonstrate that creators of Data Cards were able to discover surprising future opportunities to improve their dataset design decisions, such as considering reasons for a high percentage of unknown values and the need to create a shared understanding of lexicons used in dataset labeling during problem formulation.

In summary, our contributions are four-fold:

- We explain our multi-pronged approach in the setting of a large-scale technology company and present a typology of stakeholders that span a typical dataset lifecycle. We translate outcomes from our development methodology into corresponding objectives and principles for the creation of Data Cards to systematically reduce the knowledge asymmetries across stakeholders.

- We introduce a transparency artifact for at-scale production and research environments, **Data Cards**— structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development, and describe the content (*What information to present*), design (*How to present information*), and evaluation (*Assess the efficacy of information*) of Data Cards.
- We propose three frameworks for the construction of Data Cards that focus on information organization, question framing, and answer evaluation, respectively. Specifically, we describe OFTEn, our novel knowledge acquisition framework to arm dataset producers with a robust, deliberate, and repeatable approach for producing transparent documentation.
- We present case studies on the creation of Data Cards for a computer vision dataset and a language dataset to demonstrate their impact as boundary objects in practice, and discuss epistemic and organizational lessons learned in scaling Data Cards.

Our collective efforts suggest that in addition to comprehensive transparency artifacts[1], the creation of structured frameworks are not only beneficial in adding nuance to the dataset documentation process itself, but also transformational in introducing human-centric and responsible practices when using datasets in ML applications.

2 DEVELOPMENT METHODOLOGY

Over the course of 24 months, multiple efforts were employed to design Data Cards and its supporting frameworks, borrowing from methods in human-centered design, participatory design, and human-computer interaction. We worked with dataset and ML teams in a large technology company to iteratively create Data Cards, refining our design decisions to respond to challenges in production contexts. In parallel, we ran studies and workshops to identify opportunities and challenges in the implementation of Data Cards. In this section, we detail the various efforts and describe their impact on the development of Data Cards.

Specifically, we worked with 12 teams in a large technology company to create 22 Data Cards that describe image, language, tabular, video, audio, and relational datasets in production settings. Teams ranged in size from four to over 20 members, and were comprised of some combination of research software engineers, research scientists, data analysts and data program managers. This allowed us to observe each teams' documentation workflows, collaborative information gathering practices, information requests from downstream stakeholders, review and assessment practices. Our co-creative approach in conjunction with feedback received across other studies yielded continuous improvements in the usability and utility of each new Data Card created.

As we worked with ML dataset and model owners to produce prototypical transparency artifacts, drafts were evaluated in an external focus group with nine participants. These participants represented non-expert, technical use cases from User Experience (UX) and Human-Computer Interaction (HCI) research, Policy, Product Design & Development, Academia, and Law. Participants were asked to complete a paper-based questionnaire to reflect on their ideals of transparency, used as a basis for broader discussions on transparency. Participants were then provided with printed drafts which they annotated with their feedback. This allowed us to capture specific feedback and establish relationships across themes and topics in the artifacts. We concluded with a discussion reflecting on their use of transparency artifacts and an offline survey to capture their overall expectations. Through this focus group, we were able to arrive at a working definition and values of transparency relevant to domains within AI product life cycles. We further synthesized feedback on the transparency artifacts into an initial set of recommendations to combat common reader-side challenges, which were then offered as guidance to teams creating Data Cards.

Based on our experience in co-creating Data Cards with teams, we were able to consolidate recurring and overlapping questions into a canonical template that documents 31 different aspects of data sets. Questions that are were modality-specific were consolidated into appendable blocks, but largely left out of the canonical template. A follow-up internal MaxDiff survey (n=191) was conducted to understand the information needs in dataset documentation within our company. Through this survey, we learned the relative importance of the 31 aspects documented in a Data Card, how these vary by dataset modality and job function, and further incorporated insights into our design of Data Cards. We observed the need for a generative framework that Data Card creators could use to add or tailor question to new datasets without compromising the readability, navigability, comparability and transparency intrinsic to the Data Card.

Our internal study recruited 30 experts spanning sixteen teams within our company. Participants represented stakeholders who (a) create datasets designed for ML use cases and (b) use or review datasets for applied and foundational model development. Over the course of three days, this group engaged in various participatory activities to articulate use cases for transparency artifacts, information requirements, and strategies for evaluation of transparency artifacts. Participants were then invited to actively contribute to future discussions of Data Cards and their development as it related to the participant's specific data domains. We found that despite their deep expertise and experience, participants were unable to provide examples of exemplary documentation, but were quick to furnish 'excellent' examples of poor documentation. This pointed us to the need for a set of dimensions that can be used to assess transparency and documentation without conflating documentation with the dataset.

Further, we developed a structured participatory workshop-based approach to engage cross-functional stakeholders when creating transparent metadata schema for dataset documentation [25]. This methodology was open-sourced and tested in the data domains of human computation, geo-spatial ML, multi-modal data operations, healthcare data, community-engaged research, and large-scale multitask language models. Common to all workshops, we found that participating teams often started with an intuition about the benefits of transparency in dataset documentation. We found that teams needed to necessarily align on a shared definition of transparency, audience, and the audience's requirements as prerequisites define the content, infrastructure, and processes to scale Data Card creation. We observed organization-specific factors that can impact long-term sustainability of scaling Data Cards, such

[1]For the purposes of practicality, we use transparency artifacts as a general term to describe both Data and Model Cards [23] because of their inextricably linked nature. In this paper, we primarily focus on our insights and advances on datasets and correspondingly Data Cards, our novel contribution.

as knowledge asymmetries between stakeholders, organizational processes that incentivize the creation and maintenance of documentation, infrastructure compatibility and readiness, and communication culture across and within stakeholder groups. While a detailed discussion of our participatory methodology to developing transparency metadata schemas and survey is beyond the scope of this paper, we introduce relevant critical frameworks from our methodology.

2.1 Framing Transparency in the Context of Data Cards

Despite the diverse backgrounds of participants across studies, the shared dominant perception was that transparency artifacts were ironically opaque. The opacity in documentation, quite simply, increases when language used is technical, dense, and presumptive of a reader's background, making it difficult for non-technical stakeholders to interpret. This, in turn, leads to sub-optimal decision making, and propagates asymmetries in power structures and myopic AI data practices. Further, focus group and workshop participants described transparency as "*subjective*", "*audience-specific*" and "*contextual*". To that end, we frame our definition of transparency as *"a clear, easily understandable, and plain language explanation of what something is, what it does and why it does that"*, to emphasize the domain-agnostic and inclusive prerogative of transparency artifacts. In table 1, We present eight characteristics of transparency that are vital for a robust discussion of the benefits, values, ethics, and limitations of AI datasets. Data Cards aim to provide a single scalable, artifact that allows non-traditional stakeholders across product, policy, and research to understand aspects about datasets and how they are used to make informed decisions. We found that stakeholders review role-related topics in Data Cards with amplified scrutiny, and follow-up questions progressively increase in specificity, which suggests that transparency is attained when we establish a shared and socratic understanding of datasets based on the ability to ask and answer questions *over time*.

2.2 A Typology of Stakeholders

At first, our audience for Data Cards was fairly broad, comprising a mix of experts and non-experts. Frameworks proposed by Suresh, et al [29] have distinguished higher-level domain goals and objectives from lower-level interpretability tasks, but are limited by their epistemological framing and vast scope. We created a broad yet decomposable typology describing three stakeholders groups in a dataset's life cycle, allowing us to consider how cross-functional stakeholders engage in decision-making on the basis of a single transparency artifact.

In our typology, ***Producers*** are upstream creators of dataset and documentation, responsible for dataset collection, ownership, launch and maintenance. We observed that producers often subscribe 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. However, in practice, we find that only a few readers or ***Agents*** actually meet all these requirements.

Agents are stakeholders who read transparency reports, and possess the agency to use or determine how themselves or others might use the described datasets or AI systems. After testing prototypes and proof of concepts with different audience groups, it became clear that agents with operational and reviewer needs were distinct categories. Reviewers include stakeholders who may never directly use the dataset, but will engage with the Data Card (for e.g. reviewers or non-technical subject matter experts). Agents may or may not possess the technical expertise to navigate information presented in typical dataset documentation, but often have access to expertise as required.

Additionally, agents are distinct from ***Users***, who are individuals and representatives who interact with products that rely on models trained on dataset. Users may consent to providing their data as a part of the product experience, and require a significantly different set of explanations and controls grounded within product experiences. We therefore suggest the use of Data Card target agents with access to technical expertise, and encourage the use of alternative transparency artifacts for users that are designed exclusively for that purpose.

We further dis-aggregate these high-level groups to generate awareness and emphasize the unique decisions that each sub-group must make (Fig[3]). However, these groupings exist on a continuum and stakeholders may fall into more than one group concurrently, depending on their context. We used this typology to unearth assumptions that are often made about the rich intersectional attributes of individual stakeholders, such as expertise (e.g. novice or expert), data fluency (e.g. none to high), job roles (e.g. Data Scientist, Policy Maker), function performed vis-à-vis the data (Data Contributor, Rater), and goals or tasks (Publishing a dataset, Comparing datasets) when conceptualizing Data Cards. Usability studies across these groups revealed guidelines for the successful and appropriate adoption of Data Cards in practice and at scale. These are distilled into the following objectives for Data Cards:

2.2.1 ***O1. Consistent:*** Data Cards must be comparable to one another, regardless of data modality or domain such that claims are easy to interpret and validate within context of use. While deploying one-time Data Cards is relatively easy, we find that organizations need to preserve comparability when scaling adoption. A Data Card creation effort should solicit equitable information from all datasets.

2.2.2 ***O2. Comprehensive:*** Rather than being created as a last step in a dataset's lifecycle, it should be easy to create a Data Card concurrently with the dataset. Further, the responsibility of filling out fields in a Data Card should be distributed and assigned to the most appropriate individual. This requires standardized methods that extend beyond the Data Card, and apply to the various reports generated in the dataset's lifecycle.

2.2.3 ***O3. Intelligible and Concise:*** Readers have varying levels of proficiency[2] which affects their interpretation of the Data Card. In scenarios where stakeholder proficiency differs, individuals with the strongest mental model of the dataset become de-facto decision makers. Finally, tasks that are more urgent or challenging can reduce the participation of non-traditional stakeholders (See 3) in decisions, which are left to "the expert". This risks omitting critical perspectives that reflect the situated needs of downstream and lateral stakeholders. A Data Card should efficiently communicate to

[2]Proficiency is a combination of data fluency and domain expertise. Data fluency is described as the familiarity and comfort that readers have in working with data that is both, in or outside of their domain of expertise. The greater the comfort with understanding, manipulating, and using data, the greater the fluency. Domain expertise is defined as "knowledge and understanding of the essential aspects of a specific field of inquiry" [22] in reference to the domain of the dataset.

Table 1: Characteristics of transparency surface through participatory sessions

Transparency Characteristic	Description
Balance opposites	For example, disclosing information about AI systems without leaving creators vulnerable beyond reason, reporting fairness analyses without legitimizing inequitable or unfair systems, introducing standards for transparency that are wholly automated or become checklists.
Increase in expectations	Any information included in a transparency artifact can be expected to receive greater scrutiny.
Constant availability	Users want access to transparency information at multiple levels, even if they don't need to use it.
Require checks and balances	Transparency artifacts and their creation must be amenable to 3rd party evaluation, with the caveat that excessive transparency can open an AI system vulnerable to adversarial actors.
Subjective interpretations	Stakeholders have different definitions and unique ideas on what constitutes transparency.
Trust enabler	Accessible and relevant information about AI systems in-creases the the willingness of a data consumer or user to take a risk based on the expectation of benefits from the data, algorithms and the products they use.
Reduce knowledge asymmetries	Cross-disciplinary stakeholders are more effective when they possess a shared mental model and vocabulary to describe aspects of the AI system.
Reflects human values	It comes from both technical and non-technical disclosure about assumptions, facts and alternatives.

the reader with the least proficiency, while enabling readers with greater proficiency to find more information as needed. The content and design should advance a reader's deliberation process without overwhelming them, and encourage stakeholder cooperation towards a shared mental model of the dataset for decision-making.

2.2.4 ***O4. Explainability, Uncertainty:*** Workshop participants reported that 'known unknowns' were as important as known facets of the dataset in decision making. Communicating uncertainty along with meaningful metadata was considered a feature and not a bug, allowing readers to answer questions such as *"Is a specific analysis irrelevant to the dataset or were the results insignificant?"* or *"Is information withheld because it is proprietary or is it unknown?"*. Clear descriptions and justifications for uncertainty can lead to additional measures to mitigate risks, leading to opportunities for fairer and equitable models. This builds greater trust in the dataset and subsequently, its publishers [10].

3 DATA CARDS

Data Cards capture critical information about a dataset across its life cycle. Just as is true with every dataset, each Data Card is unique, and no single template satisfactorily captures the nuance of all datasets. In this section, we introduce our guiding principles, and elaborate on decisions towards the design, content, and evaluation of Data Cards. We introduce corresponding frameworks that allow Data Cards to be tailored but preserve the utility and intent of Data Cards.

3.1 Principles

In comparison to prior related documentation toolkits (A) that have been prescriptively adopted by producers, our novel contributions are the generative design of Data Cards as an underlying framework for transparency reporting for domain- and fluency-agnostic readability and scaling in production contexts. To meet the objectives stated above, Data Cards have been designed along the following principles:

- **P1. Flexible:** Describe a wide range of datasets such as static datasets, datasets that are actively being curated from single or multiple sources, or those with multiple modalities.
- **P2. Modular:** Organize documentation into meaningful sections that are self-contained and well-structured units, capable of providing an end-to-end description of a single aspect of the dataset.
- **P3. Extensible:** Components that can be easily reconfigured or extended systematically for novel datasets, analyses, and platforms.
- **P4. Accessible:** Represent content at multiple granularities so readers can efficiently find and effectively navigate detailed descriptions of the dataset.
- **P5. Content-agnostic:** Support diverse media including multiple choice selections, long-form inputs, text, visualizations, images, code blocks, tables, and other interactive elements.

3.2 Design and Structure

The fundamental "display" unit of a Data Card is a ***block*** which consists of a title, a question, space for additional instructions or descriptions, and an input space for answers. Answer inputs are reinforced with structure to create blocks that are specifically suited for long- or short-form text, multiple or single choice responses, tables, numbers, key value pairs, code blocks, data visualizations, tags, links, and demos of the data itself, in alignment with principles (**P1**) and (**P5**). In our templates, we iteratively introduced structures for open-ended answers, predetermined responses for multiple choice questions, and demonstrative examples where responses could be complex (Fig. 2). Producers found these assistive efforts as useful guides for setting expectations about consistency, clarity, and granularity in responses. When completed, blocks typically retained titles and answers (See Fig 1) to reduce the gulf between the experience of producers and agents.

Blocks are arranged thematically and hierarchically on a grid to enable an "*overview first, zoom-and-filter, details-on-demand*" [27] presentation of the dataset, to accomplish principle (**P4**). In our template, blocks with related questions are organized into *rows*, and rows are stacked to create *sections* using meaningful and descriptive titles (Figure 2). Each row is thematically self-contained so readers can effectively navigate multiple facets of a dataset in a Data Card. Answers increase in both detail and specificity across columns in the direction of the language in which the Data Card is written, allowing readers to find information at the appropriate fidelity for their tasks and decisions. Where appropriate, a single block may span multiple columns. Sections are vertically arranged based on functional importance in a nested hierarchy marked by section titles in the first Data Card [D]. Here, all necessary sections (dataset snapshot, motivations, extended use, collection and labeling methods) are established in order to provide greater context for interpreting sections that describe fairness-related analyses (fairness indicators, bounding box sizes). In contrast, sections in the second Data Card [E] are organized in a flat hierarchy, suggesting equal importance of all blocks. Variation within the formatting of the content communicates both denotative and connotative meaning, while preserving the fundamental unit of "blocks", illustrating principles (**P2**) and (**P3**).

3.2.1 Socratic Question-Asking Framework: Scopes. To ensure that agents with varying proficiency levels can progressively explore content with minimal barriers (principle **P4**), any new information in a Data Card needs to be introduced at multiple levels of abstraction. Further, the addition of ad-hoc blocks risks structurally compromising Data Cards for readers and producers alike, thereby reducing both, usability of design and integrity or content.

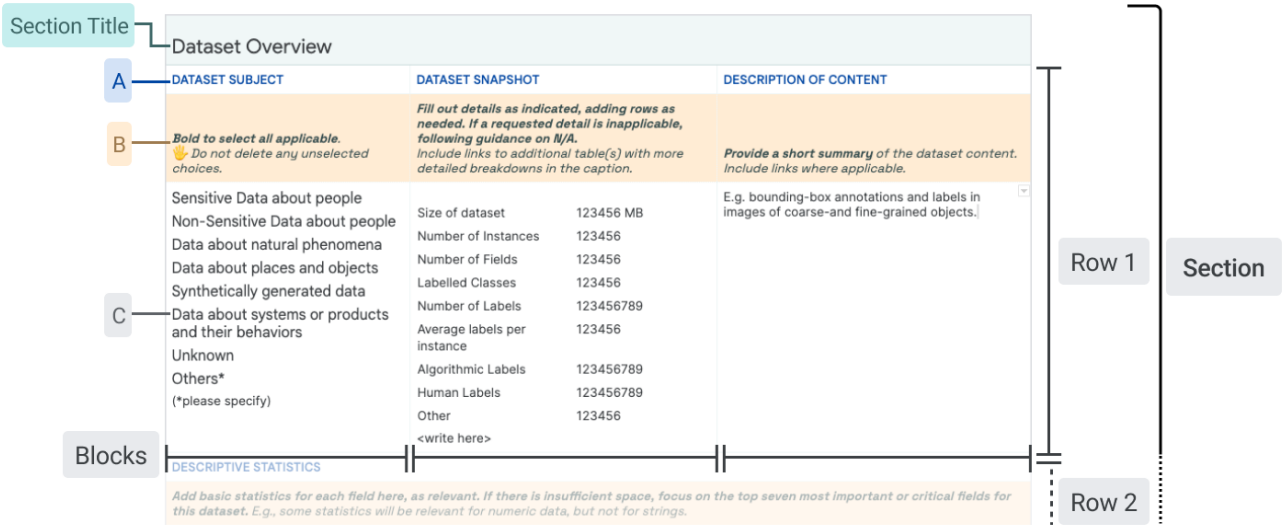


Figure 2: A Data Card Template Section: This section is titled "Dataset Overview", and contains two rows. The first row has three blocks, whereas the second row spans the entire width of the section. Blocks contain (A) A Title, (B) A prompting question, and (C) an answer input space populated with predetermined choices or suggested answer structures.

Pertinent to objectives ***O2*** and ***O3***, we provide a structured approach to framing and organizing questions to address common challenges in adapting Data Card templates for new datasets. Depending on the specificity desired, new themes are deconstructed into broad questions, which are then extrapolated into at least three questions framed at varying granularities. We characterize these as telescopes, periscopes, and microscopes. Depending on the topic documented, a Data Card may require an uneven distribution of telescopic, periscopic, or picroscopic questions. Our aforementioned row-and-column design, combined with our organization principle provides us with sufficient flexibility to intermix content hierarchy that caters to different combinations of scope types. For the purposes of demonstration, we consider the documentation of sensitive human attributes:

Telescopes provide an overview of the dataset. These are questions about universal attributes applicable across multiple datasets, for example "*Does this dataset contain Sensitive Human Attributes?*". Telescopes can be binary (*contains, does not contain*) or multiple choice (*Select all that apply: Race, Gender, Ethnicity, Socio-economic status, Geography, Language, Sexual Orientation, Religion, Age, Culture, Disability, Experience or Seniority, Others (please specify)*). These serve three specific purposes. First, telescopic questions generate enumerations or tags that are useful for knowledge management, indexing and filtering in large repository of Data Cards. Second, they introduce and set context for additional information within a row, helping readers navigate larger or more complex Data Cards. Lastly, telescopic questions introduce conditional logic to streamline the experience of filling out a Data Card. When viewed together, telescopic questions offer a shallow but wide overview of the dataset.

Periscopes provide greater technical detail pertaining to the dataset. These are questions about attributes specific to the dataset that add nuance to telescopes. For example, *"For each human attribute selected, specify if this information was collected intentionally as a part of the dataset creation process, or unintentionally not explicitly collected as a part of the dataset creation process but can be inferred using additional methods)"*. A periscopic question can ask for operational information such as the dataset's shape and size, or functional information such as sources or intentions. Responses typically look like key-value pairs, short descriptions, tables, and visualizations. Since periscopes often describe analysis results, statistical summaries, and operational metadata, they are often reproducible and can be automated wherein automating generates results that are more accurate or precise than human input.

Microscopes offer fine-grained details. These are questions about the "unobservable" human processes, decisions, assumptions and policies that shape the dataset. These elicit detailed explanations of decisions or summarize longer process documents that governed responses to the corresponding periscopic questions. For example, "*Briefly describe the motivation, rationale, considerations or approaches that caused this dataset to include the indicated human attributes. Summarize why or how this might affect the use of the dataset.*". Necessarily, answers to these questions are difficult to automate in the absence of standardized terms and operating procedures. Answers to microscopes are typically long-form text with lists and links, data tables, and visualizations.

Telescopic questions are easiest to answer, but offer relatively low utility. Periscopic questions facilitate quick assessments of suitability and relevance of the dataset, essential for simple decision-making. We observed that microscopic questions were most challenging to answer since they require articulating implicit knowledge. We find that the interpretations of a Data Card are greatly influenced by the presence or absence of these levels of abstraction. These questions enabled agents and producers alike to assess risk, plan mitigations, and where relevant, identify opportunities for better dataset creation.Together, telescopes, periscopes, and microscopes layer useful details such that numerous readers can navigate without losing sight of the bigger picture.

3.3 Content and Schema

Our initial approach was to create a single template capable of capturing the provenance, intentions, essential facts, explanations and caveats in an accessible and understandable way. In co-creating Data Cards for different types of datasets, we identified 31 broad, generalizable themes (Table 2) that comprehensively describe any dataset (**O2**). However, themes vary in

Table 2: Content themes in the Data Card template. Our content schema extends the constitution of traditional dataset documentation to include explanations, rationales, and instructions pertaining to 31 themes. We anticipate that not all themes will be uniformly relevant to all datasets or equally applicable to features within a single dataset.

(1) The publishers of the dataset and access to them	(17) The data collection process (inclusion, exclusion, filtering criteria)
(2) The funding of the dataset	(18) How the data was cleaned, parsed, and processed (transformations, sampling, etc.)
(3) The access restrictions and policies of the dataset	(19) Data rating in the dataset, process, description and/or impact
(4) The wipeout and retention policies of the dataset	(20) Data labeling in the dataset, process, description and/or impact
(5) The updates, versions, refreshes, additions to the data of the dataset	(21) Data validation in the dataset, process, description and/or impact
(6) Detailed breakdowns of features of the dataset	(22) The past usage and associated performance of the dataset (eg. models trained)
(7) Details about collected attributes which are absent from the dataset or the dataset's documentation	(23) Adjudication policies and processes related to the dataset (labeler instructions, inter-rater policy, etc.)
(8) The original upstream sources of the data	(24) Relevant associated regulatory or compliance policies (GDPR, licenses, etc.)
(9) The nature (data modality, domain, format, etc.) of the dataset	(25) Dataset Infrastructure and/or pipeline implementation
(10) What typical and outlier examples in the dataset look like	(26) Descriptive statistics of the dataset (mean, standard deviations, etc.)
(11) Explanations and motivations for creating the dataset	(27) Any known patterns (correlations, biases, skews) within the dataset
(12) The intended applications of the dataset	(28) Human attributes (socio-cultural, geopolitical, or economic representation)
(13) The safety of using the dataset in practice (risks, limitations, and trade-offs)	(29) Fairness-related evaluations and considerations of the dataset
(14)Expectations around using the dataset with other datasets or tables (feature engineering, joining, etc.)	(30) Definitions and explanations for technical terms used in the Data Card (metrics, industry-specific terms, acronyms)
(15) The maintenance status and version of the dataset	(31) Domain-specific knowledge required to use the dataset
(16) Difference across previous and current versions of the dataset	

importance on a per-task basis to stakeholders. Sections in our template (F) capture these themes, further demonstrating how they are deconstructed into sets of scopes (3.2.1). To illustrate the differences in descriptions of a theme elicited per dataset, we include two Data Cards from our case studies (4.1, 4.2) in appendix D and E respectively.

3.3.1 OFTEn Framework. Over time, we found it necessary to develop a consistent and repeatable approach to identify and add new themes from dataset life cycles in a Data Card that are reportable by everyone in the organization. Additionally, certain topics such as consent, can span entire dataset life cycles with different implications at each stage. We introduce OFTEn, a conceptual tool for systematically considering how topics promulgate across all parts of a Data Card (**P1, P3**), through detailed inductive and deductive dataset transparency investigations.

OFTEn (Table 3) abbreviates common stages in the dataset life cycle ("Origins, Factuals, Transformations, Experience, and n=1 *example*"). Though ordered, stages are loosely defined to mirror typical non-linear dataset development practices. Notably, agents' use of the dataset is considered a distinct stage in OFTEn, affording the flexibility to incorporate feedback from downstream stakeholders (dataset consumers, product users, and even data contributors). This establishes a trail to track the performance of AI systems trained and evaluated on the dataset, and exposes any caveats or limitations that potential agents should be aware of.

An OFTEn analysis of the dataset can preemptively enable the discovery of insights that would otherwise not be generally evident. Inductively, OFTEn supports activities with agents to formulate questions about datasets and related models that are important for decision-making. At its simplest, it can be visualized as a matrix in which rows represent the dataset life cycle, and columns provide prompts to frame questions (who, what, when, where, why, and how) about a given topic in the dataset's lifecycle (Table 3). Its participatory use enables reporting both dataset attributes and implicit information that can affect outcomes in real-world deployment. Deductively, we use OFTEn to assess if a Data Card accurately represents the dataset, resulting in formative effects on both, documentation and dataset. Lastly, we find that Data Cards with a clear underlying OFTEn structure are easy to expand and update. This structure allows Data Cards to capture information *over time*, such as feedback from downstream agents, notable differences across versions, and ad-hoc audits or investigations from producers or agents.

3.4 Evaluation of Data Cards

We worked with over 18 producers to understand workflows of creating and maintaining Data Cards, and conducted an interview study (n=10) to validate our observations. While a detailed report of this study is out of scope of this paper, we found that producers had a tendency to fork completed Data Cards (which described similar datasets) as a starter template instead of using the provided template. While this practice made Data Cards easier to complete, it resulted in an increase in inaccurate responses, the propagation of errors and modifications to templates in forked Data Cards. Producers would delete blocks and sections that were irrelevant to their dataset, and in specific cases, producers would semantically modify questions to suit their datasets. Though justifiable in the context of a single Data Card, these practices led to the subsequent fragmentation of forked Data Cards. Deleted but relevant questions were irrecoverable, and reconciling updates to the original template was labor-intensive. Finally, we observed that Producers resorted to answering "N/A" when they were unsure of the answer, or when uncertainty was high. These real-world constraints motivated us to identify mechanisms for assuring the quality of Data Cards, expand organizational vocabularies on uncertainty, and introduce low-barrier processes across the dataset lifecycle that can be easily adopted by organizations.

Initially, each new Data Card created was assigned two reviewers representing job functions typical to agents. Selected reviewers were always unfamiliar with the dataset, but typically fluent in manipulating data or the domain of the dataset. Despite their expertise, feedback provided on these Data Cards were observational and speculative in nature (*"The first two listed applications are commonly used and should be understood by both practitioners and laypeople, but I'm not sure about [application]*); and often not tactical enough for producers to incorporate into the Data Card. To make reviewer feedback actionable and holistic, we worked with a mix of subject matter experts, data reviewers, functional and tactical roles at our company to identify 98 concepts used to assess datasets and their documentation. From these, we excluded 13 usability and 8 user-experience related concepts, which are captured in our objectives. We then consolidated the remaining concepts into 20 clusters using affinity mapping. Clusters were

Table 3: The OFTEn framework

	Description	Themes
Origins	Various planning activities such as problem formulation, defining requirements, design decisions, collection or sourcing methods, and deciding policies which dictate dataset outcome	*Authorship, Motivations, Intended Applications, Unacceptable uses, Licenses, Versions, Sources, Collection Methods, Errata, Accountable parties*
Factuals	Statistical and other computable attributes that describe the dataset, deviations from the original plan, and any pre-wrangling analysis and investigations, including those pertaining to biases and skews	*Number of Instances, Number of Features, Number of Labels, Breakdown of subgroups, Description of features, Taxonomies of labels, Missing/Duplicates, Inclusion and exclusion criteria*
Transformations	Various operations such as filtering, validating, parsing, formatting, and cleaning through which raw data is transformed into a usable form including labeling or annotation policies, validation tasks, feature engineering and related modifications	*Rating or Annotation, Filtering, Processing, Validation, Synthetic features, Handling of PII, Sensitive Variables, Fairness Analyses, Impact Assessments, Skews & Biases*
Experience	Dataset is benchmarked or deployed in experimental, production, or research practice, including specific tasks, access training requirements, modifications made to suit the task, analyses, unexpected behaviors, limitations, caveats and comparisons to similar datasets	*Intended Performance, Unintended Application, Unexpected Performance, Caveats, Extended Use Cases, Safety of Use, Downstream Outcomes, Use & Use Case Evaluation*
N=1 (examples)	Examples in the dataset, including typical, outlier, raw and transformed examples; concrete examples or links to additional artifacts of relevance; links to guided or unguided explorers of datapoints in the dataset	*Examples or links to typical examples and outliers ; Examples that yield errors; Examples that demonstrate handling of null or zero feature values; code blocks & scripts, extended documentations, web demos*

then classified into five umbrella topics or "dimensions" that represent contextual decision-making signals used by our experts to evaluate the rigor with which a Data Card describes a dataset, and it's corresponding efficacy for the reader.

3.4.1 Dimensions. Dimensions are directional, pedagogic vectors that describe the Data Card's usefulness to the agents. They represent the different types of judgments readers might make, and yield qualitative insights into the consistency, comprehensiveness, utility, and readability of Data Card templates and completed Data Cards alike. Here, we briefly summarize these dimensions:

- **Accountability**: Demonstrates adequate ownership, reflection, reasoning, and systematic decision making by producers.
- **Utility or Use:** Provides details that satisfy the needs of the readers' responsible decision-making process to establish the suitability of datasets for their tasks and goals.
- **Quality:** Summarizes the rigor, integrity and completeness of the dataset, communicated in a manner that is accessible and understandable to many readers.
- **Impact or Consequences of Use:** Sets expectations for positive and negative outcomes as well as subsequent consequences when using or managing the dataset in suitable contexts.
- **Risk and Recommendations:** Makes readers aware of known potential risks and limitations, stemming from provenance, representation, use, or context of use. Provides enough information and alternatives to help readers make responsible trade-offs.

Reviewers with varying levels of domain and data fluency were asked to test the aforementioned dimensions, set up as a rubric for grading, during their evaluations of Data Cards *and* any associated Model Cards. Reviewers were asked independently rate the completed Data Card on each dimension, using a 5-point scale with choices *Poor*, *Borderline*, *Average*, *Good*, and *Outstanding*. In addition, they were asked to provide evidence in support of their ratings, and steps that producers could take to improve that specific rating. Reviewers found it easier to offer structured and actionable feedback using these dimensions (*"Utility or Use: Average. Evidence: Data Card provides all necessary steps for users who may wish to access the dataset, but it's hard for me to determine what use cases are suitable for this dataset. I know the dataset was collected for the purpose of evaluating the performance of the [specific model], but what does the [specific model] do? Next Steps: Provide additional examples of suitable use cases, provide additional detail on what the [specific model] does under intended use case."*). Multiple reviewers reported feeling more confident in their assessments. While these dimensions are primarily used to asses if Data Cards help readers arrive at acceptable conclusions about datasets, feedback from expert reviewers revealed specific opportunities to enhance the datasets themselves.

4 CASE STUDIES

4.1 A Computer Vision Dataset for Fairness Research

A research team 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. Sampled from Open Images [20], the dataset included 100,000 bounding boxes over 30,000 images. Each bounding box was manually annotated with perceived gender 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 parties were involved in the creation of this Data Card [12], which started after the dataset was prepared. First, the dataset authors who had deep tacit knowledge of the processes and decisions across the dataset's lifecycle. They also had explicit knowledge from extensive analysis performed for the dataset release. However, this was distributed across several documents, and the Data Card was an exercise in organizing knowledge into a "readable format" that could be consistently repeated for multiple datasets. This process occurred asynchronously over a few days.

The next group involved were internal reviewers 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, in alignment with an established set of AI Principles [24]. The reviewers recommended that the team create a Data Card. Creating the Data Card as a result of the review process revealed differences in perception across experts. For example, in the Data Card, producers noted that nearly 40% of perceived age-range labels were 'unknown'. Reviewers were unable to ascertain if this was acceptable, and subsequent conversations raised further questions about the criteria used to label a bounding box with 'unknown' perceived age-range. It was found that 'high' levels of unknowns were relatively typical to datasets in this problem space, and was attributed to the size of 30% of the bounding boxes being less than 1% of the image. As a result, producers added a custom section about bounding boxes to the Data

Card, and created additional supporting visualizations. Further, producers uncovered and iterated on additional Data Card fields for future CV datasets.

The last group involved in the creation of the Data Card were the authors of this paper, who provided human-centered design perspectives on the Data Card. Feedback was primarily geared towards uncovering agent information needs for acceptable conclusions about the accountability, risk & recommendations, uses, consequences, and quality of the dataset (3.4.1). A post-launch retrospective revealed that though the producers did not have access to dataset consumers, downstream agents reported finding the Data Card useful, and requested Data Card templates for their own use.

4.2 A Geographically Diverse Dataset for Language Translation

A team of software engineers and a product manager noticed that certain models were attentive to names to classify a person's perceived gender. Upon investigation, it was found that previous training datasets had insufficient names that belonged to a non-American geography or were uncommon in English. It was also found that model creators were making assumptions about these datasets. In response, the team decided to create a geographically diverse evaluation 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 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 [7]. 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. They specifically considered what they did not know about the dataset, their assumptions, the advantages and limitations of the dataset. In doing so, the team was forced to rethink design decisions which increased the overall timeline, but resulted in a more principled and intentional dataset of geographically diverse biographies.

The team utilized the Data Card to engage in overall clearer discussions with stakeholders. 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 found that some discussions around the Data Card were actually about the dataset, and noted the usefulness of this feedback if received during the design stage. The final Data Card describes the data selection criteria, sampling criteria, sources of fields, and emphasizes the distribution of countries by continental regions. In addition, the team was able to clearly justify reasons for not including non-binary individuals, excluding collected data, and the limitations of this dataset.

5 DISCUSSION

5.0.1 Experiences and outcomes from Case Studies. 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 alongside the dataset 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 lifecycle 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.

5.0.2 Data Cards as Boundary Objects. Data Cards are designed to embody a high degree of interpretive flexibility [21]. A single Data Card can support tasks such as conducting reviews and audits, determining use in AI systems or research, comparison of multiple datasets, reproduction of research, or tracking of dataset adoption by various groups. For example, data practitioners seeking to evaluate the quality of a dataset for benchmarking or analysis; AI practitioners determining use case suitability of a dataset for deployment in new or existing models; product managers assessing the downstream effects to make data-related decisions about model or product optimizations for the desired user experience; policy stakeholders evaluating the representativeness of a dataset in relation to end users, and the role of various agencies involved in the creating the dataset creation. Importantly, while Data Cards are able to hold a common identity across these groups, they allow stakeholders to analytically make decisions using dimensions, constructs and vocabulary that are meaningful to their own communities of practice. Data Cards are able to facilitate collaborative work across stakeholders, while supporting individual decision making without consensus.

Our design of Data Cards enables the embedding of relevant sections into transparency artifacts that describe ML models and AI systems. Conversely, sections in the Data Card are designed to capture documentation surrounding the use of datasets in ML models. This establishes a network of artifacts that stakeholders can examine when conducting fairness and accountability interrogations, and achieve overall better results for meta-problems across the domain such as knowledge transfer, dataset reusability, organizational governance, and oversight mechanisms. Data Cards, therefore, effectively act as boundary objects [28] and where relevant, boundary infrastructures.

5.0.3 Path to Adoption. Following our initial Data Card release [5], public and private organizations have since sought to adopt similar constructs ([16], [17], [3]). Within our organization, we observed an increase in *non-mandated* Data Cards created by individuals who organically came across completed Data Cards. While these speak to the utility of Data Cards as a documentation artifact, its quality and comprehensiveness depend on the rigor of the producers, the nuance in expressing uncertainty, and their knowledge of the dataset. Organizational factors include the presence of minimum or mandatory content requirements, process incentives, training materials, and infrastructure for creating and sharing Data Cards. While we propose a relatively comprehensive template for documenting datasets in Data Cards, industry-wide adoption could be spurred by agreed-upon interoperability and content standards that serve as a means for producers and agents to develop more equitable mental models of datasets.

5.0.4 Infrastructure and Automation. Critical to an organization's success is its ability to tailor Data Cards to their datasets, models, and technological stack. Knowledge management infrastructures must be connected to data and model pipelines so new knowledge can be seamlessly incorporated into the Data Card, keeping it up to date. We find that Blocks allows for easy implementation on interactive platforms (digital forms, repositories, dataset catalogs) and adaptation for non-interactive surfaces (PDFs, documents, physical papers, markdown files). While both these case studies produced static PDFs, sections and fields can be easily implemented in a browser-based user interface, configured for views tailored to different stakeholders.

Centralized repositories that can perform search-and-filter operations over hundreds of Data Cards have long-tail benefits for agents in identifying the most suitable datasets for their tasks; measurably distributing the

accountability of how datasets are used. We observed a marked preference for infrastructures that enables stakeholder collaboration and co-creation of Data Cards, linking and storage of extraneous artifacts, and the partial automation of visualizations, tables and analyses results. Interestingly, we observed that readers had strong opinions about *not* automating certain fields in the Data Card, especially when responses contain assumptions or rationales that help interpret results. Fields should be automated to guarantee accuracy and antifragility at all times, preventing the misrepresentation and the subsequent legitimizing of poor quality datasets. Implicit knowledge is articulated by providing contextual, human-written explanations of methods, assumptions, decisions and baselines. We find that adopting a co-creative approach that spans the entire dataset life cycle will result in a deliberate approach to automation in documentation.

6 CONCLUSION

We presented a framework for transparent and purposeful documentation of datasets at scale for responsible AI development, Data Cards. Our underlying approach advances the state of the art by surfacing transparency principles and establishing objectives for transparency; expanding existing paradigms of the constitution of dataset documentation; and by enabling the human-centered design of frameworks for structuring, adapting or expanding, and evaluating Data Cards. We provide an in-depth discussion each framework, and detail qualitative and anecdotal evidence for the efficacy of Data Cards towards creating responsible AI systems through two case studies. A limitation of our approach was the use of Google Docs for Data Card templates. This allowed stakeholders to collaborate and preserved a forensic history of the development of the Data Card, producers were limited to providing answers using text, tables and images. Additionally, this format prevented us from improving template usability through design and automations, a much requested feature from producers. Future work requires a more principled approach for extending and adapting Data Card templates without compromising comparability. Insights from studies call for participatory approaches that engage diverse, non-traditional stakeholders early into the dataset *and* Data Card development process. Lastly, defining quantitative measures to assess the true value of Data Cards will require adoption at both breadth and depth in the industry. To address this, further investigation is needed into the perceived and actual importance of the content of Data Cards to tasks for different stakeholder groups, and requires the expansion of user studies to a broader participant pool spanning multiple industries. Data Cards templates and frameworks encourage customized implementations that foster a culture for deep, detailed, and transparent documentation. Data Cards are capable of thoughtfully explaining the implications of datasets while highlighting unknowns appropriately. They reveal insights about inherent aspects of dataset that cannot be intrinsically determined by interacting with the dataset. Data Cards enable future industry standards of transparency and documentation that emphasize the ethical considerations of a dataset in ways that can be practically acted upon, support production and research decisions, and well-informed development of large AI models with increasingly complex dataset dependencies.

ACKNOWLEDGMENTS

We are grateful to Aybuke Turker for research contributions; Romina Stella, Candice Schumann, Reena Jana and Susanna Ricco for the Data Cards and two case studies presented in this paper; Emily Denton, Lauren Wilcox, Michael Terry, Negar Rostamzadeh, Kathy Meier-Hellstern, Meredith Morris for their feedback and expertise; Tulsee Doshi, Margaret Mitchell, Timnit Gebru, Martin Wattenberg, Fernanda Viegas, Parker Barnes, Dan Nanas, Nicole Maffeo, Will Carter, Sebastian Gehrmann, Catherine Xu, Vivian Tsai, Danielle Smalls, Anthony Keene, and Lora Aroyo for their constant guidance. We thank internal and external workshop and study participants, and attendees of Data Cards Playbook workshop at 2021 CRAFT for their participation and insightful discussions. We also thank the Center for Responsible AI and Human Centered Technology at Google Research for enabling this work. This work was jointly conducted by the Ethical AI and People + AI Research teams, funded by Google Research. The authors declare no additional sources of funding. The legal department of Google participated in the review and approval of the manuscript; and the decision to submit the manuscript for publication. Aside from the authors and their collaborators, Google had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; or preparation of the manuscript. The authors declare no other financial interests.

REFERENCES

[1] 2017. *AI Now Institute.* https://ainowinstitute.org/
[2] 2021. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).* https://facctconference.org/
[3] Joint Artificial Intelligence Center Public Affairs. 2021. Enabling AI with Data Cards. https://www.ai.mil/blog_09_03_21_ai_enabling_ai_with_data_cards.html
[4] Nuno Antunes, Leandro Balby, Flavio Figueiredo, Nuno Lourenco, Wagner Meira, and Walter Santos. 2018. 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, 188–193.
[5] Parker Barnes Anurag Batra. 2020. Open Images Extended - Crowdsourced Data Card. https://research.google/static/documents/datasets/open-images-extended-crowdsourced.pdf
[6] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261 [cs.CY]
[7] Anja Austermann, Michelle Linch, Romina Stella, and Kellie Webster. 2021. https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/Data%20Card.pdf
[8] Iain Barclay, Harrison Taylor, Alun Preece, Ian Taylor, Dinesh Verma, and Geeth de Mel. 2020. A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions. *Concurrency and Computation: Practice and Experience* (2020), e6129.
[9] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.
[10] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 401–413.
[11] Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. 2018. Working with beliefs: AI transparency in the enterprise. In *IUI Workshops*.
[12] Candice chumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. https://storage.googleapis.com/openimages/open_images_extended_miap/Open%20Images%20Extended%20-%20MIAP%20-%20Data%20Card.pdf
[13] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[14] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics* 26, 6 (2020), 3333–3361.
[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).
[16] GEM. 2022. Natural Language Generation, its Evaluation and Metrics Data Cards. https://gem-benchmark.com/data_cards
[17] HuggingFace. 2021. HuggingFace - Create a Dataset Card. https://huggingface.co/docs/datasets/v1.12.0/dataset_card.html
[18] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.
[19] People + AI Research Initiative. 2022. Know Your Data. https://knowyourdata.withgoogle.com/
[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[21] Susan Leigh Star. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, Technology, & Human Values* 35, 5 (2010), 601–617.
[22] Colleen McCue. 2014. *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.
[23] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[24] Sundar Pichai. 2018. AI at Google: our principles. *The Keyword* 7 (2018), 1–3.
[25] Mahima Pushkarna, Andrew Zaldivar, and Daniel Nanas. [n. d.]. Data Cards Playbook: Participatory Activities for Dataset Documentation. https://facctconference.org/2021/acceptedcraftsessions.html#data_cards
[26] Mahima Pushkarna, Andrew Zaldivar, and Vivian Tsai. [n. d.]. Data Cards GitHub Page. https://pair-code.github.io/datacardsplaybook/
[27] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.
[28] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.
[29] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

A RELATED DOCUMENTATION FRAMEWORKS & TOOLKITS

To standardize documentation procedures that convey performance characteristics of AI or aspects that lead to the creation and distribution of datasets, many groups have created frameworks and toolkits to support transparency in AI. Each of these efforts were developed with particular stakeholders and issues in mind. The following is a summary of some of these efforts:

- **Model Cards** is a modular, ethics-informed framework to report trained ML model details [23]. Model Cards consist of qualitative information, such as ethical considerations, target users, and use cases; as well as quantitative information, with an emphasis on model evaluation that is disaggregated (split across the different target subgroups) and intersectional (including evaluation on multiple subgroups in combinations, for example race and gender).
- **Datasheets for Datasets** is a set of questions designed to evoke information about a dataset that reflect key stages in a dataset's lifecycle [15]. Drawing critical analogies from the automobile industry, clinical trials in medicine, and the electronics industry, Datasheets for Datasets is also used as a workflow by: 1.) Dataset creators to guide their thinking during the process of creating, distributing and maintaining a dataset. 2.) Dataset consumers to decide appropriateness for task, strengths, limitations, and place in a broader system associated with the dataset documented.
- **FactSheets** is an extensive set of declaration items intended to disclose information about the creation and deployment of an AI service [6]. Modeled after a supplier's declaration of conformity (SDoC) and similar artifacts used in telecommunications and transportation to demonstrate a service's conformity to regulation, items in FactSheets include: purpose and audience; performance variation; safety and security aspects; and provenance of training data—all to gain trustworthiness of AI services.
- **Data Statements**, originally developed for documenting natural language processing systems, is a practice on how to characterize a dataset using schema elements that minimizes critical scientific and ethical issues—issues that could arise from datasets used in contexts not well suited [9]. In its original form, schema elements in Data Statements featured particular aspects of language datasets, including speech context, speaker demographic and annotator demographic—all of which were inspired by practices from the fields of psychology and medicine that require such disclosure about populations being studied.

B TYPOLOGY OF STAKEHOLDERS

Figure 3: A typology of typical stakeholders in the life cycle of datasets that we created Data Cards for, broken down by type, identifiers and tasks with example roles. We find that including non-technical and indirect stakeholders in a dataset's lifecycle during initial considerations of content and structure builds foresight for successful Data Card adoption.

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, Academics
		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) due to 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

C OFTEN FRAMEWORK AS A GENERATIVE TOOL

Table 4: In this figure, we demonstrate how it can be used to generate questions about data consent across a dataset's life cycle. During the creation of our template (Appendix F or [26]), OFTEn was used to anticipate standardization requirements and enable the forensic investigation of dataset documentation over time.

	Who	What	When	Where	Why
O	Who was responsible for setting the terms of consent?	What were the terms of consent?	When do the terms of consent expire?	Where all are the terms of consent applicable? Are there any exceptions?	Why were these specific terms of consent chosen?
F	How was consent delivered to the surveyed population?	How many data points accompanied consent?	When was the consent collected with respect to data creation or collection?	Where can the consent be accessed? How is it stored?	If at all, why were exceptions made? What happened in cases where consent was not or conditionally provided? Provided but revoked?
T	Who tracks consent?	What manipulations of the data are permissible under the given consent?	When can consent be revoked?	X	Why are said transformations in direct conflict with consent?
E	Under the terms of the consent, who all can use the dataset?	Under the terms of the consent, what are the permissible uses of the dataset?	When must consent be reacquired from individuals to sustain use of the dataset?	Geographically, where all does the consent permit dataset use?	Summarize conditions and rationales that justify the use of data without consent.
N=1	Provide an example of a consent form	Provide an example of a data point with partial consent	X	X	X

D DATA CARD FOR COMPUTER VISION DATASET

Open Images Extended - More Inclusively Annotated People (MIAP)

Dataset Download • Related Publication

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.

Authorship

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

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

FUNDING
Google LLC

FUNDING TYPE
Private Funding

DATASET CONTACT
open-images-extended@google.com

Motivations

DATASET PURPOSE(S)
Research Purposes
Machine Learning
Training, testing, and validation

KEY APPLICATION(S)
Machine Learning | Object Recognition | Machine Learning Fairness

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

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

INTENDED AND/OR SUITABLE USE CASE(S)
- **ML Model Evaluation for:** Person detection, Fairness evaluation
- **ML Model Training for:** Person detection, Object detection

Additionally:
- **Person detection:** Without specifying gender or age presentations
- **Fairness evaluations:** Over gender and age presentations
- **Fairness research:** Without building gender presentation or age classifiers

Use of Dataset

SAFETY OF USE
Conditional Use
There are some known unsafe applications.

UNSAFE APPLICATION(S)
⚠ Gender classification | Age classification

UNSAFE USE CASE(S)
This dataset **should not** be used to create gender or age classifiers. The intention of percieved gender and age labels is to capture gender and age presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender or actual age.

CONJUNCTIONAL USE
Safe to use with other datasets

KNOWN CONJUNCTIONAL DATASET(S)
- The data in this dataset can be combined with Open Images V6

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

METHOD
Object Detection

SUMMARY
A person object detector can be trained using the Object Detection API in Tensorflow.

KNOWN CAVEATS
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.

METHOD
Fairness Evalutaion

SUMMARY
Fairness evaluations can be run over the splits of gender presentation and age presentation.

KNOWN CAVEATS
There still exists a gender presentation skew towards unknown and predominantly masculine, as well as an age presentation range skew towards middle.

Data Card v2.0 • Published June 2021 • Updated Nov 2021 Page 1 of 5

Figure 4: Data Card for Computer Vision Dataset, Page 1 of 5

Open Images Extended - (MIAP)

Dataset Snapshot

PRIMARY DATA TYPE(S)

Non-Sensitive Public Data about people

DATASET SNAPSHOT

Total Instances	100,000
Training	70,000
Validation	7,410
Testing	22,590
Total boxes	454,331
Total labels	908,662
Average labels per image	9.08
Human annotated labels	All

DESCRIPTION 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*). This adds adds nearly 100,000 new boxes that were not annotated under the original labeling pipeline of the core Open Images Dataset.

Note: 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 MODALITY

Labels or Annotations

KNOWN CORRELATION(S)

- Gender presentation numbers are skewed towards predominantly perceived as `masculine` & `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`

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 Common Visual Data Foundation (CVDF) repository.

Datapoint Example

EXAMPLE OF ACTUAL DATA POINT WITH DESCRIPTIONS

Field	Value	Description
`ImageID`	`164b0e6d1fcf8e61`	The image this box lives in
`LabelName`	`/m/01g317`	Labels are identified by MIDs (Machine-generated Ids) as can be found in Freebase or Google Knowledge Graph API. Label descriptions here
`Confidence`	`1`	A dummy value, always 1
`XMin`	`0.897112`	Normalized image coordinates indicating the leftmost pixel of the annotation
`XMax`	`0.987365`	Normalized image coordinates indicating the rightmost pixel of the annotation
`YMin`	`0.615523`	Normalized image coordinates indicating the topmost pixel of the annotation
`YMax`	`0.895307`	Normalized image coordinates indicating the bottomost pixel of the annotation
`IsOccluded`	`0`	Binary value indicating if the object is occluded by another object in the image
`IsTruncated`	`1`	Binary value indicating if the object extends beyond the boundary of the image
`IsGroupOf`	`0`	Binary value indicating if the box spans a group of objects
`IsDepictionOf`	`1`	Binary value indicating if the object is a depiction and not a real physical instance
`IsInsideOf`	`1`	Binary value indicating if the image is taken from the inside of the object
`IsInsideOf`	`1`	Binary value indicating if the limage is taken from the inside of the object
`GenderPresentation`	`Predominantly Masculine`	Indicates the perceived gender presentation of the subject assessed by a third party
`AgePresentation`	`Middle`	Indicates the perceived age range of the subject assessed by a third party

Data Collection

DATA COLLECTION METHOD(S)

Derived

Vendor Collection Efforts

DATA SOURCES BY COLLECTION METHOD(S)

Images	Open Images V6
Labels	Human annotators
Bounding Boxes	Human annotators

SUMMARIES OF DATA COLLECTION METHODS

100,000 images randomly sampled from the positive set of Open Images V6, which contains approximately 9.9M images

- Training Set: 70,000 sampled from 9,011,219 images
- Testing/Validation: 30,000 sampled from 167,056 images

EXCLUDED DATA

No excluded data

DATA SELECTION CRITERIA - SCRAPING

- Images were sampled from the positive subset of training and testing/validation containing annotator-verified image lables
- Images contained at least one of five person classess (`man`, `woman`, `boy`, `girl`, or `person`)

Note: We did not include non-binary as a class label as it is not possible to label gender identity from images. Gender identity should only be used in situations where participants are able to self-report gender.

Data Card v2.0 • Published June 2021 • Updated Nov 2021 Page 2 of 5

Figure 5: Data Card for Computer Vision Dataset, Page 2 of 5

Open Images Extended - (MIAP)

Labelling Process

METHOD(S)

Human labels

LABEL TYPE(S)

Human Attributes Labels	
PerceivedGender	Human annotators
PercievedAge	
Bounding Boxes (where missing)	
rectangular box	Drawn by human annotators, computed into normalized image coordinates
IsTruncated	Object attributes annotated by human annotators to describe the bounding box
IsOccluded	
IsGroup	
IsInside	
IsDepiction	

METHOD(S) SUMMARY

Compensated workers based out of India were recruited through vendors to annotate and re-label images. 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 Box

LABEL DISTRIBUTION

Label	Original	MIAP
boxes	357,870	454,331

Above: Counts of boxes across the MIAP in comparison to the 100,000 samples from Open Images V6. For a more detailed breakdown, see our paper.

LABEL DESCRIPTION(S)

Bounding Box: A rectangular bounding box around each person in an image. Object Attributes include: is truncated, is occluded, is inside, is group, *and* is depiction.

LABELING TASK(S) OR PROCEDURE(S)

"Create the bounding box around all people"
"Label object attributes"
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

LABEL DISTRIBUTION

Label	Original	MIAP
Predominantly feminine	76,283	100,672
Predominantly masculine	143,320	174,047
Unknown gender presentation	138,267	179,612

Above: Counts of boxes for different classes of the perceived gender label across the MIAP in comparison to the 100,000 samples from Open Images V6. For a more detailed breakdown, see our paper.

LABEL DESCRIPTION(S)

Classes for the perceived gender presentation label are:

- predominantly feminine
- predominantly masculine
- unknown

LABELING TASK(S) OR PROCEDURE(S)

"Label the perceived gender presentation"
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.

Note: Gender presentation for people marked as **young** is always set to **unknown**.

LABEL TYPE

Perceived Age

LABEL DISTRIBUTION

Label	Original	MIAP
young	21,548	28,806
middle	198,055	233,674
older	no such label	9,023
Unknown	138,267	182,828

Above: Counts of boxes for different classes of the perceived age label across the MIAP in comparison to the 100,000 samples from Open Images V6. For a more detailed breakdown, see our paper.

LABEL DESCRIPTION(S)

Classes for the perceived age range label are:

- young
- middle
- older
- unknown

LABELING TASK(S) OR PROCEDURE(S)

"Label the perceived age range"
Annotators were asked to select either 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.

Data Card v2.0 • Published June 2021 • Updated Nov 2021 Page 3 of 5

Figure 6: Data Card for Computer Vision Dataset, Page 3 of 5

Open Images Extended - (MIAP)

Human Attributes

HUMAN ATTRIBUTE(S)

Age

Gender

ATTRIBUTE(S) INTENTIONALITY

`PerceivedGender`	Intended
`PercievedAge`	Intended

SUMMARY OF INTENTIONS

This data collection and annotation effort was primarily introduced to help fairness research and evaluations. The intention of perceived gender 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.

ATTRIBUTE TYPE

Perceived Gender

REPRESENTED SUBGROUPS DISTRIBUTION

`Predominantly feminine`	22.2%
`Predominantly masculine`	38.3%
`Unknown gender presentation`	39.5%

EXPECTATIONS, RISKS, & 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.

SOURCES OF SUBGROUPS

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.

ATTRIBUTE TYPE

Perceived Age

REPRESENTED SUBGROUPS DISTRIBUTION

`young`	6.3%
`middle`	51.4%
`older`	2.0%
`Unknown`	4.2%

EXPECTATIONS, RISKS, & 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.

SOURCES OF SUBGROUPS

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`.

TRADEOFFS

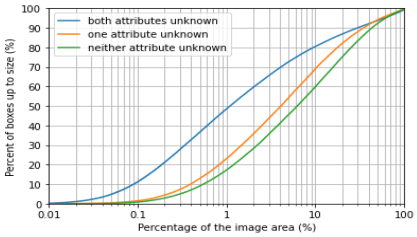
Although these labels do not represent the true age ranges of individuals in images, they are still valuable because they allow researchers to assess the performance of models across age ranges, which can ultimately lead to less biased models that work well for all users.

Additional Analysis

ANALYSIS

Box Size

BOX SIZE BY ATTRIBUTES



REASONS FOR DIFFERENCES

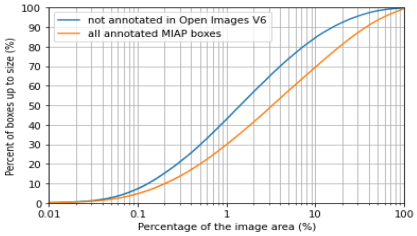
Many boxes are annotated with either unknown perceived gender presentation or perceived age range. These bounding boxes are typically smaller, corresponding to people that are either farther away or occluded in some way.

- 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.

EXAMPLES OF BOX SIZES

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

BOX SIZE FOR PREVIOUSLY MISSING ANNOTATIONS



REASONS FOR DIFFERENCES

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.

Data Card v2.0 • Published June 2021 • Updated Nov 2021 Page 4 of 5

Figure 7: Data Card for Computer Vision Dataset, Page 4 of 5

Open Images Extended - (MIAP)

License & Access

LICENSE TYPE(S)

CC-BY-SA 4.0

LICENSE BREAKDOWN

Annotations are licensed by Google LLC under CC BY 4.0 License. The images (available separately) are listed as having a CC BY 2.0 license.

CC-BY-SA 4.0

Note: We make no representations or warranties regarding the license status of each image and you should verify the license for each image yourself.

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.
- **No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

ACCESS TYPE(S)

Open Access

ACCESS COST

N/A – Open Acess

CC-BY-SA 4.0

ACCESS PREREQUISITE(S)

Read note on perceived gender presentations and percieved age presentation, and acceptable use.

ACCESS SUPPORT

Dataset Website

Research Paper

DATASET WEBSITE

https://storage.googleapis.com/openimages/web/extended.html#miap

RESEARCH PAPER

https://storage.googleapis.com/openimages/web/extended.html#miap

ACCESS DETAILS

Dataset includes bounding box annotations only. Images are accessed separately.

CITATION GUIDELINE(S)

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}
}
```

Versioning & Maintenance

VERSION STATUS

Actively Maintained

No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.

DATASET STATUS

Version	1.0
Last Updated	05/2021
First Released	05/2021

Note: Annotations were completed between late 2019 - early 2020.

MAINTENANCE PLAN

Updates to the dataset are pushed on the dataset website

Data Card v2.0 • Published June 2021 • Updated Nov 2021 Page 5 of 5

Figure 8: Data Card for Computer Vision Dataset, Page 5 of 5

E 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 9: 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 10: 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%
Philipines	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 11: Data Card for Language Translation Dataset, Page 3 of 3

F DATA CARD TEMPLATE

Dataset Name (Acronym)	Write a short summary describing your dataset (limit 200 words). Include information about the content and topic of the data, sources and motivations for the dataset, benefits and the problems or use cases it is suitable for.
DATASET LINK	DATA CARD AUTHOR(S)
Provide a link to the dataset:	*Select **one role per** Data Card Author:* *(Usage Note: Select the most appropriate choice to describe the author's role in creating the Data Card.)*
Dataset Link	**Name, Team:** (Owner / Contributor / Manager) **Name, Team:** (Owner / Contributor / Manager) **Name, Team:** (Owner / Contributor / Manager)

Figure 12: Data Card Template - The *Summary* section introduces the dataset and the authors of the Data Card.

Authorship

Publishers

PUBLISHING ORGANIZATION(S)	INDUSTRY TYPE(S)	CONTACT DETAIL(S)
Provide the names of the institution or organization responsible for publishing the dataset:	*Select **all applicable** industry types to which the publishing organizations belong:*	*Provide publisher contact details:*
Organization Name	Corporate - Tech Corporate - Non-Tech (please specify) Academic - Tech Academic - Non-Tech (please specify) Not-for-profit - Tech Not-for-profit - Non-Tech (please specify) Individual (please specify) Others (please specify)	**Publishing POC:** <Provide the name for a POC for this dataset's publishers.> **Affiliation:** <Provide the POC's institutional affiliation.> **Contact:** <Provide the POC's contact details.> **Mailing List:** <Provide a mailing list if available.> **Website:** <Provide a website for the dataset if available.>

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
Provide the names of the groups or team(s) that own the dataset:	*Provide pathways to contact dataset owners:*	*Provide the details of all authors associated with the dataset:* *(Usage Note: Provide the affiliation and year if different from publishing institutions or multiple affiliations.)*
Name of Group or Team	**Dataset Owner(s):** <Provide the names of the dataset owners> **Affiliation:** <Provide the affiliation of the dataset owners> **Contact:** <Provide the email of the dataset owner> **Group Email:** <Provide a link to the Mailing-list@server.com for the dataset owner team> **Website:** <Provide a link to the website for the dataset owner team>	Name, Title, Affiliation, YYYY Name, Title, Affiliation, YYYY Name, Title, Affiliation, YYYY Name, Title, Affiliation, YYYY

Funding Sources

INSTITUTION(S)	FUNDING OR GRANT SUMMARY(IES)
Provide the names of the funding institution(s):	*Provide a short summary of programs or projects that may have funded the creation, collection, or curation of the dataset.* *Use additional notes to capture any other relevant information or considerations.*
Name of Institution Name of Institution Name of Institution	*For example, Institution 1 and institution 2 jointly funded this dataset as a part of the XYZ data program, funded by XYZ grant awarded by institution 3 for the years YYYY-YYYY.* <Summarize here. Link to documents if available.> **Additional Notes:** <Add here>

Figure 13: Data Card Template - The *Authorship* section describes the authors of the dataset. This includes subsections on *Publishers*, which may be different from *Dataset Owners*. The *Funding Sources* subsection describes grants and programs academic, research, and industry organizations that supported the creation of the dataset from.

Dataset Overview

DATA SUBJECT(S)	DATASET SNAPSHOT	CONTENT DESCRIPTION
*Select **all applicable** subjects contained the dataset:*	*Provide a snapshot of the dataset:* *(Use the additional notes to include relevant information, considerations, and links to table(s) with more detailed breakdowns.)*	*Provide a short description of the content in a datapoint.*
Sensitive Data about people Non-Sensitive Data about people Data about natural phenomena Data about places and objects Synthetically generated data 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 Labeled Classes: 123456 Number of Labels: 123456789 Average Labels Per Instance: 123456 Algorithmic Labels: 123456789 Human Labels: 123456789 Other Characteristics: 123456 **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	<Summarize here. Include links if available> **Additional Notes:** <Add here>

DESCRIPTIVE STATISTICS

Provide basic descriptive statistics for each field.
Use additional notes to capture any other relevant information or considerations.
Usage Note: Some statistics will be relevant for numeric data, but not for strings.

Statistic	Field Name	Field Name	Field Name	Field Name	Field Name	Field Name
count						
mean						
std						
min						
25%						
50%						
75%						
max						
mode						

Above: <Provide a caption for the above table or visualization.>

Additional Notes: <Add here>

Sensitivity of Data

SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
	List fields in the dataset that contain S/PII, and specify if their collection was	*Summarize the measures or steps to*

Figure 14: Data Card Template - The *Dataset Overview* section (1/3) of the Data Card was designed as a top-level summary of the dataset that could be included within other transparency artifacts. In those cases, we encourage producers to include a link to a more complete Data Card with other sections.

Sensitivity of Data

SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
*Select **all applicable** data types present in the dataset:*	*List fields in the dataset that contain S/PII, and specify if their collection was intentional or unintentional.* *Use additional notes to capture any other relevant information or considerations.*	*Summarize the measures or steps to handle sensitive data in this dataset.* *Use additional notes to capture any other relevant information or considerations.*
User Content User Metadata User Activity Data Identifiable Data S/PII Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children's Data None Others (Please specify)	**Intentionally Collected Sensitive Data** (S/PII were collected as a part of the dataset creation process.) **Field Name** / **Description** <Field Name> / <Type of S/PII> <Field Name> / <Type of S/PII> <Field Name> / <Type of S/PII> ... / ... **Unintentionally Collected Sensitive Data** (S/PII were not explicitly collected as a part of the dataset creation process but can be inferred using additional methods.) **Field Name** / **Description** <Field Name> / <Type of S/PII> <Field Name> / <Type of S/PII> <Field Name> / <Type of S/PII> ... / ... **Additional Notes:** <Add here>	<Summarize here. Include links and metrics where applicable.> **<method>:** [description] **<method>:** [description] **<method>:** [description] **Additional Notes:** <Add here>

RISK TYPE(S)	SUPPLEMENTAL LINK(S)	RISK(S) AND MITIGATION(S)
*Select **all applicable** risk types presenting from the dataset:*	*Provide link(s) for documentation pertaining to sensitive data in the dataset:*	*Summarize the steps taken to identify and mitigate risks from PII or sensitive information.* *Use additional notes to capture any other relevant information or considerations.*
Direct Risk Indirect Risk Residual Risk No Known Risks Others (Please Specify)	**<Link Name or Document Type>:** [Link] **<Link Name or Document Type>:** [Link] **<Link Name or Document Type>:** [Link]	<Summarize here. Include links and metrics where applicable.> **<Risk type>:** [Description + Mitigations] **<Risk type>:** [Description + Mitigations] **<Risk type>:** [Description + Mitigations] **Additional Notes:** <Add here>

Figure 15: Data Card Template - (Contd., 2/3) The *Sensitivity of Dataset* and *Dataset Version and Maintenance* subsections in the *Dataset Overview* section. The *Sensitivity of Dataset* subsection describes the intentionality, handling, and risks associated with potentially sensitive fields in a dataset.

Dataset Version and Maintenance

MAINTENANCE STATUS	VERSION DETAILS	MAINTENANCE PLAN
*Select **one**:*	*Provide details about **this** version of the dataset:*	*Summarize the maintenance plan for the dataset:* *Use additional notes to capture any other relevant information or considerations.*
Regularly Updated (New versions of the dataset have been or will continue to be made available.) Actively Maintained (No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.) Limited Maintenance (The data will not be updated, but any technical issues will be addressed.) Deprecated (This dataset is obsolete or is no longer being maintained.)	**Current Version:** 1.0 **Last Updated:** MM/YYYY **Release Date:** MM/YYYY	<Summarize here. Include links and metrics where applicable.> **Versioning:** <Summarize here. Include information about criteria for versioning the dataset.> **Updates:** <Summarize here. Include information about criteria for refreshing or updating the dataset.> **Errors:** <Summarize here. Include information about how errors are triaged or handled.> **Feedback:** <Summarize here. Include information for providing feedback.> **Additional Notes:** <Add here>
	NEXT PLANNED UPDATE(S)	EXPECTED CHANGE(S)
Fill this row if this dataset is regularly updated, actively maintained or another version is planned.	*Provide details about the next planned update:*	*Summarize the updates to the dataset and/or data that are expected on the next update.* *Use additional notes to capture any other relevant information or considerations.*
	Version affected: 1.0 **Next data update:** MM/YYYY **Next Version:** 1.1 **Next Version update:** MM/YYYY	**Updates to Data:** <Summarize here. Include links, charts, and visualizations as appropriate.> **Updates to Dataset:** <Summarize here. Include links, charts, and visualizations as appropriate.> **Additional Notes:** <Add here>

Figure 16: Data Card Template - (Contd., 3/3) The *Dataset Version and Maintenance* subsection in the *Dataset Overview* section.

Example of Data Points

PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS
*Select **one**:*	*Provide link(s) to data points or exploratory demos:*	*List the fields in data points and their descriptions.* *(Usage Note: Describe each field in a datapoint. Optionally use this to show the example.)*
Image Data Text Data Tabular Data Audio Data Video Data Time Series Graph Data Geospatial Data Multimodal (Please Specify) Unknown Others (Please specify)	[Demo Link] [Typical Data Point Link] [Outlier Data Point Link] [Other Datapoint Link] [Other Datapoint Link]	**Field Name** / **Field Value** / **Description** <Field Name> / <Field Value> / <Description> <Field Name> / <Field Value> / <Description> <Field Name> / <Field Value> / <Description> ... / ... / ... **Above:** <Provide a caption for the above table or visualization if used.> **Additional Notes:** <Add here>

TYPICAL DATA POINT	ATYPICAL DATA POINT
Provide an example of a typical data point and describe what makes it typical. ***Use additional notes to capture any other relevant information or considerations.***	*Provide an example of an outlier data point and describe what makes it atypical.* ***Use additional notes to capture any other relevant information or considerations.***
<Summarize here. Include any criteria for typicality of datapoint>	<Summarize here. Include any criteria for atypicality of datapoint>
``` {'q_id': '8houtx', 'title': 'Why does water heated to room temperature feel colder than the air around it?', 'selftext': '', 'document': '', 'subreddit': 'explainlikeimfive', 'answers': {'a_id': ['dylcnfk', 'dylcj49'], 'text': ["Water transfers heat more efficiently than air. When something feels cold it's because heat is being transferred from your skin to whatever you're touching. ... Get out of the water and have a breeze blow on you while you're wet, all of the water starts evaporating, pulling even more heat from you."], 'score': [5, 2]}, 'title_urls': {'url': []}, 'selftext_urls': {'url': []}, 'answers_urls': {'url': []}} ```	``` {'q_id': '8houtx', 'title': 'Why does water heated to room temperature feel colder than the air around it?', 'selftext': '', 'document': '', 'subreddit': 'explainlikeimfive', 'answers': {'a_id': ['dylcnfk', 'dylcj49'], 'text': ["Water transfers heat more efficiently than air. When something feels cold it's because heat is being transferred from your skin to whatever you're touching....Get out of the water and have a breeze blow on you while you're wet, all of the water starts evaporating, pulling even more heat from you."], 'score': [5, 2]}, 'title_urls': {'url': []}, 'selftext_urls': {'url': []}, 'answers_urls': {'url': []}} ```
Additional Notes: <Add here>	**Additional Notes:** <Add here>

Figure 17: Data Card Template - The *Example of Data Points* section is designed to help readers interpret and first-hand explore data points in the dataset without needing to download the dataset. This improves both the use of the dataset and usability of the Data Card.

Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
*Select **one**:*	*Provide a list of key domains of application that the dataset has been designed for:* *(Usage Note: Use comma-separated keywords.)*	*List the primary motivations for creating or curating this dataset:* *(Usage Note: use this to describe the problem space and corresponding motivations for the dataset.)*
Monitoring Research Production Others (Please Specify)	*For example: `Machine Learning`, `Computer Vision`, `Object Detection`.* `keyword` , `keyword`, `keyword`	*For example:* *- Bringing demographic diversity to imagery training data for object-detection models.* *- Encouraging academics to take on second-order challenges of cultural representation in object detection.* <Summarize motivation here. Include links where relevant.>

Intended Use

DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
*Select **one**:*	*Summarize known suitable and intended use cases of this dataset.* *Use additional notes to capture any specific patterns that readers should look out for, or other relevant information or considerations.*	*Summarize known unsuitable and unintended use cases of this dataset.* *Use additional notes to capture any specific patterns that readers should look out for, or other relevant information or considerations.*
Safe for production use Safe for research use Conditional use- some unsafe applications Only approved use Others (Please specify)	**[Suitable Use Case]** : <Summarize here. Include links where necessary.> **[Suitable Use Case]**: <Summarize here. Include links where necessary.> **[Suitable Use Case]**: <Summarize here. Include links where necessary.> **Additional Notes:** <Add here>	**[Unsuitable Use Case]** : <Summarize here. Include links where necessary.> **[Unsuitable Use Case]**: <Summarize here. Include links where necessary.> **[Unsuitable Use Case]**: <Summarize here. Include links where necessary.> **Additional Notes:** <Add here>

	RESEARCH AND PROBLEM SPACE(S)	CITATION GUIDELINES
	Provide a description of the specific problem space that this dataset intends to address.	*Provide guidelines and steps for citing this dataset in research and/or production work.* *Use additional notes to capture any specific patterns that readers should look out for, or other relevant information or considerations.*

	<Summarize here. Include any specific research questions.>	**Guidelines & Steps:** <Summarize here. Include links where necessary.>
		BiBTeX:
		``
		@article{kuznetsova2020open,
		title={The open images dataset v4},
		author={Kuznetsova, Alina and Rom, Hassan and Alldrin, and others},
		journal={International Journal of Computer Vision},
		volume={128},
		number={7},
		pages={1956--1981},
		year={2020},
		publisher={Springer}
		}
		``
		
		Additional Notes: <Add here>

Figure 18: Data Card Template - The *Motivations and Intentions* section asks producers to describe their motivations for creating the dataset, as well as the intended uses of the dataset. The *Motivations* subsection sets up the domain of research or application as well as the specific problems the dataset was designed for. We encourage producers to describe *known* suitable and unsuitable use cases for their dataset in the *Intended Use* subsection since it is impossible to list every possible use case of datasets.

Access, Retention, & Wipeout

Access

ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
*Select **one**:*	*Provide links that describe documentation to access this dataset:*	*Please describe any required training or prerequisites to access this dataset.*
Internal - Unrestricted Internal - Restricted External - Open Access Others (Please specify)	[Dataset Website URL] [Github URL]	*For example,* *This dataset requires membership in [specific] database groups:* • *Complete the [Mandatory Training]* • *Read [Data Usage Policy]* • *Initiate a Data Requesting by filing [a bug]*
	POLICY LINK(S)	ACCESS CONTROL LIST(S)
	Provide a link to the access policy:	*List and summarize any access control lists associated with this dataset. Include links where necessary.* *Use additional notes to capture any other information relevant to accessing the dataset.*
	• Direct download URL • Other repository URL Code to download data `# ...`	**[Access Control List]:** <Write summary and notes here.> **[Access Control List]:** <Write summary and notes here.> **[Access Control List]:** <Write summary and notes here.> **Additional Notes:** <Add here>

Retention

	DURATION	POLICY SUMMARY
	Specify the duration for which this dataset can be retained:	*Summarize the retention policy for this dataset.*
	<Specify duration in days, months, or years.>	**Retention Plan ID:** <Write here> **Summary:** <write summary and notes here>
	PROCESS GUIDE	EXCEPTION(S) AND EXEMPTION(S)
	Summarize any requirements and related steps to retain the dataset. *Use additional notes to capture any other relevant information or considerations.*	*Summarize any exceptions and related steps to retain the dataset. Include links where necessary.* *Use additional notes to capture any other relevant information or considerations.*
	For example, *This dataset complies with [standard policy guidelines]* **Additional Notes:** <Add here>	**Exemption Code:** `ANONYMOUS_DATA` / `EMPLOYEE_DATA` / `PUBLIC_DATA` / `INTERNAL_BUSINESS_DATA` / `SIMULATED_TEST_DATA` **Summary:** <Write summary and notes here.> **Additional Notes:** <Add here>

Wipeout and Deletion

	DURATION	DELETION EVENT SUMMARY
	Specify the duration after which this dataset should be deleted or wiped out:	*Summarize the sequence of events and allowable processing for data deletion.* *Use additional notes to capture any other relevant information or considerations.*

Figure 19: Data Card Template - The *Access, Retention and Wipeout* section (1/2) is decomposed into separate subsections. The *Access* subsection details the storage locations of the dataset, as well as any pre-requisites and policies that govern access to the dataset. This is particularly important for regulated industries. The *Retention* subsection describes the retention duration and summarizes the retention policies and exceptions that are applicable to the dataset.

Wipeout and Deletion		
	DURATION	DELETION EVENT SUMMARY
	Specify the duration after which this dataset should be deleted or wiped out:	*Summarize the sequence of events and allowable processing for data deletion.* *Use additional notes to capture any other relevant information or considerations.*
	<Specify duration in days, months, or years.>	**Sequence of deletion and processing events:** <Summarize first event here.> <Summarize second event here.> <Summarize third event here.> **Additional Notes:** <Add here>
	ACCEPTABLE MEANS OF DELETION	POST-DELETION OBLIGATIONS
	List the acceptable means of deletion:	*Summarize the sequence of obligations after a deletion event.* ***Use additional notes to capture any other relevant information or considerations.***
	<Write acceptable means of deletion.> <Write acceptable means of deletion.> <Write acceptable means of deletion.>	**Sequence of post-deletion obligations:** <Summarize first obligation here.> <Summarize second obligation here.> <Summarize third obligation here.> **Additional Notes:** <Add here>
	OPERATIONAL REQUIREMENT(S)	EXCEPTIONS AND EXEMPTIONS
	List any wipeout integration operational requirements:	*Summarize any exceptions and related steps to a deletion event.* ***Use additional notes to capture any other relevant information or considerations.***
	Wipeout Integration Operational Requirements: <Write first requirement here.> <Write second requirement here.> <Write third requirement here.>	**Policy Exception bug:** [bug] **Summary:** <Write summary and notes here.> **Additional Notes:** <Add here>

Figure 20: Data Card Template - The *Access, Retention and Wipeout* section (Contd., 2/2) include a subsection on *Wipeout and Deletion* to provide guidance on the most appropriate way to delete a dataset after the retention period has expired. It also asks producers to include information about exceptions and exemptions to wipeout policies.

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
*Select **all applicable** methods used to collect data:*	*Provide a description of each collection method used.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for collection method type.)*	*Provide a description of each upstream source of data.* *Use additional notes to capture any other relevant information or considerations.*
API Artificially Generated Crowdsourced - Paid Crowdsourced - Volunteer Vendor Collection Efforts Scraped or Crawled Survey, forms or polls Taken from other existing datasets Unknown To be determined Others (Please specify)	**<Collection Type>** **Source:** <Describe here. Include links where available.> **Platform:** [Platform Name], <Describe platform here. Include links where relevant.> **Is this source considered sensitive or high-risk?** [Yes / No] **Dates of Collection:** [MMM YYYY - MMM YYYY] **Primary modality of collected data:** *Usage Note: Select one for this collection type.* Image Data Text Data Tabular Data Audio Data Video Data Time Series Graph Data Geospatial Data Unknown Multimodal (Please specify) Others (Please specify) **Update Frequency for collected data:** *Usage Note: Select one for this collection type.* Yearly Quarterly Monthly Biweekly Weekly Daily Hourly Static Others (Please specify) **Additional Links for this collection:** • [Access Policy] • [Wipeout Policy] • [Retention Policy] **Additional Notes:** <Add here>	**[Source]:** <Describe here. Include links, data examples, metrics, visualizations where relevant.> **[Source]:** <Describe here. Include links, data examples, metrics, visualizations where relevant.> **[Source]:** <Describe here. Include links, data examples, metrics, visualizations where relevant.> **Additional Notes:** <Add here>

Figure 21: Data Card Template - The *Dataset Provenance* section (1/4) describes the origin of the datasets using subsections. The *Data Collection and Sources* subsection provides an overview that describes several qualitative and procedural attributes of the collection methods and upstream sources of datapoints in the dataset.

COLLECTION CADENCE	DATA INTEGRATION	DATA PROCESSING
*Select **all applicable**:*	*List all fields collected from different sources, and specify if they were included or excluded from the dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each upstream source.)*	*Summarize how data from different sources or methods aggregated, processed, or connected.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each source OR collection method.)*
Static (Data was collected once from single or multiple sources.) Streamed (Data is continuously acquired from single or multiple sources.) Dynamic (Data is updated regularly from single or multiple sources.) Others (Please specify)	**<Source>** **Included Fields** (Data fields that were collected and are included in the dataset.) **Field Name** / **Description** <Field Name> / <Describe here. Include links, data examples, metrics, visualizations where relevant.> <Field Name> / <Describe here. Include links, data examples, metrics, visualizations where relevant.> ... / ... **Additional Notes:** <Add here> **Excluded Fields** (Data fields that were collected but are excluded from the dataset.) **Field Name** / **Description** <Field Name> / <Describe here. Include links, data examples, metrics, visualizations where relevant.> <Field Name> / <Describe here. Include links, data examples, metrics, visualizations where relevant.> ... / ... **Additional Notes:** <Add here>	**<Collection Method or Source>** **Description:** <Describe here. Include links where relevant.> **Methods employed:** <Describe here. Include links where relevant.> **Tools or libraries:** <Describe here. Include links where relevant.> **Additional Notes:** <Add here>

Figure 22: Data Card Template - Within the *Dataset Provenance* section (Contd., 2/4) captures collection cadence, integration themes, and methods of processing data by source for more complex datasets.

Collection Criteria

DATA SELECTION	DATA INCLUSION	DATA EXCLUSION
Summarize the data selection criteria. *Use additional notes to capture any other relevant information or considerations.*	*Summarize the data inclusion criteria.* *Use additional notes to capture any other relevant information or considerations.*	*Summarize the data exclusion criteria.* *Use additional notes to capture any other relevant information or considerations.*
[Collection Method or Source]: <Summarize data selection criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data selection criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data selection criteria here. Include links where available.> **Additional Notes:** <Add here>	**[Collection Method or Source]:** <Summarize data inclusion criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data inclusion criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data inclusion criteria here. Include links where available.> **Additional Notes:** <Add here>	**[Collection Method or Source]:** <Summarize data exclusion criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data exclusion criteria here. Include links where available.> **[Collection Method or Source]:** <Summarize data exclusion criteria here. Include links where available.> **Additional Notes:** <Add here>

Relationship to Source

USE & UTILITY(IES)	BENEFIT AND VALUE(S)	LIMITATION(S) AND TRADE-OFF(S)
Describe how the resulting dataset is aligned with the purposes, motivations, or intended use of the upstream source(s). *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each source type.)*	*Summarize the benefits of the resulting dataset to its consumers, compared to the upstream source(s).* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each source type.)*	*What are the limitations of the resulting dataset to its consumers, compared to the upstream source(s)?* *Break down by source type.* *(Usage Note: Duplicate and complete the following for each source type.)*
[Source Type]: <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **Additional Notes:** <Add here>	**[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **Additional Notes:** <Add here>	**[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.>

Figure 23: Data Card Template - Provenance (Contd., 3/4) the *Criteria* subsection elaborates on decisions and parameters pertaining to selection, inclusion, and exclusion of datapoints from the dataset, while the *Relationship to Source* subsection establishes the nature of upstream sources of datapoints in the dataset. Both subsections have been designed to account for multiple collection methods and upstream sources, particularly relevant where datasets have been created through aggregation or joining.

Version and Maintenance

Fill this next row if: this is not the first version of the dataset, and there is no data card available for the first version.

	FIRST VERSION	NOTE(S) AND CAVEAT(S)
	*Provide a **basic description of the first version** of this dataset.*	*Summarize the caveats or nuances of the first version of this dataset that may affect the use of the current version.* *Use additional notes to capture any other relevant information or considerations.*
	Release date: MM/YYYY **Link to dataset:** [Dataset Name + Version] **Status:** [Select one: Actively Maintained/ Limited Maintenance / Deprecated] **Size of Dataset:** 123 MB **Number of Instances:** 123456	<Summarize here. Include links where available.> **Additional Notes:** <Add here>
CADENCE	LAST AND NEXT UPDATE(S)	CHANGES ON UPDATE(S)
*Select **one**.*	*Please describe the update schedule*	*Summarize the changes that occur when the dataset is refreshed.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each source type.)*
Yearly Quarterly Monthly Biweekly Weekly Daily Hourly Static Others (Please Specify)	**Date of last update:** DD/MM/YYYY **Total Data points affected:** 12345 **Data points updated:** 12345 **Data points added:** 12345 **Data points removed:** 12345 **Date of next update:** DD/MM/YYYY	**[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **[Source Type]:** <Summarize here. Include links where available.> **Additional Notes:** <Add here>

Figure 24: Data Card Template - Provenance (4/4) In practice we find that producers find it easier to create Data Cards for new dataset or new versions of existing datasets, rather than retroactively creating data cards for previous versions. This decision has been frequently attributed to the loss of knowledge to time. The *Updates to Dataset* subsection is a part of the *Data Provenance* section, and is designed to capture nuances of the most recent updates to the dataset, and plans for future updates to the dataset.

Human and Other Sensitive Attributes

SENSITIVE HUMAN ATTRIBUTE(S)	INTENTIONALITY	RATIONALE
*Select **all attributes** that are represented (directly or indirectly) in the dataset.*	*List fields in the dataset that contain human attributes, and specify if their collection was intentional or unintentional.* *Use additional notes to capture any other relevant information or considerations.*	*Describe the motivation, rationale, considerations or approaches that caused this dataset to include the indicated human attributes.* *Summarize why or how this might affect the use of the dataset.*
Race Gender Ethnicity Socio-economic status Geography Language Sexual Orientation Religion Age Culture Disability Experience or Seniority Others (Please Specify)	**Intentionally Collected Attributes** (Human attributes were labeled or collected as a part of the dataset creation process.) **Field Name** / **Description** <Field Name> / <Human Attributed Collected.> <Field Name> / <Human Attributed Collected.> ... / ... **Additional Notes:** <Add here> **Unintentionally Collected Attributes** (Human attributes were not explicitly collected as a part of the dataset creation process but can be inferred using additional methods.) **Field Name** / **Description** <Field Name> / <Human Attributed Collected.> <Field Name> / <Human Attributed Collected.> ... / ... **Additional Notes:** <Add here>	<Summarize here. Include links, tables, and media as relevant>
	SOURCE(S)	METHODOLOGY DETAIL(S)
	List the sources of the human attributes. *Use additional notes to capture any other relevant information or considerations.*	*Describe the methods used to collect human attributes in the dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each human attribute.)*
	[Human Attribute]: Sources **[Human Attribute]:** Sources **[Human Attribute]:** Sources **Additional Notes:** <Add here>	**[Human Attribute]** **Method:** <Describe the collection method here. Include links where necessary.> **Collection task:** <Describe the task here. Include links where necessary.> **Platforms, tools, or libraries:** [Platform, tool or library]: <Write description here.> [Platform, tool or library]: <Write description here.> [Platform, tool or library]: <Write description here.> **Additional Notes:** <Add here>

Figure 25: Data Card Template - The *Human and Other Sensitive Attributes* (1/2) is of particular importance to human-centered machine learning applications and fairness analyses. Here, we encourage producers to report the rationales behind decisions to capture or include human attributes as well as various disaggregated statistics and correlations, risks and trade-offs (see Figure 26).

DISTRIBUTION(S)

Provide basic descriptive statistics for each human attribute, noting key takeaways in the caption.
Use additional notes to capture any other relevant information or considerations.
(Usage Note: Duplicate and complete the following for each human attribute.)

Human Attribute

[Human attribute]	Label or Class	Label or Class	Label or Class	Label or Class
Count	123456	123456	123456	123456
[Statistic]	123456	123456	123456	123456
[Statistic]	123456	123456	123456	123456
[Statistic]	123456	123456	123456	123456

Above: <Provide a caption for the above table or visualization.>
Additional Notes: <Add here>

	KNOWN CORRELATIONS	RISK(S) AND MITIGATION(S)
	Describe any known correlations with the indicated sensitive attributes in this dataset. *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate for each known correlation.)*	*Summarize systemic or residual risks, performance expectations, trade-offs and caveats because of human attributes in this dataset.* *Use additional notes to capture any other relevant information or considerations.* *Usage Note: Duplicate and complete the following for each human attribute.)*
	[`field_name`,`field_name`] **Description:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Impact on dataset use:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Additional Notes:** <Add here>	**[Human Attribute]** <Summarize here. Include links and metrics where applicable.> **<Risk type>:** [Description + Mitigations] **<Risk type>:** [Description + Mitigations] **<Risk type>:** [Description + Mitigations] **Trade-offs, caveats, & other considerations:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Additional Notes:** <Add here>

Figure 26: Data Card Template - *Human and Other Sensitive Attributes* (2/2)

Extended Use

Use with Other Data

SAFETY LEVEL	KNOWN SAFE DATASET(S) OR DATA TYPE(S)	BEST PRACTICES
*Select **one**:*	*List the known datasets or data types and corresponding transformations that **are safe to join or aggregate** this dataset with.*	*Summarize best practices for using this dataset with other datasets or data types.* *Use additional notes to capture any other relevant information or considerations.*
Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others (Please Specify)	**Dataset or Data Type:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Dataset or Data Type:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Dataset or Data Type:** <Summarize here. Include visualizations, metrics, or links where necessary.>	<Summarize here. Include visualizations, metrics, demonstrative examples, or links where necessary.> **Additional Notes:** <Add here>
	KNOWN UNSAFE DATASET(S) OR DATA TYPE(S)	LIMITATION(S) AND RECOMMENDATION(S)
Fill out this row if you selected "Conditionally safe to use with other datasets" or "Should not be used with other datasets":	*List the known datasets or data types and corresponding transformations that are **unsafe to join or aggregate** with this dataset.*	*Summarize limitations of the dataset that introduce foreseeable risks when the dataset is conjoined with other datasets.* *Use additional notes to capture any other relevant information or considerations.*
	Dataset or Data Type: <Summarize here. Include visualizations, metrics, or links where necessary.> **Dataset or Data Type:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Dataset or Data Type:** <Summarize here. Include visualizations, metrics, or links where necessary.>	<Summarize here. Include links and metrics where applicable.> **<Limitation type>:** [Dataset or data type, description and recommendation.] **<Limitation type>:** [Dataset or data type, description and recommendation.] **<Limitation type>:** [Dataset or data type, description and recommendation.] **Additional Notes:** <Add here>

Figure 27: Data Card Template - The *Extended Use* section (1/3) is designed to capture guidance necessary for the responsible use of the dataset, including what is known about the safety of using the dataset with other datasets and data types – as well as any limitations and recommendations.

Forking & Sampling

SAFETY LEVEL	ACCEPTABLE SAMPLING METHOD(S)	BEST PRACTICE(S)
*Select **one**:*	*Select **all applicable** acceptable methods to sample this dataset:*	*Summarize the best practices for forking or sampling this dataset.* *Use additional notes to capture any other relevant information or considerations.*
Safe to form and/or sample Conditionally safe to fork and/or sample Should not be forked and/or sampled Unknown Others (Please specify)	Cluster Sampling Haphazard Sampling Multi-stage Sampling Random Sampling Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others (Please specify)	<Summarize here. Include links, figures and demonstrative examples where available.> **Additional Notes:** <Add here>
	RISK(S) AND MITIGATION(S)	LIMITATION(S) AND RECOMMENDATION(S)
Fill out this row if you selected "Conditionally safe to fork and/or sample" or "Should not be forked and/or sampled".	*Summarize known or residual risks associated with forking and sampling methods when applied to the dataset.* *Use additional notes to capture any other relevant information or considerations.*	*Summarize the limitations that the dataset introduces when forking or sampling the dataset and corresponding recommendations.* *Use additional notes to capture any other relevant information or considerations.*
	<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **Additional Notes:** <Add here>	<Summarize here. Include links and metrics where applicable.> **<Limitation Type>:** [Description + Recommendation] **<Limitation Type>:** [Description + Recommendation] **<Limitation Type>:** [Description + Recommendation] **Additional Notes:** <Add here>

Figure 28: Data Card Template - The *Extended Use* section (Contd., 2/3) captures safe ways to join or fork the dataset, to support upstream decision making.

Use in ML or AI Systems

DATASET USE(S)	NOTABLE FEATURE(S)	USAGE GUIDELINE(S)
*Select **all applicable**:*	*Describe any notable feature distributions or relationships between individual instances made explicit.* *Include links to servers where readers can explore the data on their own.*	*Summarize usage guidelines or policies that consumers should be aware of.* *Use additional notes to capture any other relevant information or considerations.*
Training Testing Validation Development or Production Use Fine Tuning Others (Please Specify)	**Exploration Demo:** [Link to server or demo.] **<Notable Field Name>:** <Describe here. Include links, data examples, metrics, visualizations where relevant.> **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**Usage Guidelines:** <Summarize here. Include links where necessary.> **Approval Steps:** <Summarize here. Include links where necessary.> **Reviewer:** <Provide the name of a reviewer for publications referencing this dataset.> **Additional Notes:** <Add here>
	DISTRIBUTION(S)	KNOWN CORRELATION(S)
	Describe the recommended splits and corresponding criteria. *Use additional notes to capture any other relevant information or considerations.*	*Summarize any known correlations with the indicated features in this dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate for each known correlation.)*
	Train 62,563 Test 62,563 Validation 62,563 Dev 62,563 **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	`field_name`,`field_name` **Description:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Impact on dataset use:** <Summarize here. Include visualizations, metrics, or links where necessary.> **Risks from correlation:** <Summarize here. Include recommended mitigative steps if available.> **Additional Notes:** <Add here>
	SPLIT STATISTICS	
	Provide the sizes of each split. As appropriate, provide any descriptive statistics for features.	

Statistic	Train	Test	Valid	Dev
Count	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456

Above: Caption for table above.

Figure 29: Data Card Template - In the *Use in Machine Learning or AI systems* subsection of the *Extended Use* section (Contd., 3/3), producers are ask to report descriptive statistics for different training and testing splits. For wide scale adoption, we encourage the automation of these types of fields for accuracy and rigor.

Transformations

Fill this section if any transformations were applied in the creation of your dataset.

Synopsis

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED
*Select **all applicable** transformations that were applied to the dataset.*	*Provide the fields in the dataset that were transformed.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each transformation type applied. Include the data types to which fields were transformed.)*	*Provide a description of the methods used to transform or process the dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each transformation type applied.)*
Anomaly Detection Cleaning Mismatched Values Cleaning Missing Values Converting Data Types Data Aggregation Dimensionality Reduction Joining Input Sources Redaction or Anonymization Others (Please specify)	**<Transformation Type>** **Field Name** | **Source & Target** <Field Name> | <Source Field: Target Field> <Field Name> | <Source Field: Target Field> ... | ... **Additional Notes:** <Add here>	**<Transformation Type>** **Method:** <Describe the transformation method here. Include links where necessary.> **Platforms, tools, or libraries:** [Platform, tool or library]: <write description here> [Platform, tool or library]: <write description here> [Platform, tool or library]: <write description here> **Transformation Results:** <Provide results, outcomes, and actions taken because of the transformations. Include visualizations where available.> **Additional Notes:** <Add here>

Figure 30: Data Card Template - The *Transformations* section is used to describe the processes by which raw data is transformed into usable formats. Here, we first ask producers to provide a aggregate of the transformations, following which a more detailed breakdowns are collected. .

Breakdown of Transformations

Fill out relevant rows.

CLEANING MISSING VALUE(S)	METHOD(S) USED	COMPARATIVE SUMMARY
Which fields in the data were missing values? How many?	*How were missing values cleaned? What other choices were considered?*	*Why were missing values cleaned using this method (over others)? Provide comparative charts showing before and after missing values were cleaned.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>

Figure 31: Data Card Template - The *Transformations* section (Contd.). Producers are asked to include information about specific transformation applied to datasets that could potentially introduce residual or system-level risks and require oversight.

CLEANING MISMATCHED VALUE(S)	METHOD(S) USED	COMPARATIVE SUMMARY
Which fields in the data were corrected for mismatched values?	*How were incorrect or mismatched values cleaned? What other choices were considered?*	*Why were incorrect or mismatched values cleaned using this method (over others)? Provide a comparative analysis demonstrating before and after values were cleaned.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where available.>	<Summarize here. Include links where available.> Field Name | Diff <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>
ANOMALIES	**METHOD(S) USED**	**COMPARATIVE SUMMARY**
How many anomalies or outliers were detected? *If at all, how were detected anomalies or outliers handled?* *Why or why not?*	*What methods were used to detect anomalies or outliers?*	*Provide a comparative analysis demonstrating before and after anomaly handling measures.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where available.>	<Summarize here. Include links where available.> Field Name | Diff <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>

Figure 32: Data Card Template - The *Transformations* section (Contd.). Producers are asked to include information about specific transformation applied to datasets that could potentially introduce residual or system-level risks and require oversight.

AGGREGATION	METHOD(S) USED	COMPARATIVE SUMMARY
Which fields in the dataset were aggregated?	*What methods were used to aggregate the data? Include the aggregating operator. What other choices were considered?*	*Why was the data aggregated using this method (over others)? Provide comparative charts that demonstrate the choices of aggregators.*
<Summarize here. Include links where available.> **Field Name**: Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available.> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>
DIMENSIONALITY REDUCTION	**METHOD(S) USED**	**COMPARATIVE SUMMARY**
How many original features were collected and how many dimensions were reduced?	*What methods were used to reduce the dimensionality of the data? What other choices were considered?*	*Why were features reduced using this method (over others)? Provide comparative charts showing before and after dimensionality reduction processes.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>

Figure 33: Data Card Template - The *Transformations* section (Contd.). Producers are asked to include information about specific transformation applied to datasets that could potentially introduce residual or system-level risks and require oversight.

JOINING INPUT SOURCES	METHOD(S) USED	COMPARATIVE SUMMARY
What were the distinct input sources that were joined?	*What are the shared columns of fields used to join these sources?*	*Why were features joined using this method over others?* *Provide comparative charts showing before and after dimensionality reduction processes.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>
REDACTION OR ANONYMIZATION	**METHOD(S) USED**	**COMPARATIVE SUMMARY**
Which features were redacted or anonymized?	*What methods were used to redact or anonymize data?*	*Why was data redacted or anonymized using this method over others? Provide comparative charts showing before and after redaction or anonymization process.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>

Figure 34: Data Card Template - The *Transformations* Section (Contd.). Producers are asked to include information about specific transformation applied to datasets that could potentially introduce residual or system-level risks and require oversight.

OTHERS (PLEASE SPECIFY)	METHOD(S) USED	COMPARATIVE SUMMARY
What was done? Which features or fields were affected?	*What methods were used?*	*Why was this method used over others? Provide comparative charts showing before and after this transformation.*
<Summarize here. Include links where available.> **Field Name:** Count or Description **Field Name:** Count or Description **Field Name:** Count or Description	<Summarize here. Include links where necessary.> **Platforms, tools, or libraries:** **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.> **[Platform, tool or library]:** <Write description here.>	<Summarize here. Include links, tables, visualizations where available> **Field Name** | **Diff** <Field Name> | <Before: After> <Field Name> | <Before: After> ... | ... **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>
RESIDUAL & OTHER RISK(S)	**HUMAN OVERSIGHT MEASURE(S)**	**ADDITIONAL CONSIDERATIONS**
What risks were introduced because of this transformation? Which risks were mitigated?	*What human oversight measures, including additional testing, investigations and approvals were taken due to this transformation?*	*What additional considerations were made?*
<Summarize here. Include links and metrics where applicable.> **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations] **<Risk Type>:** [Description + Mitigations]	<Summarize here. Include links where available.>	<Summarize here. Include links where available.>

Figure 35: Data Card Template - The *Transformations* Section (Contd.). Producers are asked to include information about specific transformation applied to datasets that could potentially introduce residual or system-level risks and require oversight.

Annotations & Labeling

Fill this section if any human or algorithmic annotation tasks were performed in the creation of your dataset.

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTIC(S)	ANNOTATION DESCRIPTION(S)
*Select **all applicable** annotation workforce types or methods used to annotate the dataset:*	*Describe relevant characteristics of annotations as indicated. For quality metrics, consider including accuracy, consensus accuracy, IRR, XRR at the appropriate granularity (e.g. across dataset, by annotator, by annotation, etc.).* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*	*Provide descriptions of the annotations applied to the dataset. Include links and indicate platforms, tools or libraries used wherever possible.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*
Annotation Target in Data Machine-Generated Annotations Human Annotations (Expert) Human Annotations (Non-Expert) Human Annotations (Employees) Human Annotations (Contractors) Human Annotations (Crowdsourcing) Human Annotations (Outsourced / Managed) Teams Unlabeled Others (Please specify)	**(Annotation Type)** Number of unique annotations 123456789 Total number of annotations 123456789 Avg. Annotations per example 123456789 Number of annotators per example 123456789 [Quality metric per granularity] 123456789 [Quality metric per granularity] 123456789 [Quality metric per granularity] 123456789 **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Annotation Type)** **Description:** <Description of annotations (labels, ratings) produced. Include how this was created or authored.> **Link:** <Relevant URL link.> **Platforms, tools, or libraries:** [Platform, tool or library]: <Write description here.> [Platform, tool or library]: <Write description here.> [Platform, tool or library]: <Write description here.> **Additional Notes:** <Add here>

	ANNOTATION DISTRIBUTION(S)	ANNOTATION TASK(S)
	Provide a distribution of annotations for each annotation or class of annotations using the format below. *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*	*Summarize each task type associated with annotations in the dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each task type.)*
	(Annotation Type) Annotations (or Class) 12345 (20%) Annotations (or Class) 12345 (20%) Annotations (or Class) 12345 (20%) Annotations (or Class) 12345 (20%) Annotations (or Class) 12345 (20%) **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Task Type)** **Task description:** <Summarize here. Include links if available.> **Task instructions:** <Summarize here. Include links if available.> **Methods Used:** <Summarize here. Include links if available.> **Inter-rater adjudication policy:** <Summarize here. Include links if available.> **Golden Questions:** <Summarize here. Include links if available.> **Additional Notes:** <Add here>

Figure 36: Data Cards Template - The *Annotations and Labeling* section captures a variety of annotation types, including quantitative characteristics, qualitative descriptions, resulting distributions, and task or instruction summaries that affect outcomes.

Human Annotators

Fill this section if human annotators were used.

	ANNOTATOR DESCRIPTION(S)	ANNOTATOR TASK(S)
	Provide a brief description for each annotator pool performing the human annotation task. *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*	*Provide a brief description for each annotator pool performing the human annotation task.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*
	(Annotation Type) **Task type.** <Summarize here. Include links if available> **Number of unique annotators:** <Summarize here. Include links if available.> **Expertise of annotators:** <Summarize here. Include links if available.> **Description of annotators:** <Summarize here. Include links if available.> **Language distribution of annotators:** <Summarize here. Include links if available.> **Geographic distribution of annotators:** <Summarize here. Include links if available.> **Summary of annotation instructions:** <Summarize here. Include links if available.> **Summary of gold questions:** <Summarize here. Include links if available.> **Annotation platforms:** <Summarize here. Include links if available.> **Additional Notes:** <Add here>	**(Task Type)** **Task description:** <Summarize here. Include links if available.> **Task instructions:** <Summarize here. Include links if available.> **Methods Used:** <Summarize here. Include links if available.> **Inter-rater adjudication policy:** <Summarize here. Include links if available.> **Golden Questions:** <Summarize here. Include links if available.> **Additional Notes:** <Add here>

LANGUAGE(S)	LOCATION(S)	GENDER(S)
Provide annotator distributions for each annotation type. *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*	*Provide annotator distributions for each annotation type.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*	*Provide annotator distributions for each annotation type.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each annotation type.)*
(Annotation Type) <Language> [Percentage %] <Language> [Percentage %] <Language> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Annotation Type)** <Location> [Percentage %] <Location> [Percentage %] <Location> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Annotation Type)** <Gender> [Percentage %] <Gender> [Percentage %] <Gender> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>

Figure 37: Data Cards Template - Important to human computation datasets, this *Human Annotators* sub-section captures attributes where human annotators were employed.

Validation Types

Fill this section if the data in the dataset was validated during or after the creation of your dataset.

METHOD(S)	BREAKDOWN(S)	DESCRIPTION(S)
*Select **all applicable**:*	*Provide a description of the fields and data points that were validated.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each validator type.)*	*Provide a description of the methods used to validate the dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each validator type.)*
Data Type Validation Range and Constraint Validation Code/cross-reference Validation Structured Validation Consistency Validation Not Validated Others (Please Specify)	**(Validation Type)** **Number of Data Points Validated:** 12345 **Fields Validated:** Field 123456 [Count if available] Field 123456 [Count if available] Field 123456 [Count if available] **Above:** <Provide a caption for the above table or visualization.>	**(Validation Type)** **Method:** <Describe the validation method here. Include links where necessary.> **Platforms, tools, or libraries:** [Platform, tool or library]: <Write description here.> [Platform, tool or library]: <Write description here.> **Validation Results:** <Provide results, outcomes, and actions taken because of the validation. Include visualizations where available.> **Additional Notes:** <Add here>

Description of Human Validators

Fill this section if the dataset was validated using human validators

	CHARACTERISTIC(S)	DESCRIPTION(S)
	Provide characteristics of the validator pool(s). Use additional notes to capture any other relevant information or considerations.	*Provide a brief description of the validator pool(s). Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each validator type.)*
	(Validation Type) Unique validators 12345 # of examples per validator 123456 Average cost/task/ validator $$$ Training provided Y/N Expertise required Y/N	**(Validation Type)** **Validator description:** <Summarize here. Include links if available.> **Training provided:** <Summarize here. Include links if available.> **Validator Selection Criteria:** <Summarize here. Include links if available.> **Training provided:** <Summarize here. Include links if available.> **Additional Notes:** <Add here>

LANGUAGE(S)	LOCATION(S)	GENDER(S)
Provide validator distributions for each validation type. Use additional notes to capture any other relevant information or considerations. (Usage Note: Duplicate and complete the following for each annotation type.)		
(Validation Type) <Language> [Percentage %] <Language> [Percentage %] <Language> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Validation Type)** <Location> [Percentage %] <Location> [Percentage %] <Location> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Validation Type)** <Gender> [Percentage %] <Gender> [Percentage %] <Gender> [Percentage %] **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>

Figure 38: Data Card Template - Producers are expected to complete the *Validation Methods* section if a part or the entirety of the dataset was validated. This section also details attributes of human validators.

Sampling Methods

Fill out the following block if your dataset employs any sampling methods.

METHOD(S) USED	CHARACTERISTIC(S)	SAMPLING CRITERIA
*Select **all applicable** methods used in the creation of this dataset:*	*Provide characteristics of each sampling method used.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each sampling method used.)*	*Describe the criteria used to sample data from upstream sources.* *Use additional notes to capture any other relevant information or considerations.*
Cluster Sampling Haphazard Sampling Multi-stage Sampling Random Sampling Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others (Please specify)	**(Sampling Type)** Upstream Source [Write here] Total data sampled 123m Sample size 123 Threshold applied property 123k units at Sampling Rate 123 Sample Mean 123 Sample Std. Dev. 123 Sampling Distribution 123 Sampling Variation 123 Sample Statistic 123 **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	• **Sampling method:** <Summarize here. Include links where applicable.> • **Sampling method:** <Summarize here. Include links where applicable.> • **Sampling method:** <Summarize here. Include links where applicable.>

Figure 39: Data Card Template - The *Sampling Methods* section captures both quantitative metrics pertinent and qualitative summaries pertinent to sampling that may have been used in the creation of the dataset. Since not all datasets may be sampled, this section is considered conditional.

Known Applications & Benchmarks

Fill out the following section if your dataset was primarily created for use in AI or ML system(s)

ML APPLICATION(S)	EVALUATION RESULT(S)	EVALUATION PROCESS(ES)
Provide a list of key ML tasks that the dataset has been used for. Usage Note: Use comma-separated keywords.	*Provide the evaluation results from models that this dataset has been used in.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each model.)*	*Provide a description of the evaluation process for the model's overall performance or the determination of how the dataset contributes to the model's performance.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each model and method used.)*
For example: Classification, Regression, Object Detection	**(Model Name)** **Model Card:** [Link to Full Model Card] Evaluation Results Accuracy 123 (params) Precision 123 (params) Recall 123 (params) Performance metric 123 (params) **Above:** <Provide a caption for the above table or visualization.> **Additional Notes:** <Add here>	**(Model Name)** **[Method used]:** <Summarize here. Include links where available.> • **Process:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.> • **Factors:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.> • **Considerations:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.> • **Results:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.> **Additional Notes:** <Add here>
	DESCRIPTION(S) AND STATISTIC(S)	**EXPECTED PERFORMANCE AND KNOWN CAVEATS**
Duplicate this row as necessary for each model type	*Provide a description of the model(s) and task(s) that this dataset has been used in.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each model.)*	*Provide a description of the expected performance and known caveats of the models for this dataset.* *Use additional notes to capture any other relevant information or considerations.* *(Usage Note: Duplicate and complete the following for each model.)*
	(Model Name) **Model Card:** [Link to Full Model Card] **Model Description:** <Summarize here. Include links where applicable.> Model Size 123 (params) Model Weights 123 (params) Model Layers 123 (params) Latency 123 (params) **Additional Notes:** <Add here>	**(Model Name)** **Expected Performance:** <Summarize here. Include links where available.> **Known Caveats:** <Summarize here. Include links, diagrams, visualizations, and tables as relevant.> **Additional Notes:** <Add here>

Figure 40: Data Card Template - The *Known Applications & Benchmarks* section is designed to capture documentation pertaining to the use of the dataset to train or test models, for example, those that are publicly available. Producers are asked to provide a brief description of the model(s), the evaluation processes, expected performance and any known caveats that agents should be aware of.

Terms of Art

Concepts and Definitions referenced in this Data Card

Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Data Card. Use standard definitions where possible. Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your Data Card or dataset, include your interpretation as well.

[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>
[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>
[TERM OF ART]	[TERM OF ART]	[TERM OF ART]
Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>	Definition: <Write here> Source: <Write here and share link> Interpretation: <Write here>

Reflections on Data

Use this space to include any additional information about the dataset that has not been captured by the Data Card. For example, does the dataset contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please contact the appropriate parties to mitigate any risks.

[Title]	<Write notes here>
[Title]	<Write notes here>
[Title]	<Write notes here>

Figure 41: Data Card Template - The *Terms of Art* section introduces technical terms, domain-specific concepts, and acronyms that are used across the Data Card. Here, we ask producers to include any modifications or adaptations to terms to assist with interpretation in the context of the dataset. The *Reflections on Data* section is intended to be a free-form space for producers to add information not captured by the template.