Completeness of Datasets Documentation on ML/AI Repositories: An Empirical Investigation

Completeness of Datasets Documentation on ML/AI Repositories: an Empirical Investigation

Marco Rondina(✉)[0009−0008−8819−3623], Antonio Vetrò[0000−0003−2027−3308], and Juan Carlos De Martin[0000−0002−7867−1926]

Politecnico di Torino, Torino, Italy
{marco.rondina,antonio.vetro,juancarlos.demartin}@polito.it

Abstract. ML/AI is the field of computer science and computer engineering that arguably received the most attention and funding over the last decade. Data is the key element of ML/AI, so it is becoming increasingly important to ensure that users are fully aware of the quality of the datasets that they use, and of the process generating them, so that possible negative impacts on downstream effects can be tracked, analysed, and, where possible, mitigated. One of the tools that can be useful in this perspective is dataset documentation.
The aim of this work is to investigate the state of dataset documentation practices, measuring the completeness of the documentation of several popular datasets in ML/AI repositories. We created a dataset documentation schema—the Documentation Test Sheet (DTS)—that identifies the information that should always be attached to a dataset (to ensure proper dataset choice and informed use), according to relevant studies in the literature. We verified 100 popular datasets from four different repositories with the DTS to investigate which information were present.
Overall, we observed a lack of relevant documentation, especially about the context of data collection and data processing, highlighting a paucity of transparency.

Keywords: data documentation · AI transparency · AI accountability.

1 Introduction and Motivation

Machine Learning / Artificial Intelligence (ML/AI) research made great strides in recent years, and its industrial applications became increasingly pervasive within society, automating organizational processes and decisions in several fields.

Datasets are fundamental in the ML/AI ecosystem and many issues related to fairness, transparency, and accountability in ML/AI systems are rooted in the data collection and in the data processing procedures [11]. Every decision made during the workflow may contain implicit values and beliefs [21], so tracking them can improve transparency [1]. The information accompanying them plays a very significant role in uncovering data issues [5], in fostering reproducibility and auditability [13], in ensuring accountability [10], users' trust [2], and in avoiding *data cascading* effects on the entire ML/AI pipeline [20]. With documentation,

it is possible to better understand the characteristics of the training data to at least partially mitigate the risks of downstream negative effects [4]. This is particularly true for those technologies where the impact of biased results can be severe, such as Speech Language Technologies (SLT) [16]. Documentation production should be seen as an essential part of dataset production, as a place to disclose fundamental choices, in parallel with what is proposed to be documented in terms of models [15,19] or rankings [25,26].

This study focuses on the dataset documentation state of practice. The aim was to measure whether and how much relevant information about data collection and data processing procedures is present in the documentation of the most popular (and influential [12]) datasets. The research question which directed the design of the research is: **Which of the information, that should be transparent to dataset users, is present in the most popular datasets in ML/AI repositories?** In order to answer this research question, we developed a test schema to measure the completeness of dataset documentation: Section 2 will describe the construction of the DTS. Subsequently, Section 3 describes the selection of repositories and datasets. The results of the application of the DTS are presented in Section 4. Finally, limitations (Section 5), future work (Section 7) and the conclusions (Section 6) are presented. Furthermore, we provide additional materials in the online Appendices[1].

2 Documentation Test Sheet from Related Works

We built a collection of recommended information that should be present in dataset documentation to ensure a proper choice of dataset and informed use. The aim was to recognize, with a study of relevant work in the literature of dataset documentation schemas, which information are important to be present in dataset documentation to achieve transparency, accountability, and reproducibility. The goal of this schema is to measure how complete a dataset documentation is: this property is the first necessary element to be scrutinized for enabling any further analysis on further quality dimensions of documentation (e.g., correctness). We called this schema the *Documentation Test Sheet* (DTS).

2.1 Fields of Information

The list of *Test Fields* is largely based on *Datasheets for Datasets* [7], with some further insights from relevant documentation standardization proposals in the literature [9,3]. We grouped the information into 6 sections, following the categorization presented in *Datasheets for Datasets* (DfD): **1** *Motivation*, **2** *Composition*, **3** *Collection processes*, **4** *Data processing procedures*, **5** *Uses*, **6** *Maintenance*. In addition to dataset metadata, some characteristics of the

[1] The appendices available at https://doi.org/10.5281/zenodo.8052683 contain: the DTS (A), the provenance of the field of information composing it (B), the metadata of the selected datasets (C), the reading principles that guided the documentation investigation (D), the raw results (E) and additional tables and figures (F).

data were tracked in section **c** *Characteristics*. We discarded the *Distribution* section because it proved inapplicable when testing documentation of datasets already published in public repositories. The full DTS can be found in Appendix A, but the list of *Test Fields* is also available in Table 2. Further details on the motivations behind this choice, and a description of the provenance of each information field, are reported in Appendix B. One of the novelties of this work is the design of the individual *Test Fields* as concepts expressed by few words to which it is easy to answer 'yes' or 'no', depending on the presence or absence of the related information in the documentation under analysis. Some fields from the related work were collapsed in a few *Test Fields* for the sake of brevity and ease of application. We designed the DTS to be generalizable as possible to any type of documentation, so that it can be used for datasets pertaining to different areas of ML/AI.

2.2 Measurement

The other core elements of the DTS are the *Presence Check Values* and the *Presence Averages*. During the analysis of the documentation, each *Test Field* is associated with a value indicating the presence or the absence of the represented information. Specifically, the *Presence Check Value* can take on one of the following three possible values:

- 1: it is possible to retrieve the information represented by the *Test Field*;
- 0: it is not possible to retrieve the information represented by the *Test Field*;
- NA: the information represented by the *Test Field* does not apply to dataset;

The *Presence Average* represents the completeness measure of the DTS. It is obtained by averaging the *Presence Check Values* of the group of *Test Fields* under analysis such as dataset (*Dataset Presence Average*), section (*Section Presence Average*), and field (among different datasets, *Field Presence Average*).

3 Study Design

One of the novel elements of this study concerns the analysis of the information found in the very same place where the data can be accessed, instead of selecting datasets from a corpus of academic papers [17,8]. The documentation in the public repository provides information about how the dataset is actually used in practice. Since the purpose of a scholarly article is different from that of a repository, some information may have no reason for inclusion in the article, and vice versa. For this reason, the documentation being analysed is the documentation web page where data can be downloaded. Given this design choice, we first selected the repositories under analysis, as described in section 3.1. In the second step, we collected the metadata useful to perform the dataset selection, as described in section 3.2. We focused on the most popular datasets, as seen in other work [6]. This is because these datasets are the most influential ones, and therefore studying their documentation is an important step in the path towards

a deeper understanding of common documentation practices. We then collected data on the presence of information in the documentation [2].

3.1 Repositories Under Analysis

The choice of repository is a relevant decision in this study, due to the design decision to analyse the online documentation present in the same place where the data are hosted. Indeed, different repositories have different documentation and metadata schemas. Therefore, we decided to select more than one repository to avoid obtaining too specific results. We selected four well-known and commonly used repositories to capture different practices in the ML/AI community. The criteria used for the choice were: free access; the presence of popularity proxies among the metadata; the presence of hundreds of datasets. We consulted the Wikipedia *List of portals suitable for multiple types of ML applications*[3] and, as a result, the following repositories have been selected: Hugging Face(HUG), Kaggle (KAG), OpenML (OML) and UC Irvine ML Repository (UCI).

3.2 Datasets Selection

To guarantee the feasibility of the research, it was also necessary to limit the number of datasets for each repository. For this reason, we selected 25 datasets from each repository, for a total of 100 datasets to be examined. We decided to focus on the concept of *popularity*, so that we could analyse some of the most used and influential datasets: where available, the number of *downloads* was identified as the best proxy; where not available, number of *views* was identified as a good alternative. The resulting metrics used are: HUG, number of downloads (APIs); KAG, platform upvotes and then the number of downloads (APIs)[4]; OML, number of downloads (web scraping); UCI, number of views (web scraping). We eliminated any duplicates within the same or different repositories (the comparison of information about the same dataset in different repositories was not a central aim of this research). As a selection criterion between duplicates, we used the highest 'popularity'. In the case of two datasets at the same ranking position, we eventually observed whether one of them was the primary source of the other. The full list of selected datasets, together with the date of data collection, can be found in Appendix C. The principles that guided the reading of the documentation are presented in the Online Appendix D.

4 Results and Discussion

In this section, we present the results and their discussion for each of the following levels: dataset (4.1), DTS section (4.2), *Test Fields* (4.3). Additional information on the distribution and dispersion of values is included in Appendix F.

[2] For reasons of space, summary tables with raw data are presented in Appendix E

[3] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

[4] Due to the unavailability of direct download count APIs, datasets were sorted by upvotes via APIs and then sorted by download count, as presented in the results.

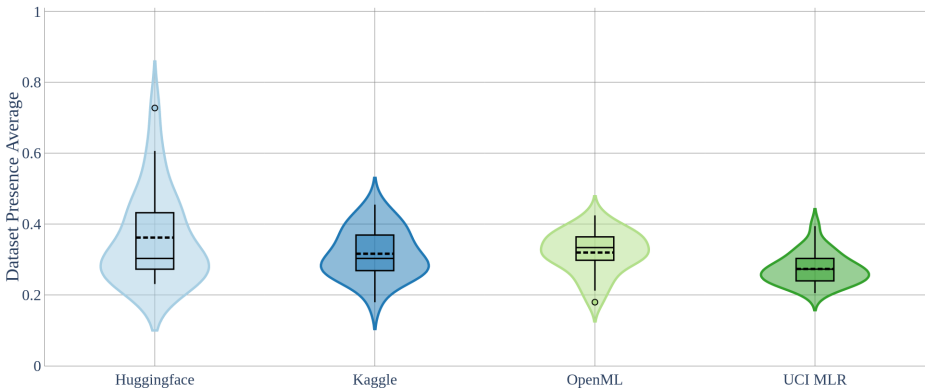

Fig. 1: Distribution of Dataset Presence Averages grouped by repository.

Table 1: Characteristics of the 100 selected datasets (25 for each repository).

Repository	**Data is people related**	**Presence of explicit target variable**	**Dataset is a sample or a reduction of a larger set**	**Recently updated**
Hugging Face	04	21	01	25
Kaggle	12	08	02	05
OpenML	11	25	07	00
UCI MLR	11	22	04	00
Total	38	76	14	30

4.1 Datasets Level

The dataset with the most comprehensive documentation was **hug16**, the *cnn dailymail* from Hugging Face (HUG). It contains over 300k unique news articles written by journalists. Its *Dataset Card* (i.e. its documentation) was comprehensive in all the different sections, and it can be considered a positive reference from the point of view of documentation practice. Figure 1 shows that overall very few datasets achieved more than 50% completeness, and variation between repositories is small. The selected datasets from HUG have the highest mean of the *Dataset Presence Average* distribution, while the ones from UCI have the lowest mean of the *Dataset Presence Average* distribution. One of the contributing factors to this result is that the three most complete documentations belong to HUG datasets.

In Table 1 it is possible to observe specific characteristics of the datasets: most datasets did not contain personal data, had an explicit target variable, and were not a sample or reduction of a larger set. All datasets updated after 1 January 2021 were considered 'Recently updated': all datasets from HUG and five datasets from KAG have been recently updated in terms of data or documentation, while all the OML and UCI datasets have not been updated in this timeframe. Additional statistics can be found in Appendix F.

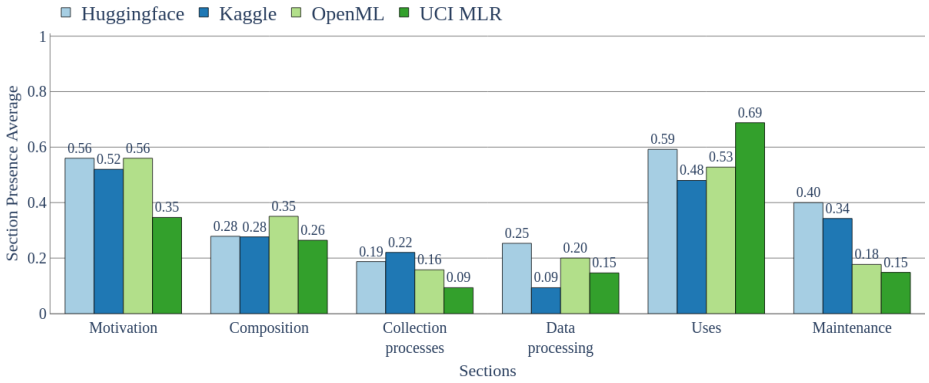

Fig. 2: *Section Presence Averages.*

4.2 Sections Level

As can be seen in Figure 2, the *Uses* section was the most complete one, followed by the *Motivation* section. Sections *Collection processes* and *Data processing procedures* had the lowest values of *Section Presence Average*. Additionally, we observed that the results of the *Maintenance* section are very different between repositories. These results suggest that the documentation of public datasets is currently utilisation-oriented, with less attention to the previous stages of the dataset construction pipeline. This aspect is also correlated with the high *Section Presence Average* of the *Motivation* section: the purpose of the dataset often encapsulates the meaning of why the data within it should be used. The low completeness of the *Composition*, *Collection processes* and *Data processing procedures* sections suggests that either little effort is devoted to describing the early stages of the dataset construction phase. Frequently, there is no information at all about these delicate phases. The failure to take into account these contextual aspects can lead to various problems in the models trained on such undocumented data [20]. Recent work devoted to partially automating the documentation process could help users to easily complete the *Composition* section [22,23,18]. Finally, the *Maintenance* results, although very variable between repositories, confirmed recent studies in the literature about the opportunities to improve documentation in datasets repositories and the lack of attention paid to what happens after the dataset is published [14,24]. As for the other aggregation level, the distributions of measurements are provided in Appendix F.

4.3 Test Fields Level

Table 2 shows the *Presence Average* value of each information field, globally and by the repository. Results show that certain documentation fields are very commonly used, such as the **2.01** *Description of the instances* (0,92), **2.02** *Number of the instances* (0,90) and **5.01** *Description of the tasks in which the dataset*

has already been used and their results (0,95). In many cases, the high level of completeness could be explained by the ability of the repository's metadata structure to promote the presence of a particular piece of information. Indeed, the information represented by these fields was very much present in repositories that structurally expose this information in the metadata schema of the repository. Conversely, it was almost completely absent in repositories that do not include such information in their metadata schema. Some examples are: **2.11** *Statistics* (HUG 0,00; KAG 1,00; OML 1,00; UCI 1,00), **5.04** *Repository that links to papers or system that use the datasets* (HUG 0,92; KAG 0,00; OML 0,00; UCI 1,00), **5.05** *Description of license and terms of use* (HUG 0,48; KAG 0,68; OML 1,00; UCI 1,00). This highlights the role played by repository hosts, who have the potential to trigger virtuous documentation practices.

In the *Motivation* section, on the one hand, it was very common to find information about the authors (the 'resource creators'), while on the other hand, it was rare to find details about who funded the creation of the dataset, important information for achieving accountability. Within the *Composition* section, basic information such as the description or number of instances was usually present. On the contrary, information about data confidentiality and dangerousness was usually absent. It is important to note that data protection laws may have been different before the datasets were made available (see Section 5 for more details). The analysis of information related to *Collection processes* pointed out, in a context of a general scarcity of details, the near total absence of specifics about ethical review processes and about analysis of potential impacts of dataset uses. With regard to the *Data processing procedures*, we observed that the 'Dataset Card' in HUG favoured the presence of (at least) some useful tags to obtain indications on the workers involved in these procedures. As already mentioned above, in terms of *Uses*, much attention on the part of dataset creators is paid to the description of the previous usage of the dataset and to the description of the recommended uses. The same cannot be said for non-recommended uses: only the documentation of a couple of HUG datasets contained this information. Surprisingly, although it was common to find details on the subject that supports or manages the dataset, the contact of the owner was rarely present. Furthermore, in terms of *Maintenance*, the DOI was quite rare and no information on the management of older dataset versions could be retrieved.

Table 2: *Field Presence Averages*: overall and for each repository.

ID	Field description	Tot	HUG	KAG	OML	UCI
1.01	*Purpose for the dataset creation*	0,57	0,64	0,52	0,68	0,44
1.02	*Dataset creators*	0,86	0,88	0,96	1,00	0,60
1.03	*Dataset funders*	0,06	0,16	0,08	0,00	0,00
2.01	*Description of the instances*	0,92	1,00	1,00	0,80	0,88
2.02	*Number of the instances*	0,90	0,92	0,72	1,00	0,96
2.03	*Information about missing values*	0,50	0,00	0,12	1,00	0,88
2.04	*Recommended data splits*	0,31	0,92	0,08	0,12	0,12

Continue on next page.

Table 2 – continued from previous page.

ID	Field description	Tot	HUG	KAG	OML	UCI
2.05	*Description of errors, noise or redundancies*	0,13	0,00	0,16	0,08	0,28
2.06	*Information about data confidentiality*	0,04	0,08	0,08	0,00	0,00
2.07	*Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases*	0,03	0,12	0,00	0,00	0,00
2.08	*Information about people involved in data production and their compensation (if people related)*	0,43	0,25	0,42	0,64	0,31
2.09	*Description of identifiability for individuals or subpopulations (if people related)*	0,15	0,50	0,17	0,09	0,08
2.10	*Description of data sensitivity (if people related)*	0,03	0,25	0,00	0,00	0,00
2.11	*Statistics*	0,50	0,00	1,00	1,00	0,00
2.12	*Pair plots*	0,00	0,00	0,00	0,00	0,00
2.13	*Probabilistic model*	0,00	0,00	0,00	0,00	0,00
2.14	*Ground truth correlations*	0,00	0,00	0,00	0,00	0,00
3.01	*Description of instances acquisition and data collection processes*	0,53	0,52	0,60	0,64	0,36
3.02	*Information about people involved in the data collection process and their compensation*	0,08	0,16	0,12	0,00	0,04
3.03	*Time frame of data collection*	0,19	0,04	0,48	0,12	0,12
3.04	*Information about ethical review processes*	0,01	0,04	0,00	0,00	0,00
3.05	*Information on individuals' knowledge of data collection (if people related)*	0,05	0,25	0,00	0,09	0,00
3.06	*Information on individuals' consent for data collection (if people related)*	0,05	0,25	0,00	0,09	0,00
3.07	*Analysis of potential impacts of the dataset and its use on data subjects*	0,00	0,00	0,00	0,00	0,00
4.01	*Description of sampling, preprocessing, cleaning, labelling procedures*	0,39	0,32	0,24	0,56	0,44
4.02	*Information about people involved in the data sampling, preprocessing, cleaning, labelling procedures and their compensation*	0,11	0,44	0,00	0,00	0,00
4.03	*Description of others possible sampling, preprocessing, cleaning, labelling procedures*	0,02	0,00	0,04	0,04	0,00
5.01	*Description of the tasks in which the dataset has already been used and their results*	0,95	0,92	1,00	1,00	0,88
5.02	*Description of recommended uses or tasks*	0,62	0,56	0,72	0,64	0,56
5.03	*Description of not recommended uses*	0,02	0,08	0,00	0,00	0,00
5.04	*Repository that links to papers or system that use the datasets*	0,48	0,92	0,00	0,00	1,00
5.05	*Description of license and terms of use*	0,79	0,48	0,68	1,00	1,00

Continue on next page.

Table 2 – continued from previous page.

ID	Field description	Tot	HUG	KAG	OML	UCI
6.01	*Information about subject supporting, hosting, maintaining the dataset*	0,84	0,36	1,00	1,00	1,00
6.02	*Contact of the owner*	0,30	0,20	0,80	0,16	0,04
6.03	*DOI*	0,09	0,24	0,04	0,08	0,00
6.04	*Erratum*	0,00	0,00	0,00	0,00	0,00
6.05	*Information about dataset updates*	0,38	1,00	0,52	0,00	0,00
6.06	*Information about management of older dataset versions*	0,00	0,00	0,00	0,00	0,00
6.07	*Information about the mechanism to extend, augment, build on, contribute to the dataset*	0,26	1,00	0,04	0,00	0,00

5 Threats to Validity and Limitations

One of the main limitations of this research is the non-scalability of the proposed procedure, which was primarily based on manual inspection of dataset documentation: the alignment of repositories metadata with the documentation fields proposed in the literature, and included in the DTS, was very poor.

The choice of repositories may have influenced the final result. However, by focusing on some of the most prominent repositories and the most popular datasets in each repository, we analysed the documentation of influential datasets. The dataset selection criteria - *popularity* - was implemented slightly differently to the different repositories, due to differences in the metadata schemas: however, the number of downloads was present in three out of four repositories, and for the remaining one we selected the most reasonable and available proxy (visualizations). In addition, popularity tends to be a proxy for longevity: this criterion may have introduced a selection bias, favouring datasets from a time when documentation was less important or emphasized and with different data protection laws. On the contrary, the lack of documentation updates on such datasets reinforces the findings of this study, i.e. poor attention/availability on dataset documentation.

Despite the fact that considerable effort has been made to make the data collection as accurate and standardized as possible, the study design, strongly based on human reading and interpretation of documentation texts, is inherently prone to the risk of interpretation errors. We controlled this threat by providing the reading principles in Appendix D.

Finally, due to lack of resources, the DTS was not tested for consistency and validation with target users: however, the information fields were all derived from documentation schemes already available in the academic literature.

6 Conclusions

We empirically investigated the state of documentation practice in the most popular datasets in the ML/AI community. A set of information that should

always be clear to the users of the datasets, in order to achieve transparency and accountability, was adapted into a *Documentation Test Sheet* (DTS) able to measure the completeness of documentation. The DTS was applied to 100 dataset documentations from Hugging Face, Kaggle, OpenML and UC Irvine MLR repositories.

This investigation brought out some relevant results about the state of practice of documentation of datasets manufacturing. First, it emerged that information related to how to use the dataset was the most present. On the contrary, maintenance over time or processes behind the data generation were very poorly documented. In general, a lack of relevant information was observed, highlighting a paucity of transparency. All these observations are even more relevant when considering that the analysis was restricted to some of the most popular and well-known datasets. Finally, the potential of repositories to help curators of datasets to produce better documentation emerged.

Altogether, these results let us hypothesize that efforts of the ML/AI community in devoting more attention to the dataset documentation process are necessary. These efforts might enable the reuse of datasets in a way that is more aware of the choices, assumptions, limitations and other aspects of their creation, and ultimately facilitating human-respectful ML/AI innovations. The proposed DTS can be an easy-to-use tool in the hands of dataset creators, maintainers, and hosts to move a further step in this direction.

7 Future Work

The first hypothesis of future work relates to increasing the number of datasets and repositories under investigation. Moreover, a complementary analysis of a selection of recent datasets could tell us if the growing awareness of data curation is bringing some results in common practice. Quantitative expansions of the research could be put investigating the feasibility of an automatic system capable of controlling the presence of information. This possibility, however, is fully dependent on the evolution of the repositories, and actions made possible by dataset hosts to standardize documentation and make it machine-readable.

From the qualitative point of view, it might be possible to expand the DTS to measure other aspects of documentation quality. For example, comparing the information found in the repositories with the information retrieved from academic articles using those datasets could reveal further insights to understand documentation practices, reduce documentation debt and possibly integrate it with additional aspects (e.g., 'sparsity' [6], dataset quality). Finally, a test with target users that also explores the differences between different types of dataset users could be useful for prioritizing DTS Test Fields according to possible users and uses.

Acknowledgements. This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022,

PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

1. Afzal, S., Rajmohan, C., Kesarwani, M., Mehta, S., Patel, H.: Data Readiness Report. In: 2021 IEEE Int. Conf. on Smart Data Serv. (SMDS). pp. 42–51 (2021). https://doi.org/10.1109/SMDS53860.2021.00016
2. Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K.N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., Varshney, K.R.: FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev **63**(4/5), 6:1–6:13 (2019). https://doi.org/10.1147/JRD.2019.2942288
3. Bender, E.M., Friedman, B.: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Trans. of the Ass. for Comp. Ling. **6**, 587–604 (2018). https://doi.org/10.1162/tacl_a_00041
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proc. of the 2021 ACM Conf. on FAccT. pp. 610–623. FAccT '21, ACM (2021). https://doi.org/10.1145/3442188.3445922
5. Boyd, K.L.: Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2), 438:1–438:27 (2021). https://doi.org/10.1145/3479582
6. Fabris, A., Messina, S., Silvello, G., Susto, G.A.: Algorithmic fairness datasets: The story so far. Data Min. Knowl. Disc. **36**(6), 2074–2152 (2022). https://doi.org/10.1007/s10618-022-00854-z
7. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III, H.D., Crawford, K.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021). https://doi.org/10.1145/3458723
8. Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., Huang, J.: Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In: Proc. of the 2020 Conf. on FAccT. pp. 325–336 (2020). https://doi.org/10.1145/3351095.3372862
9. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs] (2018). https://doi.org/10.48550/arXiv.1805.03677
10. Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., Mitchell, M.: Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In: Proc. of the 2021 ACM Conf. on FAccT. pp. 560–575. FAccT '21, ACM (2021). https://doi.org/10.1145/3442188.3445918
11. Jo, E.S., Gebru, T.: Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. Proc. of the 2020 Conf. on FAccT pp. 306–316 (2020). https://doi.org/10.1145/3351095.3372829
12. Koch, B., Denton, E., Hanna, A., Foster, J.G.: Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research (2021). https://doi.org/10.48550/arXiv.2112.01716

13. Königstorfer, F., Thalmann, S.: Software documentation is not enough! Requirements for the documentation of AI. Digital Policy, Regulation and Governance **23**(5), 475–488 (2021). https://doi.org/10.1108/DPRG-03-2021-0047
14. Luccioni, A.S., Corry, F., Sridharan, H., Ananny, M., Schultz, J., Crawford, K.: A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication. In: Proc. of the 2022 ACM Conf. on FAccT. pp. 199–212. FAccT '22, ACM (2022). https://doi.org/10.1145/3531146.3533086
15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model Cards for Model Reporting. In: Proc. of the Conf. on FAccT. pp. 220–229. FAT* '19, ACM (2019). https://doi.org/10.1145/3287560.3287596
16. Papakyriakopoulos, O., Choi, A.S.G., Thong, W., Zhao, D., Andrews, J., Bourke, R., Xiang, A., Koenecke, A.: Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In: Proc. of the 2023 ACM Conf. on FAccT. pp. 881–904. FAccT '23, ACM (2023). https://doi.org/10.1145/3593013.3594049
17. Peng, K., Mathur, A., Narayanan, A.: Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. arXiv:2108.02922 [cs] (2021)
18. Petersen, A.H., Ekstrøm, C.T.: dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R. J. Stat. Software **90**, 1–38 (2019). https://doi.org/10.18637/jss.v090.i06
19. Richards, J., Piorkowski, D., Hind, M., Houde, S., Mojsilović, A.: A Methodology for Creating AI FactSheets. arXiv:2006.13796 [cs] (2020). https://doi.org/10.48550/arXiv.2006.13796
20. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M.: "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In: Proc. of the 2021 CHI Conf. on Hum. Factors in Comput. Syst. pp. 1–15. CHI '21, ACM (2021). https://doi.org/10.1145/3411764.3445518
21. Scheuerman, M.K., Denton, E., Hanna, A.: Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2), 1–37 (2021). https://doi.org/10.1145/3476058
22. Schramowski, P., Tauchmann, C., Kersting, K.: Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? In: Proc. 2022 ACM Conf. on FAccT. pp. 1350–1361. FAccT '22, ACM (2022). https://doi.org/10.1145/3531146.3533192
23. Sun, C., Asudeh, A., Jagadish, H.V., Howe, B., Stoyanovich, J.: MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science. In: Proc. 28th ACM Int. Conf. on Inf. and Knowl. Manage. pp. 2893–2896. CIKM '19, ACM (2019). https://doi.org/10.1145/3357384.3357853
24. Thylstrup, N.B.: The ethics and politics of data sets in the age of machine learning: Deleting traces and encountering remains. Media, Culture & Soc. **44**(4), 655–671 (2022). https://doi.org/10.1177/01634437211060226
25. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H.V., Miklau, G.: A Nutritional Label for Rankings. In: Proc. 2018 Int. Conf. on Manage. of Data. pp. 1773–1776 (2018). https://doi.org/10.1145/3183713.3193568
26. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in Ranking: A Survey (2021). https://doi.org/10.48550/arXiv.2103.14000