Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Alba Irollo, Jörg Lehmann, Clemens Neudecker, Giulia Osti, Daniel van Strien

Template: Datasheet for Digital Cultural Heritage Datasets

Version 1 – September 2023

The superscripts added to section headings refer to items in the bibliography at the very end, where every section is thoroughly discussed, explained and further questions can be found. The main structure follows Gebru's (2021) and Pushkarna's templates (2022).

Motivation[b,f]

[Clearly articulate the reasons for creating the dataset and promote transparency about funding interests. Also provide a brief descriptive overview of the dataset ('at a glance'). For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organisation)? Who funded the creation of the dataset?]

Dataset Description[c]

Homepage [Add homepage URL here if available.]

Repository [E.g., if the dataset is hosted on GitHub or has a GitHub homepage, add URL here.]

Paper [If the dataset was introduced by a paper or there was a paper written describing the dataset, add URL here.]

Point of Contact [If known, name and email of at least one person the reader can contact for questions about the dataset.]

Dataset Summary[c]

[Briefly summarise the dataset, its intended use and the supported tasks. Give an overview of how and why the dataset was created. The summary should explicitly mention the languages present in the dataset (possibly in broad terms, e.g. translations between several pairs of European languages), and describe the domain, topic, genre covered, keywords, and other relevant metadata.]

Supported Tasks and Shared Tasks[c]

[For each of the tasks tagged for this dataset, give a brief description of the tag, metrics, and suggested models. Also provide reference to any shared task the dataset is part of.]

Languages[c]

[Provide a brief overview of the languages represented in the dataset. Describe relevant details about specifics of the language using e.g. BCP-47 codes.]

Composition[b,e]

[Provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks and elicit information about compliance with GDPR and copyright. What medium is the dataset (image, video, text, web archive, etc.)?[e] 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)?]

Dataset Structure[c]

[Provide one or more typical examples of data instances, a systematic description of the dataset structure or schema, and the data splits which can be used for machine learning tasks.]

Data Instances[c]

[Provide an example and brief description of a typical instance in the dataset. What metadata are available for the dataset items?]

Data Fields[c,f]

[List and describe the fields present in the dataset. Mention their data type, and whether they are used as input or output in any of the tasks the dataset currently supports. If the data has span indices, describe their attributes, such as whether they are at the character or word level, whether they are contiguous or not, etc. If the dataset contains example IDs, state whether they have an inherent meaning, such as a mapping to other datasets or pointing to relationships between data points.]

Data Splits[c]

[Describe and name the splits in the dataset if there are more than one. Describe any criteria for splitting the data, if used. If there are differences between the splits (e.g. if the training annotations are machine-generated and the development and test ones are created by humans, or if different numbers of annotators contributed to each example), describe them here. Provide the sizes of each split. If appropriate, provide any descriptive statistics for the features, such as average length. 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.]

Descriptive Statistics

[The information presented in this section should be chosen on the basis of whether it tells something valuable. Be reminded that it depends on the intended purpose of a dataset whether the numbers and metrics provided can be of use. How large is the dataset, both in cardinality and in disk storage?[e] What are its key descriptive statistics per field?]

Collection Process[b,e,f]

[Elicit information that may help researchers and practitioners to re-use this dataset or create alternative datasets with similar characteristics.]

Curation Rationale[c,e,f]

[What need motivated the creation of this dataset? What were some of the reasons underlying the major choices involved in putting it together?]

Source Data[c,f]

[This section describes the source data (e.g. news text and headlines, translated sentences, ...).]

Initial Data Collection and Normalisation[c]

[Describe the data collection process. Describe any criteria for data selection or filtering. List any keywords or search terms used. If possible, include runtime information for the collection process. If data was collected from other pre-existing datasets, link to the source here. If the data was modified or normalised after being collected (e.g. if the data is word-tokenized), describe the process and the tools used.]

Who are the source data producers?[c]

[State whether the data were produced by humans or machine generated. Describe the people or systems who originally created the data. If available, include self-reported demographic or identity information for the source data creators, but avoid inferring this information. Instead, state that this information is unknown. See Larson, 2017[d] for using identity categories as variables, particularly gender.

Digitisation Pipeline[e]

[If applicable, describe in what way digitisation presents another layer of selection of the whole of a collection available in a cultural heritage institution; state if this is not the case. Describe the motivation for digitisation (e.g. conservation, research project, or else). If applicable, provide selection criteria and metrics that demonstrate how this additional layer of selection has influenced the transformation of the original collection or dataset into the current dataset.]

Data Provenance[e,f]

[Describe data provenance including rights, licences and other obligations.]

Use of Linked Open Data, Controlled Vocabulary, Multilingual Ontologies/Taxonomies

[Describe linked open data schemes, controlled vocabularies, ontologies and taxonomies used during the establishment of the dataset.]

Version Information[f]

[Provide information on the version of the dataset, especially if a dataset grows over time. Alternatively, provide a release date and a date when the dataset was last updated as well as information on whether there was a previous version of the dataset published before. Describe how the content has been influenced by digitisation (e.g. what part of a collection has been digitised and what not).[e] Which policy drives growing or changing datasets (e.g. systematisation, subsetting for specific purposes)?]

Preprocessing and cleaning[b,f]

[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.]

Annotations[c,f]

[If the dataset contains annotations which are not part of the initial data collection, describe them in the following paragraphs.]

Annotation process[c]

[If applicable, describe the annotation process and any tools used, or state otherwise. Describe the amount of data annotated, if not all. Have the annotations been aggregated? What are the limitations of the annotation method? Describe or reference annotation guidelines provided to the annotators. If available, provide inter-annotator statistics. Describe any annotation validation processes.]

Who are the annotators?[c,f]

[If annotations were collected for the source data (such as class labels or syntactic parsers), state whether the annotations were produced by humans or machine generated. Describe the people or systems who originally created the annotations and their selection criteria, if applicable. Is socio-demographic information on the annotators available? Can the annotators be assigned to a specific population group? Is the socio-demographic data on the annotators available to the users of the dataset? If available, include self-reported demographic or identity information for the annotators, but avoid inferring this information. Instead, state that this information is unknown. See Larson, 2017[d] for using identity categories as variables, particularly gender.]

Crowd Labour[e,f]

[Describe the conditions under which the data was annotated (for example, if the annotators were crowdworkers, state what platform was used, or if the data was found, what website the data was found on). If compensation was provided, include that information here.]

Personal and Sensitive Information[c,e,f]

[Sensitive data are better not expressed through quantitative information, but are recommended to be explained narratively. Descriptions might answer the following questions: Who or what are depicted in the dataset? If the dataset depicts people, are any specific subgroups of people represented? Are any specific individuals personally identifiable? If the dataset depicts people, are any individuals still living? Does this project comply with privacy laws in the countries where it will be shared? Does copyright impact this dataset? If so, how? Does this dataset pertain to a difficult history? If so, what extra precautions are being taken? Describe other people represented or mentioned in the data. Where possible, link to references for the information given.
State whether the dataset uses identity categories and, if so, how the information is used. Describe where this information comes from (i.e. self-reporting, collecting from profiles, inferring, etc.). See Larson, 2017[d] for using identity categories as variables, particularly gender. State whether the data is linked to individuals and whether those individuals can be identified in the dataset, either directly or indirectly (i.e., in combination with other data). State whether the dataset contains other data that might be considered sensitive (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 efforts were made to anonymise the data, describe the anonymisation process.]

Uses[b,f]

[Encourage dataset creators to reflect on the tasks for which the dataset should and should not be used.]

Links to Related Datasets, Publications and Models

[If the dataset relates to other datasets, was introduced by a paper or if there is a model which has been trained on this dataset, add URL here.]

Social Impact of Dataset[c]

[Discuss some of the ways you believe the use of this dataset will impact society. The statement should include both positive outlooks, such as outlining how technologies developed through its use may facilitate new knowledge creation, and discuss the accompanying risks. These risks may range from making important decisions more opaque to people who are affected by the technology, to reinforcing existing harmful biases (whose specifics should be discussed in the next section), among other considerations. Also describe in this section if the proposed dataset contains a low-resource or under-represented language. If this is the case or if this task has any impact on underserved communities, please elaborate here.]

Discussion of Biases[c]

[Provide descriptions of specific biases that are likely to be reflected in the data, and state whether any steps were taken to reduce their impact. See Blodgett et al., 2020[a] for a general discussion of the topic in NLP. If analyses have been run quantifying these biases, please add brief summaries and links to the studies here.]

Other Known Limitations[c]

[If studies of the datasets have outlined other limitations of the dataset, such as annotation artefacts, please outline and cite them here.]

Unanticipated Uses made of this Dataset[f]

[Since the intended (and especially the unintended) use of a dataset may possibly result in issues which are particular to a dataset, it is recommended to flag them here.]

Distribution[b]

[Dataset creators should provide answers to the following points prior to distributing the dataset: distribution information, dissemination strategy, licences, DOIs, third party involvement, restrictions.]

Dataset Curators[c]

[List the people involved in collecting the dataset and their affiliation(s). If funding information is known, include it here. Datasheets should be conceived of as flexible and dynamic objects which change according to the version of a dataset documented by them. Therefore, please provide a possibility to respond to and comment on the datasheet and indicate a reliable contact to the datasheet curators in order to enable the incorporation of responses and comments into a new version.]

Licensing Information[c]

[Provide the licence and link to the licence(s) webpage(s) if available.]

Citation Information[c]

[Provide the BibTex-formatted reference for the dataset. If the dataset has a DOI, please provide it here.]

Contributions[c]

[List the people who have contributed to this dataset as well as those who have curated and published it.]

Maintenance[b,f]

[Encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers. Four suggestions to communicate this: a) Regularly Updated – New versions of the dataset have been or will continue to be made available. / b) 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. / c) Limited Maintenance – The data will not be updated, but any technical issues will be addressed. / d) Deprecated – This dataset is obsolete or is no longer being maintained. Provide a possibility to respond to and comment on the datasheet; indicate a reliable contact to the datasheet curators and information on maintenance or updating of the dataset. Be prepared to incorporate responses and comments into a new version of the datasheet where this deems sensible.]

Bibliography Used for Establishing this Template

[a]Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[b]Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

[c]Hugging Face Dataset Card Creation Guide, https://github.com/huggingface/datasets/blob/main/templates/README_guide.md

[d]Larson, B. N. (2017). Gender as a Variable in Natural-Language Processing: Ethical Considerations. *EthNLP@EACL*. https://doi.org/10.18653/v1%2FW17-1601

[e]Lee, B. C. G. (2023). The 'Collections as ML Data' Checklist for Machine Learning & Cultural Heritage. *Journal of the Association for Information Science and Technology*, 1–22. https://doi.org/10.1002/asi.24765

[f]Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1776–1826. https://doi.org/10.1145/3531146.3533231