# A Multi-language Database of Notable People


Université Sorbonne Paris Cité



*This project introduces a database of 1.8 million notable people throughout human history (3000BCE-2015AD) with a biography available in Wikipedia. Together, they represent the largest existing database about notable people : we add more than 30% individuals who had not been identified in previous works based on the English-language edition of Wikipedia only. We describe here the various approaches and procedures that we adopted to extract the relevant information from these biographies. We then discuss a series of new historical trends in human History that emerged from the analysis of this extended database.*


## Project team :


### Olivier GERGAUD

Olivier GERGAUD is a senior professor of Economics at KEDGE Business School. His main research interests are Economics of Pro-social Behavior, Cultural Economics (Celebrities), Restaurant and Wine Economics, Environmental Economics, Behavioral Finance (Hedge Funds, Betting) and Sports Economics (Cycling, Football).

### Morgane LAOUÉNAN

Morgane Laouénan is a CNRS researcher at the Centre d'Economie de la Sorbonne. She is specialized in Labor Economics and Applied Microeconomics. Her research focuses on discrimination against African immigrants in France and against African-Americans in the US. In particular, she uses both individual-level data from surveys and from the Internet to study the impact of racial prejudice on labor market and housing outcomes of minorities.

### Etienne WASMER

Etienne WASMER is Professor of Economics at Sciences Po and Co-Director of LIEPP. His main research interests cover labour economics, search theory, discrimination and human capital. His research based work among others has been published in the The American Economic Review, The Journal of the European Economic Association, The American Economic Journal (macro).

### Jean-Benoît EYMEOUD

Jean-Benoît EYMEOUD is a PhD student in the Department of Economics of Sciences Po, under the supervision of Etienne WASMER, and a LIEPP Fellow. His research focuses on the evaluation of housing and labor market public policies.


## Context

- The growing number of datasets allows us to document historical facts. Three recent approaches by Schich et al. (2014), De la Croix and Licandro (2015) and Yu et al., (2016) have particularly focused on « famous individuals ».
- We extend those three approaches in three different ways. First we compile the largest possible database of notable people. Second, we develop a semi-automatic methodology to assign occupations and gender to these individuals. Third, we collect information in **7 different languages** to address the potential Anglo-Saxon/Western bias.

## Data Collection

1. **Individual characteristics**: We analyze the source code of each Wikipedia page to extract basic information about dates and locations for birth and death, occupations, gender and citizenship.
2. **Validity of the extraction procedure**: The extraction procedure has been checked by a series of manual verifications.
3. **Types of occupation**: We define occupations at different levels of aggregation

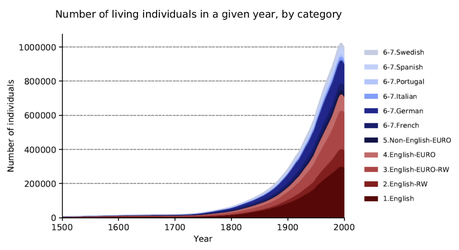
| CATEGORY A | CATEGORY B | CATEGORY C (five most frequent) |
|---|---|---|
| SPORTS | Sports | football, cricket, baseball, rugby, ice hockey |
| CULTURE | Culture core | actor, art, actress, singer, writer |
| | Culture related | journalist, television, architect, radio, design |
| ACADEMICS | Science | historian, physician, mathematician, physicist, economist |
| | Education | professor, scholar, college, academic, educator |
| GOVERNANCE EXECUTIVE | Politics | politician, representative, president, democrat, minister |
| | Law enforcement | lawyer, judge, attorney, jurist, justice |
| GOVERNANCE SYMBOLIC | Military | army, officer, soldier, militar, navy |
| | Nobility | noble, king, peer, prince, duke |
| | Religious | bishop, priest, church, clergy, theologian |
| | Explorer | inventor, explorer, settler, adventurer, developer |
| ENTREPRENEUR | Corporate | business, director, entrepreneur, pioneer, executive |
| | Worker | engineer, merchant, farmer, sailor, computer |
| Family | Family | son, child, daughter, wife, brother |
| Other | Other | recipient, beauty, convicted, philantropist, criminal |

## Partition Category of Individuals

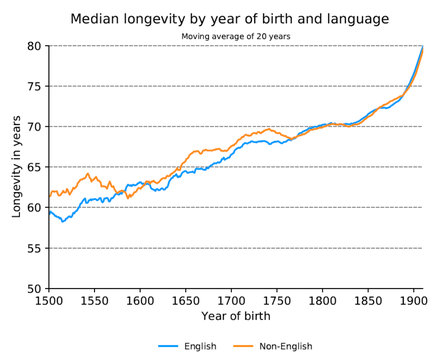
- Each Wikipedia biography is extracted in a **given language**, namely English, French, German, Italian, Portuguese, Spanish or Swedish.
- A notable individual is either:
  - 1.English (biography in English only)
  - 2.English-RW (in English + Rest of the World)
  - 3.English-EURO-RW (in English + European editions + Rest of the World)
  - 4.English-EURO (in English + European editions)
  - 5.Non-English-EURO (in European editions)
  - 6.French (biography in French only), 6.German, etc....
  - 7.French-RW (in French only + Rest of the World), etc

## Specific Comparisons

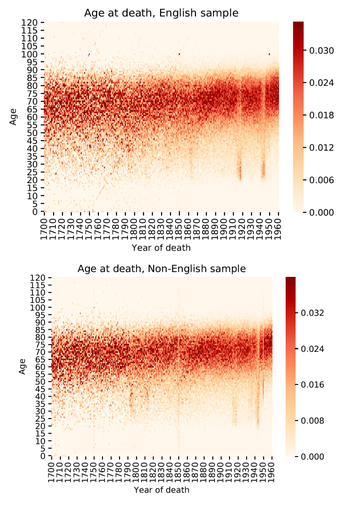
1. The structure of both samples (English vs Non-English Editions) is similar with a more than exponential evolution of the sample size over time.



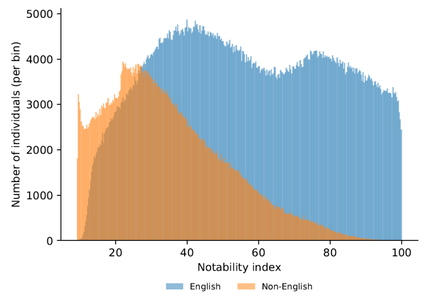

2. The average lifespan has increased by 20 years, from 60 to 75 years, between the cohort in 1400CE and the one born in 1900AD.



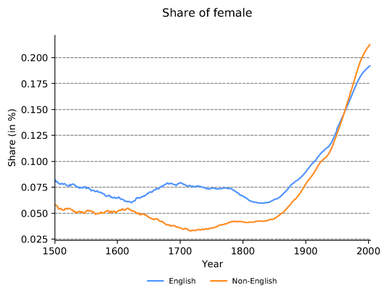

3. War episodes are noticeable as darker downward sloping lines corresponding to abnormal death rates during war periods.



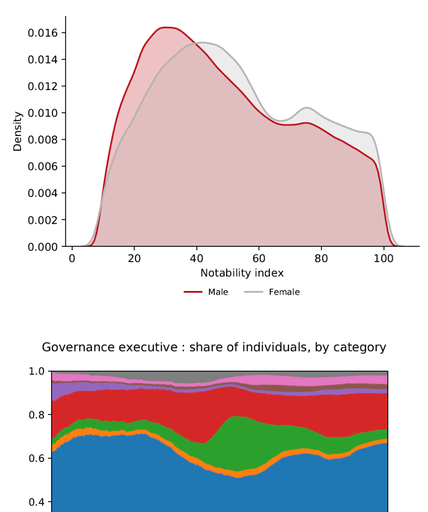

4. The English language edition contains relatively more visible individuals than specific language editions (with no equivalent in English).




5. The share of women in the database follows a U-shape pattern, with a minimum in 1800 and a maximum of 20% for the most recent cohorts.

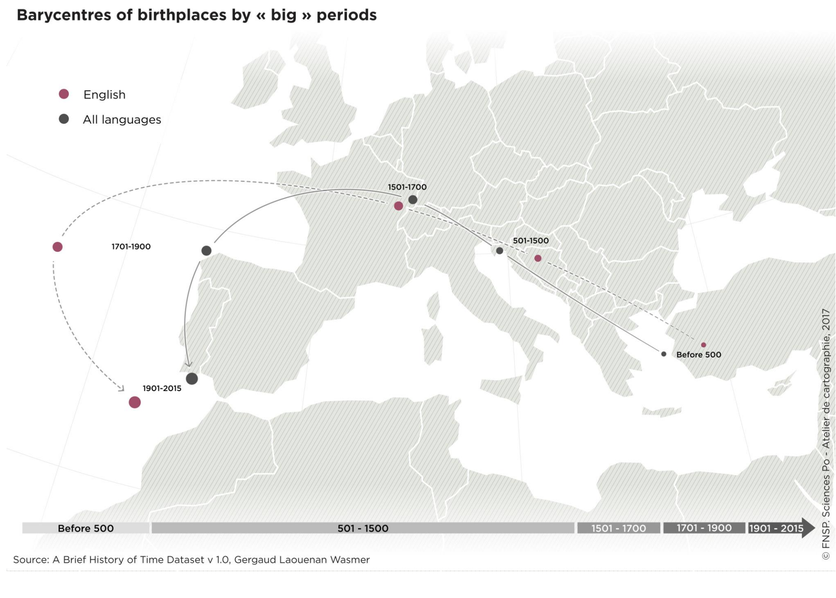



6. One can observe that women are on average more visible in Wikipedia once they are in.




7. Due to the inclusion of the French edition in the data collection, the number of politicians has been inflated significantly around year 1789 which corresponds to the beginning of the French Revolution.




SciencesPo
LABORATOIRE INTERDISCIPLINAIRE D'ÉVALUATION DES POLITIQUES PUBLIQUES

www.sciencespo.fr/liepp