Datasheet for ‘A cross-verified database of notable people, 3500BC-2018AD’*

Parth Samant

2024-12-02

Extract of the questions from Gebru et al. (2021).

Motivation

1. *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.*
    - The dataset was created to address specific questions in social science, particularly relating to gender, economic growth, and urban/cultural development,
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by a research group from multiple different universities across the world.
3. *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.*
    - This article was funded from grants by LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005–02)
4. *Any other comments?*
    - No

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

*Code and data are available at: https://github.com/samantparth/Wikipedia-Historical-Prominence-Trends

- Instances that comprise the dataset represent notable individuals with variables representing their characteristics.

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

   - Over 2 million instances of notable people on the dataset

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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - It is a sample from a larger group of 'notable people'. I do not believe it is representative of the entire sample of notable people, since there is a bias towards both modern figures and ones in certain countries.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of rankings related to Wikipedia prominence (such as the number of readers), geographic information (such as place of birth), occupational information, and other demographic information such as their name and year of birth.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - No

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Yes, many individuals have missing information, particularly information about years of death (since many people are still living).

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Yes, these are made explicit. the ranking_5criteria and the sum_ln_5criteria variables are variables that are a calculation of prominence based on 5 metrics relating to fame/prominence.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- No

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Since there are millions of entries, there are likely at least some errors. This may be related to their occupation, name, year of death, etc.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - It is self contained and does not rely on any other datasets.

11. *Does the dataset contain data that might be considered confidential (for example, 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.*

    - No

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

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, these subpopulations include age, gender, name, year of birth, year of death, and their time period. These are identified from the Wikidata website that provides basic information on notable individuals

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Yes, since the dataset contains the name of everyone on there.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.*

- No, since the data obtained is publically available.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was indirectly derived and it was validated/verified through multiple techniques. Some of these validations include removing duplications and identifying incorrect (including fictional) instances of humans.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Wikipedia and Wikidata have mechanisms for which the data can be downloaded from ther website. Most of the demographic information comes from a raw Wikidata dataset, where the researchers perform data cleaning to make it accurate.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - No

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Researchers were involved and they were compensated through the grants.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The timeframe is notable people born anywhere between 3500BC and 2018 AD.

6. *Were any ethical review processes conducted (for example, 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.*

- No

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

   - Obtained it from a publically accessible website, the Sciences-Po Dataverse

8. *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.*

   - No, they were not notified

9. *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.*

   - No

10. *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).*

    - N/A

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.*

    - No, an analysis has not been conducted on this specific dataset relating to its impact on data subjects.

12. *Any other comments?*

    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.*

   - Yes, they eliminated multiple instances of the same person and removed instances of fictional people. They also created many of their own variables that are an amalgamation of existing variables, such as a ranking of each individual by prominence.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - No

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

   - Yes, the code that they used to preprocess the data was given. The link can be found on their github: https://medialab.github.io/bhht-datascape/

4. *Any other comments?*

   - No

Uses

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

   - Other than this repository, I was unable to find other instances that used this same dataset.

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.*

   - No

3. *What (other) tasks could the dataset be used for?*

   - It can be used for:
     - evaluating bias of the amount of notable people by region
     - analysing life expectancy based on the country of origin
     - seeing which occupations notable people tend to have most often

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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - Yes, since certain characteristics may be biased (such as an overrepresentation of men, people from Europe, and people from the modern era), future uses may have a higher chance of analysing this dataset for bias.

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

- No

6. *Any other comments?*

    - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes - the dataset is publically available from Sciences Po, which is a universtiry in Paris.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed through the Sciences Po dataverse. It has a DOI: https://doi.org/10.21410/7E4/RDAG3O.

3. *When will the dataset be distributed?*

    - There is no set date for releases, but it tends to be once every few years.

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.*

    - No, the article is open access and licensed under a Creative Commons Attribution 4.0 International License

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.*

    - No

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.*

    - No

7. *Any other comments?*

    - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is hosted and maintainedby SciencesPo, with 6 authors that update the dataset every few years.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The authors can be contacted directly through email, which is contained on their GitHub website.

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

    - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - The dataset is updated every few years, and new updates are uploaded onto their github website

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.*

    - No, since the dataset obtains data from Wikipedia.

6. *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 dataset consumers.*

    - Yes, they are hosted and they are featured on their GitHub website

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 dataset consumers? If so, please provide a description.*

    - Yes, since the dataset is publically accessible to anyone, anyone can augment or extend the parametres of the dataset if they so wish. They can notify the authors by email if they believe it would be singificant for the goal of identifying notable people. The contributions do not have to be validated because of the nature of the dataset's distribution.

8. *Any other comments?*

    - No

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.