DISCUSSION PAPER SERIES

DP15852
(v. 2)

A Cross-verified Database of Notable People, 3500BC-2018AD

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

ECONOMIC HISTORY

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

LABOUR ECONOMICS

CEPR

ISSN 0265-8003

A Cross-verified Database of Notable People, 3500BC-2018AD

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

Discussion Paper DP15852
First Published 26 February 2021
This Revision 24 March 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Economic History
- International Trade and Regional Economics
- Labour Economics

A Cross-verified Database of Notable People, 3500BC-2018AD

Abstract

We add to the literature on notable individuals (famous, prominent, distinguished) in collecting first a massive amount of data from various editions of Wikipedia and Wikidata along with deduplication techniques; and then using these partially overlapping sources to cross-verify each retrieved information. This strategy results in a cross-verified database of 2.2 million individuals, including a third who are not present in the English edition of Wikipedia. An extension to 4.7 million entries is currently not recommended given the inaccuracy of the information and discrepancies between Wikidata and other sources. A non-negligible fraction of newly-added individuals were collected from non-English editions of Wikipedia. We adopt a social science approach: data collection is driven by specific social questions on gender, economic and cul- tural development and quantitative exploration of cultural trends, that we document in this paper. A sample of 100,000 individuals is available here http://medialab.github.io/bhht-datascape, together with the most recent version of this paper.

JEL Classification: N01, N9, R00

Keywords: Notable individuals, Creative Class, Urban Economics, economic history

Morgane Laouenan - morgane.laouenan@sciencespo.fr
centre d'economie de la sorbonne

Palaash Bhargava - pb2794@columbia.edu
Columbia University

Jean-Benoît Eyméoud - jeanbenoit.eymeoud@sciencespo.fr
Banque de France

Olivier Gergaud - olivier.gergaud@kedgebs.com
Kedge Business School, France

Guillaume Plique - guillaume.plique@sciencespo.fr
Medialab, Sciences Po

Etienne Wasmer - wasmer.etienne@gmail.com
NYUAD and CEPR

Acknowledgements

Corresponding authors: Morgane Laouenan and Etienne Wasmer. Financial support from LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02) and NYUAD is gratefully acknowledged. In the paper, BC refers to Before Common Era (negative calendar years) and AD to Anno Domini (positive calendar years). The original data collection and the extraction of individual characteristics were mainly done by Morgane Laouenan using Stata. Jean-Benoit Eymeoud carried out the second wave of data collection using Python. Palaash Bhargava processed the Wikidata dumps. Guillaume Plique ran algorithms to detect and remove duplicate individuals and developed the data vizualisation website. The other five authors have all contributed to the data analysis and to forthcoming follow-up papers. We thank Paul Girard and the Sciences Po Medialab for helping us collecting additional data and creating an effective visualization tool who greatly helped us improve the quality of our database, the Atelier de Cartographie at Sciences Po and in particular Thomas Ansart, Anouk Pettes and Patrice Mitrano, Sarah Asset, Simon Fredon and Nicolas Britton, Marie N'Dongue, Jordane Roussel, Marie Le Tallec, Maeva Hartmann, Cassiopeia Van den Bussche from Sciences Po, as well as Ke Shi, Ian Quinn Lutz, Anna Pustovoit, Amna Hassan, Alemayehu Mekonen Abebe, Anas Jawed, Sorin Panfile, Oleksandr Serhiyovych Petriv, Samridha Man Shrestha, Martin Smit, Karim Boudlal, Mouhamad Ba, Mate Hekfusz, Minda Belete from NYUAD, Chandan Thapa, Abhishek Nehra, Aditya Chhabra, Hema Baid, Apoorv Somanchi from DSE for expert research assistance. Special thanks to Julia Mink for expert verifications of the database, to the students of the different sections of the CORE class "5000 years of human lives" at NYUAD and the instructors Mendgi Song and Dayin Wijaya. We also thank Sascha O. Becker, Karol Borowiecki, Nicolas Baumard, David de la Croix, Michel Serafinelli as well as participants to various conferences and seminars for insightful discussions.

A Cross-verified Database of Notable People (3500BC-2018AD)*

Morgane Laouenan†, Palaash Bhargava‡,
Jean-Benoît Eyméoud§Olivier Gergaud¶,
Guillaume Plique‖, Etienne Wasmer**

March 23, 2021

Abstract

Short abstract We add to the literature on notable individuals (famous, prominent, distinguished) in collecting first a massive amount of data from various editions of `Wikipedia` and `Wikidata` along with deduplication techniques; and then using these partially overlapping sources to cross-verify each retrieved information. This strategy results in a cross-verified database of 2.29 million individuals, including a third who are not present in the English edition of `Wikipedia`. An extension to 4.7 million entries is currently not recommended given the inaccuracy of the information and discrepancies between `Wikidata` and other sources. We adopt a social science approach: data collection is driven by specific social questions on gender, economic, urban and cultural development and quantitative exploration of cultural trends, that we document in this paper, as well as geo-location of individuals. A sample of 100,000 individuals is available here https://medialab.github.io/bhht-datascape, together with the most recent version of this paper.

*Corresponding authors: Morgane Laouenan and Etienne Wasmer. This paper describes the 2.0 version of a project started in 2014 collecting and exploiting data from individuals with a biography in several language editions of `Wikipedia` under license https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License and from `Wikidata` under license https://creativecommons.org/publicdomain/zero/1.0/. This paper is a discussion paper aiming at getting comments and feedback and sharing freely a sample of 100,000 observations. The final version of the full database will be published under the terms of the Creative Commons Attribution-ShareAlike at the time of the scientific publication. Financial support from LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02) and NYUAD is gratefully acknowledged. In the paper, BC refers to *Before Common Era* (negative calendar years) and AD to *Anno Domini* (positive calendar years). The original data collection and the extraction of individual characteristics were mainly done by Morgane Laouenan using Stata. Jean-Benoît Eyméoud carried out the second wave of data collection using Python. Palaash Bhargava processed the `Wikidata` dumps. Guillaume Plique ran algorithms to detect and remove duplicate individuals and developed the data vizualisation website. The other five authors have all contributed to the data analysis and to forthcoming follow-up papers. We thank Paul Girard and the Sciences Po Medialab for helping us collecting additional data and creating an effective visualization tool who greatly helped us improve the quality of our database, the Atelier de Cartographie at Sciences Po and in particular Thomas Ansart, Anouk Pettes and Patrice Mitrano, Sarah Asset, Simon Fredon and Nicolas Britton, Marie N'Dongue, Jordane Roussel, Marie Le Tallec, Maeva Hartmann, Cassiopeia Van den Bussche from Sciences Po, as well as Ke Shi, Ian Quinn Lutz, Anna Pustovoit, Amna Hassan, Alemayehu Mekonen Abebe, Anas Jawed, Sorin Panfile, Oleksandr Serhiyovych Petriv, Samridha Man Shrestha, Martin Smit, Karim Boudlal, Mouhamad Ba, Mate Hekfusz, Minda Belete from NYUAD, Chandan Thapa, Abhishek Nehra, Aditya Chhabra, Hema Baid, Apoorv Somanchi from DSE for expert research assistance. Special thanks to Julia Mink for expert verifications of the database, to the students of the different sections of the CORE class "5000 years of human lives" at NYUAD and the instructors Mendgi Song and Dayin Wijaya. We also thank Sascha O. Becker, Karol Borowiecki, Nicolas Baumard, David de la Croix, Michel Serafinelli as well as participants to various conferences and seminars for insightful discussions.

†CNRS, Centre d'Economie de la Sorbonne and LIEPP-Sciences Po. Email: morgane.laouenan@univ-paris1.fr

‡Columbia University. Email: palaash.bhargava@columbia.edu

§Department of Economics and LIEPP-Sciences Po. Email: jeanbenoit.eymeoud@sciencespo.fr

¶KEDGE Business School and LIEPP-Sciences Po. Email: olivier.gergaud@kedgebs.com

‖médialab-Sciences Po. Email: guillaume.plique@sciencespo.fr

**NYU Abu Dhabi and LIEPP-Sciences Po. Email: ew1555@nyu.edu

Long abstract A new strand of literature has recently emerged that attempts to build the most comprehensive and accurate database of notable (famous, prominent, distinguished) individuals or important public figures. We add to this literature in collecting first a massive amount of data from various editions of `Wikipedia` and `Wikidata` along with deduplication techniques; and then using these partially overlapping sources to cross-verify each retrieved information. For some variables, `Wikipedia` adds 15% more information when it is missing in `Wikidata`. We also find that there are very few errors in the part of the database that contains the most documented individuals. We also find non trivial error rates (around 1%) in the bottom of the notability distribution, due to sparse information and classification errors or ambiguity. This either requires manual corrections for future use or a statistical treatment of these errors in statistical approaches. One therefore needs to trade-off the size of the database and the precision of the data and the adequacy of the concepts to the research questions. This strategy results in a cross-verified database of 2.29 million individuals, including a third who are not in the English edition of `Wikipedia` (the extension to 4.7 million entries is currently not recommended given the inaccuracy of the information and discrepancies between `Wikidata` and other sources) and a non-negligible fraction of newly-added individuals who played a significant role in important periods of Human history. Second, we adopt a social science approach: data collection is driven by specific social questions on gender, economic, urban and cultural development and quantitative exploration of cultural trends, that we document in this paper, as well as geo-location of individuals. This approach is used in particular to document the Anglo-Saxon bias naturally present in existing projects based on the English edition of `Wikipedia`. A sample of 100,000 individuals is available here https://medialab.github.io/bhht-datascape, together with the most recent version of this paper.

1 Introduction

Since the time Plutarch's *Parallel Lives* was written in the beginning of the second century AD and his 23 biographies have survived two thousand years (or even more ancient, the Epic of Gilgamesh dates back to 2000 BC), the task of registering famous individuals and their influence has been a recurrent field of study. Over the last few years, this task has been undertaken to a much larger scale, with a growing number of databases documenting history, allowing statistical analysis of socio-historical facts, at a scale that had never been reached so far. This paper develops a cross-verified database of 2.3 million notable individuals using several `Wikipedia` editions and `Wikidata`.

This approach was pioneered by Schich et al. (2014) who focused on lifetime mobility of 150,000 notable individuals across centuries and continents using all available birth and death data found in `Freebase`, a Google-owned knowledge database. Yu et al. (2016) also used `Freebase` and the English edition of `Wikipedia` to assemble a manually-verified database of 11,341 individuals present in more than 25 language editions of `Wikipedia` called Pantheon 1.0. For larger databases, the fast-growing verification cost currently prevents scientists from considering in a satisfactory way less prominent individuals, who may however have had a significant impact at a more local level. To overcome these limitations, larger-scale Knowledge Bases (KB) such as Yago (Tanon et al. (2020)), DBpedia and `Wikidata` were built automatically from the information found in the Infoboxes and categories of `Wikipedia`.

Such projects represent promising developments for the Social Sciences. A good example is illustrated by Serafinelli and Tabellini (2017) who analyze the historical development of cities and relate it to the emergence of a creative class using Schich et al. (2014) and 40,000 individuals, among which 22,000 creatives that are then matched to a sample of 2,200 European cities between 800 and 1,800. de la Croix and Licandro (2015) similarly built a sample of 300,000 famous people born between Hammurabi's epoch and 1879, Einstein's birth year from Index Bio-bibliographicus Notorum Hominum, to estimate the timing of improvements in longevity and its role on economic growth. Gergaud et al. (2016) use a larger set of 1,243,776 notable individuals such as artists, sportsmen, but also politicians, entrepreneurs, military people, etc. that they match to a broader set of cities. The second version of Pantheon (2.0) in Jara-Figueroa and Hidalgo (2019) also addresses long-run social science trends and requires more statistical power in including around 71,000 with 14+ biographies. The expansion of these databases rely more and more on carefully checked algorithms rather than on manual verification. As an example, a recent independent work by Nekoei and Sinn (2020) gathers 7 million entries using machine learning algorithms, with partial human verifications based on 50 to 308 observations, while our own tests rely on 5,000 observations each verified by three independent research assistants.

The goal of building a comprehensive and accurate Knowledge Base of notable individuals for immediate use in social sciences and econometric and statistical analyses complements existing approaches in several ways. First, we collect a massive amount of data that leads to several cross-verifications. It is based on multiple sources (various editions of `Wikipedia` and `Wikidata`) and deduplication techniques. The combination of `Wikipedia` and `Wikidata` brings 2.72% new birth dates, 8.16% new occupations and 17.16% new citizenships. We find that there are very a few errors in the part of the database that contains the most documented individuals. We also find non trivial error rates (around 1%) in the bottom part of the notability distribution, due to sparse information and classification errors or ambiguity. This either requires manual corrections for future use or a statistical treatment of these

errors in statistical approaches. The combination of `Wikipedia` and `Wikidata` corrects about 0.5% of errors. One therefore needs to trade-off the size of the database and the precision of the data.

Second, we adopt a social science approach: data collection is driven by specific social questions on gender, economic and cultural development and quantitative exploration of cultural trends that we document in this paper. This approach is used in particular to document the Anglo-Saxon bias naturally present in existing projects based on the English edition of `Wikipedia`.

This strategy results in a cross-verified database of 2.3 million unique individuals (we do not recommend to go beyond given the errors in the extended database of 4.7 million people. Note that there is an *exhaustive sample* of more than 6 millions biographies.). We also add a large fraction of individuals from non-English editions of `Wikipedia` who actually played a significant role in important periods of human history. There are more than 700,000 such individuals, almost a third of the database we verified.

The structure is as follows. We provide in **section 2** a framework to think about significant individuals and their role in society. Second, as explained in **section 3**, we adopt a decentralized approach to enlist significant individuals from scraping recursively across `Wikipedia` biographies instead of using centralized registries such as `Freebase`. Third, we extract in **section 4** the information on a *restricted sample* of these individuals when they have a biography in one of the seven major European editions of `Wikipedia` and then cross-verify the information from the various `Wikipedia` editions and `Wikidata`. Fourth, we document in **section 5** the historical trends and compare our approach to existing works based only on biographies present in the English edition. We verify our algorithms in **section 6** taking advantage of the manually-verified information contained in Pantheon 1.0 and with series of manual verifications of random samples, section 7 discusses extensions and applications of the database to economics.

2 A framework for social scientists

2.1 Universe described

In this social science project, we are interested in significant people insofar as they have influenced the trajectory of the society in their times or later on. For instance, Julius Caesar clearly affected its contemporaries, while Vincent van Gogh became famous long after his death. Let's call $\mathfrak{S}$ the universe of these **s**ignificant individuals, without any further discussion of what influence means - it is actually entirely specific to the question social scientists would ask, e.g. the dynamics of the arts, or science, or demography, etc. As the Van Gogh example suggests, some individuals may not have been detected as such at the time they lived, and some significant individuals may not have been detected yet in 2020.

Let us call $\mathfrak{D}$ the universe of already **d**etected individuals, for instance all those individuals above a certain visibility threshold at their time, that would be invariant over time, e.g. the so called 'elite', the top 1/100,000 famous individuals. Some of them will be part of our data collection process, from **W**`ikipedia` and **W**`ikidata` (let us call this set $\mathfrak{W}$), while others are in known sources but **u**nexploited yet, or not yet in the Wiki universe (let us denote it $\mathfrak{U}$).

Some individuals may have been significant and even remembered for a while, but are currently forgotten. We can call this set $\mathfrak{f}$. To sum up, the theoretical universe of significant individuals is decomposed as:

$$\begin{aligned}\mathfrak{S} &= \mathfrak{f} \cup \mathfrak{D} \\ &= \mathfrak{f} \cup \mathfrak{W} \cup \mathfrak{U}\end{aligned} \tag{1}$$

Our project is to try to reach $\mathfrak{S}$ by $\mathfrak{W}$ with automatic routines with two additional bets: first, unexploited archives in $\mathfrak{U}$ will gradually converge to the `Wiki` universe $\mathfrak{W}$; and forgotten individuals in $\mathfrak{f}$ will - albeit slowly - be gradually rediscovered by historians, or scholars in the arts or science disciplines.

2.2 Selection into `Wikipedia`

The selection rules for `Wikipedia` entries are category-specific. They are in particular specific for living persons, and described here: https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons, as well as general guidelines for notability https://en.wikipedia.org/wiki/Wikipedia:Notability_(people). Rules differ marginally across language editions; rules are specific to the type of human activities. It is beyond the scope of this paper to systematically discuss these rules but a few principles emerge: i) one should avoid biographies based on a unique, arbitrary source (see subsection 3.5), as in the universe of `Wikidata` only (no biography in `Wikipedia`), which adds millions of individuals from unverified sources and include homonyms and duplicates; ii) biographies of *living* persons should be used with caution and stricter criteria such as the existence in several language editions should be applied; iii) some categories are more likely to be subject to idiosyncrasies (judgment call) of contributors, in particular those related to family members, criminals, victims of accidents, athletes with no international recognition. These considerations motivate our restrictions on the sample studied throughout this paper.

A last issue affecting selection is the so-called *survival bias*, which states that we only observe the characteristics of the survivors and those could be biased - those present in the dataset but not those who may have had an impact - they would be in the set we called $\mathfrak{f}$. We can, however, approximate the rate at which people survived to make it into the final dataset in Section 4.1, under the assumption that the fraction of notable people affecting society at the time they lived is a constant of the living population at that time, and that they are forgotten at a constant rate per unit of time.[1]

2.3 A social science background

Our database spans over more than 5000 years of history and encompasses a wide range of human activities and domains of influence. We propose a classification inspired by the seminal work of Baumol and Bowen (1966) and that introduced the so-called Baumol's cost disease. We assume that the amount of global wealth Y produced by humans within a year is of two types as follows:

$$Y = Y^R + Y^E \tag{2}$$

[1] This is of course a pure thought experiment but provides an order of magnitude of the number of notable individuals we may still be missing.

Y^R is the amount of materialistic goods and services generated by the commonly-called productive sector, while Y^E is the amount of non-materialistic goods and services created in other sectors of the economy such as arts, leisure, health, sports, etc. The evolution of Y^R reflects fully the improvement in technology and science, while that of Y^E is almost insulated from it. Indeed, the famous Magic flute by Wolfgang Amadeus Mozart still requires the same amount of resources (time, performers, space, etc.) nowadays than in 1791, when it was first performed. Absent any change in preferences, the Baumol effect states that the size of the second sector should constantly decrease over time.

However, the first sector produces essential goods where the second sector produces luxury goods. This implies that the relative demand for the arts would actually increase and this prevents the arts sector to continuously decline and eventually disappear.

Additional notations can help visualize this. Materialistic goods and services are produced by combining a set of inputs $\mathcal{B}$ featuring the economy and business, which, for simplicity, include labor L and capital K. The level of total factor productivity $\mathcal{A}$ combines the various outputs of Science, Education, technology and discoveries of all kinds:

$$Y^R = \mathcal{A}f(\mathcal{B}). \tag{3}$$

Last, the consumption level C is only a fraction α of the amount of global wealth as follows:

$$C = \alpha\mathcal{C}(Y^R, Y^E), \tag{4}$$

α is typically determined by the way a society is governed by politicians, military, religious authorities, noble people: a less efficient society will divert resources for rent-seeking activities (e.g. La Porta et al. (1999). The production function in $\mathcal{C}(.,.)$ is such that the first factor is a necessity good and the second factor has a higher demand as income grows.

3 Data collection for the exhaustive sample: method

3.1 Extracting raw data from `Wikidata` and `Wikipedia`

To build the most comprehensive, reliable database about notable individuals, and organize it in a way that can be later used by social scientists and economic historians, we consider two main sources of information: `Wikidata` and the information contained in seven language editions of `Wikipedia`. In this section, we explain how we extract key information on humans present either in one or both sources, and then how we merge this information to avoid 1) dealing with duplicate individuals and 2) cross-verify the information contained in both universes.

3.1.1 `Wikipedia`

The methodology used here to collect a relevant set of biographies from the `Wikipedia` universe is an extension of Gergaud et al. (2016), with one noticeable difference. Indeed, we do not use here `Freebase`, whose website shutdown in 2016 (Chah, 2017). Instead of using a top-down approach, we use a bottom-up procedure based on `Wikipedia` categories. Categories are commonly used in `Wikipedia` to link articles under a common topic and are found at the bottom of the article page. All existing categories about individuals were systematically analyzed and scraped. Using the relevant urls corresponding to

human biographies, we scraped the 7 following Wikipedia language editions: English, French, German, Italian, Spanish, Portuguese and Swedish that we call hereafter with an abuse of language the *European editions*.[2] See Appendix A for a detailed description of the method.

3.1.2 Wikidata

In the Wikidata universe, we use the "instance of humans" Q5 category to define our sample of individuals in this universe. Information such as first name, family name, gender, country of citizenship, date of birth, date of death (if applicable), domain of influence, as well as the Q code (identifier) of the individual have been retrieved. In addition to this first set of variables, we extracted the exhaustive list of Wikipedia urls contained in each Wikidata entry.

3.2 Merging Wikidata and Wikipedia

Wikidata and Wikipedia may disagree on key information about notable individuals. Considering 7 different Wikipedia editions, instead of one, naturally increases that heterogeneity. In this paper, we use this feature to improve the reliability of each information extracted from both repositories. To do so, we develop and use in what follows a series of algorithms to i) come up with a relevant sample of humans, ii) eliminate duplicate biographies, iii) detect systematic errors contained either in Wikipedia or Wikidata and correct them.

3.2.1 Humans and non-humans

A large number of entries are wrongly identified as humans in Wikidata whereas it contains the Q5 code in Wikidata which stands for 'instance of humans'. The same issue arises in Wikipedia: there are non-humans despite being identified as humans. These entries are either biblical, mythological, fictional characters, animals, music bands, events, murder cases, etc. We use a list of expressions such as "murder of", "list of", "duos", "bands", "attack of", etc. to detect such cases. In total, we identify 20,000 entries corresponding to 16,000 questionable "pseudo-individuals" that we eliminate from our database.

3.2.2 Dealing with multiple biographies

By construction, a given individual has a maximum of eight observations in the database once merged: 7 Wikipedia biographies plus one Wikidata entry. There are several cases to consider to eliminate duplicate observations from this universe of around 6.3 million biographies/entries. The most common situation is one individual present in Wikidata and in at least one of the 7 language editions of Wikipedia with one unique identifier present in both universes. In such situations, we eliminate duplicates by checking whether i) the links present in the Wikipedia section of each Wikidata biography and ii) the Wikidata links retrieved from the "Tools" section of each Wikipedia biography, coincide or

[2]For example, a list of urls of individuals who died in 1953 can be found here: https://en.Wikipedia.org/wiki/Category:1953_deaths. The corresponding urls in the French (resp. Portuguese, Spanish, Italian, German and Swedish) edition were accessed by using "fr" (resp. "pt", "es", "it", "de", "sv") instead of "en" in all urls. In the particular case of France, Wikipedia sorted individuals by month-year of birth and death, so the loop for scraping individual biographies was adjusted accordingly to cope with this monthly frequency.

not.[3]

3.2.3 Treating duplicates

In addition, a given individual may have more than one biography either a) within the same `Wikipedia` edition or b) across different editions without any explicit link between these duplicate biographies or c) in `Wikidata` and/or `Wikipedia` under different Q codes (identifiers). It is indeed very likely that several biographies coexist for a given individual as the number of contributors in that community is high and the information widespread. In this context, some contributors may fail to detect (or just forget to check) the work done by others previously about the individual they are interested in. For example, Sarendy Vong has two separate biographies in the English edition: https://en.Wikipedia.org/wiki/Sarendy_Vong and https://en.Wikipedia.org/wiki/Vong_Sarendy. Both contributors ignored one another when creating their biography. Appendix A.1 provides further details about this methodology. As a result we remove 0.7% of individuals (34,562/4,678,040). Although we cannot claim we eliminated all duplicates, we adopted a conservative approach in multiplying the number of tests we have run and rather eliminated false duplicates than keeping true duplicates.

3.3 Exhaustive sample: descriptive statistics

3.3.1 `Wikipedia`: comparing language editions

The 5 most popular language editions in `Wikipedia` in number of biographies are, in decreasing order, 1) English, 2) German, 3) Japanese, 4) French and 5) Russian. Considering the German edition after English adds 340,913 individuals (of which 259,013 individuals have one unique biography in German). It is worth noting here that these 340,913 individuals would have been ignored without considering the German edition. Using next the Japanese edition brings 179,466 new individuals. Following the same logic of recursive elimination, we add 155,391 undetected individuals with the French edition, and 156,728 individuals with the Russian edition. The full ranking is available in Table 7 in Appendix A.4 and provides a better sense of the relative importance of each language edition.

3.3.2 Main sample characteristics

Overall, we gather information about 4,678,040 distinct notable individuals out of 6,291,767 biographies extracted from 7 different `Wikipedia` language editions and `Wikidata`. There are 2,291,817 individuals (49%) with at least one biography in `Wikipedia` along with a `Wikidata` entry. There are 2,386,223 notable individuals (51%) with a `Wikidata` entry but no existence in the 7 European `Wikipedia` language editions.

Table 1 collapses these famous individuals in 6 groups to help us figure out the respective contributions of both `Wikidata` and `Wikipedia`. In subsequent columns, we display similar statistics for different profiles of individuals according to their popularity in `Wikipedia`: at least 1 biography among the 7 European language editions (described above) and at least 14 language editions (Pantheon 2.0). Each row corresponds to a specific language group: *English* (individuals with at least one biography

[3] In some rare cases, the number of editions varies from one source to another, for the same individual, or two different individuals are associated to the same `Wikidata` code. Manual modifications are then necessary to get rid of these confounding observations. Further details about these different steps are available from authors upon request.

in the English edition), *Western* (individuals with one biography in a language part of the Western world but not English), *Eastern* (individuals with an edition that is part of the eastern world but not English/Western), *Eurasia-Arabia* (individuals with an edition that is part of the Eurasia-Arabia world but not English/Western/Eastern), *Southern and natives* (individuals with an edition that is part of the Southern/natives world but not English/Western/Eastern/Eurasia-Arabia), and `Wikidata` only (individuals with one `Wikidata` entry but no existence in the `Wikipedia` universe).

The exhaustive sample (Table 1, column 1 contains 1,578,917 individuals in the first block (English), 1,327,543 in the second block (Western non-English). The third, fourth and fifth profiles correspond to 342,783; 223,213 and 4,249 notables from the Eastern, Eurasia-Arabia and Southern/natives worlds respectively. The last group which is made of individuals with one `Wikidata` entry only represents around 1/4th of individuals.

The information present in existing databases are based on the English edition of `Wikipedia` only (Schich et al. (2014) and Yu et al. (2016)). This project, by considering other language editions and `Wikidata`, almost triples the universe of notable individuals as follows: Western languages (+27.0%), Eastern languages (+7.3%), Eurasian-Asian languages (+4.8%) and Southern & native languages (0.1%), `Wikidata` (+25.7%). In adding these new language editions, the total number of newly introduced individuals is sizeable: 3,099,123. Examples of the most famous individuals present in English are Barack Obama, Cristiano Ronaldo or Albert Einstein. Other language blocks naturally include lesser known individuals like French film director Olivier Nakache (*Western*), Japanese actor Ryō Iwamatsu (*Eastern*), scientist Qayum Nasıyri (*Eurasia-Arabia*), former writer Boerneef (*Southern & natives*) and art historian Martin Hardie in `Wikidata` *only*. Table 8 in Appendix shows the 5 most famous individuals in each recursive language blocks. These individuals have achieved a lower level of international recognition but may have had a significant impact at the local level, in their own country, region or town. In the rest of the paper, we explore how these "newly added" individuals differ from those present in the English recursive block on a few dimensions such as year of birth/death, gender, occupation or citizenship.

3.4 Measuring notability

The large number of individuals in the database is the main distinctive feature of our work. Some of them, such as Barack Obama, have achieved international recognition over the course of their life. while most of these notable individuals are celebrated more locally, in their home country most of the time. Illustrations of this are Louis Bérard, chief engineer in charge of the construction of the Canal Saint-Martin in Paris and Georg Christian Kessler, famous and influential German entrepreneur in the 18th century who had a major influence in the wine industry both in France and Germany. None of them had a biography in English in 2018 but Bérard has a biography in French and Kessler one in German that are included in the current database. Yet, these less-known individuals may have contributed in a significant way to the social, economic or cultural development of their community (city, region or country). It is therefore important to document such individual notability profiles as well. Research will compare the relative contribution of each profile on economic variables such as city growth, GDP, etc.

To accurately measure the relative influence of each individual present in our database, we build a

Table 1: **Marginal contribution of different blocks of `Wikipedia` language editions**

	All `Wikipedia` & `Wikidata`		At least one European edition		At least 14 editions	
	Freq.	%	Freq.	%	Freq.	%
	Exhaustive sample (section 3)		**Restricted sample** (sections 4 and 5)		Pantheon 2.0	
`Wikipedia` *editions (by recursive language blocks)*						
English (En)	1,578,917	33.8	**1,547,174**	67.5	76,139	99.7
Western (We)	1,327,543	28.4	**744,643**	32.5	240	0.3
Eastern (Ea)	342,783	7.3	0	0.0	0	0.0
Eurasia - Arabia (EuAr)	223,213	4.8	0	0.0	0	0.0
Southern & natives (Sn)	4,249	0.1	0	0.0	0	0.0
`Wikidata` *only*	1,201,335	25.7	0	0.0	0	0.0
Total	4,678,040	100	**2,291,817**	100	76,379	100

Notes. English (*En*): individuals present in the English edition; Western (*We*): individuals absent from the *En* edition but present in *We* editions; Eastern (*Ea*): individuals absent from the *En* & *We* blocks but with at least one biography in editions of the *Ea* block; Eurasia-Asia (*EuAr*): individuals absent from the previous blocks (*En*, *We*, *Ea*) but present in at least one *EuAr* edition. Southern & natives (*Sn*): individuals absent from the other blocks (*En*, *We*, *Ea*, *EuAr*) but present in at least one edition of the *Sn* block. `Wikidata` *only* includes individuals with a `Wikidata` biography only. In columns, we display the total headcount figures (number and share) in various samples: all individuals regardless of their `Wikipedia` profile (exhaustive sample); individuals with at least 1 biography among the 7 European editions; individuals with 14+ biographies in the `Wikipedia` universe.

synthetic notability index using five dimensions:

1. the number of `Wikipedia` editions of each individual;
2. the length, i.e total number of words, found in all available biographies that we collected.[4] It sums to zero in the total absence of a `Wikipedia` biography;
3. the average number of biography views (hits) for each individual between 2015 and 2018 in all available language editions[5] or zero in the absence of a `Wikipedia` biography;
4. the number of items retrieved from `Wikipedia` or `Wikidata` for birth date, gender and domain of influence. The intuition here is that the more famous, notable, the individual, the more documented his/her biographies;
5. the total number of external links (sources, references, etc.) from `Wikidata`.

We then determine the quantile values from each dimension and add them all to define our notability measure that is used to compare/rank individuals over time and across space. Arguably, it is harder to compare individuals born in different centuries and we will not present any of these comparisons, since it favors individuals born most recently.

Table 2 shows the 10 most renowned individuals per big period and gender. Other projects, including Yu et al. (2016), use the number of biographies in `Wikipedia` to determine such rankings. Given that we have to deal with many individuals present in only one edition, we use other metrics such as biography length and biography views to figure out a ranking out of this enlarged set of individuals.

[4]We consider here the 7 European language editions that we parsed. In case an individual does not have a biography in, say Swedish, the number of words for that edition is set equal to zero.

[5]We used for that task the following API available in https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews.

Table 2: **Most famous historical figures: breakdown by period of death - or contemporaneous - and by gender**

Death (AD)	*Women*
Before 500	Catherine of Alexandria, Saint Cecilia, Cleopatra, Helena (empress), Hypatia, Livia, Saint Lucy, Messalina, Sappho, Agrippina the Younger
501-1500	Eleanor of Aquitaine, Joan of Arc, Clare of Assisi, Hildegard of Bingen, Mary of Burgundy, Isabella of France, Elizabeth of Hungary, Murasaki Shikibu, Catherine of Siena, Bridget of Sweden
1501-1750	Anne, Queen of Great Britain, Catherine of Aragon, Anne Boleyn, Elizabeth I of England, Lady Jane Grey, Isabella I of Castile, Mary I of England, Mary, Queen of Scots, Catherine de' Medici, Teresa of Ávila
1751-1900	Marie Antoinette, Jane Austen, Emily Dickinson, Catherine the Great, Thérèse of Lisieux, Ada Lovelace, George Sand, Mary Shelley, Maria Theresa, Mary Wollstonecraft
1901-1979	Agatha Christie, Marie Curie, Anne Frank, Judy Garland, Frida Kahlo, Marilyn Monroe, Florence Nightingale, Édith Piaf, Queen Victoria, Virginia Woolf
1980-2019	Diana, Princess of Wales, Marlene Dietrich, Audrey Hepburn, Katharine Hepburn, Whitney Houston, Grace Kelly, Elizabeth Taylor, Mother Teresa, Margaret Thatcher, Amy Winehouse
Alive	Beyoncé, Hillary Clinton, Lady Gaga, Selena Gomez, Angelina Jolie, Madonna (entertainer), Rihanna, J. K. Rowling, Britney Spears, Meryl Streep

Death (AD)	*Men*
Before 500	Paul the Apostle, Aristotle, Marcus Aurelius, Julius Caesar, Cicero, Alexander the Great, Augustine of Hippo, Jesus, Plato, Socrates
501-1500	Dante Alighieri, Thomas Aquinas, Francis of Assisi, Avicenna, Charlemagne, Genghis Khan, Muhammad, Petrarch, Marco Polo, Rumi
1501-1750	Johann Sebastian Bach, Christopher Columbus, René Descartes, Galileo Galilei, Henry VIII of England, Martin Luther, Michelangelo, Isaac Newton, William Shakespeare, Leonardo da Vinci
1751-1900	Ludwig van Beethoven, Charles Darwin, Vincent van Gogh, Victor Hugo, Abraham Lincoln, Karl Marx, Wolfgang Amadeus Mozart, Napoleon, Friedrich Nietzsche, George Washington
1901-1979	Charlie Chaplin, Winston Churchill, Albert Einstein, Sigmund Freud, Mahatma Gandhi, Che Guevara, Adolf Hitler, John Lennon, Pablo Picasso, Elvis Presley
1980-2019	Muhammad Ali, David Bowie, Fidel Castro, Salvador Dalí, Stephen Hawking, Michael Jackson, Nelson Mandela, Ronald Reagan, Frank Sinatra, Andy Warhol
Alive	Woody Allen, George W. Bush, Bill Clinton, Bob Dylan, Eminem, Paul McCartney, Lionel Messi, Barack Obama, Arnold Schwarzenegger, Donald Trump

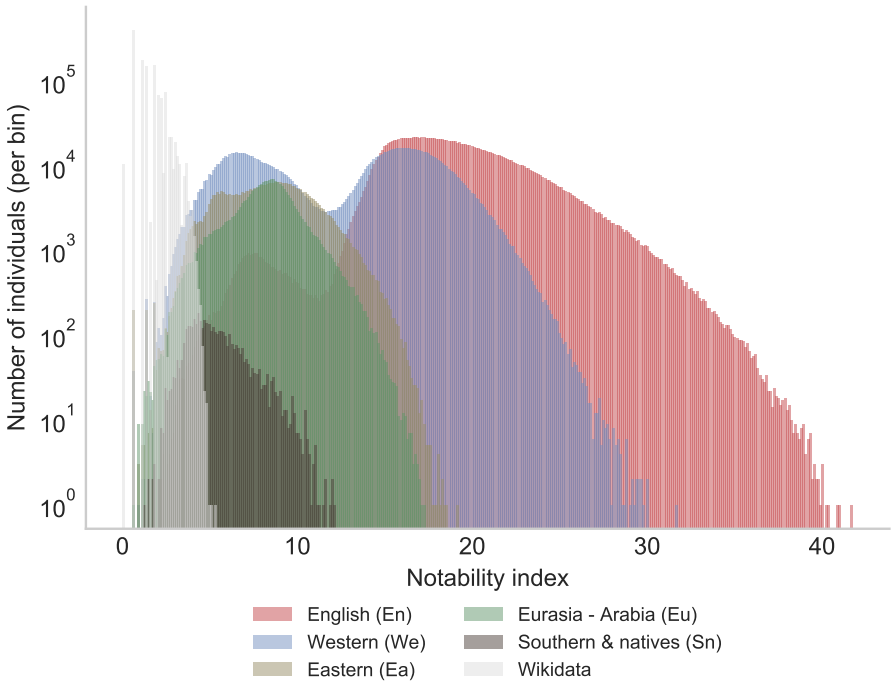
Notes. 10 most famous historical figures by period (of death year or contemporaneous) and gender. The ranking is computed from a notability index described in the text.

Figure 1: **Cloud of the most famous individuals in the database**

Notes. Size is proportional to relative notability level. The cloud focuses on the 3,000 most visible individuals (0.06% of the exhaustive sample). Colors represent the domain of influence defined later in the text (see Figures 5 and 6 for labels, e.g. green is culture, red is politics, blue is academia, etc.).

Figure 2 shows the dispersion across individuals in this notability index. The chart shows the distribution by recursive language block. With no surprise, the English edition contains relatively more visible individuals than the other language blocks. For the Western language block, the distribution is bi-modal, both peaks being quite low in terms of notability with respect to the overall distribution of notability and dominated by individuals from the English edition. For the other language blocks, the notability index is even lower.

Figure 2: **Density distribution of notability by language blocks**



Notes. Exhaustive sample (4.7 million individuals). English (*En*): individuals present in the English edition; Western (*We*): individuals absent from the *En* edition but present in *We* editions; Eastern (*Ea*): individuals absent from the *En* & *We* blocks but with at least one biography in editions of the *Ea* block; Eurasia-Asia (*EuAr*): individuals absent from the previous blocks (*En*, *We*, *Ea*) but present in at least one *EuAr* edition. Southern & natives (*Sn*): individuals absent from the other blocks (*En*, *We*, *Ea*, *EuAr*) but present in at least one edition of the *Sn* block. `Wikidata` **Only** includes individuals with a `Wikidata` biography only.

3.5 Sources of information

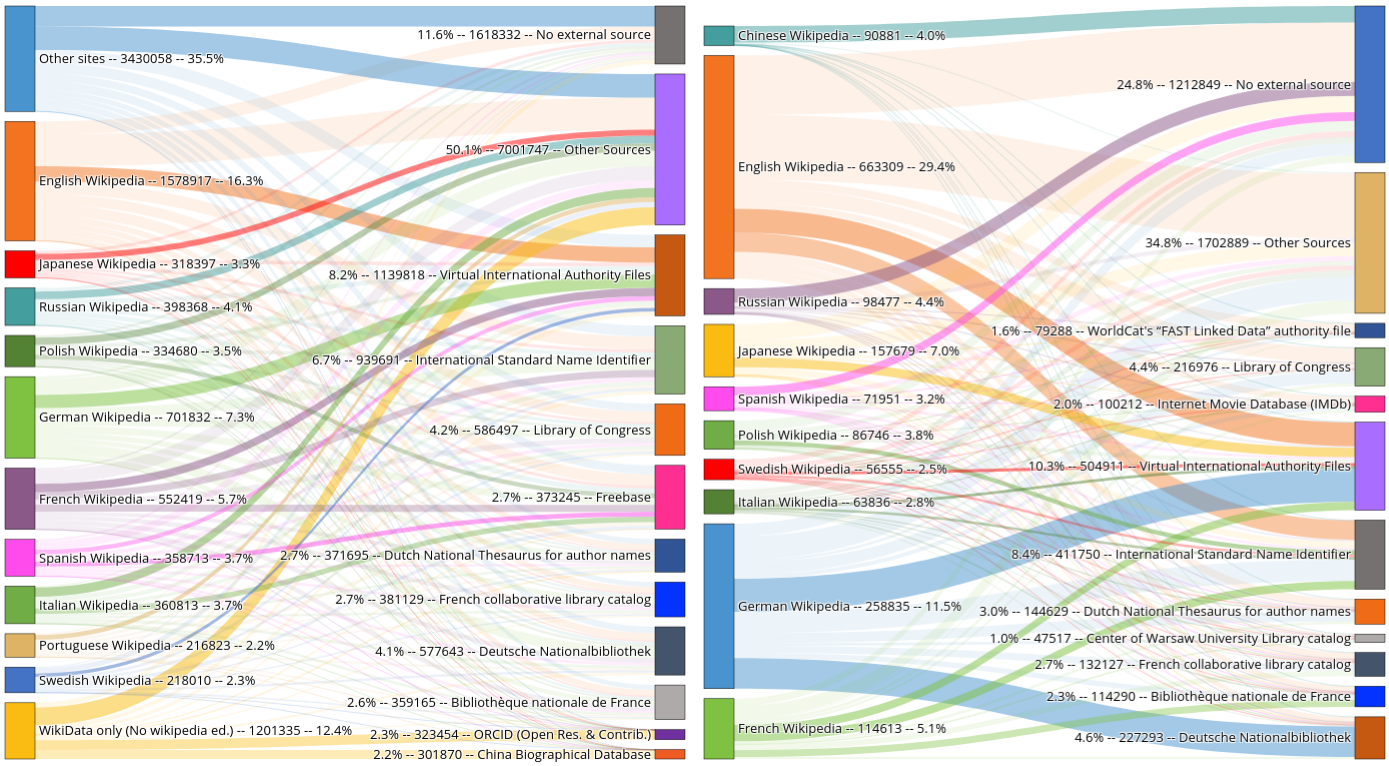
The restricted sample is clearly dominated by the German edition which is derived from two main sources: i) VIAF and ii) the Deutsche National Bibliothek. The latter source is not intensively used in the other `Wikipedia` editions, while the former is often mentioned in the main `Wikipedia`. One learns from these graphs that the efforts to inflate the number of biographies in some editions would require the construction and development of national repositories.

Figure 3 represents the correspondence between editions and source of information related to individuals. The left panel represents the exhaustive sample while the right panel is restricted to individuals with one unique biography in the `Wikipedia` universe. Only a small fraction of biographies (11.6%) does not report any specific source. In the matter, the most frequent sourced mentioned are, in decreasing order: the Virtual International Authority File (VIAF)[6], Freebase, the Deutsche National Bibliothek, followed by two French sources (Bibliotheque Nationale de France and a French collaborative library catalogue). One can see that the large number of German individuals stems from two main sources, the

[6]The VIAF is a joint project of several national libraries operated by the Online Computer Library Center (OCLC).

VIAF and the Deutsche National Bibliothek. The latter source does not provide many links to other `Wikipedia` editions, while the former also bring individuals with a single biography in English.

Figure 3: **Relation between the most frequent `Wikipedia` editions and sources**



Notes. Exhaustive sample (left panel) and sample restricted to individuals with only one biography in `Wikipedia` (right panel). We consider the 10 most frequent external sources and 10 most popular `Wikipedia` language editions, and merge the rest into other sources.

4 Extracting information for a restricted sample

In this section, we use a restricted sample of cross-verified biographies with at least one page in the European language edition of `Wikipedia`. This allows for a systematic comparison with `Wikidata` for the following demographic characteristics: birth/death dates, gender, domain of influence and citizenship. Therefore we can cross-verify the respective information and detect systematic errors contained either in `Wikipedia` or in `Wikidata` and correct them.

We use this sample to highlight the difference between the English edition (used in existing projects) and the other editions and measure a possible Anglo-Saxon bias. As displayed in column 2 of Table 1, this sample contains information about 2,291,817 notable individuals with at least one biography among the universe of `Wikipedia` editions we considered (7 European language editions enumerated above) plus a `Wikidata` entry.

4.1 Demographics: birth, death and gender

We consider all personal (he/she) and possessive (his/her) pronouns present in the first part of each available `Wikipedia` biography to figure out a gender for each individual. In case both masculine and feminine pronouns are detected, we use the one that appears first in the biography to determine the individual's gender. The latter gender information, based on the most frequent type of pronouns found in `Wikipedia`, is then compared with that extracted from `Wikidata` when reported. We keep the gender information present in `Wikidata` when both sources contradict each other. In case the information is missing in one universe, we use the gender found in the other source when available.

Regarding dates, when the birth and death information is present in both sources, we determine first the most frequent year found in the different (up to 7) language editions of `Wikipedia` for the individual, and compare it with the year extracted from `Wikidata` when reported. In case they are different, we give more credit to the information coming from `Wikidata`. The reason being that the information has been processed by someone when creating the `Wikidata` biography. For a significant number of individuals especially from ancient times, the exact year is not available, we use the century, millennium, circa or decade information when available to estimate it. We build the relevant time intervals and use the middle of the interval as a proxy for birth/death year. In addition, there are a few cases of super centenarians, allegedly known as having lived more than 120 years.[7] Overall, the exact date of birth (death) is known for more than 90% of cases (see Table 4 for exact numbers, and we are able to impute 4% of new birth dates and 14% additional death dates.[8] Table 6 in Appendix A.3 reports the list of the eldest people in the exhaustive database.

We next calculate the number of living individuals in a given year in the database. This is represented on Figure 4. There is an exponential evolution of the sample size over time. However, the ratio of the number of notable individuals in the restricted sample over the world population (estimates from Kremer (1993) and Manning (2008)) is not constant: as illustrated in Appendix Figure 18, the ratio of the number of famous people alive at a given point in time to the world population increased from 500BC to 1950. The ratio was around 1 over 250,000 in the Antiquity and is 1 over 3,000 in 1950. Under the assumption that there is a constant fraction of "famous individuals" throughout history, one can calculate the rate at which these were forgotten. Details of the calculation can be found in Appendix A.3. Under the assumption that people are forgotten at rate bt, the fraction of famous people forgotten relative to total population is $1-(1-b)^T$ where T is distance to present, which is 15.2% each century, or 56.1% after 500 years, or 80,8 % after 1000 years. Therefore, it is likely that a large number of individuals who should have been remembered given their achievement are currently not listed in the database, again under the assumption that the detection threshold of the recent period should have been applied since the start of our sample.

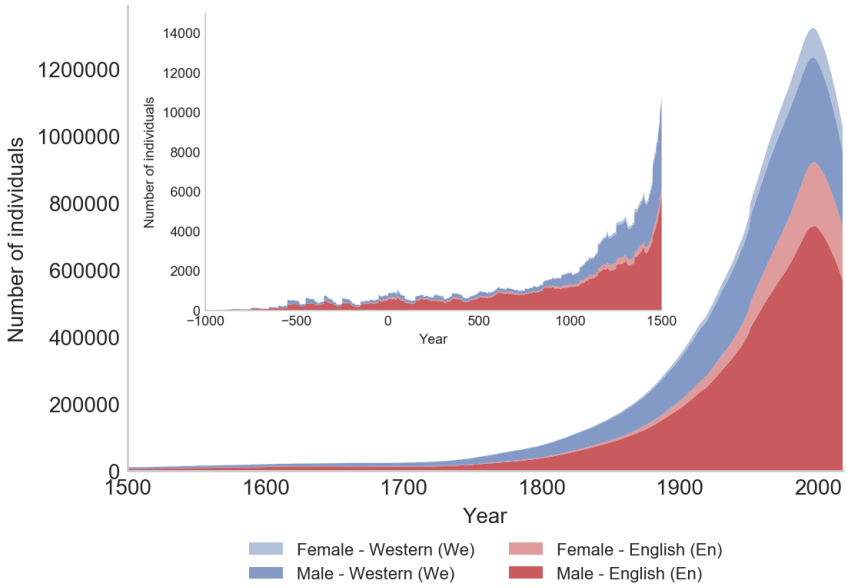
Figure 4 also shows that the structure of the language blocks in `Wikipedia` (English vs. Western blocks). In Appendix (Fig. 19), we also show a similar evolution for Eastern, Eurasia-Arabia, Southern

[7] We verified manually, and when not an error from our algorithm, kept them all but caution should be exercised. The only case where we deleted the birth date is Trailanga (1607-1887).

[8] When the information is available for either birth or death dates, we estimate longevity for the time period, gender, domain of influence and region, and predict the missing date of birth or death based on estimated longevity. When we have no information on both birth and death dates, no imputation is possible and we exclude individuals from all graphs with a time dimension, although many of them are from the 20th and 21st century.

& natives language blocks from the exhaustive dataset.

Figure 4: **Time evolution of the number of individuals present in the database in a given year**



Notes. Restricted database (at least one `Wikipedia` edition among the 7 European languages analyzed), imputed lifetime if missing birth or death. **English (*En*)** language blocks include individuals with at least one biography in English in `Wikipedia`; **Western non-English (*We*)** includes individuals with a `Wikipedia` biography in at least one of the Western languages but absent from *En*. See Table 1 for precise definitions of these groups and sub-groups. Individuals with more than one biography account for one observation to avoid double counts.

4.2 Domains of influence and occupations

We proceed the same way as for dates, and systematically compare the information collected from `Wikipedia` and `Wikidata` to figure out the domain of influence, in short, either the domain of influence, the sector of activity or the field where notable individuals are meant to have had an influence. We use a frequentist approach and use probabilities to determine in which domain an individual had its main influence.

Extraction procedure

In `Wikipedia`, keywords related to the domain of influence are found in the first part of most biographies after verbal groups such as "was a"/"is a"/"was the"/"is the". We first parse the English edition to detect keywords in a list of 1911 occupations and select the first three keywords. In most cases, these correspond to a well-referenced occupation such as pianist, engineer, politician, general, etc. To give an example, we collected three different keywords for Ray Charles from the following part of sentence: "*Ray Charles was an American singer, songwriter, musician*". We also consider the other language editions using the set of keywords extracted from the English edition translated into French, Spanish, Italian, Swedish, German and Portuguese. The maximum number of keywords collected for a single individual is 21 (3

keywords per edition). We did not find any keywords for 44,808 individuals (2% of individuals with a `Wikipedia` biography).

`Wikidata` also contains information about domains of influence. On average, we find 1.3 occupation per `Wikidata` entry; 99.5% individuals have less than 6 occupations. Most of the time, the number of reported occupations is one. In this universe, Ray Charles: `Wikidata` is classified as *musician, singer, composer, pianist, singer-songwriter, saxophonist, vocalist, arrangement, jazz musician*. Following this model discussed in section 2, we group the list of 1911 identified keywords in five large categories of occupations, and a sixth residual category. We also split the categories into sub-categories as follows (in parenthesis we report how they match the theoretical concepts of section 2):

- **Discovery/Science** *(contributes mostly to total factor productivity $\mathcal{A}$)*:
  - Academia (Research, Historian, Physician, Scientist, Academic, etc.)
  - Explorer (Engineer, Explorer, Inventor, Sailor, Pioneer, etc.)
- **Culture** *(contributes mostly to immaterial goods and services Y^E)*
  - Core (Actor, Writer, Painter, Singer, Music, etc.)
  - Periphery (Journalist, Architect, Model, Designer, Presenter, etc.) *(may also contribute to business/econ inputs $\mathcal{B}$)*
- **Leadership** *(contributes mostly to the organization of society and rent-seeking activities α)*
  - Politics (Politician, Activist, Revolutionary, Trade unionist, Minister, etc.)
  - Military (Military, Officer, Commander, Soldier, Army, etc.)
  - Law (Lawyer, Diplomat, Judge, Jurist, Civil service)
  - Nobility (Aristocrat, Noble, King, Sovereign, Monarch, etc.)
  - Religious (Priest, Prelate, Rabbi, Missionary, Bishop, etc.)
  - Corporate Leadership (Business, Entrepreneur, Bank, Merchant, Manager, etc.) *(may also contribute to $\mathcal{B}$)*
- **Sports/Games** (Football, Player, Sport, Baseball, Basket, etc.) *(contributes mostly to Y^E but also to $\mathcal{B}$)*
- **Other**
  - Worker (Farmer, Librarian, Musher, Bookseller, Printer, etc.) *(contribute mostly to $\mathcal{B}$)*
  - Family (Son, Daughter, Child, Wife of, Father, etc.)
  - Misc. (Esperantist, Criminal, Convict, Killer, Philanthropist, etc.)

The most frequent generic keywords collected for each individual are shown in parenthesis. The Discovery/Science category includes individuals with a scientific and/or educational background. In Culture, we find people celebrated for their achievements in the core cultural industries (art, music, performing and visual arts, film, museums, gallery and photography) and/or some more peripheral cultural industries (TV, radio, media, sound, advertising, architecture, design and fashion) as suggested by Throsby (2008). In Leadership, we find either famous politicians, military or religious figures, lawyers and nobles, to which we add entrepreneurs although in the theory sketched above, they are identified as contributing to inputs $\mathcal{X}$. The Other category includes some family-related roles or honorific titles, that are not occupations (son/daughter of, wife/husband of, father/mother of, brother/wister of, etc.). It also includes activities related to smaller businesses such as farmer, bookseller, printer, chefs, craftsmanship and more exotic ones such as, criminals, philanthropists, victims of crime or accidents, etc. that came from the scraping `Wikipedia` biographies.

Allocation into domains of influence

First, we determine the most recurring sub-category (mode) from either `Wikipedia` and `Wikidata` and compare outcomes from both sources. The sub-category detected is consistent in a vast majority of cases.[9] In case the information is missing in one universe, we use the sub-category found in the other source when available. When sub-categories are available in both sources, we give a preference to `Wikidata`, under the assumption that its information is more structured and less subject to errors. Moreover, we report a second domain of influence (the second most frequent one), based on the full list of domains identified from `Wikidata` and all `Wikipedia` pages. We set a threshold of 25% for keeping this second occupation, rationalized after a pilot test which minimizes the number of errors under the constraint of preserving a fair amount of true positives.[10] A good illustration is Napoleon Bonaparte who is referenced in two main domains: "Politics" and also "Military". Another example is Ronald Reagan, famous first for his prominent role in American Politics in the 80's and also known as an actor.

Figure 5 describes the relative importance of domains of influence in the restricted sample described above. Interestingly, the four most popular domains are, in decreasing order: Culture (30.6%), Sports/Games (27.7%), Leadership (27.0%), Discovery/Science (11.9%) and Other (2.1%). People that are most remembered in `Wikipedia` have specialized in sectors producing non-materialistic goods and services and thus contribute to Y^E and much less to Y^R. In the Culture category, the vast majority of individuals with a biography had or still have a career in one of the many core cultural industries (26.4%) and more sporadically in the peripheral cultural industries (4.2%). The Leadership category is dominated by Politicians (14.2%), Military (3.2%), Lawyers (3.2%), Religious faces (3.1%), Corporate leaders (2%) and Nobles (1.7%). The academic sector (11.1%) dominates the Discovery/Science category. Explorers correspond to a minor part of this category (1.1%).

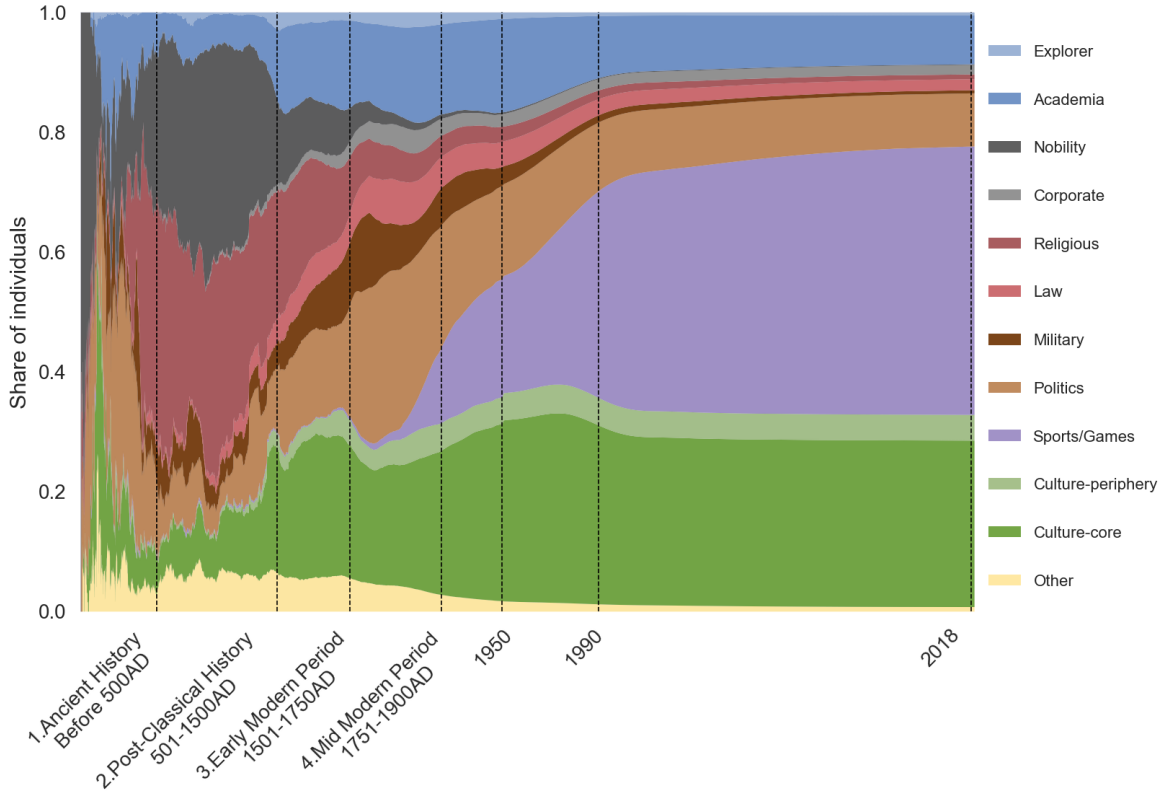
[9] See for instance Table 5 in subsection 6.4 for the systematic comparison of sources.

[10] See the Appendix for more details.

Figure 5: **Sunburst: relative importance of the main occupations and domains of influence**

Notes. Restricted database (at least one `Wikipedia` edition among the 7 European languages analyzed).

Figure 6: **Share of individuals present in the database, breakdown by domain of influence**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Imputed life time when missing.

Figure 6 shows the evolution of the share of these different domains over time. We see clearly on this graph the prominent role played by religious figures and the military until the beginning of the 15th century. Subsequent periods show the emergence of the cultural class and the political sector. One last remark is the rapid rise in the 20th century then the dominant role of the Sports sector after 1950.

4.3 Citizenship

For citizenships, we collect and then compare the information available from both `Wikipedia` and `Wikidata`. In `Wikipedia`, we may collect up to two citizenships per language edition, based on the current name of the country therefore summing to a maximum of 14 possible citizenships, obviously generally very similar across language editions. We then keep the two most frequent citizenships across all language editions we parsed, and then compare them to `Wikidata`. The comparison group in `Wikidata` is a long list (up to 10 possible citizenships at a very detailed level). We first aggregate them into meaningful entities that are not necessarily time-invariant: e.g. entities such as "Allemagne", "Berlin", etc. are clubbed under Germany, "Ting Dynasty", "Ming Dynasty", "Song Dynasty", etc. are clubbed under Old regimes in China, "Russian Empire", "Tsardom of Russia", "Grand Duchy of Moscow", etc. are clubbed under Old Russia, "First French Republic", "Second French Republic", "Third French

Republic", etc. are clubbed under France. For individuals living in the modern state, we retain the name of the entity as is, for e.g. for a person who lives in the current state of China, the regime mentioned is China. We also keep three empires in these aggregate groups (Holy Roman Empire, Roman Empire and Soviet Union), given the size and geographical expanse of these political entities (it would be impossible to associate them with a single modern country).[11] We then aggregate these entities over time to match efficiently with `Wikipedia`, e.g. Old Regimes in China, Ancient China, People's Republic of China, etc. all get clubbed under China.

In 95% of cases, `Wikidata` and `Wikipedia` gave the same citizenship. When they instead contradict each other, we adopt a decision rule detailed in Appendix B.2 giving higher priority to information present in the Infoboxes from `Wikipedia` and `Wikidata`. To classify individuals under the old regimes v.s. the current state, we exploit the information on acquisition of sovereignty of modern state, details of which are provided in the Appendix. We use information available on the collapse of empires such as the Holy Roman Empire or the Soviet Union to correctly classify individuals into these supranational entities. We also treat the case of new nation states that emerged after the collapse of these empires, based on the disaggregated information from either `Wikipedia` and `Wikidata`. Finally, a fraction of individuals may have several citizenships and we report two of them if relevant.

5 Documenting key historical periods

Here, we show that the additional editions we analyzed beyond English provide a broader and finer coverage of some key periods in human history. We now illustrate how the inclusion of more language editions is useful.

5.1 Politics and the emergence of modern democracy

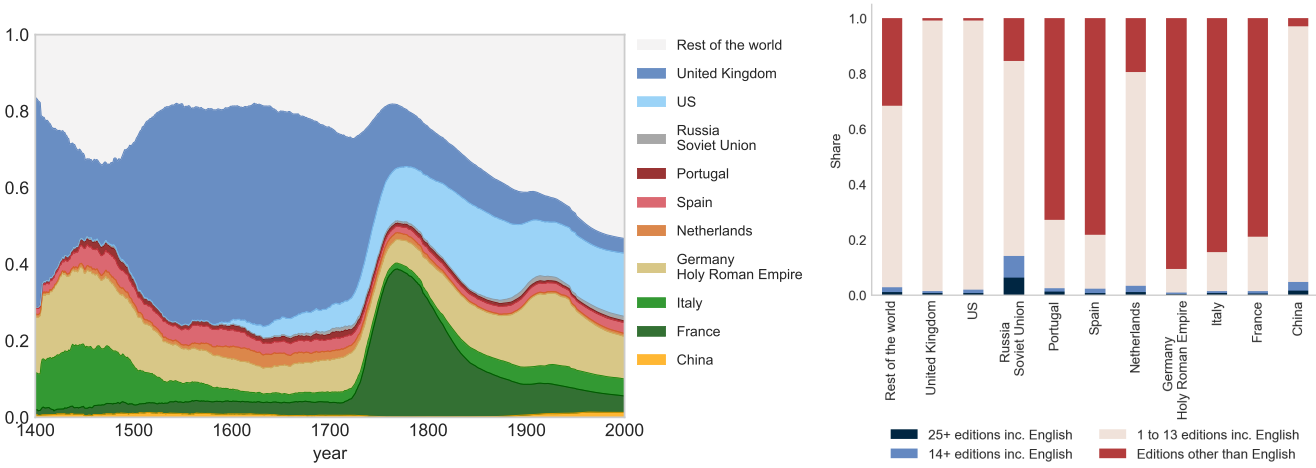
Figure 7 illustrates the rise of American politics in the second half of the 18th during the creation of the new independent nation of the USA, as shown in the left panel of Figure 7. The same holds true for France: at that time, the share of French and US politicians in our database is rising in a significant way around and after the end of the 18th century. While the emergence of modern democracy in those countries is noticeable on the graph, the main feature here is the dark blue part in the 16th century in British Islands, which saw the gradual integration of the four kingdoms of England, Wales, Scotland and Ireland. Many historians consider that central elements of modernity, of capitalism, industry, and science, originated in early modern Britain (circa 1500-1750).[12] In the right panel, we display the relative importance of different language editions other than English - see e.g. the importance of red bars featuring individuals not present in English `Wikipedia`. This is particularly large for French individuals (80% individuals), and true for other citizenships including Italy (+70%), Germany (+85%), Portugal (+70%) and Spain (+65%).[13]

[11] In some rare cases, we skipped extremely unfrequent citizenships such as Chenla Kingdom, Sultanate of Hobyo, Lordship of Carpi etc. that are reported as missing citizenships. With this procedure, we have aggregated 99.996% of cases.

[12] E.g. Adrian Johns from the University of Chicago, see `http://home.uchicago.edu/~johns/earlymodernbritain.htm`

[13] For instance, the French edition of `Wikipedia` contains a set of biographies for local historical figures such as delegates of the Estates General, an assembly that played a central role during that troubled period of time in France. Auguste-Louis-Dominique Delpech (French, 1818 - 1880, medical doctor and politician) is one of them. This scientist and practitioner

Figure 7: **Evolution of the number of individuals associated with "politics", 1400-2000AD**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: `Wikipedia` profile: share of individuals with biographies in a) 25+ editions including English, b) 14+ editions including English, c) 1 to 13 editions including English, d) One or more editions than English), breakdown by citizenship.

5.2 The age of exploration and discoveries

Figure 8 (left panel) shows the share of individuals in the *Explorer Inventor Developer* category by citizenship. It illustrates the end of the chinese exploration period which lasted until the 15th century followed by the European age of discovery and explorations conducted by Portugal and Spain in the 15th and 16th centuries.

These two European waves of exploration are imperfectly covered by the English edition. In the right panel, the fraction of these individuals (red bar) not included in the reference edition ranges between 10% and 15% for Portugal and Spain, respectively. The Chinese exploration benefits from a better coverage in the English edition. The contribution of the present project is even more striking when one considers the share of individuals not present in existing projects (pink and red bars). At the margin, we add almost 75% and 80% of the Portuguese and Spanish individuals of the category here. Here again, many significant individuals have been added as shown in Appendix B.3.

The dark green peak on the graph corresponds to the French Renaissance period (between the 15th and early 17th centuries). This important period of human history is not covered in a satisfactory way in the English edition again as well as in existing projects. The share of individuals retrieved with the

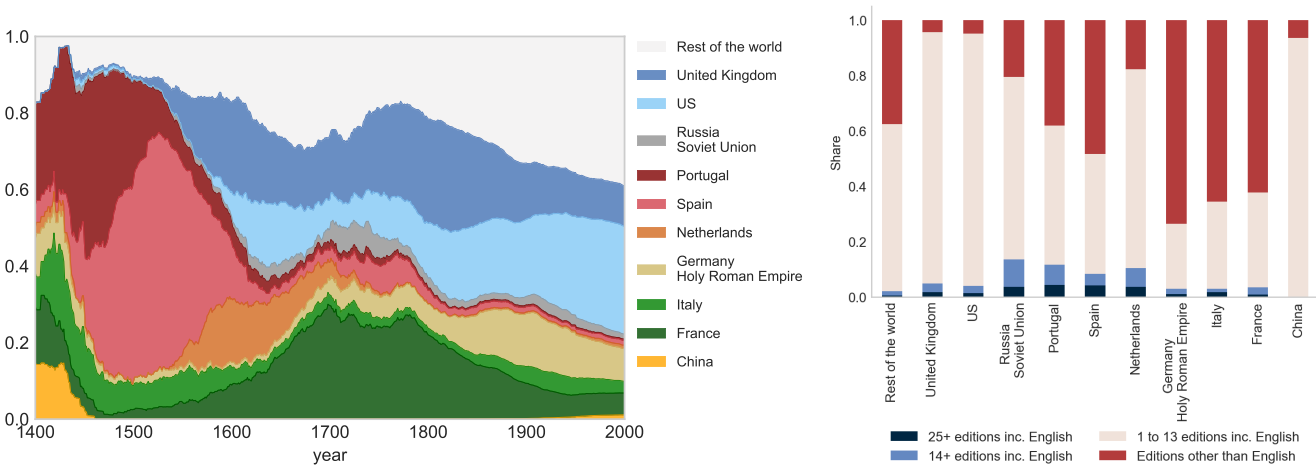
published extensively on epidemics in general and on Cholera in particular; a work for which he received the Montyon Prize among other prestigious distinctions. He became member of the French "Imperial Academy of Medecine" in 1864 for his contribution to the medical field, a noticeable distinction. Later in the career, he became a politician and played a role in the political life of Paris. To date, this French notable individual has three biographies in `Wikipedia` (Danish, French and Swedish) but no existence in the English edition. Including him in our database might contribute to disseminate his knowledge and contribution. See more discussion in section B.3.

French citizenship in editions other than English is at around 25% (red bar). If we add to this the fraction of individuals ignored so far in existing databases (65%) we can conclude that we improve the coverage by almost 90%.

A similar pattern is observed for Italians earlier during the Italian renaissance period which spans from the end of the 14th century (Trecento) to the early 16th century (Cinquecento). Indeed, in the figure, a thick band of light green is visible over the entire period. It becomes thinner thereafter.

The emergence of the US is also noticeable over the 20th century, when many inventors and creative people of all kinds were attracted thanks to a favorable immigration system aligned with a dynamic academic sector. We indeed observe this increasing share (light blue) of the US during that period in the category. These individuals all have a biography in English, but were missed by more selective databases: around 90% or more of these individuals were present in less than 14 `Wikipedia` editions.

Figure 8: **Evolution of the number of "explorer, inventor developer", 1400-2000AD**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: `Wikipedia` profile: share of individuals with biographies in a) 25+ editions including English, b) 14+ editions including English, c) 1 to 13 editions including English, d) One or more editions than English), breakdown by citizenship.

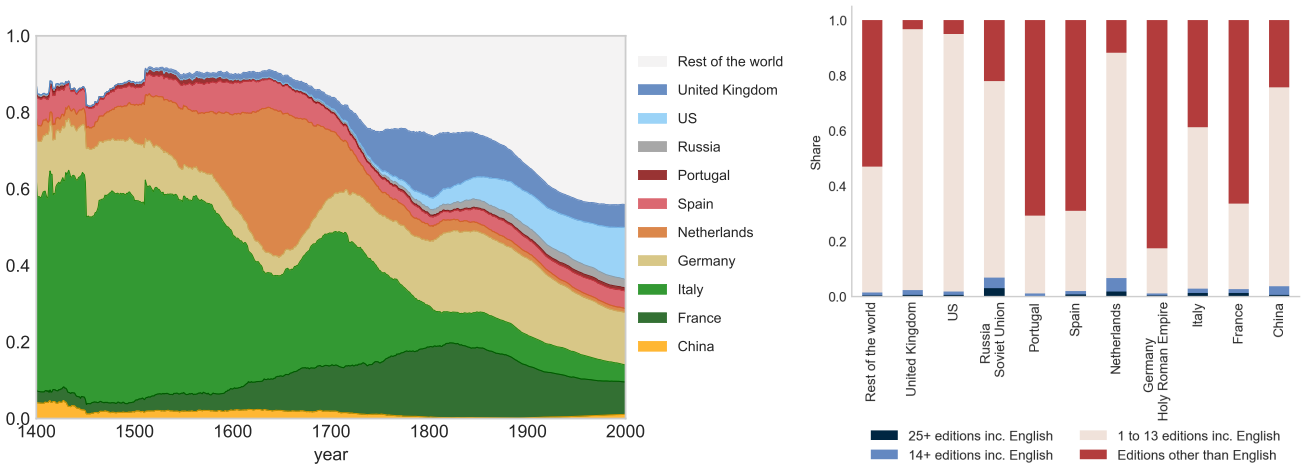
5.3 Arts: the Quattrocento and the Dutch Golden Age

We also expand the coverage of notable individuals in the arts sector. Two important artistic movements clearly stand out from the left panel of Figure 9. The Italian Quattrocento first, which corresponds to the left part of the large green area, that dominates most of the period ranging from the early 15th century till approximately the beginning of the 18th century. The Quattrocento was one of the most important periods of European art and culture. It is referred to the first phase of the movement known as Renaissance. This period is followed by three other important periods in Italian Art history: Cinquecento (1500s), Mannerism (1527 to 1580), Baroque (1600 - 1750) and Rococo Art (1699 - 1780). In the matter, the contribution of the Italian edition is sizeable. The vertical bars in the right panel show that more than 95% of Italian painters were added and absent from existing databases.

The second period corresponds to the Dutch Golden Age which competes with the Italians all over the seventeenth century. The Golden Age in Dutch History is a period spanning from 1581 to 1672, in which Dutch trade, science, and art and the Dutch military were ranked among the most powerful and influential in the world. The first part of the period analyzed is characterized by the Eighty Years' War, which ended in 1648. The number of individuals added with a Dutch citizenship who are not in the English edition of `Wikipedia` is larger than 90%.

At the end of the period, we acknowledge the rise of US modern painting. Here again, extending the scope of the database to less-known individuals proves quite useful. Examples of famous individuals who make their first appearance in a knowledge base are many. Pietro Paolo Vasta (1697-1760, Painter) is an Italian painter and one of the most emblematic renowned member of Sicilian Baroque movement which evolved on the island of Sicily.

Figure 9: **Evolution of the number of individuals associated with "painter", 1400-2000AD**

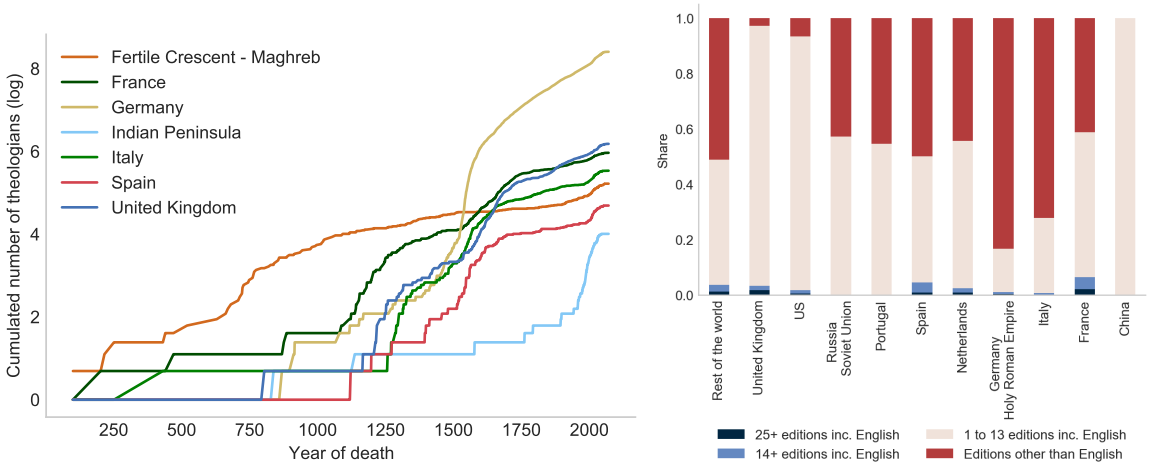


Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: `Wikipedia` profile: share of individuals with biographies in a) 25+ editions including English, b) 14+ editions including English, c) 1 to 13 editions including English, d) One or more editions than English), breakdown by citizenship.

5.4 Religion and theologians: Hegira and Reformation

Figure 10 shows the cumulated number of *theologians* in the restricted sample. Three patterns emerge: the rise of Islamic theologians in the fertile crescent following the Hegira in 622, the steady growth of Christian theologians in Europe until the break related to the Protestant reform (1517). After that period, the number of Protestant theologians rose at an exponential rate in the restricted sample. The role played by the additional editions we considered is important here as well as shown by the large red bars for most citizenships. This brings information about this specific category of notable individuals who played a central role in the History of civilizations.

Figure 10: **Evolution of the number of individuals associated with "theologians", 100-2000AD**

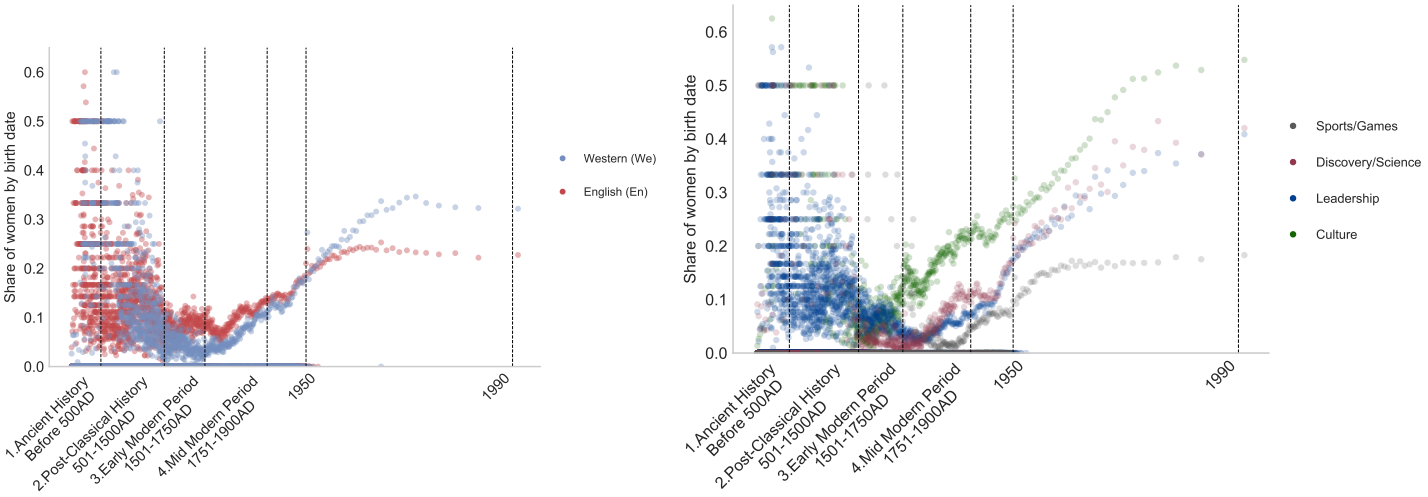


Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: `Wikipedia` profile: share of individuals with biographies in a) 25+ editions including English, b) 14+ editions including English, c) 1 to 13 editions including English, d) One or more editions than English), breakdown by citizenship, y-axis in natural logs.

5.5 From Dark Ages to the Modern age: the rise of notable women

Figure 11 reports the time evolution of the share of female individuals. It exhibits a U-shape pattern with a local minimum at around 5-10% of female individuals around 1750. At the end of the observation period, the share of female individuals is between 20% (Africa) and 30% (Asia and Oceania). The main reason for this rising share in the database is the growing influence of two categories, Sports and Culture, that are much more balanced across genders than categories such as Governance/Executive and Academics. If the share of women is higher in the English edition over the entire sample (see statistics in Table 3 infra.), the Western non-English edition is correcting the bias better after 1950, with a sample average close to 27% of women, as opposed to a 22.5% in the English edition.

Figure 11: **Share of women (left, by continent, right, by domain) , 1000BC-2000AD**

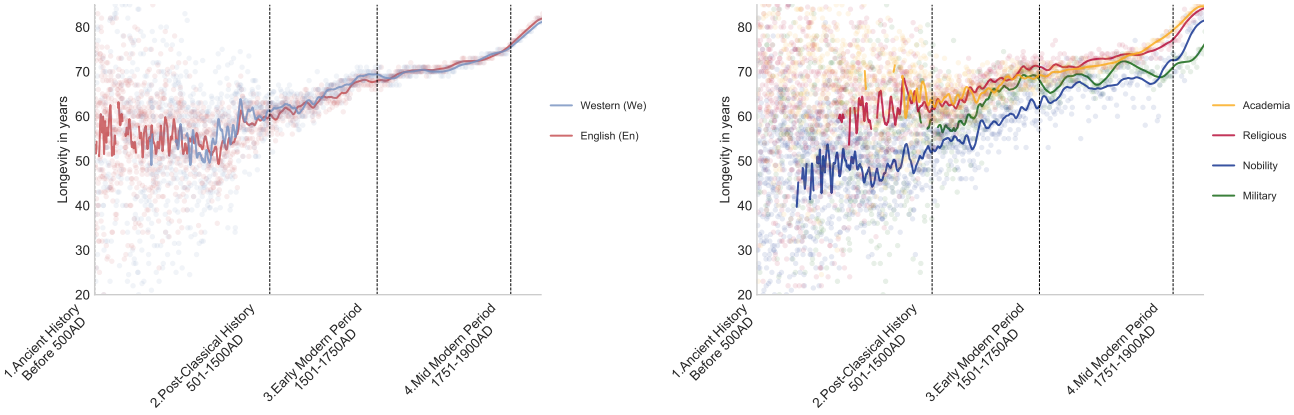


Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed).

5.6 Longevity: war and peace

The evolution over time of median longevity is shown in Figure 12. It was computed as the difference between death year and birth year when available. As found in de la Croix and Licandro (2015), we observe steady improvements in longevity of the cohorts born after 1600. However here we do not observe noticeable differences across language editions (left panel). Similarly to de la Croix and Licandro (2015), we also find (right panel) that the evolution over time of median longevity is lower for individuals in military and nobility domains, compared to academia and religious domains. Concerning nobility, the death of noble infants drives down the median life expectancy.

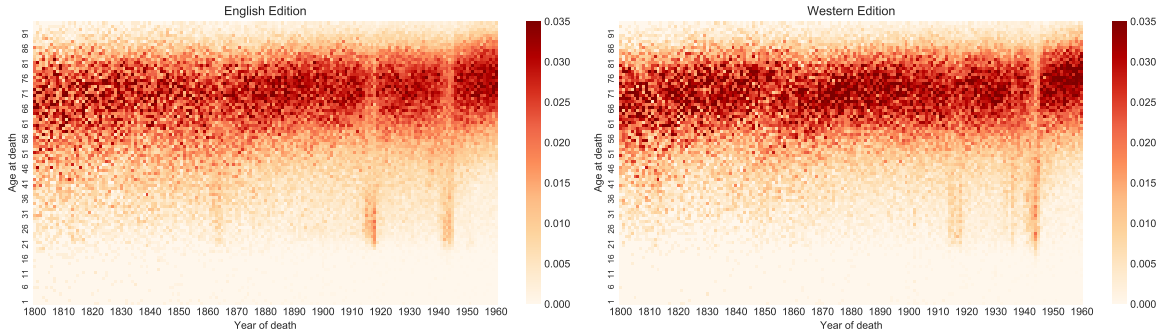
Figure 12: **Longevity, 1000BC-2000AD**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Dots correspond to median life expectancy at birth, solid lines represent moving average over 20 years when observations are available.

Figure 13 represents the age at death for individuals in the sample on the period 1700-1960. On the left panel, war episodes are noticeable as darker downward sloping lines corresponding to abnormal death rates during war periods with even darker points for young generations further exposed in those conflicts (from the right to the left, WWII, WWI, the American civil war, etc.). Schich et al. (2014) provided a detailed timeline of historical events based on a similar change in age of death (Figure 4, page 561), with a comparison of `Ngrams` intensity per period and the frequency of death. One observes the trace of the American civil war on the left chart but not on the right chart. Instead, one observes a small trace on the right around 1936 which corresponds to the Spanish civil war.

Figure 13: **Age at death on English (left) and Western non-English editions (right), 1800-2000AD**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). In both panels, a vertical line corresponds to the distribution of the age at death for a given date. The observed colors discontinuity illustrates wars episodes: American Civil War, First World War, Spanish Civil War, and Second World War.

5.7 Summary: English vs. Western (non-English) editions

The content of existing databases on notable individuals have so far been compiled from the English edition of `Wikipedia` exclusively. Working with the English edition was back then, quite a natural choice as English is still, to date, the largest edition in the `Wikipedia` universe with 1,579,940 different biographies. In this section, we tried to provide details on the addition of non-English editions.

Taking stock, Table 3 provides basic descriptive statistics based on two different samples for birth date, domain of influence, gender and citizenship: a) individuals present in the English edition (and possibly in other editions) and b) individuals absent from the English edition that we call here the Western non-English sample.

It is interesting to note that individuals in the English sample were born on average more recently than those included in the Other editions sample (first three columns). Their main domains of influence are also quite different. Sports, for example, is more prevalent in the English sample while Culture dominates in the Western non-English sample. The fraction of female individuals varies marginally. It is slightly higher in the English sample (18%) than in the other editions sample (15%). The most popular citizenships detected in the English sample are American, UK, Canadian, French versus German, French and Swedish in the Western non-English sample.

Table 3: **Sample statistics: breakdown by language edition (English vs Western non-English)**

`Wikipedia` (recursive lang. editions)	Birth year (percentile) 10	50	90	Occupation %	Female %	Citizenship %
English	1821	1946	1988	SP:34,CLT1:24.1,POL:12.8,ACAD:9.1	17.7	US:25.2,UK:13.6,CA:3.9,FR:3.5
Western non-English	1788	1928	1979	CLT1:31.7,POL:15.6,SP:15,ACAD:14.6	15.3	DE_O:14.8, DE:13.3,FR:13,SE:7.6

Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). This table provides some summary statistics (birth date, domain of influence, share of females, citizenship) on two samples: English versus Non-English. SP = Sports/Games, CLT1 = Culture-Core, POL = Politics, ACAD = Academia, US = United-States, UK = United Kingdom, FR = France, CA = Canada, DE = Germany, SE = Sweden, DE_0 = Germany (Former political/geographical entity).

Results:

1. Historical periods specific to a country are quantitatively better covered by the addition of the Western non-English edition. The French edition of `Wikipedia` allows to better document the emergence of politicians during the French Revolution; the German edition allows to better document the German reformation and the emergence of the Prussian empire, the Spanish and Portuguese editions improve the coverage of the Age of Explorations.
2. The granularity of the database allows to focus on rare occupations such as theologians or on events such as the emergence of journalism.
3. The share of women is substantially higher in the Western non-English editions of `Wikipedia` after 1950, but overall 1.4 percentage point below.
4. The American Civil War is visible in the English edition but not in the Western non-English edition, and the contrary holds for the Spanish civil war.

5. The Western non-English editions focus more on culture and politics and less on sport, and are less centered on individuals from the U.K. and the U.S. and more on Continental Europe, in particular Germany, France and Sweden.

6 Data records, discussion and limitations

This section provides some discussion concerning the bias of our restricted sample detailed above: notable individuals with at least one biography among the universe of `Wikipedia` editions we considered (7 European language editions) plus a `Wikidata` entry.

6.1 Data records

We provide here a random sample of 100,000 individual records or about 5% of the entire database to the Editor and the referees and will grant a full access to readers after publication. A record is made of information retrieved from one or several language editions of `Wikipedia` as described in section 2. Each entry is characterized by the following set of variables:

- *name*: full name of the individual;
- *group_wikipedia_editions*: partition category of the individual (from 1 to 8 as described in section 2.2.);
- *birth*: birth date of the individual (either reported or estimated);
- *death*: death date of the individual (either reported or estimated);
- *level1_main_occ & level2_all_occ & level2_main_occ & level2_second_occ & level3_all_occ & levelC_main_occ & freq_main_occ & freq_second_occ*: set of eight variables for the main domain of influence of each individual in three layers (level 1: 6 groups; level 2: 15 sub-groups and the frequency we impute to the second domain if multiple domains); level 3: keywords collected to assign a domain;
- *gender*: gender of the individual;
- *area_of_rattachment and 2*: first and if needed, second citizenship (or equivalent concept) of the individual with a reference to the current political regime or to the former political regime;
- *number_wiki_editions*: number of different `Wikipedia` editions;
- *wiki_readers_2015_2018*: number of page views in all `Wikipedia` editions (information retrieved in 2015-2018);
- *ranking_visib_5criteria*: computed from 5 variables: number of `Wikipedia` editions, non-missing biographic information, length of pages, and hits of pages. An alternative ranking based on the sum of the log of these variables plus one is in *sum_visib_ln_5criteria*;
- *bplo1 & dplo1 & bpla1 & dpla1 & birthplace_name & deathplace_name*: longitude and latitude of birthplace and deathplace and name. To be used with caution, the accuracy of these variables was not verified.

6.2 Western bias

In this paper, we introduce a multi-language database of notable individuals. We use seven language editions of `Wikipedia` and `Wikidata` to assemble a list of 4,678,040 individuals. Extending the search procedure to six additional languages and to `Wikidata` reduces significantly the Anglo-Saxon bias. We include more Latin-American biographies from the Spanish edition. Adding `Wikidata` in our procedure allows us to include individuals with Arabic, Chinese, Hindi, Russian, Japanese, etc, dimensions. Meanwhile, though our strategy now collects all

individuals included in the famous encyclopedia, it is based on the extraction of biographies in the Western European language editions of Wikipedia. This leads to 2 main drawbacks. First, we cannot exploit the non-Western language editions to cross-verify information on individual characteristics. Second, since we don't collect the number of words in each edition, we are not able to include that measure in the index of notability.

6.3 Missing information

There is some missing information concerning the main variables: birth date, gender, domain of influence, citizenship and birthplace (collected only from Wikidata). Table 4 documents the share of missing information depending on the language group English versus Western non-English.

The first column contains the proportion of individuals with no birth date information (exact or approximate). The proportion is quite low for the Western non-English group (around 2%) while the rate becomes four times larger for individuals in the English group (around 8%). The proportion of missing death date is not included in this table as around 50% of notable individuals are still alive. In Columns (2) to (4), the proportion of individuals with no gender, occupation, and citizenship is very low in both samples. Since we focus on the restricted sample (detailed above), it makes sense to have low rates of missing information as individuals with a Wikidata entry only are excluded. People included in both groups (English and Western non-English) are the most famous ones and therefore missing information on their basic characteristics is unlikely. In column (5), the proportion of individuals with no birthplace is quite high for both samples (34% for the English group and 29% for the Western non-English group). This is mainly explained by the fact that this information is collected only from Wikidata and not from Wikipedia pages (for the moment). Finally, the last column contains the proportion of individuals with all information: birth date, occupation, gender, citizenship or birthplace. In both groups, the rate is quite high with 64% for the English group while it is around 69% for the Western non-English group. It is not reported in this table, but all of the individuals included in this restricted sample have at least one information among these variables (for both groups).

Table 4: **Freq. missing information vs. complete profile**

Wikipedia (by. language group)	Birth %	Gender %	Occupation %	Citizenship %	Birthplace %	| vs. Complete profile %
English	8.15	0.09	0.42	2.31	34.25	64.07
Non-English	2.27	0	1.2	2.38	28.87	69.39

Notes. Restricted sample (at least one Wikipedia edition among the 7 European languages analyzed. First 5 columns: % missing observations; last column: % with no missing observation among these 5 variables.

6.4 Comparison between Wikipedia editions and Wikidata

The strength of our methodology comes from the fact that we use 7 editions of Wikipedia and Wikidata. Since we often have more than one entry for most individual characteristics, we combine both sources in order to cross-verify the maximum number of information. Although rare, it happens the information is different in both sources. Table 5 provides the number of observations for which the variable is reported in both sources (first row) and its share in the sample (second row). The third and fourth rows give the share of discrepancy between Wikipedia and Wikidata for each variable. Concerning dates, the share of exact mismatch is very low, with 2% for birth and 1.6% for death. When we restrict to differences of more than 10 years for birth and death dates

(fourth row), the share of discrepancy decreases to 0.23% for birth and 0.38% for death. The rate of discrepancy for gender is also very low with 0.51% of cases.

The information on the domain of influence is different (different sub-category detected) in a larger number of cases (11%). It mainly corresponds to close sub-categories such as e.g. Culture-core vs. Culture-periphery or Politics vs Military, or Nobility vs Family. To take this into account, in the fourth row, we compute the rate of discrepancy within large categories (Discovery/Science, Culture, Leadership, Sports/Games, Other) and we observe that the rate is much lower (7.2%).

The last column shows that the share of discrepancy concerning citizenship is at around 10%. This figure is mainly explained by the fact that `Wikidata` provides the citizenship based on the old regime while `Wikipedia` gives the modern citizenship. For instance, Martin Luther's country of citizenship is "Electorate of Saxony" (which belongs to the Holy Roman Empire) in his `Wikidata` biography while his English `Wikipedia` biography says he "was a German professor of theology".

The fifth and sixth rows provide some figures on the contribution of `Wikipedia` pages. It gives the number of observations (fifth row) and the frequency (sixth row) that we retrieve from `Wikipedia` only, as the information is missing from `Wikidata`. This shows that the combination of `Wikipedia` and `Wikidata` adds a substantial fraction of observations to the database. For instance, `Wikipedia` adds 2.72% of birth dates, 1.31% of death dates, 0.67% of gender, 8.16% of occupations and 17.16% of citizenships. The rates are quite low concerning the dates and gender because missing information from `Wikidata` on these basic characteristics is unlikely.

Finally, the last two rows give the number of cases with discrepancy between `Wikipedia` & `Wikidata` and for which we choose the source is `Wikipedia`.

Table 5: `Wikipedia` & `Wikidata`: **similarity and differences**

	Birth	Death	Gender	Occupation	Citizenship
P and D coincide (# Obs)	2,011,546	921,231	2,239,532	1,941,246	1,748,981
P and D coincide (%)	87.77	40.20	97.72	84.70	76.31
P and D differ (conservative %)	2.14	1.64	0.51	11.04	9.93
P and D differ (flexible %)	0.23	0.38	n.a. (*)	7.17	n.a. (*)
D (missing) and P (available) (# Obs)	62,263	30,070	15,372	186,951	393,198
D (missing) and P (available) (%)	2.72	1.31	0.67	8.16	17.16
Mismatch (and source=P) (# Obs)	1,032	484	0	9,528	19,591
Mismatch (and source=P) (%)	0.05	0.02	0.00	0.42	0.85

Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). This table provides some summary statistics on similarity and differences between the sources `Wikipedia` (P) and `Wikidata` (D). P and D differ (flexible %) difference in terms of birth/death years between P & D lower than 10 years (for columns Birth and Death), or difference between large categories between P & D (column Occupation). Conservative means that an exact match is required (for years of birth and death or for level 2 occupations). (*): n.a. = distinction not applicable.

6.5 Manual checks

We ran several tests to check the accuracy of our dataset. We asked three teams of students in three different cultural areas countries (France, the UAE and India) to compare information provided for the main variables from our dataset to information present in Wikipedia/Wikidata. They had to report mistakes on 6 different pieces of information (exact or approximate date of birth and death, gender, main occupation and possibly secondary occupation, citizenship and possibly secondary citizenship). The 5 different possible outcomes are distinguished in 2 parts: 1) Information not reported in our database: a) "No info. in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in the dataset, b) "Info. updated since data collection" means the information is today included in `Wikipedia`/`Wikidata` but was not present in the dataset at the time of collection; 2) Information reported in our database: a) "Correct" means no error, b) "Error" means certain error, c) "Other case" means possible error (for instance historical controversy, several sources diverging or information updated since data collection).

A pilot wave on 5 sub-samples of 500 individuals each, with a 20% overlap, lead to a percentage of errors below 1% except for the first and second occupation, that is above 1%. The final test is based on a larger sample: 10 students received a sub-samples of 1000 individuals each, with full overlap so that every individual would be checked twice. The total sample of 5000 individuals verified contains the top 1000 individuals of the database, 500 individuals with two pages in `Wikipedia` at least (they are entirely with the top quartile of visibility the database and the median individual in that sample was the in the 92th centile of the visibility distribution), 3,000 individuals from the databases with a equal split in the 4 quartiles, and finally 500 individuals with no `Wikipedia` page, and only a `Wikidata` entry.[14]

When the outcome variable was consistent between the two independent research assistants, we kept their outcome. When they were not consistent due to a difference in appreciation, we had it verified by a PhD student who checked the reason for the discrepancy and arbitrated between the two research assistants.[15]

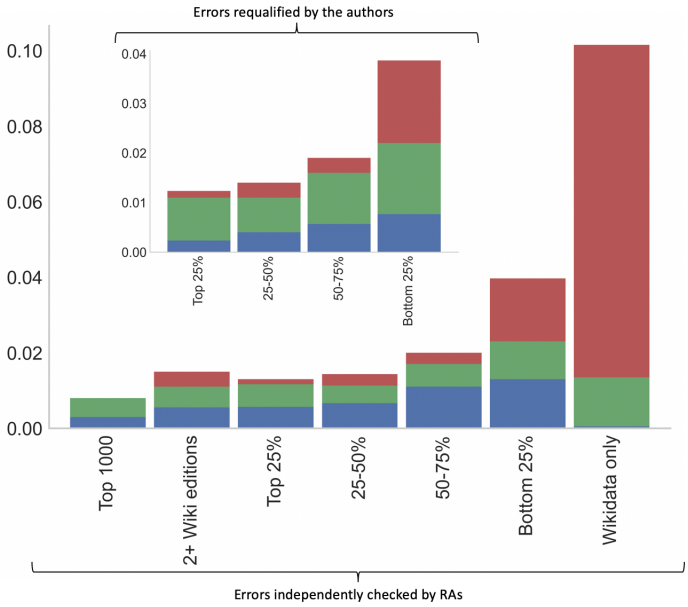
Our verification allows for a clear distinction between the errors due to the algorithm and the errors due to the evolving nature of Wikipedia. We report in Figure 14 the distribution of errors for the different samples (top panel). The error rate is increasing for individual with less detailed biographies, as expected, going from less than 0.3% for the top 1000 individuals and 0.5% for the most documented individuals in the top quartile to almost 1% in the bottom of the distribution. There is also a large number of missing information in the `Wikidata` only sample.

We finally looked at the errors detected independently by the RAs, and requalified some of them after a last cross-examination. Most of the errors detected by the RAs were actually due to a slight discrepancy - typically one year - between the birth date we gave in the database and the sources from `Wikipedia` or `Wikidata` that were themselves conflicting with each other (e.g two pages would give respectively 1944 and 1943 and we had reported 1943). So the numbers reported overestimate the true errors by a factor 3 to 6. We display the requalified numbers in the smaller chart in the top panel. We also report in green the fraction of cases where the information was not available in our database but became available after our data collection. The bottom panel reports the shares of correct information and the share of cases where the information was not available in our database and was still missing in the sources.

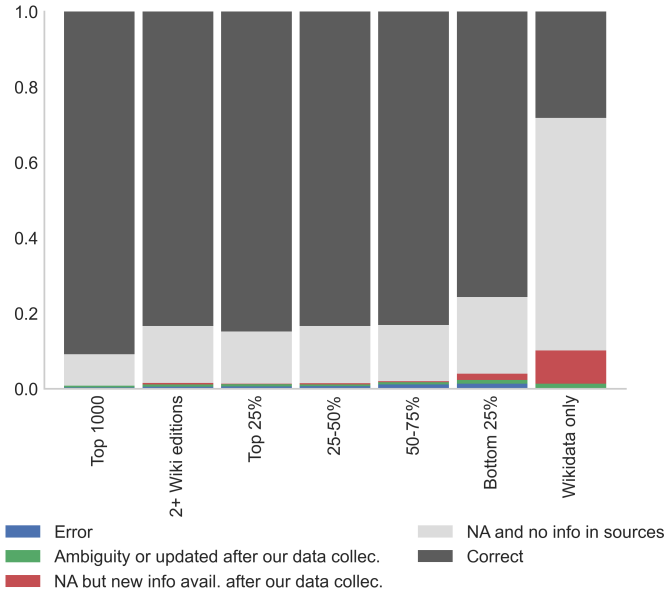
[14] See Appendix C for more details on the instructions given to the RAs. In addition, one individual was dropped, because it was selected twice - in quartile 2 and the universe of two `Wikipedia` pages at least, so the total number of verified individuals in 4,999.

[15] See Table 9 in Appendix C where we report the rates of discrepancy by variable.

Figure 14: **Manual verifications: summary statistics**



Notes. Top panel, larger graph: the statistics verified independently twice by a team of 10 RAs and cross-verified by a PhD researcher. Focus on errors, imprecisions or updates since our data collection. Smaller graph on top: requalification of some of the errors by the authors. Vertical axis represents the percentage of errors divided by the number of observations including NA (themselves approx. 20% of total number of observations). The level 0.005 on these graphs represents .5% of all observations and .5/(1-0.2)=0.625% of all non-missing observations.



Notes. Bottom panel: correct information, missing, errors, imprecisions or updates since our data collection, verified independently twice by a team of 10 RAs and cross-verified by a PhD researcher.

See also Appendix C Tables 10 to 14, which provides the summary statistics concerning manual checks for each of the following variables: birth date, death date, gender, first and second occupations, first and second citizenships, for all sub-samples.

7 Summary and Outreach

Summary Our paper complements the existing literature in five ways. First, we introduce a novel technique to define the universe of notable individuals using the "categories" present in the bottom part of all `Wikipedia` biographies. This contrasts with former methods relying exclusively on centralized lists such as `Freebase`. We then compare the list of individuals compiled from the information collected in `Wikipedia` with that retrieved using the "Instance of humans" Q5 code of `Wikidata`. This bottom-up, transversal, approach results in a enlarged sample of 6,291,767 biographies corresponding to 4,678,040 unique individuals.

Second, we parse seven major European editions of `Wikipedia` (English, French, German, Italian, Portuguese, Spanish, Swedish) to document and reduce the Anglo-Saxon bias naturally present in existing projects based on the English edition only. In `Wikipedia`, we consider the information contained in the Infobox but not only. We also parse the content of the biography itself to complement our database when the Infobox is not yet available. This is quite often the case for less prominent individuals.

Third, we consider `Wikipedia` and `Wikidata` as complementary sources of information. Combining `Wikidata` and several editions of `Wikipedia` has two major advantages: i) it adds relevant information not present yet in `Wikidata`, ii) it improves the precision of both `Wikidata` and `Wikipedia`. This cross-verification technique proved quite powerful and decreased the number of flaws or mistakes still contained in each universe due to natural human error coding. The implementation of this verification technique naturally reduces the sample size as a significant fraction of individuals are not present in any of the seven language editions we parsed. These individuals, excluded at this stage from the analysis, were detected from `Wikidata` and might be considered in the future as we plan to parse other important editions such as Urdu, Chinese, Russian, Japanese, just to name a few. The restricted sample however concerns a satisfactory number of different individuals as we managed to cross-verify the information concerning 2,291,817 individuals.

Fourth, we provided the data to discuss in a quantitative way several major historical periods or events such as the French Revolution, the rise of modern democracies (e.g. in the US and UK), The Quattrocento, the Dutch golden age, the Age of exploration clearly benefit from the inclusion of the information contained in biographies written in the native language of notable people. It is therefore useful to embed less famous individuals, in particular those with less than 14 biographies in `Wikipedia`. This information, by lack of language proficiency has not been translated into any other language edition by `Wikipedia` contributors whose language is national. Another advantage of local contributors is their privileged access to the relevant information extracted from the local archives and/or books published in French. Ignoring these notable characters and this specific information could result in a bias in later studies.

Fifth, we manually-verified the content of 5,000 biographies. This alternative verification exercise confirms that it is better to focus on the most documented biographies and be careful with `Wikidata` pages when they do not correspond to a `Wikipedia` entry.

Several research directions are in order. First, we will develop the gender dimension to quantify the under-representation of women in `Wikipedia`, around 18%, with a minimum in the XIX's century in all domains of influence. This shows that recent projects such as the "let's fill the Wiki gender gap" are important and necessary.[16]

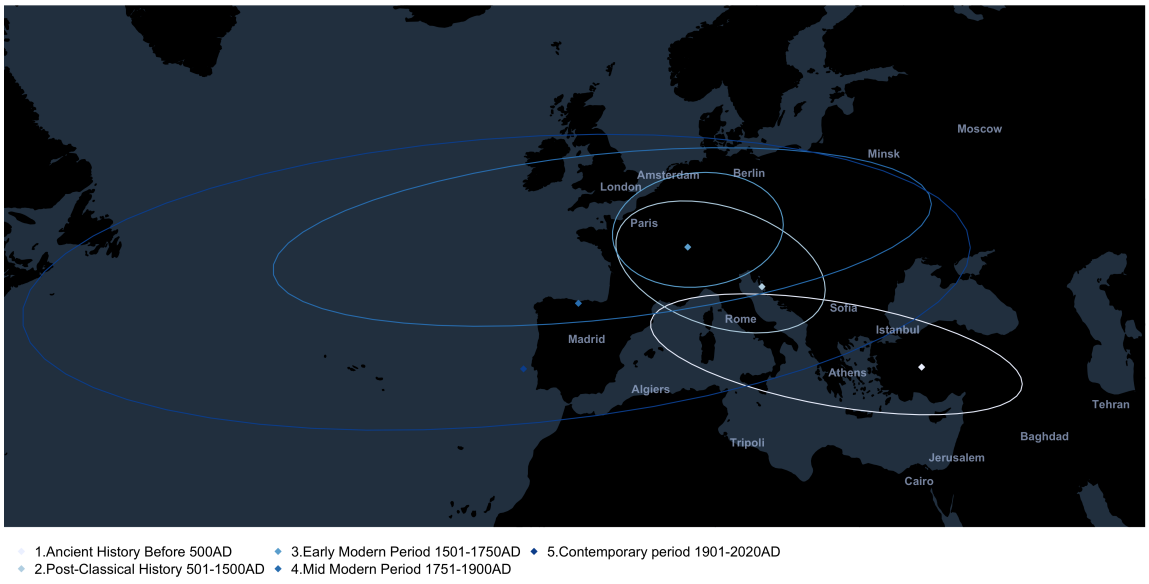
[16] https://www.wikimedia.ch/fr/event/workshop-lets-fill-the-wiki-gender-gap-1/

Second, collecting and analyzing biographies in language editions other than English seems therefore an important task to study key historical facts and economic phenomena with greater accuracy and future work should tackle major editions (Chinese, Arabic, Russian, Japanese, etc.).

Outreach To add a final word, our database can be used to provide a new level of granularity and historical depth to several fields, provided that selection is taken care of, following the discussion of subsection 2.2. Selection is a common issue in most empirical works and requires a specific treatment based on the question asked. Mincerian equations estimations require a selection equation into employment, firms data with a size criteria e.g. more than 10 employees should take care of this selection too, work with historical data must deal with endogeneity of preservation, work on cross-national OECD data must take care of the endogeneity of the OECD membership. The treatment of endogeneity of notable individuals will need to be specific to the question asked and the nature of the data.

The database is geolocalized and therefore can follow the evolution of geographic centers, as represented in Figure 15, which represents the ellipse of the variance-covariance matrix of longitudes and latitudes of birthplaces at different periods (40% of the sample population lies within each ellipse).

Figure 15: **Barycenters and ellipses of covariance matrix of lon.&lat. of birthplaces**



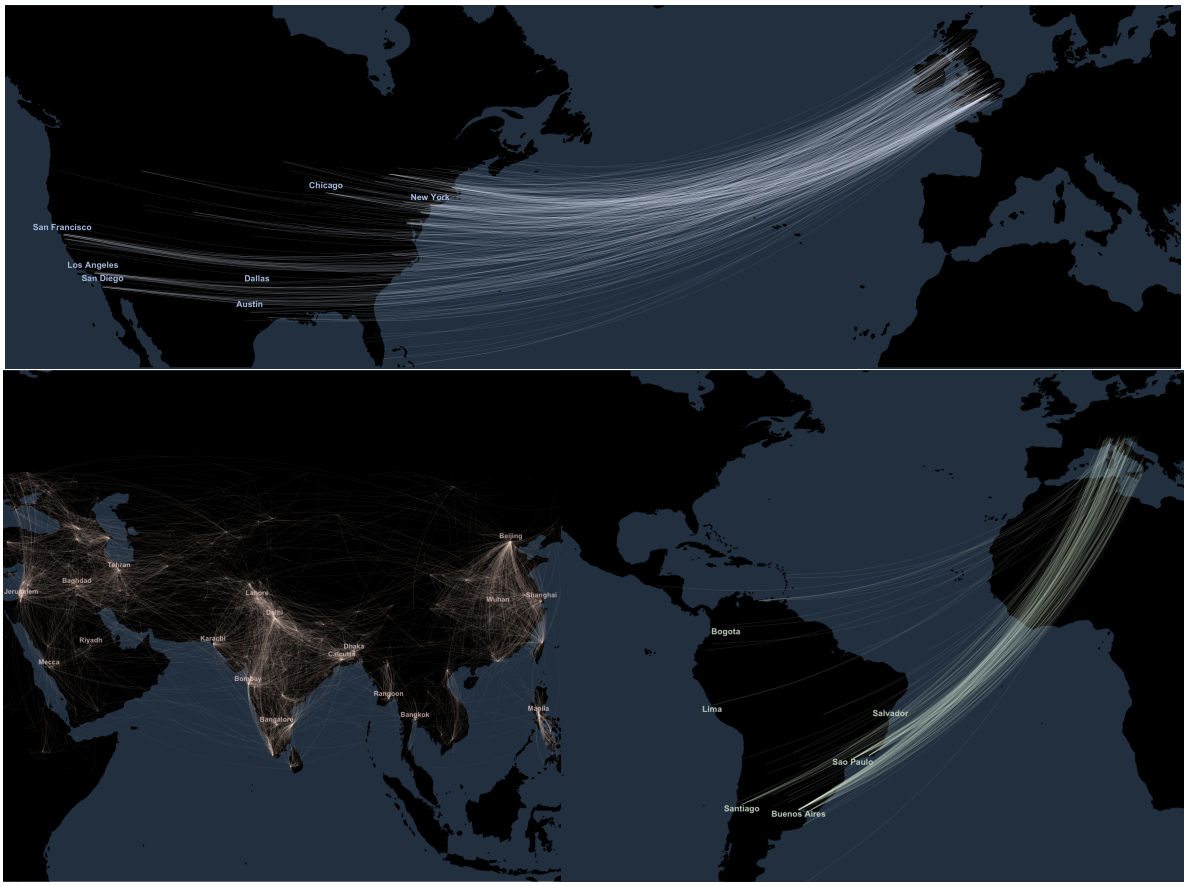
Notes. Ellipses are based on the variance-covariance matrix of longitudes and latitudes for a given period, with a threshold of 40%

First of the following non-exhaustive list is urban economics, to test the quantitative implications of the model of cities attracting talents (Behrens et al., 2014), the role of the creative class to enlighten the causation chains (Moretti, 2012; Florida, 2005; Serafinelli and Tabellini, 2017), the impact of transport infrastructure development on geographic mobility (Baum-Snow, 2007; Donaldson and Hornbeck, 2016), or the existence of gravity forces in trade to uncover ancient cities as trade places (Barjamovic et al., 2019), thanks to our measures of the notability and our geo-localized data on birth and deathplaces and their evolution over centuries.

Many gender issues can be revisited with greater historical depth and finer geographical scale. For example, the importance of role models in the dynamics of women's empowerment thanks to alternative measures of

notability and the use of geo-coded birthplaces (Beaman et al., 2012; Bertrand, 2020), the importance of the local environment enabling women to become inventors (Bell et al., 2019; Hunt et al., 2013), the convergence of male and female occupations over time (Goldin, 2014). More generally, many gender issues can be revisited with greater historical depth and finer geographical scale.

Figure 16: **Migrations charts, selected countries and domains of influence**



Notes. Top chart: birth to death flows of academics from the UK to North America. Bottom chart: on the right, birth to death flows of notable individuals in culture from Italy to South America; on the left, birth to death flows (all categories) within a rectangular area encompassing Asia including Central Asia, Western Asia and the southern part of Eurasia). All curves connect the city of birth and of death of individuals.

The database can similarly be used to study several long-run issues. For instance, to the extent that the attractiveness of cities for scientists and artists are a proxy for economic growth and the notable people in law, governance and administration capture the quality (either positively or negatively) of institutions, the database can be used to investigate the role of institutions on growth (Mokyr, 2010; Acemoglu and Johnson, 2005; Glaeser et al., 2004). Implications of unified growth theory, as the replacement of physical by human capital (Galor and Moav, 2004; Galor, 2011), as well as neo-Schumpeterian theories and the role of innovation and competition (Aghion et al., 2005; Aghion and Howitt, 2008), thanks to our account of scientists and innovators per country. The same is true for detecting evidence of cultural transmission mechanisms through naming (Fryer and Levitt, 2004; Alesina and Giuliano, 2015; Bisin and Verdier, 2001), or to deepen our understanding of the causes and

consequences of key historical events such as the Protestant reformation along the lines of pioneering works (Ekelund et al. (2002); Becker et al. (2016); Cantoni (2015), etc.) as well as the role and determinants of culture in the wealth of nations (Gorodnichenko and Roland (2017); Roland (2016); Platteau and Peccoud (2011)).

Research in demography economics for instance on longevity in space and time (de la Croix and Licandro, 2015) can be enriched thanks to our account of birth and death dates by field. Issues on international and domestic migration can also be revisited from a more longitudinal perspective. For example, Zelinksy's theory of mobility transition as explored in Dao et al. (2018), the theory and evidence on brain drain as explored in Beine et al. (2001), the birthplace diversity and economic prosperity as explored in Alesina et al. (2016) and the relationship between future outcomes and birthplace as identified by Chetty et al. (2014) can be revisited from a historical perspective, thanks to our geocoded birth and death locations and occupation variables. Domestic migration within the United States and immigration to the United States and its links to economic prosperity as explored by Abramitzky et al. (2021) or extra-place based mortality as in Finkelstein et al. (2019) can be revisited with the use of geocoded birth and death locations in our dataset, with but possibly even without matches with the historical US censuses. See Figure 16 for an overview of selected migration flows across the Atlantic, and within Asia, mapped from birth and death places.

To sum up, the database is suited to explore gender, demography, urban development, cultural transmission, human capital, growth, institutions, thanks to its time and spatial coverage and granularity.

References

Abramitzky, R., L. Boustan, E. Jacome, and S. Perez (2021): "Intergenerational Mobility of Immigrants in the United States over Two Centuries," *American Economic Review*, 111, 580–608.

Acemoglu, D. and S. Johnson (2005): "Unbundling institutions," *Journal of political Economy*, 113, 949–995.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005): "Competition and innovation: An inverted-U relationship," *The quarterly journal of economics*, 120, 701–728.

Aghion, P. and P. W. Howitt (2008): *The economics of growth*, MIT press.

Alesina, A. and P. Giuliano (2015): "Culture and Institutions," *Journal of Economic Literature*, 53, 898–944.

Alesina, A., J. Harnoss, and H. Rapoport (2016): "Birthplace diversity and economic prosperity," *Journal of Economic Growth*, 21, 101–138.

Barjamovic, G., T. Chaney, K. Coşar, and A. Hortaçsu (2019): "Trade, merchants, and the lost cities of the bronze age," *The Quarterly Journal of Economics*, 134, 1455–1503.

Baum-Snow, N. (2007): "Did Highways Cause Suburbanization?" *The Quarterly Journal of Economics*, 122, 775–805.

Baumol, W. and W. Bowen (1966): *Performing Arts: The Economic Dilemma*, New York: MIT Press.

Beaman, L., E. Duflo, R. Pande, and P. Topalova (2012): "Female leadership raises aspirations and educational attainment for girls: A policy experiment in India," *science*, 335, 582–586.

Becker, S., S. Pfaff, and J. Rubin (2016): "Causes and consequences of the Protestant Reformation," *Explorations in Economic History*, 62, 1–25.

Behrens, K., G. Duranton, and F. Robert-Nicoud (2014): "Productive cities: Sorting, selection, and agglomeration," *Journal of Political Economy*, 122, 507–553.

Beine, M., F. Docquier, and H. Rapoport (2001): "Brain drain and economic growth: theory and evidence," *Journal of Development Economics*, 64, 275–289.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019): "Who becomes an inventor in America? The importance of exposure to innovation," *The Quarterly Journal of Economics*, 134, 647–713.

Bertrand, M. (2020): "Gender in the twenty-first century," in *AEA Papers and proceedings*, vol. 110, 1–24.

Bisin, A. and T. Verdier (2001): "The Economics of Cultural Transmission and the Dynamics of Preferences," *Journal of Economic Theory*, 97, 298–319.

Cantoni, D. (2015): "The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands," *Journal of the European Economic Association*, 13, 561—598.

Chah, N. (2017): "Freebase-triples: A Methodology for Processing the Freebase Data Dumps," *ArXiv*, abs/1712.08707.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014): "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States," *The Quarterly Journal of Economics*, 129, 1553–1623.

Dao, T. H., F. Docquier, C. Parsons, and G. Peri (2018): "Migration and development: Dissecting the anatomy of the mobility transition," *Journal of Development Economics*, 132, 88–101.

de la Croix, D. and O. Licandro (2015): "The longevity of famous people from Hammurabi to Einstein," *Journal of Economic Growth*, 20, 263–303.

Donaldson, D. and R. Hornbeck (2016): "Railroads and American Economic Growth: A "Market Access" Approach," *The Quarterly Journal of Economics*, 131, 799–858.

Ekelund, Jr, R. B., R. F. Hébert, and R. D. Tollison (2002): "An economic analysis of the protestant reformation," *Journal of Political Economy*, 110, 646–671.

Finkelstein, A., M. Gentzkow, and H. L. Williams (2019): "Place-based drivers of mortality: Evidence from migration," Tech. rep., National Bureau of Economic Research.

Florida, R. (2005): "The rise of the creative class and how it's transforming work, leisure, community and everyday life (Paperback Ed.)," .

Fryer, R. and S. Levitt (2004): "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics*, 119, 767–805.

Galor, O. (2011): *Unified growth theory*, Princeton University Press.

Galor, O. and O. Moav (2004): "From physical to human capital accumulation: Inequality and the process of development," *The Review of Economic Studies*, 71, 1001–1026.

Gergaud, O., M. Laouenan, and E. Wasmer (2016): "A Brief History of Human Time: Exploring a database of 'notable people'," Sciences Po Economics Discussion Papers 2016-03, Sciences Po Departement of Economics.

Glaeser, E. L., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2004): "Do institutions cause growth?" *Journal of economic Growth*, 9, 271–303.

Goldin, C. (2014): "A grand gender convergence: Its last chapter," *American Economic Review*, 104, 1091–1119.

Gorodnichenko, Y. and G. Roland (2017): "Culture, Institutions, and the Weath of Nations," *The Review of Economics and Statistics*, 99, 402–416.

Hernández, M. A. and S. J. Stolfo (1995): "The merge/purge problem for large databases," in *Proceedings of the 1995 ACM SIGMOD international conference on Management of data - SIGMOD '95*, San Jose, California, United States: ACM Press, 127–138.

Hunt, J., J.-P. Garant, H. Herman, and D. J. Munroe (2013): "Why are women underrepresented amongst patentees?" *Research Policy*, 42, 831–843.

Jara-Figueroa, Christian, Y. A. and C. Hidalgo (2019): "How the medium shapes the message: Printing and the rise of the arts and sciences," *PLOS One, 14(2): e0205771.*

Kremer, M. (1993): "Population growth and technological change: One million BC to 1990," *The Quarterly Journal of Economics*, 108, 681–716.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. Vishny (1999): "The quality of government," *The Journal of Law, Economics, and Organization*, 15, 222–279.

Levenshtein, V. I. (1966): "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, 10, 707.

Manning, S. (2008): "Year-by-year world population estimates: 10,000 BC to 2007 AD," *Historian on the warpath*, 12.

Mokyr, J. (2010): *The Enlightened economy an economic history of Britain 1700-1850*, Yale University Press.

Moretti, E. (2012): *The new geography of jobs*, Houghton Mifflin Harcourt.

Nekoei, A. and F. Sinn (2020): "Human Biographical Record (HBR)," *Available at SSRN.*

Platteau, J.-P. and R. Peccoud (2011): *Culture, Institutions, and Development: New Insights into an Old Debate*, Routledge.

Pollock, J. J. and A. Zamora (1984): "Automatic spelling correction in scientific and scholarly text," *Communications of the ACM*, 27, 358–368.

Postel, H. J. (1969): "Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse," *IBM-Nachrichten*, 19, 925–931.

Roland, G. (2016): "Culture, Institutions and Development," *Working Paper.*

Schich, M., C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing (2014): "A network framework of cultural history," *Science*, 345, 558–562.

Serafinelli, M. and G. Tabellini (2017): "Creativity over time and space," *Available at SSRN 3070203.*

Tanon, T. P., G. Weikum, and F. Suchanek (2020): "YAGO 4: A Reason-able Knowledge Base," in *European Semantic Web Conference*, Springer, 583–596.

Throsby, D. (2008): "The concentric circles model of the cultural industries," *Cultural trends*, 17, 147–164.

Yu, A. Z., S. Ronen, K. Hu, T. Lu, and C. A. Hidalgo (2016): "Pantheon 1.0, a manually verified dataset of globally famous biographies," *Nature - Scientific data*, 3, 150075.

Online Appendix: supplementary tables and figures

A Data collection details and processing for the exhaustive sample

A.1 Removing duplicates: details

Dealing with possible duplicates is not an easy task as we need to separate these cases from real homonyms, i.e. individuals sharing exactly the same name and first name. We use a total of eleven methods, all detailed in the following box, ranging from string normalization, phonetic encoding and string distance metrics to identify likely duplicate pairs that we eventually decide to merge by manually checking their respective `Wikipedia` biographies. In order to reduce the number of candidates, which is prohibitively large in our database, we determine a score for each candidate based on some additional features such as common birth or death dates, the citizenship and domains of influence retrieved from these questionable biographies. This helps us discard candidate pairs which were not duplicates.

We then construct a score ranging from 0 to 1 which corresponds to the likelihood for a set of biographies to correspond to the same individual. A score above 0.75 for 4 criteria and above 8 for the remaining two identifies a 'cluster' of individuals who have a high probability of being the same person; and we kept the person with the highest number of available biographical information. We identify 34,562 true duplicates, that is 0.7% of the total number of biographies (34,562/4,678,041).

We use the following methods to remove duplicates:

1. Connected components solving: sometimes links between `Wikipedia` biographies are not mutual. It is therefore possible, by gathering connected components of the page lowercase names' graph, to find suitable duplicate pairs.
2. Aggressive string normalization: by normalizing hyphens, underscores, solving url encoding and dropping non-alphanumeric characters, one can find more suitable duplicate pairs.
3. Unicode standardization: some languages, such as English, do not handle accentuated characters very well and tend to avoid using them. By standardizing unicode characters to plain ascii, it is possible to match similar names in two different languages. It is also possible to match names written in other alphabets thusly.
4. String fingerprinting: there is a large variety of ways to write the same name. It is not rare, for instance, to see Asian names written in the incorrect order by occidental clerks. String fingerprinting is a method which applies a set of transformations to a string to normalize order, redudancy and case so one can match similar-looking strings.
5. Squeezed string fingerprinting: same as before except that we will "squeeze" consecutive duplicate letters into a single one. For instance, the name "Brettner" would become "Bretner". This follows the observation that double letters tend not to be well-respected across variants of the same name.
6. Small tokens filtering: small tokens composed of only one or two characters, such as "de" or "of", and stopwords tend to be frequently forgotten in names. Filtering them will produce some more matches.
7. Rusalka phonetic encoding: by producing a symbolic phonetic representation of the considered names, one is often able to match different transliterations or spellings.

8. Sorted neighborhood using the omission key and Levenshtein distance less than or equal to one: string distances such as the Levenshtein distance are very useful to find similar-looking strings. Unfortunately, a naive approach to collect pairs of duplicates in a dataset results in quadratic processing time. While this is acceptable for tiny datasets, it is not for millions of names. The sorted neighborhood method can approximate pairwise computations by considering that if you order strings using a specific key beforehand then similar pairs have a high probability of being close in the sorted list. A fixed-size window is then slided across the sorted list where pairwise distances are computed and similar pairs reported. We first chose to use the omission key, a string's key leveraging the frequency to which characters are omitted when misspelling words, to sort our dataset before proceeding to find pairs having a very low Levenshtein distance.
9. Sorted neighborhood using the skeleton key and Levenshtein distance less than or equal to one: same as before but using a different key, the skeleton key, leveraging the way words tend to be misspelled in the English language, i.e. misspelled consonants are frequently not the first ones.
10. Cologne phonetic encoding: this phonetic encoding targets specifically German and similar languages and is a good complement to the Rusalka one. Its precision is very low however since it tends to approximate sounds a lot.
11. Sorted neighborhood using the skeleton key and Levenshtein distance less than or equal to two.

    Further references: see Postel (1969); Pollock and Zamora (1984); Hernández and Stolfo (1995); Levenshtein (1966).

A.2 Data collection using categories

We develop a methodology based on the information found in the categories of `Wikipedia` to approach the universe of notable individuals. We scraped individuals from a particular procedure based on categories. Categories are present in the bottom part of most biographies. These independent `Wikipedia` objects contain lists of individuals (and their associated urls) who have one feature in common such as such as: birth date, death date, domain of influence, etc. In a first stage, we harvest all links available in the "Living People" (https://en.Wikipedia.org/wiki/Category:Living_people) category of the English edition. In a second stage, we explore additional categories such as "Possibly living people", "Deaths (resp. birth) by year", "Deaths (resp. birth) by decades", "Deaths (resp. birth) by centuries" and "Deaths (resp. birth) by millennium", etc. to collect more urls. Last, we parse the following list of categories to detect individuals that were not identified in the previous stages: "Date of birth missing", "Date of birth unknown", "Date of death missing", "Date of death unknown", "Year of birth missing", "Year of birth unknown", "Year of death missing", "Year of death unknown", "Place of birth missing", "Place of birth unknown", "Place of death missing", "Place of death unknown".

A.3 Oldest registered entries and comparison with world population estimates

The first registered human in our database was born around 430,000 BC.[17]. The second oldest entry, 11,000 BC, is a skull of a Paleo-Indian woman discovered in Mexico city in 1959. Three other famous skeletons[18] follow. The first individual with a social status comes next in 6th position in our database. Pesho, was "*chief, who lived ca. 7000-7500 years ago in territory of modern Bulgaria and known for his rich tomb (sic)* " The first notable individual, in the sense of his prominent role in history, comes next in 7th position. Ny-Hor, born between 4000 BC and 3001 BC, was *"a king in the Egyptian predynastic period, and known as Her or Hor (Horus), that*

[17]"Cranium 17", an ancient hominid skull.

[18]Namely the Kolebjerg Man (8000 BCE), Loschbur-Fra (8000 BCE), the Frau von Bäckaskog (7000 BCE).

is, "the Falcon", and his monarchy is established in Nekhen (later Hieraconpolis)." Interestingly, this Pharaoh has a biography in French, Arabic, German, Russian, Italian, Portuguese and a few other languages but, as of June 2019, not in English. We arbitrarily decided to officially start our database with this political figure. Later on, there would be more famous individuals such as the king of Tyre Deleastartus and the Assyrian kings Puzur-Ashur I and III (circa 2000 BC and 1500 BC respectively), the Chinese empire chancellor Yi Yin (born 1648 BC) and the sixth king of Babylon Hammurabi who died in 1750 BC. In total, our database contains 330 individuals who have lived before Hammurabi, the sixth king of Babylon, who is the oldest registered notable individual in de la Croix and Licandro (2015).

Table 6: Oldest individuals in the exhaustive database

Name	`Wikidata` Code	Birth Min	Birth Max	Death Min	Death Max
Cranium 17	Q41330363	-430000	-429001	-430000	-429001
Femme de Peñon	Q1988410	-11000	-10001	-11000	-10001
Koelbjerg Man	Q455750	-8000	-7001	-8000	-7001
Loschbur-Fra	Q25583326	-8000	-7001	-8000	-7001
Frau von Bäckaskog	Q6981339	-7000	-6001	-7000	-6001
Pesho	Q29510353	-5000	-4001	-5000	-4001
Ny-Hor	Q268647	-4000	-3001	-4000	-3001
Mummia del Similaun	Q171291	-3345	-3345	-3255	-3255
Menes	Q189574	-3200	-3200	-3100	-3100
Hat-Hor	Q577451	-3150	-3150	-3095	-3095
Frau von Luttra	Q179281	-3125	-3125	-3100	-3100
Djer	Q152375	-3000	-2901	-3000	-2901
Djet	Q151828	-3000	-2901	-3000	-2901
Merneith	Q230548	-3000	-2901	-3000	-2901
Teti I.	Q153154	-3000	-2901	-3000	-2901
Den (pharaoh)	Q151822	-3000	-2901	-2995	-2995
Semerkhet	Q151805	-3000	-2901	-2960	-2960
Nefer (Hofzwerg)	Q1800518	-3000	-2901	-2900	-2801
Iblul-Il	Q4202987	-3000	-2001	-3000	-2001
Mann von Porsmose	Q1726357	-3000	-2001	-3000	-2001

Notes. Exhaustive sample (4.7 million individuals). The birth and death min and max are based on the precision of the related dates: millenia, centuries.

One can compare the evolution of the living notable persons in our database to world population and world GDP. We use two sources for population, Kremer (1993) and Manning (2008), who are extremely close to each other over the most recent period we consider (1000BC to now). World GDP estimates also comes from Kremer (1993). The series are represented in Figure 17, in log, and the x-axis is either linear or a log of the calendar year. All series grow, but the notable population in the database drops before 500BC; the world population increases fast in the last two centuries, but the population of notable people increases faster and is more in line with GDP growth.

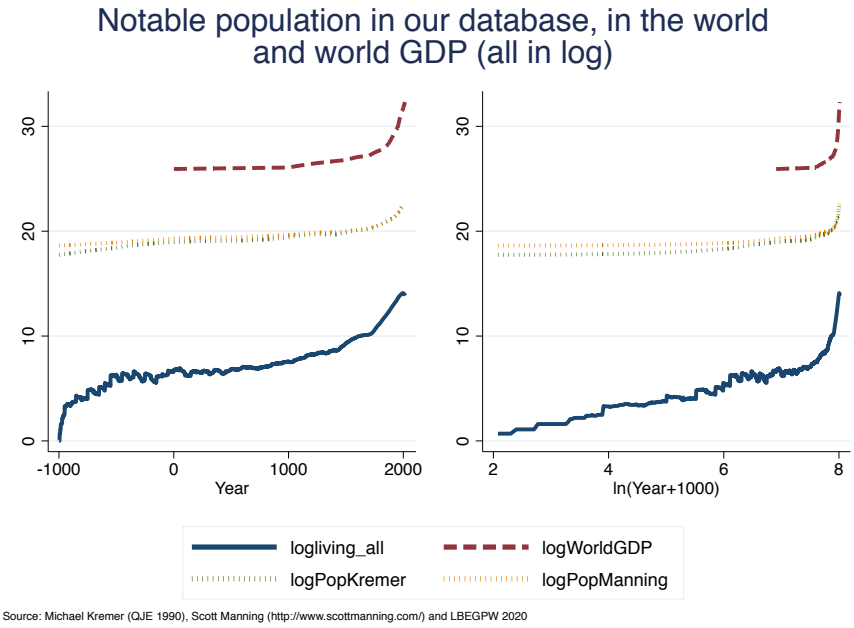
Figure 18 provide the split over 4 sub-periods of the ratio of world population to the population of notable individuals. Our database population contains approximately one person out of 250 000 before 500AD, the ratio then declines continuously until one out of 50 000 in 1500AD, still declines yet reaches a local maximum in 1700 due to a larger mortality in our database, and reaches a minimum of 1 over 3200 in 1950. Afterwards, the ratio goes up again due to fewer people in the database: famous people become so later in their career and more currently born young people will enter the database in the next decades but are not identified yet.

We next run a regression of the log of the ratio of the population in the database to the world population over time. More precisely, denoting by t the calendar time and $lnX(t)$ the log defined above in each year, we estimate

$$lnX_t = a + b \times (2018 - t)$$

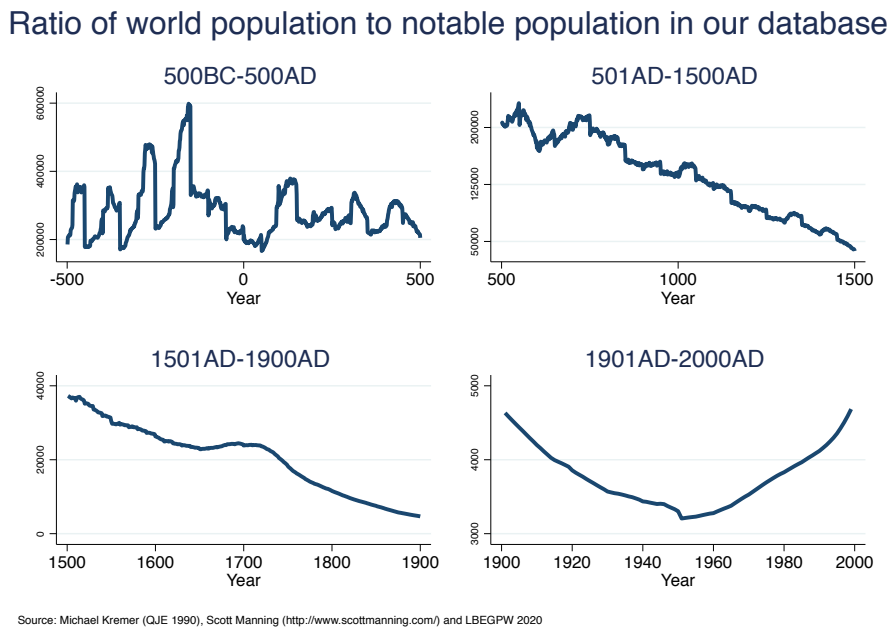
The coefficient b is negative and tells us how an additional year of distance to present times leads to a percentage decline in the number of famous people relative to the world population at that time. We find on the restricted dataset that $b = -.0016465$ with a s.d of .0000146. The rate at which the fraction of famous people declines after T periods is therefore $1 - (1 - b)^T$ which is 15.2% each century, or 56.1% after 500 years, or 80.8% after 1000 years.

Figure 17: **Time evolution of GDP, world population and population in the database**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Individuals with more than one biography account for one observation to avoid double counting.

Figure 18: **World population relative to the number of notable individuals**



Notes. Restricted sample (at least one `Wikipedia` edition among the 7 European languages analyzed). Individuals with more than one biography account for one observation to avoid double counting.

A.4 Structure of the database across language editions in `Wikipedia`

In this section, we describe the recursive structure of the database. We list, following an iterative elimination process, the most popular `Wikipedia` editions in decreasing order. For instance, once all individuals with a biography in the English edition have been removed we find 340,913 individuals absent from this edition but present in the German edition, of which 259,013 have a unique biography in this language, etc.

Table 7: Marginal contribution of each language edition

1-25 top language editions				26-50 next language editions			
Edition	Region	#unique	#total	Edition	Region	#unique	#total
English	En	663 930	1 579 940	Slovenian	We	12 193	12 902
German	We	259 013	340 913	Lithuanian	We	11 941	12 305
Japanese	Ea	157 707	179 466	Azerbaijani	EuAr	11 499	12 366
French	We	114 820	155 391	Persian	EuAr	10 441	10 888
Russian	EuAr	98 514	156 728	Romanian	We	10 434	10 644
Chinese	Ea	90 896	97 177	Greek	We	9 699	10 046
Polish	We	86 789	98 407	Indonesian	Ea	9 304	10 883
Spanish	We	72 013	97 540	Esperanto	We	7 953	8 125
Italian	We	63 944	70 263	Armenian	EuAr	7 795	7 852
Swedish	We	56 578	65 719	Kazakh	EuAr	7 178	7 213
Dutch	We	46 215	50 114	Thai	Ea	7 003	7 276
Portuguese	We	43 351	44 446	Galician	We	6 817	6 921
Ukrainian	We	36 886	38 003	Vietnamese	Ea	5 659	5 671
Finnish	We	30 909	31 715	Serbian	We	5 162	6 804
Czech	We	26 670	29 073	Slovak	We	4 693	4 707
Catalan	We	25 866	27 103	Croatian	We	4 484	6 861
Arabic	EuAr	25 585	29 227	Basque	We	3 911	3 935
Korean	Ea	25 000	25 377	Albanian	We	3 791	3 812
Hungarian	We	24 531	33 334	Luxembourgish	We	3 536	3 588
Norwegian (Bokmål)	We	23 433	26 944	Malay	Ea	3 237	4 029
Hebrew	EuAr	17 915	18 540	Belarusian	We	3 174	4 547
Bulgarian	We	17 729	19 847	Latvian	We	3 156	3 198
Estonian	We	15 443	15 793	Haitian	We	2 927	2 933
Danish	We	14 769	14 983	Tagalog	Ea	2 700	2 756
Turkish	EuAr	13 248	14 020	Hindi	Ea	2 573	2 883

51-75 next language editions				76-100 next language editions			
Edition	Region	#unique	#total	Edition	Region	#unique	#total
Afrikaans	South	2 322	2 333	Bashkir	EuAr	519	521
Telugu	East	2 045	2 056	Swahili	South	503	505
West Frisian	West	2 034	2 055	Tongan	South	503	504
Georgian	EuAr	1 992	2 024	Bosnian	West	490	493
Welsh	West	1 865	1 905	Burmese	East	403	404
Marathi	East	1 729	1 734	Chuvash	EuAr	367	370
Tatar	EuAr	1 723	1 844	Kurdish	EuAr	347	364
Bengali	East	1 633	1 652	Piedmontese	West	304	319
Icelandic	West	1 601	1 602	Amharic	South	300	301
Kirghiz	EuAr	1 590	1 604	Alemannic	West	272	278
Norwegian (Nynorsk)	West	1 541	1 544	Occitan	West	262	265
Tamil	East	1 508	1 517	Faroese	West	261	265
Volapük	West	1 366	1 369	Scots	West	255	256
Sakha	EuAr	1 360	1 365	Central Bicolano	East	224	226
Macedonian	West	1 210	1 214	Asturian	West	216	217
Urdu	East	1 085	1 154	Nepali	East	210	243
Tajik	EuAr	914	923	Yiddish	EuAr	210	211
Low Saxon	West	732	733	Gujarati	East	208	214
Latin	West	611	619	Pashto	EuAr	183	185
Mongolian	East	591	592	Aragonese	West	181	182
Breton	West	587	596	Malagasy	East	181	184
Cantonese	East	570	571	Irish	West	175	181
Uzbek	EuAr	566	570	Sicilian	West	164	166
Oriya	East	557	559	Limburgish	West	133	136
Walloon	West	523	535	Scottish Gaelic	West	128	129

Notes. Exhaustive sample (4.7 million individuals). The acronyms *We, Ea, EuAr, Sn* are defined in Table 1, and correspond to groups of language edition of `Wikipedia`. Numbers in this table slightly differ from numbers in Table 1 in that these are based on language editions as per `Wikidata`. In Table 1 instead, we used language editions as they appear in `Wikipedia` biographies, which is more relevant for our data extraction based on the 7 language editions of `Wikipedia`. In addition, English in this table includes Old English and Simplified English.

We also report in Table 8 the most famous individuals in the different language editions in recursive order, e.g. an individual in the Eastern language edition does not have a biography in English nor in any of the Western language block.

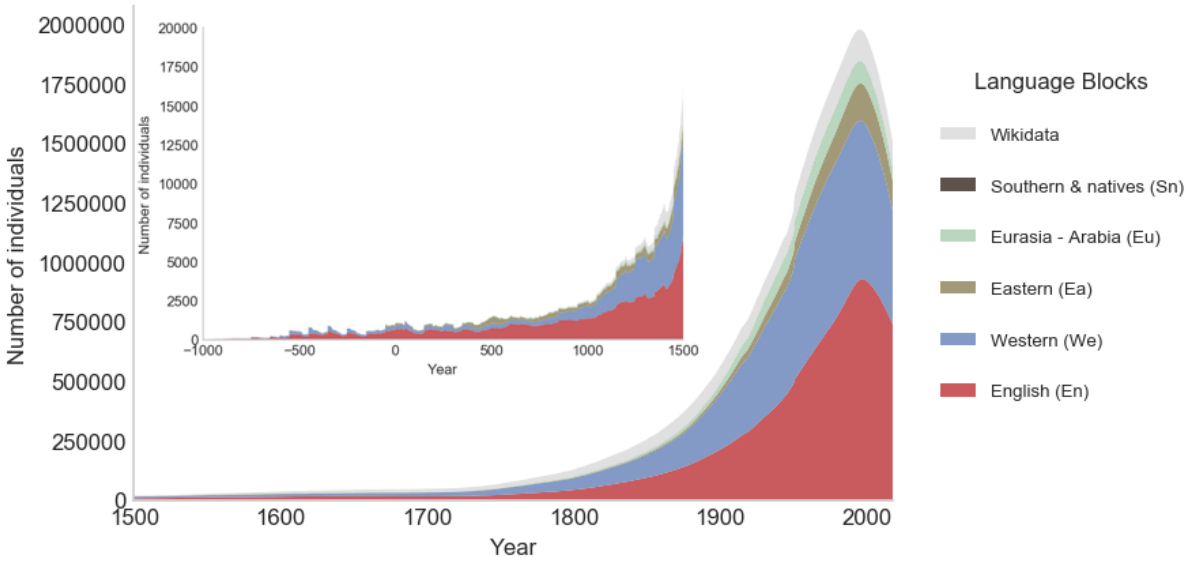
Table 8: **Visibility index: top 5 individuals in each recursive language block**

Wikipedia *(recursive lang. blocks)*	
English	Barack Obama, Donald Trump, Leonardo da Vinci, Adolf Hitler, Albert Einstein
Western	Blas de Otero, Anita Blonde, Olivier Nakache, Kristina Rose, Sophie Dee
Eastern	Qays Ibn al-Mulawwah, Husain Wāi Kāšifī, Gorō Kishitani, Ryō Iwamatsu, Miyu Takeuchi
Eurasia - Arabia	Aşık Paşa-yı Velî, Erdal Tosun, Roma Acorn, Georgiy Mirskiy, Qayum Nasıyri
Southern and natives	Boerneef, Pieter Pieterse, Jan Blohm, Frank Rautenbach, Tolla van der Merwe
`Wikidata` *only*	Martin Hardie, Lilly Wachowski, John Charles Robinson, Caspar Luyken, Ernest Henri Griset

Notes. Most famous individuals by recursive language block, e.g. absent from above language blocks, exhaustive database.

A.5 Evolution of the sample size in the exhaustive database

Figure 19: **Time evolution of the number of individuals present in the database in a given year**



Notes. Exhaustive database. English (*En*): individuals present in the English edition; Western (*We*): individuals absent from the *En* edition but present in *We* editions; Eastern (*Ea*): individuals absent from the *En* & *We* blocks but with at least one biography in editions of the *Ea* block; Eurasia-Asia (*EuAr*): individuals absent from the previous blocks (*En*, *We*, *Ea*) but present in at least one *EuAr* edition. Southern & natives (*Sn*): individuals absent from the other blocks (*En*, *We*, *Ea*, *EuAr*) but present in at least one edition of the *Sn* block. **`Wikidata` only** includes individuals with a `Wikidata` biography only. See Table 1 for precise definitions of these groups and sub-groups. Individuals with more than one biography account for one observation to avoid double counting.

B Data processing for the restricted sample

B.1 Details on the allocation into domains of influence

The easy cases are when `Wikipedia`'s and `Wikidata`'s keywords characterizing an occupation or a domain of influence converge towards two identical modal occupations across sources. When this information diverges, we generally give more credit to `Wikidata`. We however make an exception to this rule when there is a tie between the modes in `Wikidata` and instead a clear, unique, mode in `Wikipedia`. In this case, we favor `Wikipedia`. In the more problematic case in which both `Wikidata` and `Wikipedia` give several modes, we pool all keywords together and determine the mode from this combined list.

B.2 Details on the definition and creation of citizenships

The easy cases are the ones where the information on citizenship from `Wikidata` and `Wikipedia` match. When they instead contradicted each other, we retained information from `Wikipedia` if and only if it matched with a time invariant citizenship from `Wikidata`; or if the information obtained from `Wikipedia` was present in the Infobox of at least one language edition scraped; otherwise we assigned the citizenship from `Wikidata`. The reason why we give more credit to `Wikidata` here is that the code written to extract this information in `Wikipedia` may make more mistakes as it needs to crawl the entire content of the biography to detect one or several citizenships that do not necessarily belong to the individual. Lastly, in case the citizenship information is absent from one universe, we use the most frequent citizenship(s) found in the other universe.

Next, we matched citizenships and political entities at the time of the individuals life, using information on the creation of modern states to determine whether we should assign the individuals to the new or the old regime of the country. The old regimes also encompass all political entities broadly situated in the current geographical location of the modern state. For e.g.: Erstwhile colonies under the British empire such as India get divided under Old regimes of the country (India) vs India based on their independence date. The Mughal Empire, the Chola and Chela Kingdom get classified under Old regimes of India too. In cases where both birth and death dates of the individual are before (after) the date of foundation of the modern state, we assign the individual to old regime (modern regime). When the birth date is before and the death date is after the foundation of the modern state, we assign the individual to the new regime if and only if the modern state was explicitly mentioned as one of the citizenships in the disaggregated information collected from `Wikidata` in the first step, otherwise we assign her to the old regime. The citizenship for individuals assigned to the old regime reads as Old_(before_year_xx)_YY where xx refers to the threshold year used to demarcate the old regimes from the modern state and YY refers to the name of the modern state.[19]

A large number of individuals have two citizenships, either because they are true bi-nationals (e.g. Indian and US citizens) or because the country they were born in, disappeared or separated from a larger entity (for example, Bosnia and Herzegovina from Yugoslavia in the 90's). We therefore decided to report up to two citizenships in the database for a better coverage.

The thresholds used to demarcate old political and geographical regimes from the modern state for each nation state are available at: List of sovereign states by date of formation.

B.3 Further details on relevant individuals for historical periods with no biography in English editions

- Around the French Revolution, there are several notable political figures, not documented yet but present in our database, include Louis-Alexandre-Céleste d'Aumont, Anne-Christian de Montmorency-

[19] Akbar's (the Mughal emperor) citizenship would read as Old_(before_year_1947_AD)_India

Luxembourg, both members of the Chamber of Peers. The Chambre des Pairs in French was the upper house of the French parliament from 1814 to 1848. One last example is Pierre-Charles-Louis Baudin (1748-1799, Politician, French) who has two biographies, one in English and one in French only although he played a significant role during the French revolution as President of the National Convention in 1795. Eugène Chevandier de Valdrome (French, 1810-1878) offers a good illustration of the marginal contribution of the French edition. He was a politician (Interior Minister, political party leader, etc.) on the one hand and an influential business man on the other side. To date, he does not have any existence in the English edition.

- In the set of biographies in German within the same category, one similarly finds many interesting profiles. The 18th century in Germany corresponds to the early modern period and the political rise of the Prussian empire and more precisely from 1713 to 1740, a period during which King Frederick William I, also known as the "Soldier King", established a highly centralized state.
- Manuel da Silva Passos (Portuguese, 1801–1862) is a good example of notable individual, as illustrated by his quite extensive biography in the Portuguese edition, absent from the English edition until very recently and present in our database. Indeed, his biography was translated into English in 2019 only. It is still much shorter and less informative than its Portuguese equivalent. Passos Manuel was a Portuguese jurist and politician, and considered as one of the most notable personalities of 19th-century Portuguese Liberalism.
- As regards to inventors not in English `Wikiepdia`, one can mention Caspar Schmalkalden (1616-1673, explorer, topographer, military, writer): he was a German citizen who contributed in a significant way to the exploration of South America and the East indies. Although his contribution is acknowledged by a long and very detailed biography in the German edition, this explorer has no equivalent in the English edition to date. One can also cite François de Taxis (1459-1517, business man), member of the Tasso family, founder of the European postal service. He is to date curiously absent from the English edition despite his important contribution to the development/invention of one of the most important communication services of all time. Andrea Bianco (XVth century, sailer and cartographer) became famous after he published in 1436 his Altas which is considered a significant contribution to the age of exploration. He is present in 7 `Wikipedia` editions but still does not have a biography in English.

C Test protocol

C.1 Pilot

10 RAs from Science Po Paris, NYUAD and Delhi School of Economics received a sample of 1000 individuals to test.

C.1.1 Instructions

- We will give you 1000 individuals from various notability levels, and ask you to check and validate or report mistakes on 6 different pieces of information: exact or approximate date of birth and death; gender; main occupation and possibly secondary occupation; citizenship or equivalent concept for earlier periods of history.
- You will be asked to report the verification in the google sheet next to each information. “Correct” means no error, “Error” means certain error, “Missing” means the information is included in `Wikipedia`/`Wikidata` but not present in the dataset, “Other case” means possible error. Judgment is required from you.
  - For instance, if there is a historical controversy and several sources differing, report “Other case” unless there is an obvious mistake in our database.

  - It will be particularly the case for the retained citizenship that is sometimes selected among a list of ten or more different geographical areas, kingdom, franchised cities, duchy, caliphate etc., the borders of which evolved during the life of the individual.
  - The information on birth (and death if relevant) is sometimes approximated by $birth\_min$ or $birth\_max$ (by $death\_min$ or $death\_max$). For instance, someone only known for being born in the 12th century will be reported as $birth\_min = 1101$ and $birth\_max = 1201$ and $birth\_b = N/A$

- Description of variables
  - $birth\_b$ = date of birth (exact)
  - $death\_b$ = date of death (exact)
  - $birth\_min\_b$ = minimum date of birth (intervals because approximation)
  - $birth\_max\_b$ = maximum date of birth (intervals because approximation)
  - $death\_min\_b$ = minimum date of death (intervals because approximation)
  - $death\_max\_b$ = maximum date of death (intervals because approximation)
  - $gender\_b$ = gender
  - $final\_occupation$ = 1st final occupation (Level 2)
  - $freq\_1stoccu$ = Frequency associated to 1st occupation (Level 2)
  - $final\_second\_occup$ = 2nd final occupation (Level 2)
  - $freq\_2ndoccu$ = Frequency associated to 2nd occupation (Level 2)
  - $keyword\_used$ = keyword used to define 1st final occupation
  - $area1\_of\_ratt$ = 1st Citizenship (distinction current/former country)
  - $area2\_of\_ratt$ = 2nd Citizenship (distinction current/former country)
  - $euro7\_editions$ = availability of 7 European language Editions
- Cross-verification: A part of the sample is common to other research assistants to assess the accuracy of your work. There will be an end of contract reward of up to 12.5% of the contract for the quality of the work.
- Remember that the goal is neither to minimize nor to maximize the number of spotted errors but to detect and provide a fair assessment of the quality of the database. Keep all your comments and suggestions on the spreadsheet as it may be requested by editors of scientific journals. In case of doubt, report "Other case" as explained above, and the reason for the doubt about the information contained in the database.

At the end of the pilot, we looked at the various errors detected. In particular, as regards to occupations, we noticed that when the frequency of the second occupation was below 0.25, there was a large proportion of errors; we decided to set this as a threshold, since it preserves many true positive regarding the second occupation.

C.2 Final test

See the text.

C.2.1 Instructions, final set

- We will give you 1000 individuals from various notability levels, and ask you to check and validate or report mistakes on 6 different pieces of information: exact or approximate date of birth and death; gender; main occupation and possibly secondary occupation; citizenship or equivalent concept for earlier periods of history.
- You will be asked to report the verification in the google sheet next to each information. "Correct" means no error, "Error" means certain error, "Missing" means the information is included in `Wikipedia`/`Wikidata` but not present in the dataset, "Other case" means possible error. Judgment is required from you.
  - For instance, if there is a historical controversy and several sources differing, report "Other case" unless there is an obvious mistake in our database.

  - It will be particularly the case for the retained citizenship that is sometimes selected among a list of ten or more different geographical areas, kingdom, franchised cities, duchy, caliphate etc., the borders of which evolved during the life of the individual.
  - The information on birth (and death if relevant) is sometimes approximated by $birth\_min$ or $birth\_max$ (by $death\_min$ or $death\_max$). For instance, someone only known for being born in the 12th century will be reported as $birth\_min = 1101$ and $birth\_max = 1201$ and $birth\_b = N/A$

- Description of variables
  - $birth\_b$ = date of birth (exact)
  - $death\_b$ = date of death (exact)
  - $birth\_min\_b$ = minimum date of birth (intervals because approximation)
  - $birth\_max\_b$ = maximum date of birth (intervals because approximation)
  - $death\_min\_b$ = minimum date of death (intervals because approximation)
  - $death\_max\_b$ = maximum date of death (intervals because approximation)
  - $gender\_b$ = gender
  - $final\_occupation$ = 1st final occupation (Level 2)
  - $freq\_1stoccu$ = Frequency associated to 1st occupation (Level 2)
  - $final\_second\_occup$ = 2nd final occupation (Level 2)
  - $freq\_2ndoccu$ = Frequency associated to 2nd occupation (Level 2)
  - $keyword\_used$ = keyword used to define 1st final occupation
  - $citizenship\_1\_b$ 1st Citizenship (no distinction current/former country)
  - $citizenship\_2\_b$ 2nd Citizenship (no distinction current/former country)
  - $year_creation_state$1 and 2: year of the creation of the modern state in $area1_of_ratt$1 or 2
  - $euro7\_editions$ = availability of 7 European language Editions.

- Cross-verification: A part of the sample is common to other research assistants to assess the accuracy of your work. There will be an end of contract reward of up to 12.5% of the contract for the quality of the work.
- Remember that the goal is neither to minimize nor to maximize the number of spotted errors but to detect and provide a fair assessment of the quality of the database. Keep all your comments and suggestions on the spreadsheet as it may be requested by editors of scientific journals. In case of doubt, report "Other case" as explained above, and the reason for the doubt about the information contained in the database.

Table 9: **Manual verifications: discrepancy between RAs for each variable**

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Mismatch (# Obs)	142	69	75	207	467	217	128
(%)	2.84	1.38	1.50	4.14	9.34	4.34	2.56

Notes. This table provides the numbers and rates of discrepancy, when independent RAs did not report the same outcomes among Correct, Error, Missing, Other case for the same individual. The first row gives the number and the second row gives the frequency.

Table 10: **Manual verifications: summary statistics**
Sample: mix of sub-samples

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Information not reported in our database							
No info in sources	12.12	53.45	2.04	3.66	64.13	6.76	92.24
Info. updated since collec.	0.90	0.44	3.42	2.76	2.26	1.02	1.14
Information reported in our database							
Correct	84.68	45.01	94.44	92.50	28.21	91.30	5.96
Error	0.64	0.22	0	1.00	5.04	0.80	0.56
Other case (ambiguity or info. updated since collec.)	1.66	0.88	0.10	0.08	0.36	0.12	0.10
Total # cases	4,999	4,999	4,999	4,999	4,999	4,999	4,999

Notes. Test sample on a mix of the exhaustive and restricted database (at least one `Wikipedia` edition among the 7 European languages analyzed) with over sampling, see text. This table provides some summary statistics on manual checks. The different possible outcomes are" "No info in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in our dataset, "Info updated since data collection" means the information is included in `Wikipedia`/`Wikidata` but not present in our dataset; "Information Correct" means no error, "Error" means certain error, "Other case" means possible error (for instance historical controversy, several sources differing or information updated since data collection). These checks have been conducted by the 10 RAs and cross-verified by a PhD researcher.

Table 11: **Manual verifications: summary statistics**
Sample: 1+ Europ. `Wikipedia` eds.

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Information reported in our database							
No info. in sources	7.80	55.10	0	0.27	61.27	1.17	95.10
Info. updated since collec.	0.70	0.37	0.10	0.50	1.43	0.83	1.03
Information not reported in our database							
Correct	89.00	43.50	99.77	97.60	29.97	96.87	3.43
Error	0.87	0.27	0	1.53	6.83	0.97	0.37
Other case (ambiguity or info. updated since collec.)	1.63	0.77	0.13	0.10	0.50	0.17	0.07
TOTAL	3,000	3,000	3,000	3,000	3,000	3,000	3,000

Notes. Test sample on the restricted database (at least one `Wikipedia` edition among the 7 European languages analyzed) with over sampling of the top and of the bottom, see text. This table provides some summary statistics on manual checks. The different possible outcomes are: "No info in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in our dataset, "Info updated since data collection" means the information is included in `Wikipedia`/`Wikidata` but not present in our dataset; "Information Correct" means no error, "Error" means certain error, "Other case" means possible error (for instance historical controversy, several sources differing or information updated since data collection). These checks have been conducted by the 10 RAs and cross-verified by a PhD researcher.

Table 12: **Manual verifications: summary statistics**
Sample: top 1000 indiv.

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Information not reported in our database							
No info. in sources	0.20	33.13	0.00	0.00	55.66	0.00	80.28
Info. updated since collec.	0,00	0.00	0.00	0.00	6.11	0.00	1.90
Information reported in our database							
Correct	98.00	66.17	100.00	100.00	36.54	99.30	16.32
Error	0.30	0.20	0.00	0.00	1.70	0.70	1.40
Other case (ambiguity or info updated since collec.)	1.50	0.50	0.00	0.00	0.00	0.00	0.10
TOTAL	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Notes. Test sample on the top 1000 most notable of the database. This table provides some summary statistics on manual checks. The different possible outcomes are: "No info in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in our dataset, "Info updated since data collection" means the information is included in `Wikipedia`/`Wikidata` but not present in our dataset; "Information Correct" means no error, "Error" means certain error, "Other case" means possible error (for instance historical controversy, several sources differing or information updated since data collection). These checks have been conducted by the 10 RAs and cross-verified by a PhD researcher.

Table 13: **Manual verifications: summary statistics**
Sample: 2+ Europ. `Wikipedia` eds.

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Information not reported in our database							
No info in sources	3.80	56.80	0.00	0.00	66.40	0.00	91.80
Info. updated since collec.	0.40	0.60	0.00	0.20	0.40	0.40	1.20
Information reported in our database							
Correct	93.60	41.80	100.00	99.20	26.80	98.80	6.20
Error	0.60	0.20	0.00	0.60	6.00	0.80	0.60
Other case (ambiguity or info updated since collec.)	1.60	0.60	0.00	0.00	0.40	0.00	0.20
TOTAL	500	500	500	500	500	500	500

Notes. Test sample on the subset of the restricted database (at least **two** `Wikipedia` editions among the 7 European languages analyzed). This table provides some summary statistics on manual checks. The different possible outcomes are: "No info in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in our dataset, "Info updated since data collection" means the information is included in `Wikipedia`/`Wikidata` but not present in our dataset; "Information Correct" means no error, "Error" means certain error, "Other case" means possible error (for instance historical controversy, several sources differing or information updated since data collection). These checks have been conducted by the 10 RAs and cross-verified by a PhD researcher.

Table 14: **Manual verifications: summary statistics**
Sample: `Wikidata` only, no `Wikipedia`

	Birth	Death	Gender	1st Occupation	2nd Occupation	1st Citizenship	2nd Citizenship
Information not reported in our database							
No info. in sources	70.20	80.80	20.40	35.00	96.00	60.60	99.40
Info. updated since collec.	4.40	1.60	33.60	24.40	1.40	4.80	0.20
Information reported in our database							
Correct	23.20	15.00	45.80	40.20	2.40	34.40	0.20
Error	0.00	0.00	0.00	0.20	0.00	0.00	0.00
Other case (ambiguity or info. updated since collec.)	2.20	2.60	0.20	0.20	0.20	0.20	0.20
TOTAL	500	500	500	500	500	500	500

Notes. Test sample on the those with **no `Wikipedia`** edition among the 7 European languages analyzed). This table provides some summary statistics on manual checks. The different possible outcomes are: "No info in sources" means the information is not included in `Wikipedia`/`Wikidata` nor in our dataset, "Info updated since data collection" means the information is included in `Wikipedia`/`Wikidata` but not present in our dataset; "Information Correct" means no error, "Error" means certain error, "Other case" means possible error (for instance historical controversy, several sources differing or information updated since data collection). These checks have been conducted by the 10 RAs and cross-verified by a PhD researcher.