A cross-verified database of notable people, 3500BC-2018AD

Nature Scientific Data *9, no. 1 (2022): 1-19*

Morgane Laouenan, Palaash Bhargava, Jean-Benoît Eyméoud,
Olivier Gergaud, Guillaume Plique & Etienne Wasmer

Link to paper

Note on birth and death locations update
Jan. 2023

By Minda Belete[1] and Palaash Bhargava[2][3]

Summary

This document presents the steps of a minor update of the database A Cross-verified Database of Notable People 3500BC-2018AD, Nature Scientific Data 9, no. 1 (2022): 1-19. The updated database is available here as well as the codes used and intermediate datasets (https://doi.org/10.21410/7E4/YLG6YR). The document details a few corrections we made to two variables: **the birth place, and the death place when the person is not alive**. We also incorporated the information about new deaths since our initial data collection (2018), now updated as of August 1st, 2022.

Most of the changes come from new information in Wikidata which was unavailable in the previous round of data collection. In our 2022 paper, we created a cross-verified database of 2.29 million famous individuals. We update birth (resp. death) locations for 7% (4%) of these individuals (from the cross-verified database) for whom information on birth location was previously unavailable. Additionally, we update information on birth (resp. death) locations when the information was available and has been updated or corrected in the meantime. The latter changes come either from an improvement in the precision of birth place (resp. death place) reported, or from a change in the geo coordinates of the location in Wikidata. These changes tend to be small, representing a median distance of approximately 1 km. Only 4% (resp. 1.6%) of individuals have a change greater than 3 km. A few updates represent a relatively more substantive change in location: the distance between the old reported place and the new reported place becomes 76 km (resp. 21 km) for 0.9% (resp. 0.4%) of individuals.

We also provide updates and corrections for the extended database of 4.6 million individuals that was not cross-verified but reported for completeness in the original paper.[4] This database is available here.

We recommend using this updated version of the database if one wishes to use birth and death places for their analysis.

Our correction relies on the universe of Wikidata to inform our choice of the location reported. A manual check of a sample of the top 500 individuals (by notability index defined in the paper) and of a random

[1] New York University, Abu Dhabi. Email: mindabelete@nyu.edu

[2] Columbia University, Email: pb2794@columbia.edu

[3] Corresponding Author

[4] See Appendix for further details on the statistics corresponding to the extended database.

sample of 5,000 random individuals in our restricted database indicates a residual error rate of approximately 0.5%.

Details of correction procedure

We start the collection of information from the Wikidata dump dated August 1, 2022. We parse information from the dump file directly to create a smaller json file containing all humans (Q5). We extract information on birth and death place (associated to property P19 and P20 respectively) for each human being along with their date of birth and date of death. We also update the date of death for individuals who have died between our last data collection date (October 2020) and August 1, 2022. The raw information on birth and death places are stored in the form of Qcodes which are then converted into readable format using the steps mentioned below.

We compile a list of all birth and death locations from the dump and identify several instances that they might be associated with, such as hospital (Q15917), city (Q515), neighborhood (Q123705), state (Q7275) etc. We compile a list of these instances and scrape information on entries associated with such instances from the Wikidata dump. We additionally use WikiData client services to append information about missing entries. Kindly refer to the original Read me file available on the database website for further information about this step.

We utilize the latitude and longitude of the location entry and append it to the reported birth / death place. In case the coordinates are not directly available, we look for the administrative territory of the location and assign those coordinates to the birth / death place. In some cases, we get multiple locations as birth or death places for an individual. We explain in detail below how we choose the most appropriate location in that case.

In order to report the most appropriate birth / death location, we adopt the following iterative methodology. Over the course of our data compilation in the last 6 years, we have collected information from WikiData in 2018, 2020 and 2022. For the purpose of this document, we refer to them as three different datasets. For each individual, we compare the locations reported across three different datasets and choose the most appropriate location to report. Primarily, we rely on the level of precision with regards to coordinate location (called *precision level*, henceforth), the level / size of location reported (called *hierarchical precision,* henceforth) and the distance between locations reported across time.

Step 1: Assign precision level within a dataset and decide on the most appropriate location within the dataset.

We define the precision level of a location as follows:

- If the exact coordinates of the location are available, we label the location precision 'Precise',
- If the administrative territory's coordinates are available instead of the exact location coordinates (for e.g. city of the hospital an individual is born in), we report those coordinates and label them 'Imprecise level 1'.
- If the coordinates of the administrative territory are not available but the coordinates of the administrative territory of the administrative territory are available (location of location), we

report those and call it ‘Imprecise level 2’. We do not iterate further (location of location of location), however, if we have not been able to fetch coordinates after the first two iterations but the location is still present (without location coordinates), we call the location precision, ‘Imprecise level >=3’, we do not report the location coordinates but still report the name of the original location.

In case of multiple locations for an individual within the same dataset, we pick the most precise location. When locations have a precision tie, we report them all.

Step 2: Assign precision based on a hierarchy of locations within a dataset and decide on the most appropriate location.

Precision based on hierarchy is defined as follows: while scraping information about locations from the WikiData dump, we compiled a list of instances associated with a location. These instances are descriptions of the location. They tell us whether a specific location is a city, country, river, building, etc. We manually determine a hierarchy among those different instances, for e.g. city is more precise than a country, building is more precise than a city or a district, etc. In case we are left with multiple locations after step 1, we pick the most precise location based on the defined hierarchy within a dataset.

Steps 1 and 2 can only be carried out for datasets corresponding to 2020 and 2022. As for the data collected in 2018, we only extracted a single location for each individual.

Step 3: Generate Haversine distances between all possible locations left across three datasets post steps 1 and 2.

Once we have carried out steps 1, 2 and 3, we classify individuals into the following mutually exclusive and exhaustive categories. We report the most precise location for each individual (after comparing the three datasets) using a unique decision rule associated with each separate category. Information for an individual is not necessarily available in each dataset. We exhaustively use all the information available over time. The categories and the corresponding decision rules are:

Based on the information availability and type for birth (death) location, we first categorize individuals into four different sets:

1. **Info - single (IS):** information available in the 2018 dataset and at best only single locations are available across all three datasets
2. **Info - multiple (IM):** information available in the 2018 dataset and multiple locations are present in either the 2020 or 2022 dataset
3. **No info - single (NIS):** no information available in the 2018 dataset and at best single locations are available across the 2020 and 2022 datasets
4. **No info - multiple (NIM):** no information available in the 2018 dataset and multiple locations are present in either the 2020 or 2022 dataset

We further divide the 4 sets into the following subsets and implement the following decision rule:

Subsets of Info-Single:

- **IS0:** Information is present in all three datasets (2018, 2020, and 2022) and all of them match. We report information from 2022.
- **IS1:** Information is present in all three datasets and the distance between all collected locations is lower than 5km. We report information from 2022.
- **IS2:** Information is present in all three datasets, the maximum distance between all collected locations is more than 5kms. In this case, we apply a majority rule if applicable, i.e. if 2 out of 3 location coordinates are identical[5] then we report the latest location which is in majority. For example, in case location coordinates from 2018 and 2022 match after rounding but are different from 2020, we use 2022.
- **IS3:** Information is missing in either 2020 or 2022 but not in both. The location available in 2020 (2022) is within 5km from the 2018 location. We report the latest available location.
- **IS4:** Information exists in all three datasets and one of the locations (in 2020 or 2022) is less than 10 km away from all other locations. We report the latest location satisfying this criterion.
- **IS5:** Information is missing in both 2020 and 2022. We report the 2018 location.
- **IS6:** Information is missing in 2020 or 2022 but not in both and distances from the 2018 location are larger than 10km. We report the latest available location.
- **IS7:** Information exists in all three datasets and all locations in the latest year are at least >10 km away from the locations reported in the older datasets . We report the most precise location available between 2020 and 2022.
- **IS8:** Information exists in all three datasets (there is a precision tie) and all locations in the latest year are >10 km away from the locations reported in the older datasets. We report the latest location which satisfies the minimum distance criteria (i.e. it has the smallest distance to previously reported locations (previously for 2020 being 2018 dataset and for 2022 being 2020 and 2018 dataset).

Subsets of Info-Multiple:

- **IM0:** Information exists in all three datasets and at least one location matches between 2018, 2020 and 2022, we report that location.[6]
- **IM1:** Information exists in 2018 and 2020 (2022) but not 2022 (2020) and one of the locations from the latter dataset matches the 2018 location. We report the matched location.
- **IM2:** Information exists across all three datasets but there is no location that is consistent across the three datasets. We use the majority rule, i.e. if a location from 2022 matches the 2018 location or any location from 2020 then we report that one, else if a location from 2020 matches the 2018 location, we report that location from 2020. For example, consider a case where the coordinates for 2018 birthplace are (-8.408,40.201), the coordinates for 2020 birthplaces are

[5] To ensure that we do not treat locations that are only possibly 1-2 kms apart as separate locations, we round the location coordinates to the second decimal place.

[6] In case of multiple locations within the same dataset, we round up the coordinates to the second decimal place.

[(40.211,-8.429), (38.717,-9.167)], and the 2022 birthplaces are [(40.211,-8.429), (38.708,-9.139)]. In totality, we have 5 locations (1 from 2018 and 2 from 2020 and 2022 each). However, there is no unique location that is consistent across all three years. In this case, since the first location in 2022 matches the first location in 2020, the first location of 2022 is chosen as the final location.

- **IM3:** We generate distances from the newly detected locations to previously reported locations and report the latest location with the minimum distance to all previously reported locations.

Subsets of No Info-Single:

- **NIS0:** Information on 2020 and 2022 matches. We report 2022.
- **NIS1:** Information is available only in one of the datasets (2020 or 2022). We report whatever is available.
- **NIS2:** Information is available in 2020 and 2022 but one is more precise than the other. We report the most precise location.
- **NIS3:** Information is available in 2020 and 2022 and the precision level of locations across time match. We report the 2022 location.

Subsets of No Info-Multiple:

- **NIM0:** at least one location in 2020 and 2022 match (after rounding to 0.01). We report the matched 2022 location.
- **NIM1:** Information is available only in one dataset (i.e. either 2020 or 2022 is missing but not in both). We randomly select one location that is available.
- **NIM2:** Locations across time do not match. We report the most precise which satisfies the minimum distance criteria (as described under IM3). If the precision level of locations matches, we report the location that satisfies the minimum distance criteria.

Once this classification is complete, we apply one final override rule. In case the final location reported is different from the location in 2022 but has the same or worse precision level in comparison to 2022 (in terms of hierarchy based on instances), we replace the final reported location with the location present in 2022. This is done under the assumption that the most recent information which doesn’t differ in its precision level is more likely to be correct than older existing similar information .

The new version of the database only reports the updated birth and death locations. If one is interested in obtaining the interim information such as the classification, 2020 and 2022 locations, raw locations extracted from Wikidata, please contact the corresponding authors of the original manuscript.

Appendix: Full statistics for the extended database

Overall, for 379,060 (193,758) individuals in the extended dataset the new reported birth (resp. death) place differs from the birth (resp. death) place we reported originally. Of this, 330,919 (87%) individuals did not have a birth place assigned to them initially and were only assigned a birth place in the new iteration. Similarly, 174,942 (90%) individuals did not have a death place assigned to them initially and were only assigned a death place in the new iteration.[7][8] Amongst these individuals, for 50% of the cases, the new birth (resp. death) location is within a 23 km (9 km) radius of the old location. This distance becomes 547 km (444 km) at the 90th percentile for the subset with different birth (resp. death) locations when we order these observations by the distance between old and new location.

Summary Statistics and Distance between old and new reported places:

Breakdown of counts and frequency of individuals under each classification subset

	Birth Locations		**Death Locations**	
Classification	Count	% age of total observations	Count	% age of total observations
Subset: Info-Single				
IS0	1,481,451	62.017	538,177	56.082
IS1	439,593	18.402	208,679	21.746
IS2	115,122	4.819	31,106	3.241
IS3	537	0.022	175	0.018
IS4	3105	0.130	731	0.076
IS5	5,168	0.216	1,301	0.136
IS6	1,079	0.045	142	0.015
IS7	53	0.002	8	0.001
IS8	4,761	0.199	1,050	0.109
Subset: Info-Multiple				
IM0	5,211	0.218	2,116	0.221
IM1	17	0.001	5	0.001
IM2	1,717	0.072	1,175	0.122
IM3	48	0.002	11	0.001

[7] The new additions include the recently deceased.

[8] The numbers slightly differ when we compare location coordinates instead of location names. The number of individuals who have the same birth (death) location as before but different birth (death) coordinates than the one originally reported is 487,897 (216,903), However, this is driven by a change in coordinates for the respective city / location due to better data availability or change in the geo-center of the location as identified by WikiData. This happens for 16,314 (8070) birth (death) locations and in 90% of the cases, the new location coordinates lie within a 27 km radius of the old coordinates.

Subset: No Info-Single				
NIS0	85,886	3.595	38,792	4.042
NIS1	219,780	9.201	123,771	12.898
NIS2	2,964	0.124	538	0.056
NIS3	21,570	0.903	11,320	1.180
Subset: No Info-Multiple				
NIM0	246	0.010	123	0.013
NIM1	420	0.018	313	0.033
NIM2	18	0.001	2	0.000
NIM3	35	0.001	83	0.009
Total	**2,388,781**	**100**	**959,618**	**100**

Note: Frequency and counts of individuals under each subset of classification detailed in the aforementioned text. These classification numbers are reported prior to the manual and 2022 corrections we implemented as a final override. For information on the classification codes mentioned in column 1, please refer to the text above. Column 2 and 3 report statistics for birth places and Column 4 and 5 report statistics for death places.