# Datasheet for dataset “add dataset name here”

Questions from the [Datasheets for Datasets](https://arxiv.org/abs/1803.09010) paper, v7.

Jump to section:

* [Motivation](#motivation)
* [Composition](#composition)
* [Collection process](#collection-process)
* [Preprocessing/cleaning/labeling](#preprocessingcleaninglabeling)
* [Uses](#uses)
* [Distribution](#distribution)
* [Maintenance](#maintenance)

## Motivation

*The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.*

### For what purpose was the dataset created?

*Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

### Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

### Who funded the creation of the dataset?

*If there is an associated grant, please provide the name of the grantor and the grant name and number.*

## Composition

*Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU’s General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.*

### What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

*Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

### How many instances are there in total (of each type, if appropriate)?

### Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

*If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

### What data does each instance consist of?

*“Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

### Is there a label or target associated with each instance?

*If so, please provide a description.*

### Is any information missing from individual instances?

*If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

### Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?

*If so, please describe how these relationships are made explicit.*

### Are there recommended data splits (e.g., training, development/validation, testing)?

*If so, please provide a description of these splits, explaining the rationale behind them.*

### Are there any errors, sources of noise, or redundancies in the dataset?

*If so, please provide a description.*

### Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ non-public communications)?

*If so, please provide a description.*

### Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

*If so, please describe why.*

### Does the dataset relate to people?

*If not, you may skip the remaining questions in this section.*

### Does the dataset identify any subpopulations (e.g., by age, gender)?

*If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

### Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

*If so, please describe how.*

### Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

*If so, please provide a description.*

### Any other comments?

## Collection process

*[T]he answers to questions here may provide information that allow others to reconstruct the dataset without access to it.*

### How was the data associated with each instance acquired?

*Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

### What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

*How were these mechanisms or procedures validated?*

### Who was involved in the data collection process (e.g., students, contractors) and how were they compensated (e.g., how much these people were paid)?

### Over what timeframe was the data collected?

### Were any ethical review processes conducted (e.g., by an institutional review board)?

*If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

### Does the dataset relate to people?

*If not, you may skip the remainder of the questions in this section.*

### Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

### Were the individuals in question notified about the data collection?

*If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

### Did the individuals in question consent to the collection and use of their data?

*If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

### If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

*If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

### Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

*If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

## Cleaning

*The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.*

### Was any cleaning of the data done (e.g, removal of instances, processing of missing values)?

*If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

### Was the “raw” data saved in addition to the cleaned data (e.g., to support unanticipated future uses)?

*If so, please provide a link or other access point to the “raw” data.*

## Uses

*These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.*

### Has the dataset been used for any tasks already?

*If so, please provide a description.*

### Is there a repository that links to any or all papers or systems that use the dataset?

*If so, please provide a link or other access point.*

### What (other) tasks could the dataset be used for?

### Is there anything about the composition of the dataset or the way it was collected and cleaned that might impact future uses?

*For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

### Are there tasks for which the dataset should not be used?

*If so, please provide a description.*

## Distribution

### Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

*If so, please provide a description.*

### How will the dataset will be distributed (e.g., zip file, website, GitHub)?

*Does the dataset have a digital object identifier (DOI)?*

### Will the dataset be distributed under a copyright or other intellectual property (IP) license ?

## Maintenance

*These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.*

### Who is supporting/hosting/maintaining the dataset?

### How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

### Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

*If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

### If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

*If so, please describe these limits and explain how they will be enforced.*

### Will older versions of the dataset continue to be supported/hosted/maintained?

*If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

### If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

*If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*