Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles (ABOUT ML)

Appendix: Compiled List of Documentation Questions

This is an incomplete list of documentation proposals/questions/formats from a selection of published academic papers.
In the public comment process, please feel free to suggest additional papers and questions for PAI to add to this list as a resource.
In the future, the ABOUT ML project may take on a multistakeholder process of surfacing common questions and creating a synthesized single list.

Fact Sheets (Arnold et al 2018)

Source: Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv preprint arXiv:1808.07261. https://arxiv.org/abs/1808.07261

A few examples of items a FactSheet might include are:

- What is the intended use of the service output?
- What algorithms or techniques does this service implement?
- Which datasets was the service tested on? (Provide links to datasets that were used for testing, along with corresponding datasheets.)
- Describe the testing methodology.
- Describe the test results.
- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service?
- Are the service outputs explainable and/or interpretable?
- For each dataset used by the service: Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?
- Does the service implement and perform any bias detection and remediation?
- What is the expected performance on unseen data or data with different distributions?
- Was the service checked for robustness against adversarial attacks?
- When were the models last updated?

Data Sheets (Gebru et al 2018)

Source: Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.https://arxiv.org/abs/1803.09010

Motivation

- 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 this 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.
- Any other comments?

Composition

- 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.
- 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 existed 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.
- 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

- 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?
- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
- 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.
- 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.
- Any other comments?

Preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, 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 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.
- Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
- Any other comments?

Uses

- 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 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?
- Are there tasks for which the dataset should not be used? If so, please provide a description.
- Any other comments?

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., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
- When will the dataset be distributed?

- 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.
- 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.
- 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.
- Any other comments?

Maintenance

- Who is supporting/hosting/maintaining the dataset?
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
- Is there an erratum? If so, please provide a link or other access point.
- 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.
- Any other comments?

Model Cards (Mitchell et al 2018)

Source: Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM.
https://arxiv.org/abs/1810.03993

Model Details. Basic information about the model.

- Person or organization developing model
- Model date

- Model version
- Model type
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
- Paper or other resource for more information
- Citation details
- License
- Where to send questions or comments about the model

Intended Use. Use cases that were envisioned during development.

- Primary intended uses
- Primary intended users
- Out-of-scope use cases

Factors. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.

- Relevant factors
- Evaluation factors

Metrics. Metrics should be chosen to reflect potential real world impacts of the model.

- Model performance measures
- Decision thresholds
- Variation approaches

Evaluation Data. Details on the dataset(s) used for the quantitative analyses in the card.

- Datasets
- Motivation
- Preprocessing

Training Data. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.

Quantitative Analyses

- Unitary results
- Intersectional results

Ethical Considerations

Caveats and Recommendations

A "Nutrition Label" for Privacy (Kelley et al 2009)

Source: Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009, July). A nutrition label for privacy. In Proceedings of the 5th Symposium on Usable Privacy and Security (p. 4). ACM. http://cups.cs.cmu.edu/soups/2009/proceedings/a4-kelley.pdf

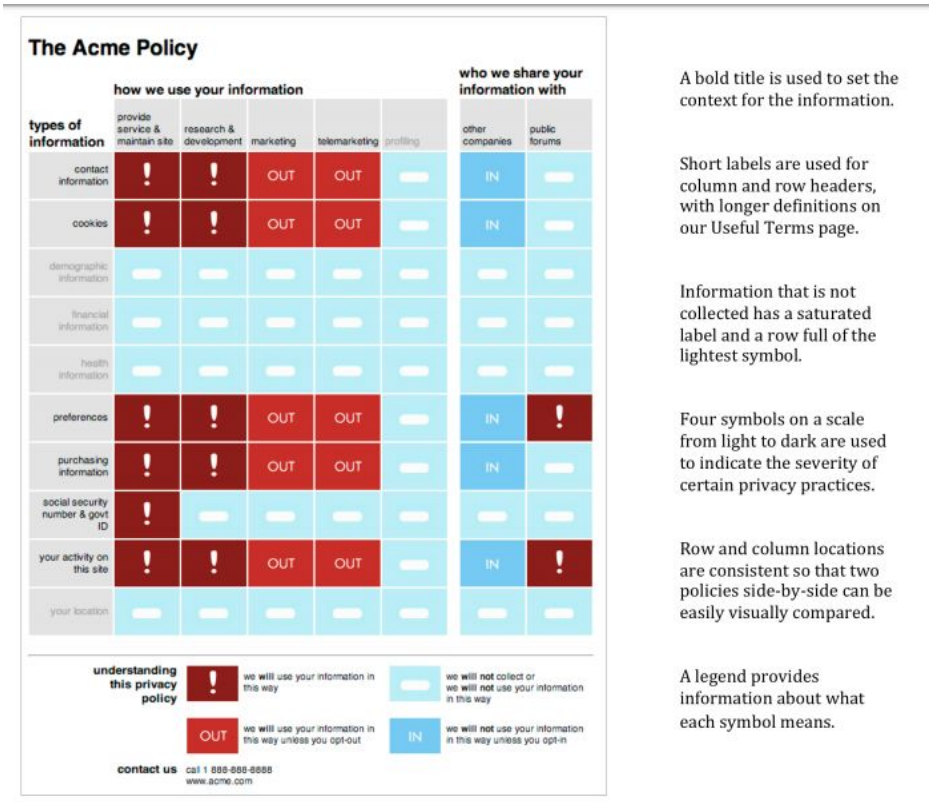

Figure 5. Our *proposed Privacy Nutrition Label.* This label is the one we tested in the second focus group and the laboratory study.

The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards (Holland et al 2019)

Source: Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. arXiv preprint arXiv:1805.03677. https://arxiv.org/abs/1805.03677

- What is the relevant metadata collected with the dataset?
- Detail source, author contact information and version history

- Ground truth correlations: linear correlations between a chosen variable in the dataset and variables from other datasets considered to be “ground truth”

Module Name	Description	Contents
Metadata	Meta information. This module is the only required module. It represents the absolute minimum information to be presented	Filename, file format, URL, domain, keywords, type, dataset size, % of missing cells, license, release date, collection range, description
Provenance	Information regarding the origin and lineage of the dataset	Source and author contact information with version history
Variables	Descriptions of each variable (column) in the dataset	Textual descriptions
Statistics	Simple statistics for all variables, in addition to stratifications into ordinal, nominal, continuous, and discrete	Least/most frequent entries, min/max, median, mean,.etc
Pair Plots	Distributions and linear correlations between 2 chosen variables	Histograms and heatmaps
Probabilistic Model	Synthetic data generated using distribution hypotheses from which the data was drawn - leverages a probabilistic programming backend	Histograms and other statistical plots
Ground Truth Correlations	Linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth", such as Census Data	Heatmaps

Table 1. Table illustrating 7 modules of the Dataset Nutrition Label, together with their description, role, and contents.

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science (Bender and Friedman 2018)

Source: Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.https://aclweb.org/anthology/papers/Q/Q18/Q18-1041/

A. CURATION RATIONALE Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? This can be especially important in datasets too large to thoroughly inspect by hand. An explicit statement of the curation rationale can help dataset users make inferences about what other kinds of texts systems trained with them could conceivably generalize to.

B. LANGUAGE VARIETY Languages differ from each other in structural ways that can interact with NLP algorithms. Within a language, regional or social dialects can also show great variation (Chambers and Trudgill, 1998).

The language and language variety should be described with:

- A language tag from BCP-479 identifying the language variety (e.g. en-US or yue-HantHK)
- A prose description of the language variety, glossing the BCP-47 tag and also providing further information (e.g. English as spoken in Palo Alto CA (USA) or Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin)

C. SPEAKER DEMOGRAPHIC Sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with speaker demographic characteristics (Labov, 1966), as speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). Transfer from native languages (L1) can affect the language produced by non-native (L2) speakers (Ellis, 1994, Ch. 8). A further important type of variation is disordered speech (e.g. dysarthria).

Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socio-economic status
- Number of different speakers represented
- Presence of disordered speech

D. ANNOTATOR DEMOGRAPHIC What are the demographic characteristics of the annotators and annotation guideline developers? Their own 'social address' influences their experience with language and thus their perception of what they are annotating.

Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socio-economic status
- Training in linguistics/other relevant discipline

E. SPEECH SITUATION Characteristics of the speech situation can affect linguistic structure and patterns at many levels. The intended audience of a linguistic performance can also affect linguistic choices on the part of speakers. The time and place provide broader context for understanding how the texts collected relate to their historical moment and should also be made evident in the data statement.

Specifications include:

- Time and place
- Modality (spoken/signed, written)

- Scripted/edited v. spontaneous
- Synchronous v. asynchronous interaction
- Intended audience

F. TEXT CHARACTERISTICS Both genre and topic influence the vocabulary and structural characteristics of texts (Biber, 1995), and should be specified.

G. RECORDING QUALITY For data that includes audio/visual recordings, indicate the quality of the recording equipment and any aspects of the recording situation that could impact recording quality.

H. OTHER There may be other information of relevance as well (e.g. the demographic characteristics of the curators). As stated above, this is intended as a starting point and we anticipate best practices around writing data statements to develop over time.

I. PROVENANCE APPENDIX For datasets built out of existing datasets, the data statements for the source datasets should be included as an appendix.

Example: Google modified image results for the search term 'ceo' based on constraints set by their declared principle on diverse representation