Title of your data set  
*Data documentation template adapted by Hannes Datta.   
Originally based on Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.*

# 1. Motivation[[1]](#footnote-1)

***1.1*** *What primary research question, business problem, or knowledge gap motivated the creation of this dataset? How does the dataset offer insights into new phenomena, contribute to developing new models, or streamline gathering essential information? Why is this dataset valuable to the broader research community or industry stakeholders? Please provide a description about your research context.*

***1.2*** *the various websites and APIs you assessed relevant to your data context, why did you choose your specific data source? Discuss the research fit, efficiency of resource use, and any other factors that made it emerge as the best choice. What extraction methods did you consider, and why did you choose the method you used? Were alternatives to web scraping evaluated? How did you ensure the scope of your data context was appropriate to maintain validity and identify any other valuable information that might be relevant? Please motivate why you selected this particular data source.*

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

***1.4*** *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.*

# 2. Composition

***2.1*** *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.*

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

***2.3*** *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 (e.g., to mitigate algorithmic interference). 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).*

***2.4*** *At what frequency was the data collected?*

***2.5*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features/operationalized variables? In either case, please provide a description.*

***2.6*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? For example, if you submit multiple data files, can they be merged meaningfully? If so, please describe how these relationships are made explicit.*

***2.7*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

***2.8*** *In collecting this data, how did you ensure research validity is balanced with the technical feasibility of collecting the data as well as any associated legal or ethical risks?*

***2.9*** *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.*

***2.10*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

***2.11*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

***2.12*** *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.*

***2.13*** *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.*

***2.14*** *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.*

# 3. Data inspection

***3.1*** *Please provide meaningful summary statistics and plots. For example, the number of units per entity, means/SD for continuous variables, or frequency distributions for categorical variables. This part of the documentation is intended to illustrate (the richness of) the collected data.*

***3.2*** *Is any information missing from individual instances? If so, please describe 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.*

# 4. Collection Process

***4.1*** *Can you describe your technical extraction plan in such a way that another researcher or team could replicate your data collection process?*

***4.2*** *Why did you choose a particular data extraction technology over others (for example, why did you opt for Selenium over BeautifulSoup for website scraping, or a specific package instead of self-coded requests for APIs)?*

***4.3*** *If you encountered technical challenges during the scaling of your data collection, how did you resolve them? Please provide a clear explanation of the debugging process.*

***4.4*** *What technical obstacles did you face during the data extraction process, and how did you overcome them?*

***4.5*** *What measures or monitoring systems were in place to ensure and validate the quality of the extracted data? Can you describe how these monitoring systems functioned?*

***4.6*** *Can you specify the infrastructure you used for the deployment and execution of your data collection?*

***4.7*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Please provide meaningful summaries, possibly referencing timestamps from log files.*

***4.8*** *Where was the data stored during the collection process? Detail any specific storage mechanisms or locations you utilized.*

***4.9*** *Who was involved in the data collection process (e.g., students, crowd workers, contractors), and how were they compensated (e.g., how much were crowd workers paid)?*

***4.10*** *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.*

***4.11*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

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

***4.13*** *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 other- wise reproduce, the exact language of the notification itself.*

***4.14*** *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.*

***4.15*** *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).*

***4.16*** *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.*

# 5. Preprocessing, cleaning, labeling

***5.1*** *Did you perform any pre-processing during the data extraction process? If yes, please provide specific examples and explain the reasoning behind each on-the-fly pre-processing step.*

***5.2*** *After collecting the data, what additional pre-processing steps were undertaken?*

***5.3*** *Were any measures implemented to ensure privacy, such as anonymizing user data? Please describe the methods used.*

***5.4*** *How did you address and clean out any implausible or erroneous observations in the dataset?*

***5.5*** *Did you modify the data structure for long-term storage, like rearranging the dataset or renaming columns for clarity? If so, provide details on these changes and their rationale.*

***5.6*** *What potential threats or biases could arise from your pre-processing steps? Please elaborate on any risks associated with the modifications made to the data and how they might impact the dataset's integrity or utility.*

***5.7*** *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

***5.8*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

# 6. Uses

***6.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

**6.2** 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.

***6.3*** *What (other) tasks / research projects could the dataset be used for? Provide a set of potential research questions or ideas for research projects.*

***6.4*** *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled 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?*

***6.5*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

**You can answer the next questions in case you decide to make the dataset public yourself.  
If you prefer to keep the data only for grading purposes within the course, you can delete the following sections.**

# 7. Distribution

***7.1*** *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.*

***7.2*** *How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? For example, you can upload your project to OSF.org, or make it available via Zenodo.org.*

***7.3*** *When will the dataset be distributed?*

***7.4*** *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

***7.5*** *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

***7.6*** *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

# 8. Maintenance

***8.1*** *Who will be supporting/hosting/maintaining the dataset?*

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

***8.3*** *Is there an erratum? If so, please provide a link or other access point.*

***8.4*** *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)?*

***8.5*** *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.*

***8.6*** *Will older versions of the dataset continue to be sup- ported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

***8.7*** *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.*

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)