

Codebook

Subject Traceability Content Analysis

= structured content analysis
* = unstructured thematic analysis

Documents

Categories in this section relate to high-level descriptions of both the coding process and the documents being coded.

Documents Analyzed *

The documents analyzed for each database. The main purpose of this column is to make it easy to track and find the documents being analyzed, particularly when cross-checking and quality check each other's coding. Any and all links to the following should be included:

- Paper PDFs
- Websites
- Contracts

Coder

Who coded the specific database.

Sample Strategy

How the database was sampled. There are 4 options:

- Random Sample
- Top 10 Most Cited Datasets
- exposing.AI
- Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 317 (October 2021), 37 pages.
<https://doi.org/10.1145/3476058>

Dataset Information

Human Subject Information

Does the data contain images of real humans?

Whether the dataset contains images of real human beings. Any synthetic or GAN-generated human beings do not count as a “yes.” This question is used to check whether the dataset meets the criteria for inclusion in the sample. Any “no” answers will be removed from the sample and replaced.

Are humans the main subject of the database?

While all datasets should be humans, they are not necessarily the main subject of the dataset. Use your best determination based on the documentation and purpose of the dataset to decide if the main subject of the dataset are humans.

If the task is something like Face Recognition, it is very likely that the answer is yes. If the task is something like Object Recognition, it is likely the answer is no.

Options: Yes or No

Are real human faces visible and plausibly identifiable to the human eye?

Use your own determination to decide whether human faces are identifiable to the human eye. In cases where humans were captured using drone footage, it may be likely that, while humans are featured, their faces are not plausibly identifiable.

Options: Yes or No.

Basic Dataset Information

Retracted

Whether the dataset was officially retracted by the authors.

Options: Yes, No, N/A if it cannot be found.

Face, Body, Non-Corporeal Task

Whether the intended task of the dataset was specific to face, body, or non-corporeal (non-human) tasks. Non-human tasks may still include humans, they are just not the main focus of the intended use of the dataset.

Task(s) Intended *

What task(s) authors intended the database to be used for which are enumerated in the paper(s)? Separate tasks with semicolons (;) for ease of parsing.

Use of Data in Model

Whether the dataset is used in a model in the original documentation or proposed solely as a dataset.

Options: Yes or No

What type of organization do the authors come from?

The type of affiliation of each author on the original dataset paper. Separate by semicolon.

Options (as derived through inductive coding): Private Company; University; Government

Funding *

Copy and paste any funding information from the original dataset paper.

Funding Organization Type

Funding type of each organization, not each grant. Separate by semicolon. N/A for no funding.

Options (as derived through inductive coding): Nonprofit; University; Private Company; Military; Government.

Domain

Motivations

The motivation the authors cite when creating the dataset (e.g., for research, for policing). This will likely be in the introduction or conclusion of the text. Separate by semicolon. Domains will be normalized after inductive coding.

Use Cases *

What real world uses the authors are imagining this database would be used for? Separate tasks with semicolons (;) for ease of parsing. Use cases will be used to normalize motivating domains.

Motivations Thematic *

How do the authors describe the intended domain(s) in the documents? Paste quotes from documentation into this section.

Contribution

Categories in this section describe the intended contribution of the dataset in comparison to prior datasets.

Comparison to Prior Datasets

If the dataset has any comparisons to prior datasets, list the prior datasets here, separated by a semicolon.

How do authors describe this database's relation to previous databases? *

If the authors compare their dataset to prior datasets, how do they do so? Feel free to copy the text here.

Limitations

Technical Limitations

Do the authors include limitations/challenges of the database itself?

Whether the authors included the limitations or challenges faced in creating the database itself (not related to models or validation). This relates to technical challenges of the database, such as gathering the data.

Options: Yes/No.

If the database has limitations/challenges, what are they? *

If the answer to the previous question was yes, how do the authors describe the limitations/challenges to the database?

If the answer to the previous question was no, write N/A.

Ethical Limitations

Do the authors explain or enact any explicit privacy considerations (of either collection or or use of the database)?

Whether the authors included privacy considerations of the database explicitly. This can include privacy consideration in collection of the data or of use of the data. For example, whether the authors decided not to collect data for privacy or reasons, or limit certain uses of the data for privacy reasons.

For example:

- Any statements of privacy of the participants
- Any statements of privacy of annotators
- Any statements explicitly stating faces were blurred for privacy reasons

Options: Yes or No.

Do the authors explain or enact any explicit ethical considerations (of either collection or or use of the database)?

Whether the authors have a specific explanation or section regarding ethical trade offs.

Options: Yes or No.

If there are ethical or privacy considerations, what are they? *

If the answer to the previous question was yes, how do the authors describe the ethical considerations of the database?

If the answer to the previous question was no, write N/A.

Data Licensing

Did the authors describe the limitations of use?

Whether the authors include some limitation to how the database can be used. This is more often on the website than in the research paper. This could include a license database users must fill out or a statement outlining the terms of service. Include both examples which require a license before accessing the database and which do not (e.g., there is a terms of use, but you can still download the database without explicitly agreeing to it).

Options: Yes, No, or N/A (when the database cannot be found).

If the authors describe limitations of use, what are they? *

If the answer to the previous question was yes, how do the authors describe the limitations of use? Copy the text from the website, paper, or license.

If the answer to the previous question was No or N/A, write N/A.

What kind of license is used? *

If there is a license, what kind of license is it? This may be up to interpretation, as the authors might not explicitly state any type of license. N/A for no license.

Examples may include:

- Terms of Use: Terms one must agree to to use the dataset
- Access Agreement: A form that must be filled out to access the dataset
- Creative Commons: A license under Creative Commons policies.

What terms does the license discuss?

If there is a license or terms of use, list the type of limitation. Items will be normalized by inductive analysis on prior question. Separated by semicolons.

Do the authors require a specific citation when using the data?

Whether the authors explicitly require a paper to be cited when using the dataset.

Options: Yes, No, N/A if license cannot be found

Do the authors block access to the data unless a license is filled out and submitted, either by email or a form?

If you cannot access the data without filling out a form or contacting someone, then the answer would be “yes.” Otherwise, it would be “no.” If you can access some part of the data freely but not all: “yes and no.” N/A if the dataset cannot be found or has been retracted.

Data licensing notes *

Any interesting notes to keep in mind about the licensing. For example, whether the data can just be downloaded, whether you have to mail in a form, etc.

Data Collection

Data Subjects

Who is featured in the dataset?

What type of person is featured in the dataset?

Options:

- Regular People
- Public Figures
- Celebrities
- Fictional Characters
- Unknown

Use best judgment to determine the difference between celebrities and public figures. Data authors will usually indicate in their own language how they define these, but there may be some overlap. For example, politicians would likely be considered public figures, not celebrities.

What type of data is featured? *

More details on who is featured in the dataset, as well as aspects of the data such as:

- Is it face or body,
- Is it images or videos,
- What are people doing?
- What are the annotations used?

Data Sources

Where was the data collected from?

List where exactly the data was collected from. Separate all sources with semicolons. Sources will be normalized with spelling and terminology after the list is complete.

E.g., Google; Wikipedia; Prior dataset

Data collection groups

From the answers from the above question, normalize categories of where the data came from into groups. Datasets may use multiple data collection groups. Separated by semicolons.

- Web search engines
- Video sharing sites

- Prior Datasets
- Online photo albums
- Public spaces
- Social media
- Unknown
 - Either the source is not named or it is too vague to determine (e.g., “the internet”)
- Informational websites
- Public records
- User rating websites
- Stock image websites
- Other
 - All other sources which do not fit into the above categories (e.g., paintings, niche/specific websites)

Was the data collected from prior datasets?

Whether the dataset reuses data from prior datasets, in all or in part.

Options: Yes or No

If from prior datasets, did the authors discuss the original data license?

If the answer to the previous question was yes, indicate whether the authors mention anything about the licensing governing the original data. It can be from any source (e.g., paper, website).

Options: Yes, No, N/A. If the answer to the previous question was No or N/A, put N/A. Put N/A if you cannot find the dataset (the authors may have previously mentioned the license but it may be lost now).

Was the data collected from public websites or other public data sources (like public records)?

Whether the data in the dataset was collected from websites online or public data sources like public records.

Options: Yes or No

If from public websites, did the authors discuss the original data license?

If the answer to the previous question was yes, indicate whether the authors mention anything about the licensing governing the original data. It can be from any source (e.g., paper, website).

Options: Yes, No, N/A. If the answer to the previous question was No or N/A, put N/A. Put N/A if you cannot find the dataset (the authors may have previously mentioned the license but it may be lost now).

If they discussed the data license, how? *

For those datasets which used data from:

- Prior datasets
- Public websites
- Public records

and also discuss the original license of those data, indicate what about the license was mentioned. Feel free to copy the explanation from the text here.

Was the data collected by authors or participants in real world public settings? *

Whether the data was collected in real world settings, such as a campus or a public street. This may happen in a few ways:

- The authors themselves went out and collected images of people
- The authors hired people to collect images
- The authors use their own collection of personal images which include public settings.

Options: Yes or No

Thematic Data Collection *

Paste any documentation about the data collection process, including any contextual information about the questions in this section.

Ethical Considerations

Copyright Owner Agency

Do the authors state whether the copyright owners could have images removed?

When the authors collected data from prior websites, did the authors state in any of their materials whether the original copyright owners could have their images removed from the dataset?

Options:

For datasets collected from public websites or records, yes or no.

If the answer to the question about public websites or records was no, N/A.

Subject Agency

Do the authors state whether those featured in the database can opt-out post collection?

Options: Yes or No

Do the authors state whether consent was given by the subjects?

Options: Yes or No

Do the authors state whether the study went through an IRB process (or international equivalent)?

Options: Yes or No.