

# scientific data



OPEN

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

DATA DESCRIPTOR

Morgane Laouenan<sup>1,2</sup>✉, Palaash Bhargava<sup>3</sup> , Jean-Benoît Eyméoud<sup>2</sup>, Olivier Gergaud<sup>2,4</sup>, Guillaume Plique<sup>5</sup> & Etienne Wasmer<sup>6,2</sup>✉

A new strand of literature aims at building the most comprehensive and accurate database of notable individuals. We collect a massive amount of data from various editions of Wikipedia and Wikidata. Using deduplication techniques over these partially overlapping sources, we cross-verify each retrieved information. For some variables, Wikipedia adds 15% more information when missing in Wikidata. We find very few errors in the part of the database that contains the most documented individuals but nontrivial error rates in the bottom of the notability distribution, due to sparse information and classification errors or ambiguity. Our strategy results in a cross-verified database of 2.29 million individuals (an elite of 1/43,000 of human being having ever lived), including a third who are not present in the English edition of Wikipedia. Data collection is driven by specific social science questions on gender, economic growth, urban and cultural development. We document an Anglo-Saxon bias present in the English edition of Wikipedia, and document when it matters and when not.

### Background & Summary

Since Plutarch's *Parallel Lives* in the beginning of the second century AD and his 23 biographies, or the Epic of Gilgamesh that further dates back to 2000BC, the task of registering famous individuals and their influence has been a recurrent field of study. It is both a source of fascination for the general public and a resource for scholars as it captures aspects of societies in different centuries. Over the last few years, this task has been undertaken at 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 approach was pioneered in digital humanities<sup>1</sup> with a focus 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. Freebase and the English edition of Wikipedia was also used to assemble a manually-verified database of 11,341 individuals present in more than 25 language editions of Wikipedia called Pantheon 1.0<sup>2</sup>. 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<sup>3</sup> or Wikidata were built automatically from the information retrieved from the Infobox and multiple categories of Wikipedia.

Such projects represent promising developments for the Social Sciences. The database we document here, thanks to its time and spatial coverage and its granularity, is suited to explore gender differences, demography, urban development, cultural transmission, human capital, growth, institutions. See sub-section *Outreach* for an exhaustive list. A good example is a study of the historical development of cities<sup>4</sup>. The authors relate it to the emergence of a creative class using 40,000 individuals, among which 22,000 creatives that are then matched to a sample of 2,200 European cities between 800 and 1800AD<sup>1</sup>. A sample of 300,000 famous people born between Hammurabi's epoch and 1879, Einstein's birth year from Index Bio-bibliographicus Notorum Hominum was built to estimate the timing of improvements in longevity and its role on economic growth<sup>5</sup>. The second version of Pantheon (2.0)<sup>6</sup> also addresses long-run social science trends and offers more statistical power in including around 71,000 with 14+ biographies.

While there has always been a debate regarding the accuracy of Wikipedia as a reliable source in its early years<sup>7</sup> pointing out more errors than in traditional Encyclopedia, we rejoin other judgements that the information present is highly reliable. In fact, the main problem arises from omissions<sup>8</sup>, that will naturally decrease as

<sup>1</sup>CNRS, Université Paris 1 Panthéon-Sorbonne, Centre d'Economie de la Sorbonne, Paris, 75013, France. <sup>2</sup>LIEPP, Sciences Po, Paris, 75007, France. <sup>3</sup>Department of Economics, Columbia University, New York, 10027, USA. <sup>4</sup>KEDGE Business School, Paris, 75012, France. <sup>5</sup>Médialab, Sciences Po, Paris, 75007, France. <sup>6</sup>New York University Abu Dhabi, Social Science Division, Abu Dhabi, 129188, UAE. ✉e-mail: [morgane.laouenan@univ-paris1.fr](mailto:morgane.laouenan@univ-paris1.fr); [ew1555@nyu.edu](mailto:ew1555@nyu.edu)

Wikipedia expands. Our dataset must be seen as an evolving object. Our contribution is primarily a methodological one, how to make use of the information available and as accurately as possible. Our next goal is to build a comprehensive and accurate Knowledge Base of notable individuals for immediate use in the social sciences and in particular econometric and statistical analyses. It 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 Wikipedia editions and Wikidata) and deduplication techniques. Combining Wikipedia and Wikidata brings 2.72% new birth dates, 8.16% new occupations and 17.16% new citizenships. We find that there are a limited number of 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 an appropriate statistical treatment of these errors in statistical approaches. Combining Wikipedia and Wikidata allows to correct about 0.5% of all detected errors. We build and make accessible a database with a limited number of errors. The scope of the final database does not cover all Wikipedia, but only when the content could be cross-verified between different language editions and Wikidata. Second, data collection is driven by specific social science questions on gender, economic, growth, urban and cultural development. We 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 2.29 million unique individuals (an elite of 1/43,000 of human being having ever lived) among which 30% come from the 6 non-English editions of Wikipedia, a significant improvement over earlier works that have only focused on English Wikipedia only. We verified our algorithms with 5,000 manual checks. Error rates vary by sources and notability levels. We therefore recommend to limit the database to individuals with at least two sources and not to use an exhaustive database of 4.7 million people. The large number of individuals in the database is the main distinctive feature of our work. Lesser-known individuals may have contributed in a significant way to the social, economic or cultural development of their community (city, region or country) in fostering innovation and growth in their cluster. To accurately measure the relative influence of each individual present in our database, we build a notability index extending earlier work<sup>2</sup> and allowing to study different groups of the distribution of notability. Note that the ranking should not be used to compare pairs of individuals but rather to illustrate the differences in demographics or sectors across quantiles of notability.

## Methods

**Universe described.** Social scientists are interested in significant people insofar as they have influenced the trajectory of the society in their times or later on. Let us call  $\mathcal{S}$  the universe of these significant 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. Let us call  $\mathcal{D}$  the universe of already detected 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 famous individuals. Some of them will be part of our data collection process, from Wikipedia and Wikidata (let us call this set  $\mathcal{W}$ ), while others are in known sources but unexploited yet, or not yet in the Wiki universe (let us denote it  $\mathcal{U}$ ). Some individuals may have been significant and even remembered for a while, but are currently forgotten. We can call this set  $\mathcal{F}$ . To sum up, the theoretical universe of significant individuals is decomposed as:

$$\mathcal{S} = \mathcal{F} \cup \mathcal{D} = \mathcal{F} \cup \mathcal{W} \cup \mathcal{U}$$

We try to reach  $\mathcal{S}$  by  $\mathcal{W}$  with automatic routines. In the future, unexploited archives in  $\mathcal{U}$  will gradually converge to the Wiki universe  $\mathcal{W}$ , while forgotten individuals in  $\mathcal{F}$  will - albeit slowly - be gradually rediscovered by historians, or scholars in the arts or science disciplines. Therefore, further editions of this database will cover more individuals as more information will become available in Wikipedia and Wikidata in the future.

**A social science background.** Social scientists develop models to illustrate causal mechanisms. As a result, we need to collect data in a way that fits the typology of these models. To encompass a wide range of human activities and domains of influence, we follow the seminal work on the so-called Baumol's cost disease<sup>9</sup>. This refers to a situation in which a sector without productivity gains (such as performing arts until the beginning of the XXth century) faces higher wages due to productivity gains in other sectors. We assume that the amount of global wealth produced by humans in a given year is of two types,  $Y^M$ ,  $Y^I$ , where  $Y^M$  is the amount of materialistic goods and services generated by the commonly-called productive sector, and  $Y^I$  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  $Y^M$ , unlike  $Y^I$  mostly made of labor-intensive activities, will benefit from science innovations and technological progress: the famous Magic flute by Wolfgang Amadeus Mozart still requires the same amount of resources (time, performers, space, etc.) nowadays as 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, implying that the relative demand for the arts would actually increase and this prevents the arts sector from continuously declining 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 and capital. The level of

total factor productivity  $\mathcal{A}$  combines the various outputs of science, education, technology and discoveries of all kinds denoted by  $\mathcal{B}$ :

$$Y^M = \mathcal{A}f(\mathcal{B}). \quad (1)$$

while the second component  $Y^I$  mostly depends on the number of famous individuals in creative activities. Last, the consumption vector  $\vec{C} = (C^M, C^I)$  is only a fraction  $(1 - \tau)$  of the amount of global wealth as follows:

$$\vec{C} = \vec{C}[(1 - \tau)(Y^M + Y^I)], \quad (2)$$

where  $(1 - \tau)$  is typically determined by the way a society is governed by politicians, military, religious authorities, noble people, and  $\tau$  is an administration tax, featuring more or less efficient societies and resources diversion<sup>10</sup>. The demand function  $\vec{C}$  is such that the first dimension ( $M$ ) is a necessity good and the second dimension ( $I$ ) gets a higher demand as income and wealth grow. This framework will help us understand long-run trends in the relative importance of different fields among notable individuals: each broad occupation defined later in the data will be associated with one of the theoretical concepts from the paragraphs above.

**Preview of the data.** The database is organized as follows: each row corresponds to one individual, with a name and the unique identifier (section *Extracting raw data from Wikidata and Wikipedia*), the birth date and gender (section *Demographics: birth, death and gender*), an imputation if missing, a range if known from a decade or century, up to two fields called domains of influence and an estimated weight for each domain (section *Domains of influence and occupations*), birthplace and death place, longitude and latitude (section *Data records*), a proxy for citizenship (section *Citizenship*) and a notability index calculated from 5 variables (section *Measuring notability*). The sample choice is described in section *Extracting biographic information from a restricted sample*.

**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, we consider two main sources of information: information from Wikidata and information contained in 7 language editions of Wikipedia. We extract key information on humans present either in one or both sources, and then merge this information to avoid 1) dealing with duplicate individuals and 2) cross-verify the information contained in both universes.

**Wikipedia.** This paper uses 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 page. All existing categories about individuals were systematically analyzed and retrieved by scraping. 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 the *European editions*. Within the sample of individuals with at least one biography in the 7 language editions, 52.8% have 1 biography, 18.1% have 2 biographies (cumulative 70.9%), 8.7% have 3 biographies (cumulative 79.6%), 5% have 4 biographies (cumulative 87.8%) and 12.2% have 5 biographies or more. See Appendix A for a detailed description of the method.

**Wikidata.** In the Wikidata universe, we use the “instance of humans” Q5 category to define our sample of individuals. 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 are retrieved. In addition to this first set of variables, we also extract the exhaustive list of Wikipedia urls contained in each Wikidata entry, to deal with multiple biographies as explained below. We also collect the sources behind the Wikidata page, used in Figure S7 in Appendix.

**Merging Wikidata and Wikipedia.** Wikidata and Wikipedia contradict each other 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.

**Humans and non-humans.** A large number of entries are wrongly identified as humans in Wikidata because they contain the Q5 code which stands for ‘instance of humans’ but some may not be 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. Let us illustrate the difficulty with a few examples. <https://www.wikidata.org/wiki/Q656193> is an example of a page of a real human in the Czech edition of Wikipedia linked to fictional characters, yet as of the time of data extraction, wrongly classified as a human. A slightly more ambiguous case is that of <https://www.wikidata.org/wiki/Q1278485>, a possibly fictional character subject of a story in Ancient Rome - and in any event the pages describe the person

	All Wikipedia & Wikidata		At least one European edition		At least 14 editions	
	Freq.	%	Freq.	%	Freq.	%
	Exhaustive sample		Restricted sample		Pantheon 2.0	
Wikipedia editions (by recursive language groups)						
English (En)	1,578,917	33.8	<b>1,547,174</b>	67.5	76,139	99.7
Western (We)	1,327,543	28.4	<b>744,643</b>	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	<b>2,291,817</b>	100	76,379	100

**Table 1.** Marginal contribution of different groups of Wikipedia language editions. 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* groups but with at least one biography in editions of the *Ea* group; Eurasia-Asia (*EuAr*): individuals absent from the previous groups (*En*, *We*, *Ea*) but present in at least one *EuAr* edition. Southern & natives (*Sn*): individuals absent from the other groups (*En*, *We*, *Ea*, *EuAr*) but present in at least one edition of the *Sn* group. 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.

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

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

as a fictional character, while she is classified as human in Wikidata. In total, we identify 20,000 entries corresponding to 16,000 questionable “pseudo-individuals” that we eliminate from our database.

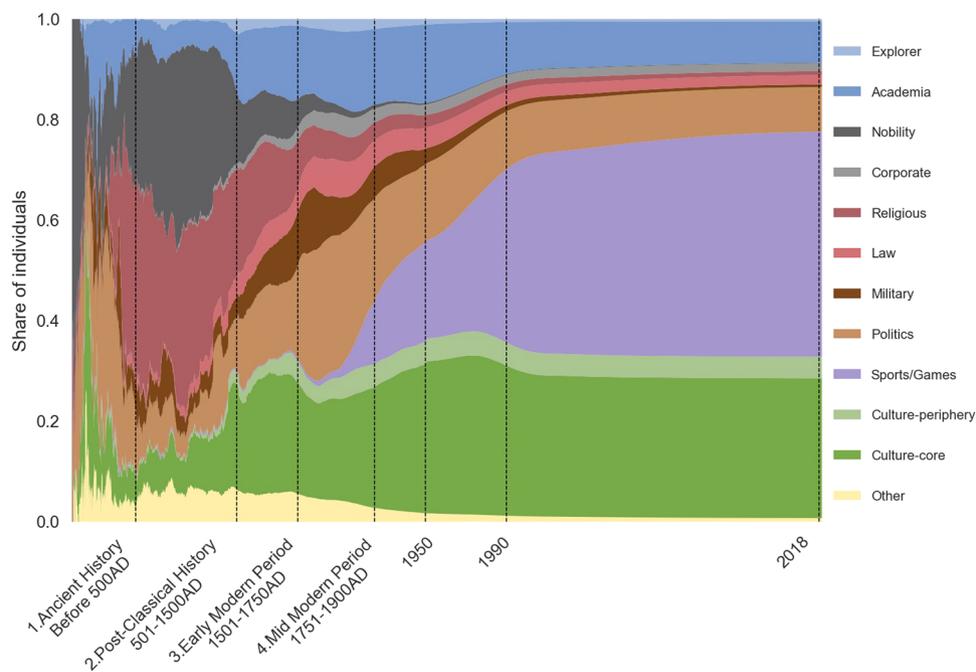
*Dealing with multiple biographies.* By construction, a given individual has a maximum of eight observations in the database once merged: 7 Wikipedia biographies plus one Wikidata entry. There are several cases to consider to eliminate duplicate observations from this universe of around 6.3 million biographies/entries. The most common situation is one individual present in Wikidata and in at least one of the 7 language editions of



**Fig. 1** Sunburst: relative importance of the main occupations and domains of influence. Notes. Cross-verified, restricted database (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The occupations are defined in Section *Domains of influence and occupations*.

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 correspond to each other and ii) the Wikidata links retrieved from the “Tools” section of each Wikipedia biography correspond to each other; there are indeed sometimes inconsistencies between these sources. We wrote a batch file to manually modify about 2000 Wikidata identifiers out of 4.7 millions not available at the time of scraping.

*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), due to the large number of contributors. Some of them may fail to detect or ignore the work of others. As an example, Sarendy Vong has two separate biographies in the English edition: [https://en.Wikipedia.org/wiki/Sarendy\\_Vong](https://en.Wikipedia.org/wiki/Sarendy_Vong) and [https://en.Wikipedia.org/wiki/Vong\\_Sarendy](https://en.Wikipedia.org/wiki/Vong_Sarendy). Appendix A.2 provides further details about this methodology. We eventually removed 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.



**Fig. 2** Share of individuals present in the database, breakdown by domain of influence. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. For a given year, the number of living individuals is calculated by summing up all individuals such that  $birth\_date \leq year \leq death\_date$ . When not available, the date of birth (resp. death) is estimated from the estimated average longevity over the period. The following log transformation has been applied to the time axis:  $year \rightarrow 8.5 - \log(2019 - year)$ .

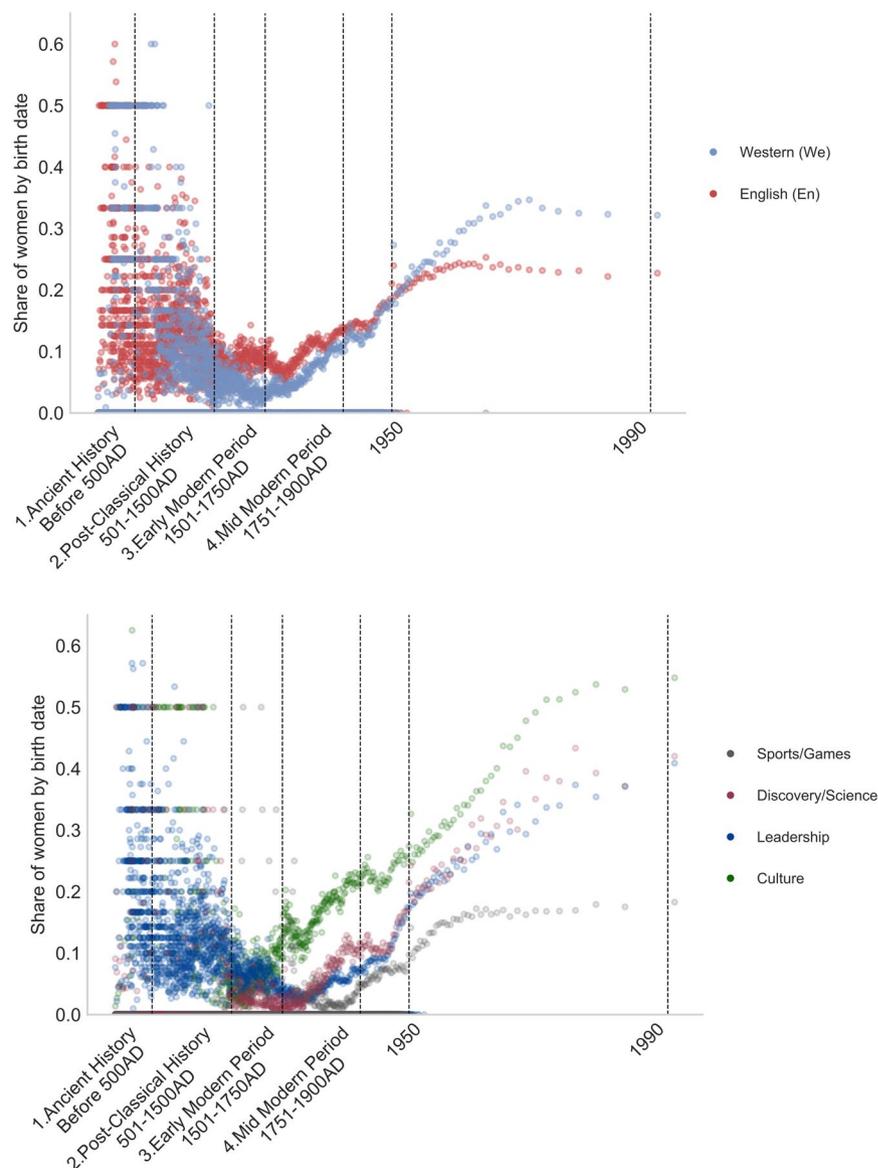
**Exhaustive sample: descriptive statistics.** *Wikipedia: comparing language editions.* The 5 most popular language editions in Wikipedia in number of biographies are, in decreasing order, 1) English, 2) German, 3) Japanese, 4) French and 5) Russian. Considering the German edition after English adds 340,913 individuals (of which 259,013 individuals have a unique biography in German). It is worth noting here that these 340,913 individuals would have been ignored without considering the German edition. In the exhaustive sample, the Japanese edition brings 179,466 new individuals, the French edition adds 155,391 individuals, the Russian edition adds 156,728 individuals. The full list of incremental additions is available in Table S2 in Appendix A.5 and provides a better sense of the relative importance of each language edition.

**Main sample characteristics.** Overall, we gather information about 4,678,040 distinct and unique notable individuals out of their 6,291,767 multiple 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. In addition, 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 contribution of both Wikidata and Wikipedia. In subsequent columns, we display similar statistics for different individual profiles according to their popularity in Wikipedia, in particular those with at least one 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, not based on a linguistic classification but mostly on geography to capture the socioeconomic groups formed around empires and trade areas: *English* (individuals with at least one biography 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). Table S2 in Appendix shows the diversity of languages and the number of *new* individuals brought by language editions. However, we will only consider those individuals from pages we could analyze, e.g. with a biography in one at least of the 7 language editions analyzed.

The exhaustive sample (Table 1, column 1) contains 1,578,917 individuals in the first group (English), 1,327,543 in the second group (Western non-English). The third, fourth and fifth groups 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 all individuals.

The information present in existing databases are based on the English edition of Wikipedia only<sup>1,2</sup>. This project, by considering other language editions and Wikidata, almost triples the universe of notable

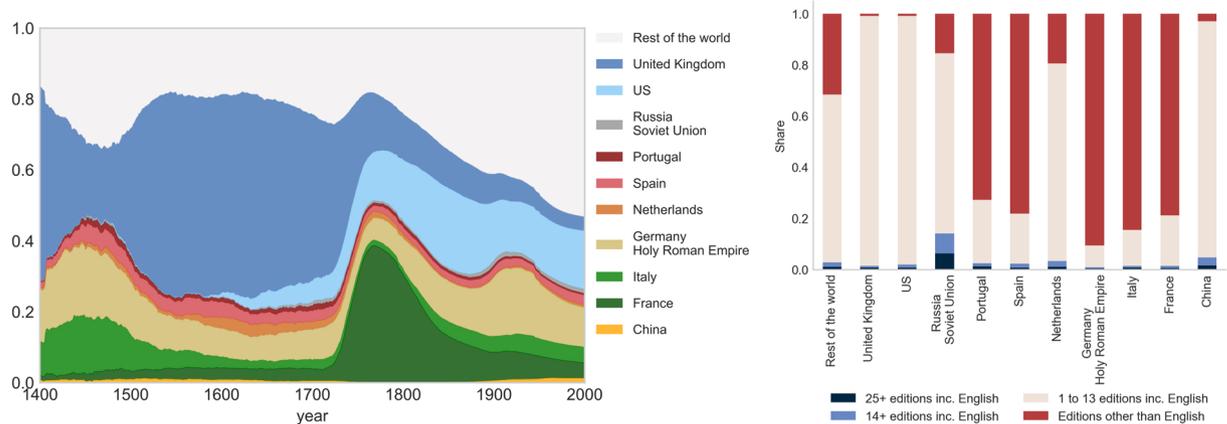


**Fig. 3** Share of women (top, by continent, bottom, by domain), 1000BC–2000AD. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. Western refers to Western, non-English editions of Wikipedia analyzed. The occupations are defined in Section *Domains of influence and occupations*. The following log transformation has been applied to the time axis:  $year \rightarrow 7.6 - \log(1991 - year)$ . The “Other” category is not displayed in the bottom chart due to its limited size.

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%). Considering 6 additional language editions results in 3,099,123 newly detected in the exhaustive sample.

**Measuring notability.** Many individuals have low visibility and sparse information. To disentangle the most visible from the less visible, we build a synthetic notability index using five dimensions to figure out a ranking for this broader set of individuals. These dimensions are:

1. the number of Wikipedia editions of each individual;
2. the length, i.e. total number of words found in all available biographies. It is equal to zero for individuals with just one Wikidata entry and no biography in Wikipedia;
3. the average number of biography views (hits) for each individual between 2015 and 2018 in all available language editions, using an API available in <https://wikitech.wikimedia.org/wiki/Analytics/AQS/PA-geviews> or zero in the absence of a Wikipedia biography;



**Fig. 4** Evolution of the share of individuals per citizenship identified as “politics”, 1400–2000AD. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The occupations are defined in Section *Domains of influence and occupations*. For a given year, the number of living individuals is calculated by summing up all individuals such that  $birth\_date \leq year \leq death\_date$ . When not available, the date of birth (resp. death) is estimated from the estimated average longevity over the period. Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: 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.

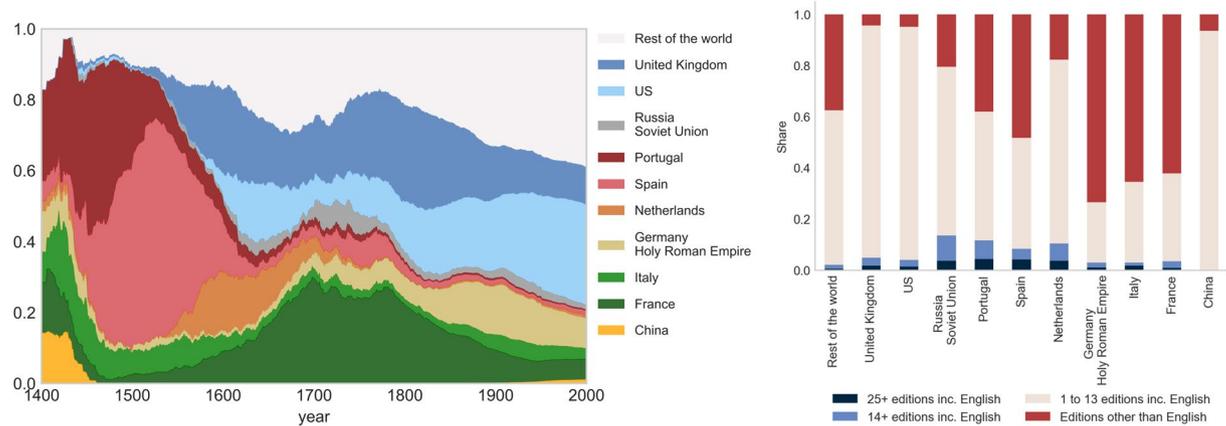
- the number of non-missing items retrieved from Wikipedia or Wikidata for birth date, gender and domain of influence. The intuition here is that the more notable the individual, the more documented his/her biographies will be;
- the total number of external links (sources, references, etc.) from Wikidata.

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 tends to favor individuals born recently<sup>2</sup>. Alternative choices for the weights of the sub-components can be made depending on the question that is tackled. We also reiterate that even within centuries or within domains of influence, pairwise comparisons make little sense and the notability index is more of a weight to statistically assess the influence or identify large groups of influence (e.g. by quintile of notability index). This being said, Table 2 shows the 10 most renowned individuals per period and gender. Examples of the most famous individuals present in the English edition are Barack Obama, Cristiano Ronaldo or Albert Einstein. Figure S5 in Appendix displays a cloud of the top 3000 individuals and suggests biases in gender (very few women), fields (large representation of culture and notability against for instance business) and historical periods (contemporaneity bias) that we address next. Other language groups naturally include lesser-known individuals like French film director Olivier Nakache (*Western*), Japanese actor Ryō Iwamatsu (*Eastern*), scientist Qayum Nasiryri (*Eurasia-Arabia*), former writer Boerneef (*Southern & natives*) and art historian Martin Hardie present in Wikidata only. Table S3 in Appendix shows the 5 most famous individuals in each recursive language group. 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 city. In the rest of the paper, we explore how these “newly added” individuals differ from those present in the English edition on a few dimensions such as year of birth/death, gender, occupation or citizenship. In particular, the Section Historical trends will systematically show the contribution of different samples (e.g. Fig. 3 or the right part of Fig. 4 and subsequent Figures).

**Extracting biographic information from a restricted sample.** As reported in Appendix Section B.3, information about lesser-known individuals is quite often sparse and does not allow to apply our cross-verification technique in a satisfactory way to detect systematic errors contained either in Wikipedia or in Wikidata and correct them. This is the reason why we decided to focus on a sub-sample of individuals with at least one biography among the 7 European language editions of Wikipedia we selected.

We use this sample to highlight the difference between the English edition (used in existing projects) and the other editions and document a potential 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.

**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



**Fig. 5** Evolution of the number of share of individuals per citizenship identified as “explorer, inventor, developer”, 1400–2000AD. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The occupations are defined in Section *Domains of influence and occupations*. For a given year, the number of living individuals is calculated by summing up all individuals such that  $birth\_date \leq year \leq death\_date$ . When not available, the date of birth (resp. death) is estimated from the estimated average longevity over the period. Left panel: Most popular citizenships (share of living individuals on the vertical axis, year on the horizontal axis). Right panel: 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.

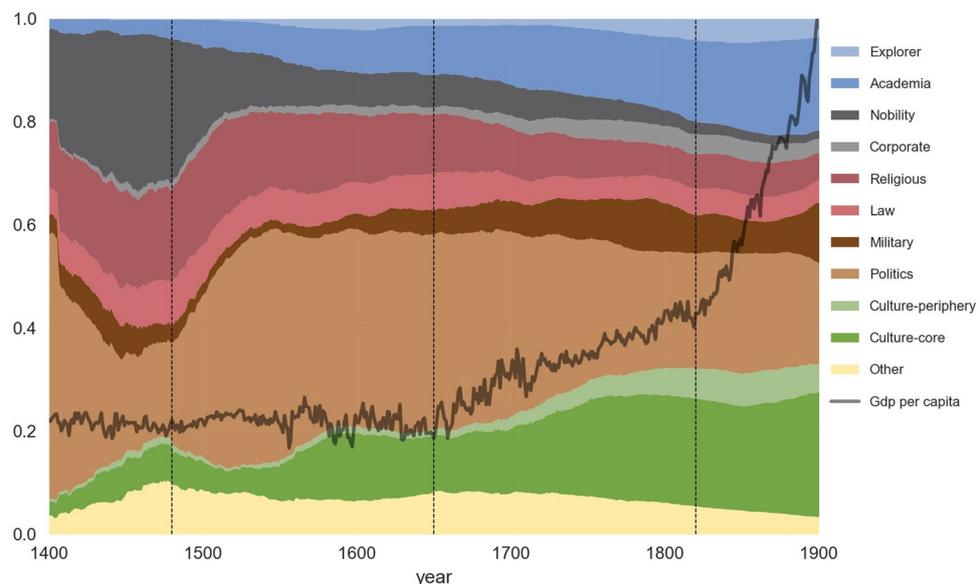
determine the individual’s gender. We then compared it to the one extracted from Wikidata, when reported. A fraction of 0.03% of the restricted database has an “Other” gender (609 individuals), corresponding mostly to transgender persons (e.g. <https://www.wikidata.org/wiki/Q57585712>) or non-binary persons (e.g. <https://www.wikidata.org/wiki/Q55237966>).

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.

When the birth and death information is present in both sources, we first determine the most recurring year among the different available (up to 7) language editions of Wikipedia, and compare it with the year extracted from Wikidata when reported. When there is a conflict between different sources, we give priority to Wikidata. See Appendix A6 for details. There are a few cases of super centenarians, allegedly known as having lived more than 120 years. The only case where we deleted the birth date is Trailanga (1607–1887).

We next calculate the number of living individuals in a given year in the database. This is represented in Appendix Figure S3. There is an exponential evolution of the sample size over time. The ratio of the number of notable individuals in the restricted sample over the world population (from two different estimates<sup>11,12</sup>) is not constant: as illustrated in Appendix Figure S2, 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 the fraction of “famous individuals” is constant throughout history, one can calculate the rate at which these were forgotten. Details of the calculation can be found in Appendix A.4. Under the assumption that people are forgotten at rate  $b$ , the fraction of famous people forgotten relative to total population is  $1 - (1 - b)^T$  where  $T$  is distance to present. The value of  $b$  is estimated in Appendix and corresponds to 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. In Appendix Figures S3 and S4, we show the time evolution of the sample size and its distribution across the English, Western, Eastern, Eurasia-Arabia, Southern & natives language groups from the exhaustive dataset.

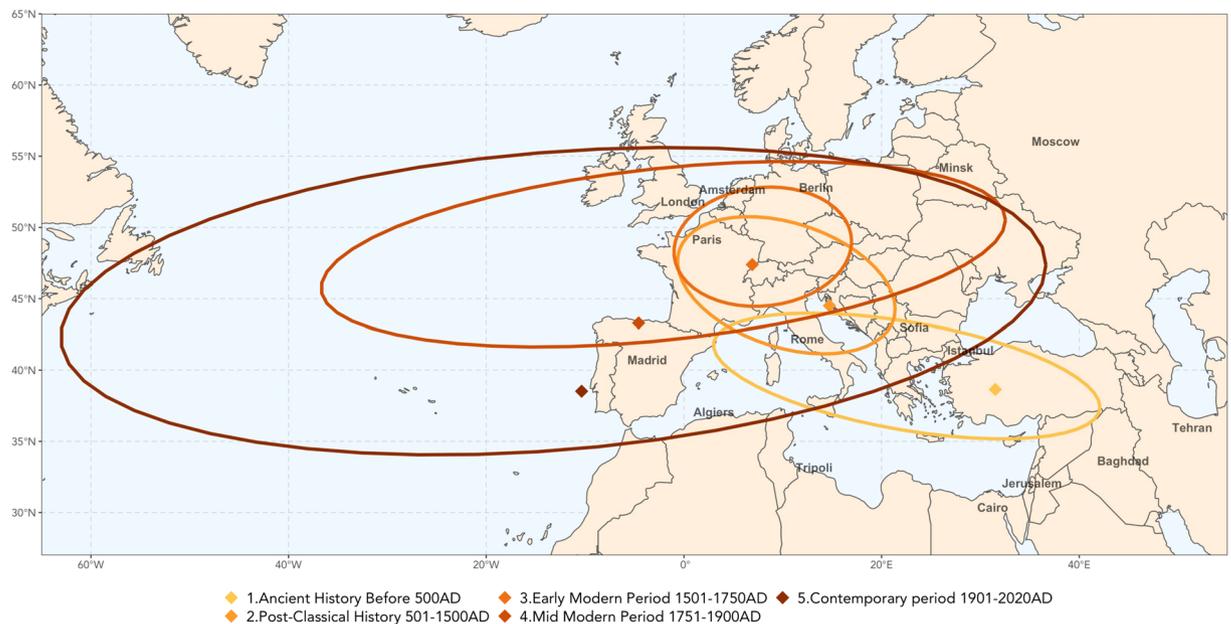
*Domains of influence and occupations.* We proceed the same way as for dates, and systematically compare the information collected from Wikipedia and Wikidata to figure out the domain of influence, the sector of activity or the field where notable individuals are meant to have had an impact. The list of occupations is reported below. The choice of the categories is based on the theory representation developed in the above Section “A social science framework”, differentiating producers of material/commercial goods and services, the producers of ideas or art, and the producers of administration services, security and rules. The choices turn out to be largely consistent with other typologies for most occupations<sup>2</sup> and culture<sup>13</sup>. The “Other” section regroups individuals whose notability originates from i) a direct link with notable individuals e.g. the children of a ruler who themselves did not rule ii) an occupation that is not classified in culture (e.g. librarian) or iii) only for their negative notorious acts (e.g. criminal, serial killer). When multiple occupations can be extracted from



**Fig. 6** Share of individuals in the database, breakdown by domain of influence, and GDP per capita, Country = United Kingdom. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The vertical axis represents the share of individuals per occupation and GDP per capita (source Maddison database) normalized to 1 in year 1900. Occupations are defined in Section *Domains of influence and occupations*. For a given year, the number of living individuals is calculated by summing up all individuals such that  $birth\_date \leq year \leq death\_date$ . When not available, the date of birth (resp. death) is estimated from the estimated average longevity over the period. The vertical bars in 1650 and 1820 correspond to the two changes in the GDP per capita, the second one being associated with the Industrial Revolution. The vertical line in 1480 corresponds to the increase in the number of notable individuals classified in Politics and the decrease in the corresponding Nobility category. Restricted sample (at least one Wikipedia edition among the 7 European languages analyzed). Comparison with GDP per capita. Imputed life time when missing.

a maximum of 7 sources, we adopt a frequentist approach, using the mode and the frequencies of competing occupations to select a main one and sometimes a secondary one, as discussed in Appendix B.1.

- **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  $Y^I$ )
  - 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 reduce the administration cost  $\tau$ ) or raise it through rent-seeking activities
  - 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  $Y^I$  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.)



**Fig. 7** Barycenters and ellipses of covariance matrix of longitude and latitude of birthplaces per period of human history. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*, and a known birthplace. The barycenter is computed as the arithmetic mean of the longitudes and latitudes of the birthplaces. The function `stat_ellipse` of the package `ggplot2` of R is used to draw Gaussian ellipses in which 40% of the population is concentrated.

Figure 1 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  $Y^I$  and to a much lower extent to  $Y^M$ . 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%).

Figure 2 shows the time evolution of the share of these different domains. We clearly see on this graph the prominent role played by religious figures and the military until the beginning of the 15th century, a slow rise in Academia, Politics and Culture afterwards, and the rapid rise of a new entertainment category, Sport and Game over the 20th century that even become dominant after 1950. This trend corresponds to another trend, the secular rise in leisure and the decline in hours worked over the last two centuries<sup>14</sup> due to a combination of technical progress sustaining living standards and leisure and many other immaterial goods being luxury goods.

**Citizenship.** For citizenships, we collect and then compare the information available from both Wikipedia and Wikidata. From Wikipedia, we retrieve up to two citizenships per language edition, based on the current name of the country therefore summing to a maximum of 14 possible citizenships. We keep the two most frequent citizenships of this pool and compare them to Wikidata. The comparison group from Wikidata retains a list of up to 10 citizenships. We group these citizenships 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 modern entity, for e.g. for a person who lives or have lived under today’s People’s Republic of China, the reported citizenship is simply China. We also keep three empires, Holy Roman Empire, Roman Empire and Soviet Union in these aggregate groups. We skipped extremely unfrequent citizenships such as Chenla Kingdom, Sultanate of Hoby, Lordship of Carpi etc. that were not aggregated with our main citizenship (0.004% of those with a citizenship only).

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. We use information available on the collapse of empires such as the Holy Roman Empire or the Soviet Union to correctly classify individuals



**Fig. 8** Migration charts, selected countries and domains of influence. Notes. Cross-verified, restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*, and both a known birthplace and deathplace. 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.

into these supranational entities. Finally, a fraction of individuals may have several citizenships and we report two of them if relevant.

**Historical trends.** Additional editions we analyzed beyond English provide a broader and finer coverage of some key periods in human history and social trends. We review here only a small number of examples of interest for the sake of conciseness. With a few emblematic historical events in mind, we ask whether the database could quantitatively illustrate these periods.

*From Dark Ages to the Modern age: the rise of notable women.* Figure 3 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 ranges 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 period (see statistics in Table S4), the Western non-English edition further mitigates the bias after 1950, with a sample average close to 27% of women, as opposed to a 22.5% in the English edition.

*Politics and the emergence of modern democracy.* Figure 4 documents the number of individuals associated with politics. This illustrates the rise of American politics in the second half of the 18th century at the time with the birth of the independent nation of the USA, as shown in the left panel of Fig. 4. 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 this 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

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

**Table 3.** Freq. missing information vs. complete profile. Notes. Restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The occupations are defined in Section *Domains of influence and occupations*. First 5 columns: % missing observations; last column: % with no missing observation among these 5 variables.

consider that central elements of modernity, of capitalism, industry, and science, originated in early modern Britain (circa 1500–1750)<sup>15,16</sup>. 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 the English edition of Wikipedia. This is particularly large for French individuals (80% individuals), and true for other citizenships including Italians (+70%), Germans (+85%), Portuguese (+70%) and Spanish (+65%) individuals. For this specific occupation (politics), restricting the sample to English pages would drastically change the relative importance of citizenships, for instance almost entirely removing French politicians and considerably expanding English and US ones.

*The age of exploration and discoveries.* Figure 5 (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. 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 and therefore in existing projects. The share of individuals retrieved with the 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 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.

*Conjecture: the role of science and politics in escaping the Malthusian trap.* Theories of long run economic growth<sup>17,18</sup> explain that the main reasons for escaping the Malthusian trap (where improvement due to technical progress allows population increases but no change in income per capita) are due to the accumulation of *latent* factors such as human capital and the evolution of political regimes around 1750AD in Western countries. A first insight could be derived from Fig. 2 where, in the two centuries before that year, one observes a rise of Academia and Politics, and a decline in Religion and Nobility. In Fig. 6, we take a closer look at the UK, the country leading the Industrial Revolution. The figure shows two distinct regimes of GDP per capita growth after the initial Malthusian stagnation: a first economic expansion in the late 17th and 18th century, that follows a century of a latent relative expansion of “Politics” with Nobility declining instead; a second regime of the Industrial Revolution of the early 19th century, which is preceded by the latent relative expansion of science and the arts. An overlook at other countries (unreported) suggests that similar patterns emerge with a particularly strong role of the composition of the leadership category which contains Nobility, Religion, Military and Politics. Previous economic history work<sup>19</sup> offers a detailed review of the arguments in favor of the emergence of new technologies and the respective role of Enlightenment vs. trade and profitability argument to explain the economic take-offs.

*Geolocalisation.* Using longitudes and latitudes of birth and death places, one can track the evolution of the global geographical centers of the database. Fig. 7 represents the ellipse of the variance-covariance matrix of longitudes and latitudes of birthplaces at different periods of time (40% of the sample population lies within each ellipse). Fig. 8 represents the birth to death flows and generalizes early pioneering works<sup>1,2</sup>. The higher granularity of our database allows us to cover less explored destinations or smaller geographical divisions. Going from the 150,000 individuals in earlier studies to 2.3 million individuals in our case allows us to cover migration outflows of, e.g. artists who left Italy to emigrate in numbers to South America, or scientists who left the UK to live in North America. We can also represent intra-Asian migration flows which is mostly within countries contrary to inter-country European flows.

	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
<b>D (missing) and P (available) (%)</b>	<b>2.72</b>	<b>1.31</b>	<b>0.67</b>	<b>8.16</b>	<b>17.16</b>
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

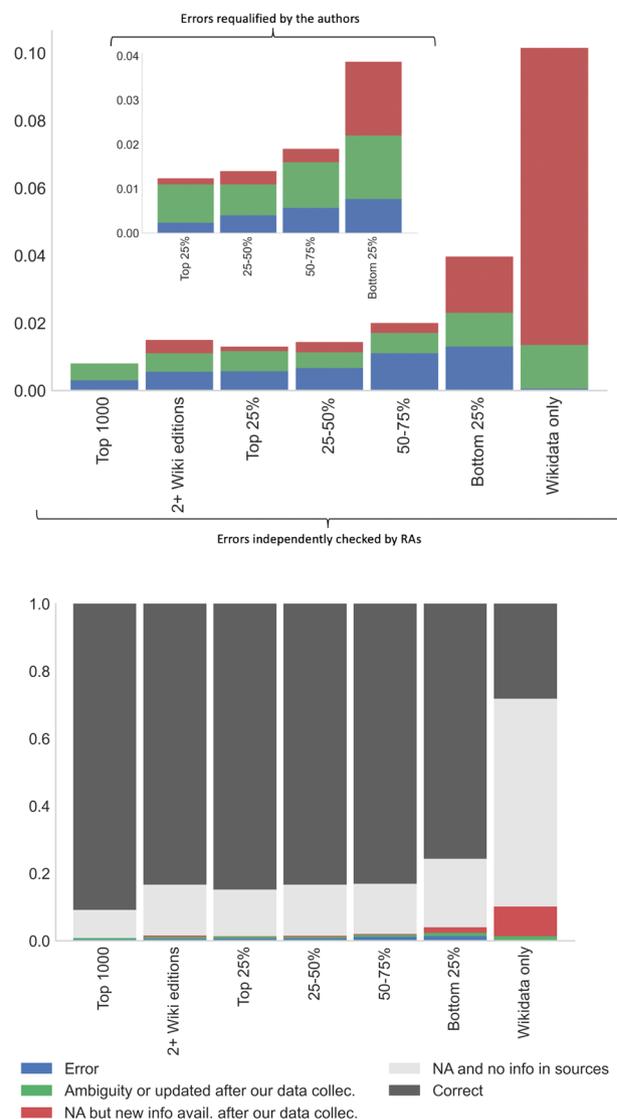
**Table 4.** Wikipedia & Wikidata: similarity and differences. Notes. Restricted sample (at least one Wikipedia edition among the 7 European languages analyzed), see Section *Extracting biographic information from a restricted sample*. The occupations are defined in Section *Domains of influence and occupations*. 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.

**Outreach.** Our database can be used to provide a new level of granularity and historical depth to several fields. In urban economics, one may test the quantitative implications of the model of cities attracting talents<sup>20</sup>, the role of the creative class to enlighten the causation chains<sup>4,21,22</sup>, the impact of transport infrastructure development on geographic mobility<sup>23,24</sup>, or the existence of gravity forces in trade to uncover ancient cities as trade places<sup>25</sup>. Many gender issues can be revisited with greater historical depth and finer geographical scale, like 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<sup>26,27</sup>, the importance of the local environment enabling women to become inventors<sup>28,29</sup>, the convergence of male and female occupations over time<sup>30</sup>. 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<sup>15,31,32</sup>. It offers testable implications of the unified growth theory, such as the replacement of physical by human capital<sup>17,18</sup>, as well as neo-Schumpeterian theories and the role of innovation and competition<sup>33,34</sup>, thanks to our account of scientists and innovators per country. The same is true for detecting evidence of cultural transmission mechanisms through naming<sup>35–37</sup>, 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<sup>38–40</sup> as well as the role and determinants of culture in the wealth of nations<sup>41–43</sup>. Research in demography economics for instance on longevity in space and time<sup>5</sup> can be enriched thanks to our list of birth and death dates by field. Issues on international and domestic migration can also be revisited from a more longitudinal perspective. For example, Zelinsky's theory of mobility transition<sup>44</sup>, the theory and evidence on brain drain<sup>45</sup>, the birthplace diversity and economic prosperity<sup>46</sup> and the relationship between future outcomes and birthplace<sup>47</sup> can be revisited from a historical perspective, thanks to the 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<sup>48</sup> or extra-place based mortality<sup>49</sup> 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. Beyond these economic approaches, and without exhaustivity, political scientists can use the database to analyze the evolution of political ideas and the way they interact with the development of democracy; given the mass of data, historians may identify more easily new key protagonists of most historical periods; similarly, sociologists may identify individuals who played or are still playing a role in subculture; demographers could use the demographic information to compare life expectancy of the subgroup of notable individuals in different regions or to the general population.

### Data Records

The cross-verified dataset is available at the publicly accessible Sciences-Po Dataverse: <https://doi.org/10.21410/7E4/RDAG3O><sup>50</sup>. 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 A (\textit{English} edition of \textit{Wikipedia}) to F (\textit{Wikidata} only) as described in Table 1.);
- *birth*: birth date of the individual (either reported or estimated);
- *death*: death date of the individual (either reported or estimated);
- *level1\_main\_occ* & *level2\_main\_occ* & *level2\_second\_occ* & *level3\_all\_occ* & *level3\_main\_occ* & *freq\_main\_occ* & *freq\_second\_occ*: set of seven variables for the main domain of influence of each individual in three



**Fig. 9** 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 0.5% of all observations and  $0.5/(1-0.2) = 0.625\%$  of all non-missing observations. 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 S6 to S10) 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.

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 1 and 2*: first and if needed, second citizenship of the individual with a reference to the current country or to a former political regime;
- *ranking\_visib\_5criteria*: computed from 5 variables: number of Wikipedia editions, non-missing biographic information, length of pages, hits of pages and total number of external links. An alternative ranking based on the sum of the log of these variables plus one is in *sum\_visib\_In\_5criteria*;
- *number\_wiki\_editions*: number of different Wikipedia editions;
- *non\_missing\_score*: total number of non-missing items retrieved from Wikipedia or Wikidata for birth date, gender and domain of influence;
- *total\_count\_words*: total number of words in all biographies from Wikipedia;

- *wiki\_readers\_2015\_2018*: average per year number of page views in all Wikipedia editions (information retrieved in 2015–2018);
- *total\_noccur\_links*: total number of external links (sources, references, etc.) from Wikidata;
- *bplo1 & dplo1 & bpla1 & dpla1 & birthplace\_name & deathplace\_name*: longitude, latitude and name of birthplace, deathplace. To be used with caution, the accuracy of these variables has not yet been verified.

The exhaustive dataset is available at the publicly accessible Sciences-Po Dataverse: <https://doi.org/10.21410/7E4/YLG6YR><sup>51</sup> This dataset is not cross-verified and should not be used directly or under the full responsibility of users. It includes all the variables detailed above. The following five variables:

- *exhaustive\_level1\_main\_occ & exhaustive\_level2\_main\_occ & exhaustive\_level2\_second\_occ & exhaustive\_area1\_of\_rattachment & exhaustive\_area2\_of\_rattachment* correspond to the variables
- *level1\_main\_occ & level2\_main\_occ & level2\_second\_occ & area1\_of\_rattachment & area2\_of\_rattachment* in the restricted dataset.

## Technical Validation

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.

**Missing information.** Table 3 documents the share of missing information of the main variables, 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 this 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. These low rates of missing information result from our initial choice to work on a restricted sample of individuals with at least one biography in Wikipedia. People present in both groups (English and Western non-English) are the most famous ones and very few information are missing for these individuals. In column (5), the proportion of individuals with no birthplace is quite high in both samples (34% for the English group and 29% for the Western non-English group). Finally, the last column contains the proportion of individuals for whom no information is missing. In both groups, the rate is quite high with 64% in the English group while it is at around 69% in the Western non-English group.

**Comparison between Wikipedia editions and Wikidata.** The strength of our methodology comes from the fact that we use 7 editions of Wikipedia and Wikidata in order to cross-verify the maximum number of information. Although the convergence rate is quite high, it happens in rare occasions that Wikipedia editions and Wikidata contradict each other on some dimensions. Table 4 reports the number of individuals for whom one information (e.g. birth year) has been detected in both sources (row #1) and its share in the sample (row #2). The third and fourth rows give the discrepancy rate between Wikipedia and Wikidata on each dimension. Concerning dates, the share of exact mismatch is very low (row #3): 2% (birth) and 1.6% (death). Larger differences (10+ year difference) are even less frequently observed (row #4): 0.23% (birth) and 0.38% (death). The discrepancy rate for gender is also very low at around 0.51%.

The information on the domain of influence is not consistent in a larger number of cases (11%). Most of the time this is due to different sub-categories detected that are usually close to each other, e.g. Culture-core vs. Culture-periphery or Politics vs Military, or Nobility vs Family. To get rid of such minor differences we compute similar discrepancy rates considering the following set a broader categories: Discovery/Science, Culture, Leadership, Sports/Games, Other. As expected, the discrepancy rate shrinks to 7.2% (row #4) for occupations. Mean discrepancy rates for citizenships are at around 10% (last column). Such high rates mostly reflect the fact that Wikidata reports the regime in force at the time when the individual lived while Wikipedia reports more often the current regime information. For instance, Martin Luther's country of citizenship is "Electorate of Saxony" (which belongs to the Holy Roman Empire) in his Wikidata biography (<https://www.wikidata.org/wiki/Q9554>) while his English Wikipedia biography ([https://en.wikipedia.org/wiki/Martin\\_Luther](https://en.wikipedia.org/wiki/Martin_Luther)) says he "was a German professor of theology".

Subsequent rows (#5 and #6) illustrate the importance of considering different sources: they show how frequently one information is found in Wikipedia while absent from Wikidata. These statistics are useful to figure out the marginal contribution of Wikipedia which is sizeable. Indeed, combining Wikipedia and Wikidata adds a substantial fraction of observations to the database. For instance, Wikipedia brings in 2.72% additional birth dates, 1.31% additional death dates, 0.67% additional gender information, 8.16% additional occupations and 17.16% additional citizenships. While the contribution of Wikipedia appears limited on birth (+2.72%) and death dates (+1.31%) as such basic information are most often reported in Wikidata, Wikipedia proves much more useful to complement Wikidata on occupations (+8.16%) and citizenships (+17.16%). The last two rows report the number of cases with discrepancy between Wikipedia & Wikidata and for which we selected Wikipedia, for a small proportion of cases.

**Manual checks.** We ran several tests to check the accuracy of our dataset. We hired three teams of students in three areas (from France, the UAE and India) representing very diverse cultures to compare the information contained in our final database with the information present in Wikipedia/Wikidata. We asked them to report mistakes on 6 different dimensions (exact or approximate birth and death dates, gender, main occupation and possibly secondary occupation, citizenship and possibly secondary citizenship). The 5 different possible outcomes are as follows: 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 included in the current version of Wikipedia/Wikidata but was not present in this universe 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 contradicting each other or information updated since data collection).

A pilot wave on 5 sub-samples of 500 individuals each, with a 20% overlap, led 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-sample of 1000 individuals each, with exact overlap so that every individual would be checked twice. The total sample of 5,000 individuals thus verified contains the top 1,000 individuals of the database, 500 individuals with two biographies in Wikipedia at least (these are all present in the top quartile of the visibility distribution and the median individual in that sample was in the 92th centile of the visibility distribution), 3,000 individuals from the database with the 4 quartiles equally represented, and finally 500 individuals with no Wikipedia biography, and only a Wikidata entry. See Appendix C for more details on the instructions given to the RAs. In addition, one individual was dropped, because it had been selected twice. To conclude, we manually verified the information for 4,999 individuals.

An information is deemed valid when two independent research assistants assessed its correctness. For cases where their assessment diverged, we asked a PhD researcher to double check the information and try to determine the origin of that discrepancy. See Table S5 in Appendix C where we report the various discrepancy rates for each variable.

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 Fig. 9 the distribution of various cases for the different samples (top panel). The error rate is increasing for individuals with shorter 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.

The authors finally verified the errors detected independently by the different teams of RAs and the PhD researcher, and requalified some of them after a final cross-examination. Most of the errors detected by the RAs were actually due to a slight discrepancy - typically one year - between the birth date contained in the database on one side and what was found in our reference sources. Most of the time Wikipedia and Wikidata were themselves conflicting on these cases (e.g two biographies would report respectively 1944 and 1943 and we had reported 1943). So the numbers reported tend to overestimate the true number of errors by a factor 3 to 6. We display the number of errors after we requalified them 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 both reference sources.

**Western bias.** In this paper, we introduced a multi-language database of notable individuals with the use of 7 language editions of Wikipedia and Wikidata to assemble a list of 4,678,040 individuals. This significantly reduced the Anglo-Saxon bias, but not all. Two main drawbacks remain. First, we did not exploit the non-Western language editions to cross-verify information on individual characteristics. Second, we did not collect the number of words beyond these 7 language editions: they enter in the notability index, but this index cannot be considered as global, resulting in a Western-world bias in notability measures. This is however partly compensated by the use of the total number of hits for all Wikipedia editions and not only 7, in our aggregate notability measure.

**Errors in Wikipedia.** The accuracy of Wikipedia being not perfect<sup>7,8</sup>, our data is as good as the source data, but our approach adds new possibilities: to cross-check across different language editions and reduce errors when possible.

### Usage Notes

The full data set will be on our dedicated website here <https://medialab.github.io/bhht-datascape>. The data are available both in.csv and Stata (.dta) format. The restricted database was cross-verified. It is subject to CC-BY-SA licensing. The intermediate files as well as the exhaustive database are not cross-verified and should not be used directly or under the full responsibility of users.

### Code availability

Code to replicate all analyses presented here, the intermediate datasets and the exhaustive dataset are available at the publicly accessible Sciences-Po Dataverse: <https://doi.org/10.21410/7E4/YLG6YR><sup>51,52</sup>.

Received: 28 July 2021; Accepted: 9 May 2022;

Published online: 09 June 2022

## References

- Schich, M. *et al.* A network framework of cultural history. *Science* **345**, 558–562 (2014).
- Yu, A. Z., Ronen, S., Hu, K., Lu, T. & Hidalgo, C. A. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data* **3**, 1–16 (2016).
- Tanon, T. P., Weikum, G. & Suchanek, F. Yago 4: A reason-able knowledge base. In *European Semantic Web Conference*, 583–596 (Springer, 2020).
- Serafinelli, M. & Tabellini, G. Creativity over time and space. *Journal of Economic Growth* (2022).
- de la Croix, D. & Licandro, O. The longevity of famous people from Hammurabi to Einstein. *Journal of Economic Growth* **20**, 263–303 (2015).
- Jara-Figueroa, C., Yu, A. Z. & Hidalgo, C. A. How the medium shapes the message: Printing and the rise of the arts and sciences. *PLOS ONE* **14**, 1–14 (2019).
- Rector, L. H. Comparison of wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review* **36**, 7–22 (2008).
- Brown, A. R. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics* **44**, 339–343 (2011).
- Baumol, W. & Bowen, W. *Performing Arts: The Economic Dilemma* (MIT Press, New York, 1966).
- La Porta, R., Lopez-de Silanes, F., Shleifer, A. & Vishny, R. The quality of government. *The Journal of Law, Economics, and Organization* **15**, 222–279 (1999).
- Kremer, M. Population growth and technological change: One million bc to 1990. *The Quarterly Journal of Economics* **108**, 681–716 (1993).
- Manning, S. Year-by-year world population estimates: 10,000 bc to 2007 ad. *Historian on the warpath* **12** (2008).
- Throsby, D. The concentric circles model of the cultural industries. *Cultural trends* **17**, 147–164 (2008).
- Aguiar, M. & Hurst, E. Measuring trends in leisure: The allocation of time over five decades. *The Quarterly Journal of Economics* **122**, 969–1006 (2007).
- Mokyr, J. *The Enlightened economy an economic history of Britain 1700-1850* (Yale University Press, 2010).
- Allen, R. C. *The British industrial revolution in global perspective* (Cambridge University Press, 2009).
- Galor, O. & Moav, O. From physical to human capital accumulation: Inequality and the process of development. *The Review of Economic Studies* **71**, 1001–1026 (2004).
- Galor, O. *Unified growth theory* (Princeton University Press, 2011).
- Crafts, N. Explaining the first industrial revolution: two views. *European Review of Economic History* **15**, 153–168 (2011).
- Behrens, K., Duranton, G. & Robert-Nicoud, F. Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy* **122**, 507–553 (2014).
- Moretti, E. *The new geography of jobs* (Houghton Mifflin Harcourt, 2012).
- Florida, R. *The rise of the creative class and how it's transforming work, leisure, community and everyday life* (Basic Books, 2005).
- Baum-Snow, N. Did Highways Cause Suburbanization? *The Quarterly Journal of Economics* **122**, 775–805 (2007).
- Donaldson, D. & Hornbeck, R. Railroads and American Economic Growth: A “Market Access” Approach. *The Quarterly Journal of Economics* **131**, 799–858 (2016).
- Barjamovic, G., Chaney, T., Cosar, K. & Hortaçsu, A. Trade, merchants, and the lost cities of the bronze age. *The Quarterly Journal of Economics* **134**, 1455–1503 (2019).
- Beaman, L., Duflo, E., Pande, R. & Topalova, P. Female leadership raises aspirations and educational attainment for girls: A policy experiment in india. *science* **335**, 582–586 (2012).
- Bertrand, M. Gender in the twenty-first century. In *AEA Papers and proceedings* **110**, 1–24 (2020).
- Bell, A., Chetty, R., Jaravel, X., Petkova, N. & Van Reenen, J. Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly Journal of Economics* **134**, 647–713 (2019).
- Hunt, J., Garant, J.-P., Herman, H. & Munroe, D. J. Why are women underrepresented amongst patentees? *Research Policy* **42**, 831–843 (2013).
- Goldin, C. A grand gender convergence: Its last chapter. *American Economic Review* **104**, 1091–1119 (2014).
- Acemoglu, D. & Johnson, S. Unbundling institutions. *Journal of political Economy* **113**, 949–995 (2005).
- Glaeser, E. L., La Porta, R., Lopez-de Silanes, F. & Shleifer, A. Do institutions cause growth? *Journal of economic Growth* **9**, 271–303 (2004).
- Aghion, P., Bloom, N., Blundell, R., Griffith, R. & Howitt, P. Competition and innovation: An inverted-u relationship. *The quarterly journal of economics* **120**, 701–728 (2005).
- Aghion, P. & Howitt, P. W. *The economics of growth* (MIT press, 2008).
- Fryer, R. & Levitt, S. The causes and consequences of distinctively black names. *Quarterly Journal of Economics* **119**, 767–805 (2004).
- Alesina, A. & Giuliano, P. Culture and institutions. *Journal of Economic Literature* **53**, 898–944 (2015).
- Bisin, A. & Verdier, T. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory* **97**, 298–319 (2001).
- Ekelund, R. B. Jr, Hébert, R. F. & Tollison, R. D. An economic analysis of the protestant reformation. *Journal of Political Economy* **110**, 646–671 (2002).
- Becker, S., Pfaff, S. & Rubin, J. Causes and consequences of the protestant reformation. *Explorations in Economic History* **62**, 1–25 (2016).
- Cantoni, D. The economic effects of the protestant reformation: Testing the weber hypothesis in the german lands. *Journal of the European Economic Association* **13**, 561–598 (2015).
- Gorodnichenko, Y. & Roland, G. Culture, institutions, and the wealth of nations. *The Review of Economics and Statistics* **99**, 402–416 (2017).
- Roland, G. *12. Culture, institutions, and development*, 414–448 (Princeton University Press, 2020).
- Platteau, J.-P. & Peccoud, R. *Culture, Institutions, and Development: New Insights into an Old Debate* (Routledge, 2011).
- Dao, T. H., Docquier, F., Parsons, C. & Peri, G. Migration and development: Dissecting the anatomy of the mobility transition. *Journal of Development Economics* **132**, 88–101 (2018).
- Beine, M., Docquier, F. & Rapoport, H. Brain drain and economic growth: theory and evidence. *Journal of Development Economics* **64**, 275–289 (2001).
- Alesina, A., Harnoss, J. & Rapoport, H. Birthplace diversity and economic prosperity. *Journal of Economic Growth* **21**, 101–138 (2016).
- Chetty, R., Hendren, N., Kline, P. & Saez, E. Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics* **129**, 1553–1623 (2014).
- Abramitzky, R., Boustan, L., Jacome, E. & Perez, S. Intergenerational Mobility of Immigrants in the United States over Two Centuries. *American Economic Review* **111**, 580–608 (2021).
- Finkelstein, A., Gentzkow, M. & Williams, H. Place-based drivers of mortality: Evidence from migration. *American Economic Review* **111**, 2697–2735 (2021).
- Laouenan, M. *et al.* A Brief History of Human Time - Cross-verified Dataset. *Sciences Po Dataverse*, <https://doi.org/10.21410/7E4/RDAG3O> (2022).
- Laouenan, M. *et al.* A Brief History of Human Time - Codes & Datasets. *Sciences Po Dataverse*, <https://doi.org/10.21410/7E4/YLG6YR> (2022).
- Gergaud, O., Laouenan, M. & Wasmer, E. A brief history of human time: Exploring a database of ‘notable people’. *Sciences Po Economics Discussion Papers 2016-03*, *Sciences Po Department of Economics* (2016).

## Acknowledgements

In social science, working papers are shared prior to submission to collect comments from colleagues and peers. A slightly longer version of this text was initially released as a working paper CEPR DP 15852, March 2021. 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). Previous work by some of the authors of this manuscript used `Freebase` and the English edition of `Wikipedia` to build a dataset of 1,243,776 notable individuals that they match to a broader set of cities. 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, Djellel Difallah, Michel Serafinelli, Guido Tabellini, Oded Galor, David Weil, Alexander Yarkin, Stelios Michalopoulos, Louis Putterman, Guillaume Blanc, Markus Poschke, Fabian Lange, Camille Hémet, Bryan Waterman as well as participants to various conferences and seminars for insightful discussions.

## Author contributions

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 visualisation website. The other five authors have all contributed to the data analysis and to forthcoming follow-up papers. Olivier Gergaud, Morgane Laouenan and Etienne Wasmer were involved in the tests of the datasets. Etienne Wasmer coordinated the project. All authors reviewed the manuscript. Morgane Laouenan and Etienne Wasmer are corresponding authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-022-01369-4>.

**Correspondence** and requests for materials should be addressed to M.L. or E.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022