Datasheet for 'A dataset'*

subtitle

Julia Kim

April 2, 2024

Abstract

This supplementary datasheet accompanies our replication paper. Following the standardised format proposed by Gebru et al. (2021), this datasheet documents the motivation, composition, collection process, recommended uses, maintenance and distribution of our cleaned dataset. Its objective is to facilitate better communication between the dataset creator and consumer, and to encourage the statistics community to prioritise transparency and accountability.

Table of contents

1 Acknowledgement

All of the following questions are extracted from Gebru et al. (2021).

2 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 enable analysis of the main determinants of governmental decisions about the listing of endangered species under the Endangered Species Act (ESA).

*Code and data are available at: https://github.com/julia-ve-kim/US_Climate_Change_Biodiversity; Minimal replication on Social Science Reproduction platform available at https://doi.org/10.48152/ssrp-tpay-ac71.

- This dataset was lightly adapted from the publicly available dataset provided in the replication package by Moore et al.(2022), which contained all necessary information as to $60,481$ endemic species of the United States, their taxonomic classification, ESA listing status, conservation status, scientific and common standardised n-gram frequency and genus size. Changes in the cleaning process, documented in the scripts folder, include removing (1) ten columns not of interest to the analysis, (2) long tails of common and scientific n-grams, (3) n-grams with suspicious ratios and (4) "extinct" and "prob. extinct" from the NatureServe conservation status variable.

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 primarily created by F. Moore, A. Stokes and X. Dong of the Department of Environmental Science & Policy at the University of California Davis as well as M. Conte of the Department of Economics at Fordham University. It was lightly adapted by myself, Julia Kim, an undergraduate student of the Department of Physics at the University of Toronto.
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.*
   - The work of Moore et al. (2022) was funded by NSF CNH2-S: Understanding the Coupling between Climate Policy and Ecosystem Change. My own work is not funded.
4. *Any other comments?*
   - Note that Moore et al. (2022) obtained their data from a variety of government and scientific sources, including NatureServe, Google Ngram, Integrated Taxonomic Information Service (ITIS), National Centre for Biotechnology Information (NCBI) and the United States Fish & Wildlife Service (USFWS).

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.*
   - Each instance denotes a species endemic to the United States. It gives the species' taxon, binary listing status, standardised common and scientific n-gram frequencies and the size of its genus.
2. *How many instances are there in total (of each type, if appropriate)?*
   - The dataset contains 60,481 instances.
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).*
   - The dataset is a sample of instances from the larger dataset created by Moore et al. (2022), available in their replication package as well as in the raw data folder of our GitHub repository. The larger dataset contains $64,583$ instances of the same kind of data, prior to processing: it contains additional rows with long tails of common and scientific n-grams, n-grams with suspicious ratios or species with "extinct" or "prob. extinct" NatureServe conservation status.
   - The sample is not intended to be representative of the larger dataset. It selects only the rows (and columns) whose values were considered valid or of main interest for this analysis.

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.

- As described previously, each instance contains the species' taxon, NatureServe conservation status, standardised common and scientific n-gram frequencies, the size of its genus and binary listing status.
- The species' taxon is one of nine possible taxonomic classes, as determined by NatureServe, of which there are five vertebrate species – amphibians, birds, fishes, mammals and reptiles – and four non-vertebrate species – plants, invertebrates, fungi and protists.
- Its conservation status, also assigned by NatureServe, is one of the following five: critically imperiled, imperiled, vulnerable, apparently secure and secure.
- Its standardised common and scientific n-gram frequencies give (transformed) Google Ngram frequencies of the common and scientific name of the species, respectively, as determined by case-insensitive searchers in Google Books' English corpus of 2019.
- Its genus size gives the number of species in that species' genus, as collected by the Integrated Taxonomic Information Service (ITIS) and the National Centre for Biotechnology Information (NCBI).
- Its binary listing status is set to 1 if the species is listed under the ESA at the time of the assessment, and to 0 otherwise.

5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - there is no label or target associated with each instance. In the dataset provided by Moore et al. (2022), each instance was accompanied by the specific name of the species. This we removed during cleaning, because it was not necessary for our analysis.
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.*
   - There are no missing information from the taxon, listed or status column.
   - The number of entries missing from the scientific and common n-gram columns, as well as the genus size column, is 93, 260 and 7122, respectively. In the case of n-grams, there were some cases where a name failed to return valid data between the years 1800 and 2016 in the Google Books Ngram Viewer (which only offers frequencies for words and phrases that occur in at least 40 books), as discussed by Moore et al. (2022). In the case of genus size, there are some species for which this information was simply not yet available in either the ITIS or NCBI repositories.
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.*
   - There is a mild relationship between the scientific and common n-grams. Figure 2 of our main paper shows the distribution of the standardised common and n-gram frequencies per taxon, wherein we can observe a mild correlation between them.
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.*
   - There are no data splits used in this analysis.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no redundancies in the dataset, as every instance corresponds to a unique species.
   - There are likely sources of error in every variable in our dataset, mostly due to measurement. These are comprehensively discussed in the Discussion section of our main paper. Most importantly, variables associated with the greatest amount of error are likely to be the scientific and common n-grams. The primary sources of their error are discussed by Orwant (2014) as being:

   (1) optical character recognition (OCR) errors, particularly where the type has faded, the pages are dirty, or the typeface is hard to OCR,

   (2) sample bias in the collection, whereby we can only count the words in books we scan, and

we can only scan the books we can find,

(3) metadata errors in library records, whereby a book is incorrectly classified as belonging to the wrong time period.

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.*
    - The dataset is a slight modification of the analysis dataset provided in the replication package of Moore et al. (2022), of which a copy is contained in the raw data folder of my GitHub repository.
    - Like the original dataset, it relies on data from various scientific and government sources. These are described, as follows:

    (1) The taxon and conservation status data are obtained from NatureServe, an authoritative source of biodiversity data throughout the United States (NatureServe 2024). The database is dynamic, being continuously updated by botanists, zoologists and ecologists, with inputs from scientists at heritage and conservation programs (NatureServe 2024). The assessment of species is revised on a periodic basis, for which we use the most recent assessment conducted between 1985 and 2019, which will remain constant over time. Since NatureServe methodology is characterised by the FAIR principles (NatureServe 2024), it is guaranteed that NatureServe will continue to support the dataset and make it publicly available on-line. NatureServe also provides official archival versions of the dataset. One can access its full data repository here.

    (2) The scientific and common name Google Ngrams data are provided by the Google Books Ngram Viewer's English corpus of 2019. It is guaranteed that the English corpus of 2019 will exist, and remain constant over time. This is so far the first corpus, consisting of over 16 million books published between 1470 and 2019 (Michel et al. 2011), so there are no official archival versions yet available. The English corpus of 2019 is freely available on-line to the public here.

    (3) Measures of the number of species per genus is collected from two automated databases, ITIS and NCBI, through the package `taxize` (Chamberlain and Szocs 2013) in `R` (R Core Team 2023). It is guaranteed that their data will continue to exist and be publicly available on-line. Both provide archival versions of their dataset, free for use by the public. One can access ITIS and NCBI's full data repository at the links provided here and here, respectively.

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.*
    - No.
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.*
    - The dataset contains information and name of natural, extant species. None of these

species, however, are human persons.

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.
16. *Any other comments?*
    - None.

3 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 dataset for each instance was acquired from the analysis dataset made publicly available by Moore et al. (2022).
    - The NatureServe, NCBI and ITIS data were all observed indirectly, being collected by other field-workers and experts employed by these organisations. The Google Ngram Viewer data was directly observed and obtained by Moore et al. (2022), who performed case-insensitive searchers in Google Books' English corpus of 2019 for all scientific and common names in the NatureServe data sets.
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?*
    - Moore et al. (2022) are to be accredited for having collected the data for this replication analysis.
    - They obtained NatureServe's assessment data through the NatureServe data repository here, the Ngram frequencies by performing case insensitive searchers for all scientific and common names in the NatureServe dataset between the years 1800 and 2016, and the genus size from NCBI and ITIS data repositories through the `taxize` (Chamberlain and Szocs 2013) package in `R` (R Core Team 2023).
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - As discussed previously, the dataset was obtained from a slightly larger analysis dataset available in the replication package of Moore et al. (2022). Only cleaning was performed to eliminate unneeded rows and columns from the dataset for my analysis; no sampling strategy was required.
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)?*
    - Moore et al. (2022) are to be accredited for having collected the data for this replication analysis.
    - Professors F. Moore and X. Dong are employed by the University of California Davis, whilst Professor M. Conte is employed by Fordham University. A. Stokes is a graduate student, supported by the University of California Davis, who as of 2023 receive salaries around $5,000 monthly (Lambert 2023).
    - The taxonomic classification and NatureServe conservation status are determined by experts, who collect information from field surveys, taxonomic treatments and other scientific publications, whose salaries are estimated to be between 20-50 US dollars an hour

(NatureServe 2019).
    - The ITIS and NCBI data curate taxonomic information from submissions made by taxonomy services, specialists, and researchers working on primary literature in the field (ITIS et al. 2023; Schoch et al. 2020). These experts collect and analyse morphological and molecular data from thousands of species, and compare their values to references in relevant phylogenetic literature, in order to develop robust and consistent taxonomic classifications (Smith et al. 2006). It is not clear how much these workers were compensated for the data we used in our anaysis.
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.*
    - We used NatureServe's most recent assessment of taxonomic class and conservation status dataset, whose data were collected between 1985 and 2019.
    - We used Google Books' English corpus of 2019, whose data were collected between 2010 (at the time of the creation of Google Ngram Viewer tool) and 2019.
    - We used the NCBI and ITIS dataset of 2018, whose data were collected between 2013 and 2018.
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.*
    - There is no ethical review process conducted for the information included in our dataset.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - I did not collect the data directly. It was obtained from the replication package of Moore et al. (2022), who did not include their raw data, only their analysis data.
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.*
    - Not applicable: there is no information on any individual in the dataset, only on extant, non-human species.
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.*
    - Not applicable.
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).*
    - Not applicable.
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.
12. *Any other comments?*
    - None.

4 Preprocessing, cleaning, labelling

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, process-*

ing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Yes, there were four important steps in the cleaning process: these included

(1) removing ten columns not of interest to the analysis, being the `code`, `family`, `order`, `name`, `status_global`, `ngram_common_flag`, `evdist`, `ge`, `edge` and `probs` variables,
(2) setting standardised n-gram frequencies larger than 10 to `NA`,
(3) removing n-grams with suspicious ratios, being defined as instances for which the ratio of standardised common to scientific n-grams exceeded 10,

(4) removing "extinct" and "prob. extinct" from the NatureServe conservation status variable.

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.*
   - Yes, the raw data is included in the raw data folder of our GitHub repository
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 open-source statistical programming software `R` (R Core Team 2023) was used to clean the data.

4. *Any other comments?*
   - In particular, the `tidyverse` (Wickham et al. 2019) package of `R` (R Core Team 2023) was required during the cleaning process.

5 Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - Yes, the original dataset, prior to cleaning, was used in the paper by Moore et al. (2022). We are conducting a replication of this paper.
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.*
   - My GitHub repository is linked here.
   - The replication package of Moore et al. is linked here.
3. *What (other) tasks could the dataset be used for?*
   - The dataset is used for understanding the main determinants of governmental decisions in listing an endangered species under the ESA. It may also be used for visualising the distribution of endangerment levels, n-gram frequencies and (mean) genus sizes per taxon.
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?*
   - None.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - The dataset should not be used to make normative claims about government choices or to attempt to value species. To avoid this, our paper makes a strictly positive study of govenrment choices, seeking to reveal preferences through actual decisions.
6. *Any other comments?*
   - None.

6 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.*
   - No.
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 will be distributed via the GitHub repository.
   - There is no DOI to our dataset.
3. *When will the dataset be distributed?*
   - The dataset will be distributed on April 2nd, 2024.
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.*
   - The dataset will be distributed under the MIT 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.*
   - None.
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?*
   - None.

7 Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The dataset will be hosted in my GitHub repository. I, Julia Kim, will be responsible for maintaining the dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - I may be contacted through my e-mail, juliaym.kim@mail.utoronto.ca.
   - The contact information of the authors who created the original dataset, from which my own is but a slight modification, as discussed previously, follow:
     - F. Moore: fmoore@ucdavis.edu
     - M. Conte: mconte7@fordham.edu
     - X. Dong: xdgli@ucdavis.edu
     - Note the e-mail address of A. Stokes was not included in the author information of the paper by Moore et al. (2022).
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)?*
   - After April 18th, 2024, the dataset will not further be updated to reflect any new changes.
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.

- The dataset does not relate to people.

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.*
    - The sole version of the dataset that will be maintained is the one in the cleaned folder of my GitHub repository.
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.*
    - Partially. If others wish to extend/augment/build on/contribute to the dataset, then they must fork my publicly available GitHub repository and freely modify the data on their own platform. Changes to the forked repository, however, will not change the original work contained in my own repository.
    - Should I wish, in the future, to extend my own work, then GitHub provides an option to add collaborators to my public repository, who are granted the rights to extend/augment/build on/contribute to the dataset directly.
8. *Any other comments?*
    - None.

References

Chamberlain, Scott, and Eduard Szocs. 2013. "Taxize - Taxonomic Search and Retrieval in r." *F1000Research* 2: 191. https://f1000research.com/articles/2-191/v2.

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.

ITIS, S. Alexander, A. Hodson, D. Mitchell, D. Nicolson, T. Orrell, and D. Perez-Gelabert. 2023. "The Integrated Taxonomic Information System." In *Catalogue of Life Checklist (Version 2023-10-31)*, edited by O. Bánki, Y. Roskov, M. Döring, G. Ower, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, et al. ITIS. https://doi.org/10.48580/dfgnm-4ky.

Lambert, E. 2023. "Salary Scales and Information." UC Davis Graduate Studies. https://grad.ucdavis.edu/understanding-your-student-salary.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph Porter Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.

Moore, Frances C, Arianna Stokes, Marc N Conte, and Xiaoli Dong. 2022. "Noah's Ark in a Warming World: Climate Change, Biodiversity Loss, and Public Adaptation Costs in the United States." *Journal of the Association of Environmental and Resource Economists* 9 (5): 981–1015.

NatureServe. 2019. "NatureServe Tax Forms." NatureServe. https://www.natureserve.org/sites/default/files/2021-07/natureserve_2019_990_pd_signed.pdf.

———. 2024. "Unlocking the Power of Science to Guide Biodiversity Conservation." https://www.natureserve.org/.

Orwant, Jon. 2014. "What Are Some Sources of Errors in Google Ngram Viewer Results?" *Quora*. https://www.quora.com/What-are-some-sources-of-errors-in-Google-Ngram-Viewer-results.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schoch, Conrad L, Stacy Ciufo, Michael Domrachev, Carol L Hotton, Sujatha Kannan, Rita Khovanskaya, Detlef Leipe, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database: The Journal of Biological Databases and Curation* 2020. https://doi.org/10.1093/database/baaa062.

Smith, Alan R., Kathleen M. Pryer, Eric Schuettpelz, Petra Korall, Harald Schneider, and Paul G. Wolf. 2006. "A Classification for Extant Ferns." *Taxon* 55 (3): 705–31. https://doi.org/10.2307/25065646.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.