Collaborative research between:

Historians, economists, demographers

Archivists

Experts in machine learning

Lionel Kesztenbaum
and the SocFace team

5th Conference of the
European Society of Historical Demography
Radboud University Nijmegen
08/30/2023

SocFace:

The local face of social change: one century of French social structure seen from the ground, 1836–1936

Collecting, processing, transcribing, and organizing all French individual census lists from 1836 to 1936 (20 censuses, metropolitan France only).

SocFace aims at producing a complete microdatabase of all individuals who lived in France between 1836 and 1936 and at using it to study social change in the long run.

Why Socface? (1) General context

- Microdata are increasingly important for research in quantitative social sciences: economics, history, sociology, demography...
- Explosion in quality of automated writing recognition (especially manual writing) and treatment of images.

Mass production of microdata at the national level

Develop new methods to extract individual-level data from a very large set of archival document images

Disseminate individual information produced to the general public, genealogists, and researchers

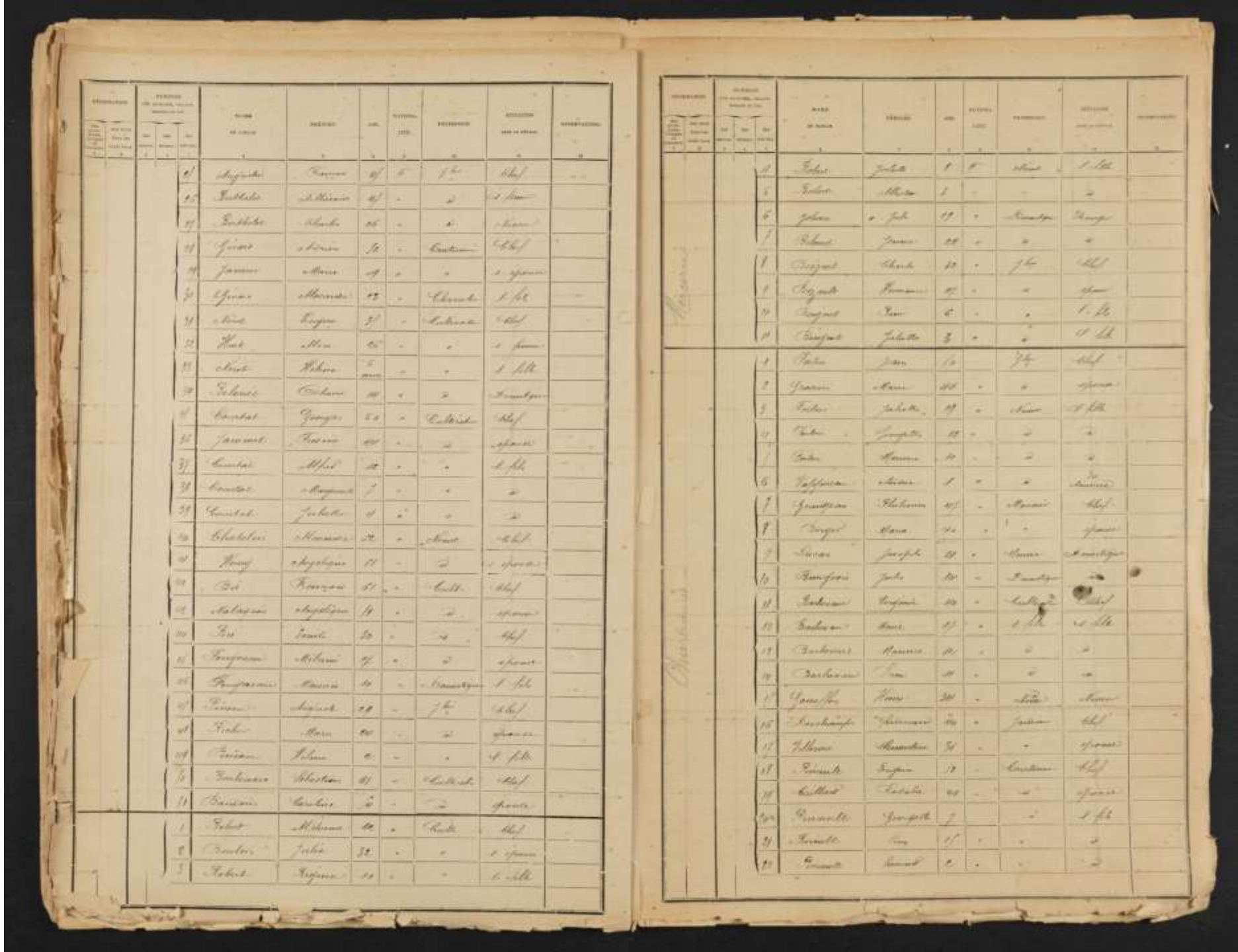
Why Socface? (2) Local context

- The decisive contribution of historical longitudinal data
  - Part of Europe and Asia: sources are natively longitudinal – population registers.
  - US, Canada, UK (and part of Europe): linked census.
- As of now, France is lagging behind
  - Many monographic works on specific areas, e.g., Le Creusot 1836-1886 (Bourdelais & Demonet).
  - A (unique) nation-wide sample: the TRA dataset. Sampling at 1/1000.
  - Need to combine the two: have individual and longitudinal data for France as a whole.
- Gap in knowledge:
  - In space: between national aggregates and local monographies.
  - In time: between the French Revolution and the 1960s, especially between 1919 and 1962.

Liste nominative:

La-Ferté-Saint-Aubin

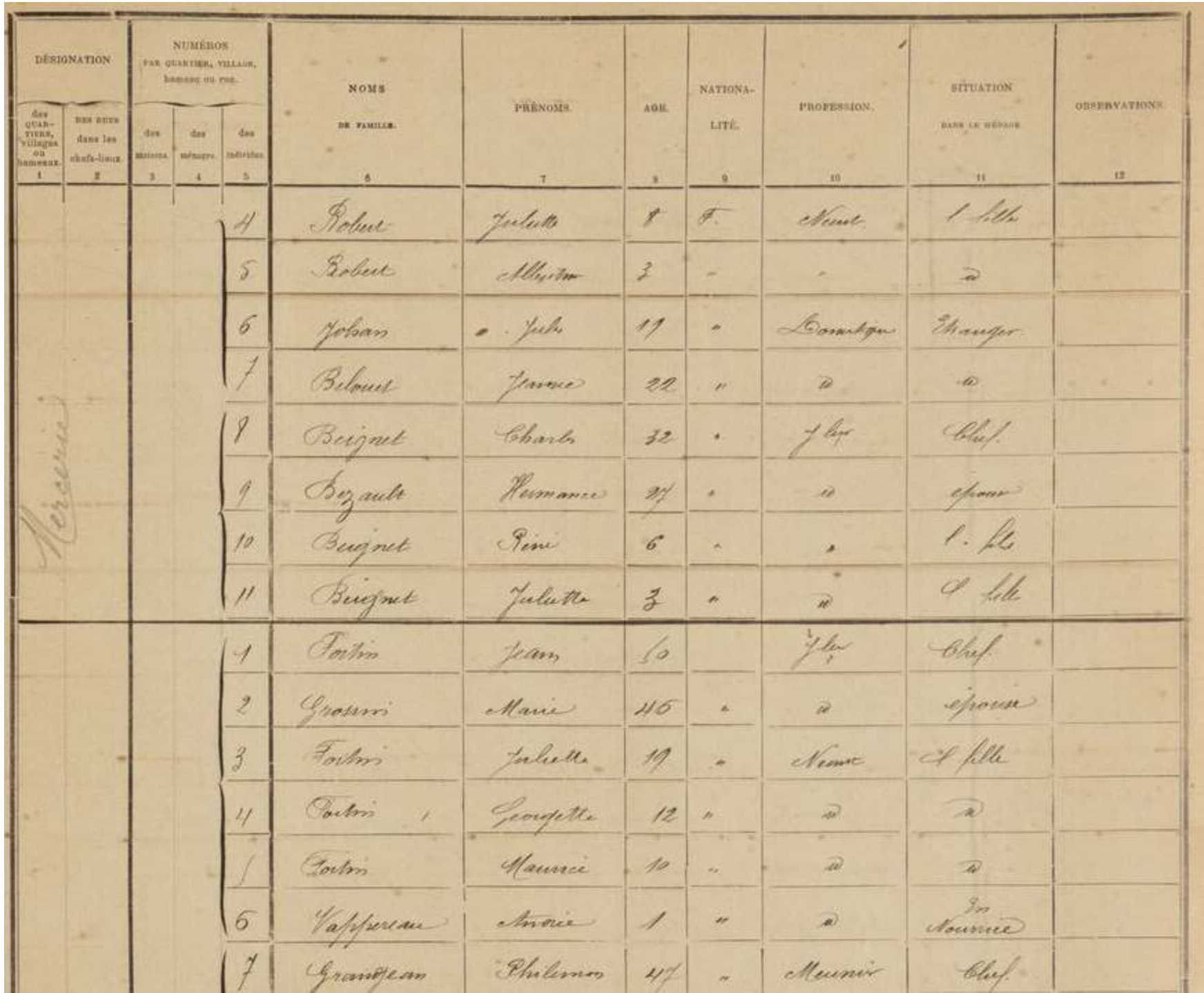
Loiret, 1896



Liste nominative:

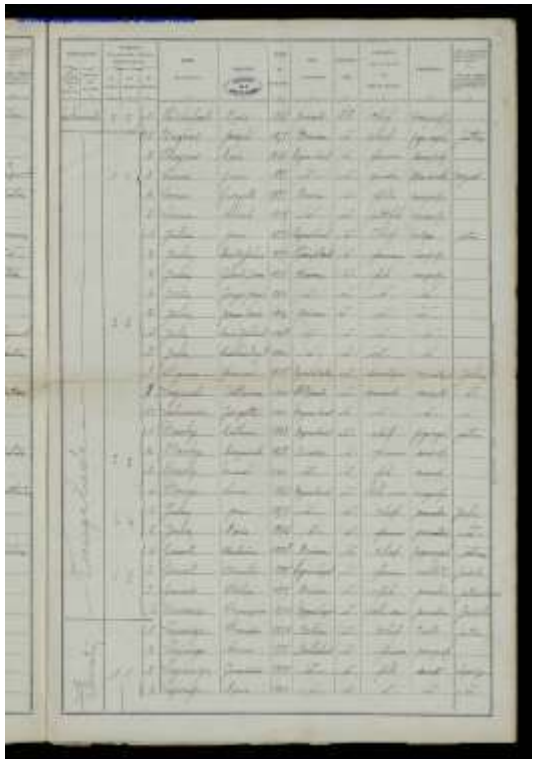
La-Ferté-Saint-Aubin

Loiret, 1896



DÉSIGNATION		NUMÉROS			NOMS	PRÉNOMS	AGE.	NATIONALITÉ.	PROFESSION.	SITUATION	OBSERVATIONS.
des quartiers, villages ou hameaux.	des rues dans les chefs-lieux	des maisons.	des ménages.	des individus.	DE FAMILLE.					DANS LE MÉNAGE	
1	2	3	4	5	6	7	8	9	10	11	12
				4	Robert	Juliette	8	F.	Néant	1 fille	
				5	Robert	Alberte	3	-	-	id	
				6	Jobard	Jules	19	"	Domestique	Etranger	
				7	Belmont	Jeanne	22	"	id	id	
				8	Beignet	Charles	32	"	Jardinier	Chef	
				9	Bezault	Romaine	27	"	id	épouse	
				10	Beignet	Rémi	6	"	"	1. fils	
				11	Beignet	Juliette	3	"	id	2 fille	
				1	Poitin	Jean	60		Jardinier	Chef	
				2	Grosset	Marie	46	"	id	épouse	
				3	Poitin	Juliette	19	"	Néant	1 fille	
				4	Poitin	Georgette	12	"	id	id	
				5	Poitin	Maurice	10	"	id	id	
				6	Vasseur	Marie	1	"	id	Sa Nourrice	
				7	Grandjean	Philemon	47	"	Meunier	Chef	

Why using *listes nominatives*?



- A standard, abundant, and quite simple source.
- A source that is (relatively) stable over time.
- Already digitalized by many archival depositories.
- A national, uniform source.
- Allows to build a database of France as a whole (almost…).

→ An ideal source for scaling up HTR.

→ A source that matters only at a (very) large scale.

Challenges and obstacles

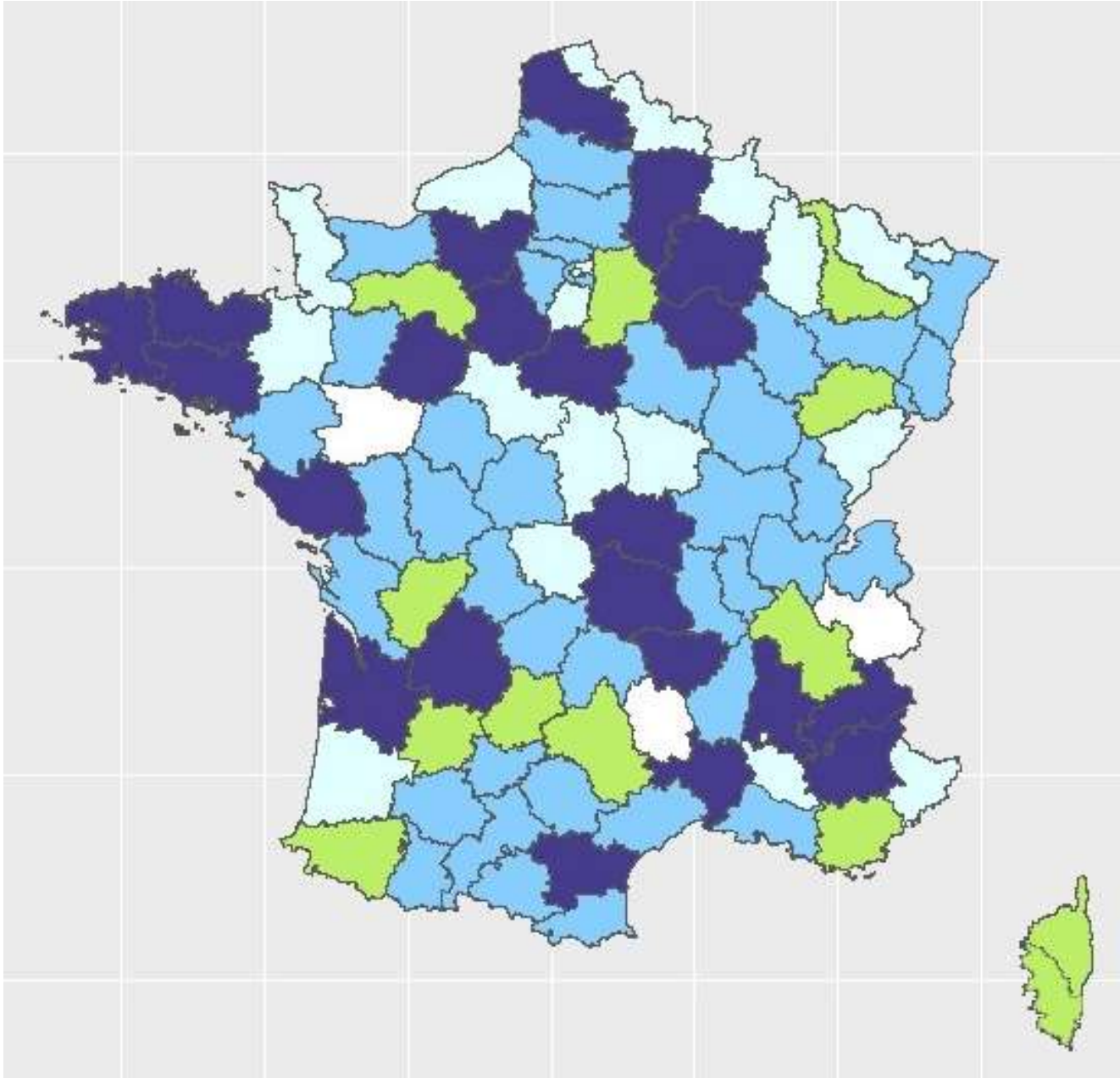
- Data recollection
  - Original sources and images are located at the département (100 of them) and municipality levels.
  - High heterogeneity in terms of conservation over time and space.
- Text transcription
  - Huge quantity of different writers, with different practices.
  - Very important heterogeneity of type of information entered, especially for abbreviation ('idem').
- Linking individuals across time and space
  - Important gap in information: whole areas are missing for some periods, no collective dwellings, etc.
  - Limited information on individuals (e.g., often only one first name).
- Social Science
  - Limited socio-economic information: only occupation, rather fragile.
  - Important gaps: no Paris (until 1926), missing areas, etc.

→ A common challenge: size, millions of images, hundreds of millions of records…

Collecting images

State of the project

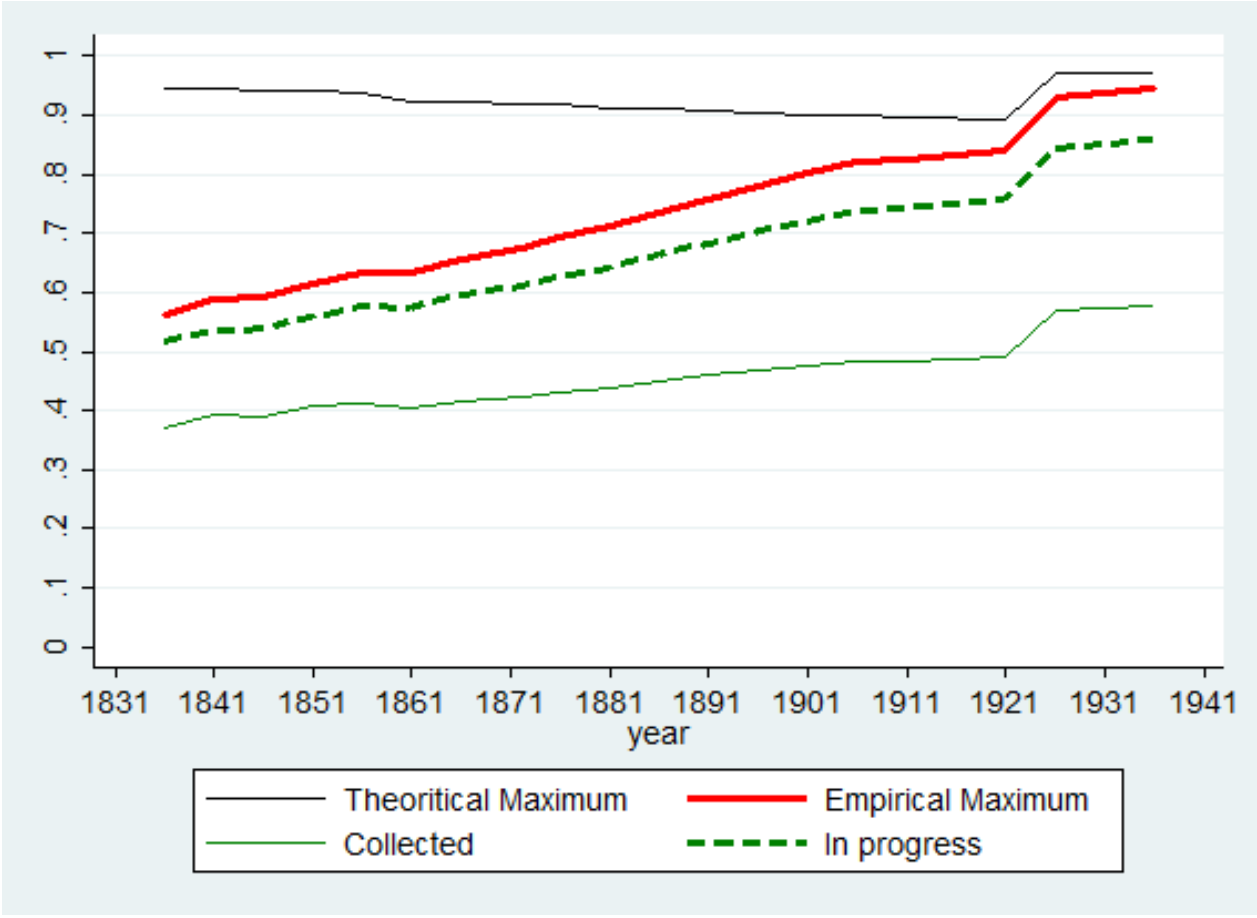
Images being collected	18
Images collected	41
Not digitalized	3
Not participating yet	13
Processing Images	22



A matter of size

- Theory: 20 census x 36 000 municipalities over 100 years
  - Between 10 and 20 million images.
  - Between 400 and 700 million entries (individual records).
- Reality (as of today):
  - Around 7 million images.
  - Around 300 million entries.
- Reality (in the end, rough estimates):
  - Around 12 million images.
  - Around 580 million entries.

Population coverage by SocFace (share of metropolitan France)



Challenges and obstacles

- Data recollection
  - Original sources and images are located at the département (100 of them) and municipality levels.
  - High heterogeneity in terms of conservation over time and space.
- Text transcription
  - Huge quantity of different writers, with different practices.
  - Very important heterogeneity of type of information entered, especially for abbreviation ('idem').
- Linking individuals across time and space
  - Important gap in information: whole areas are missing for some periods, no collective dwellings, etc.
  - Limited information on individuals (e.g., often only one first name).
- Social Science
  - Limited socio-economic information: only occupation, rather fragile.
  - Important gaps: no Paris (until 1926), missing areas, etc.

→ A common challenge: size, millions of images, hundreds of millions of records…

Challenges as objectives

- Data recollection
  - SocFace aims at giving a full picture of the situation of census conservation in France.
  - This also act as an incentive for archives to expand their collection, improve it, and digitalize it.
- Text transcription
  - Diversity of cases is not just a question of writing, but also in habits, practices, and so on.
  - Strong justification for collaboration between historians, demographers and ML experts.
- Linking individuals across time and space
  - Specific features of French census may help linking (e.g., maiden name for women).
  - Need to assess how HTR produce data affect quality of linking.
- Social Science
  - Allow to focus on understudied part of the country (far from Paris and other prominent areas).
  - Database will form the basis for other studies, using other sources.

The Socface Pipeline (1) Collecting and analyzing images

Archives (*départements* & cities)

Original images

By Commune & census year

Spatial structure (communes, population…)

Page splitting

Page classification

Summaries

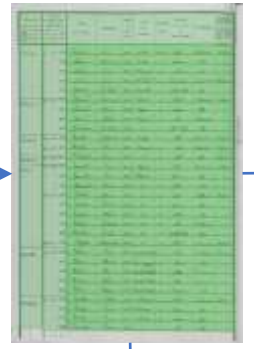
Listes

Varia

The Socface Pipeline (2) Building the database

Listes

Structure analysis



Household structure

Vocabularies/ frequencies

Individual linkage

Census reconstruction

Database

Random sample

Annotations

Collaborative transcriptions

Text transcription & entity recognition

Individual entries

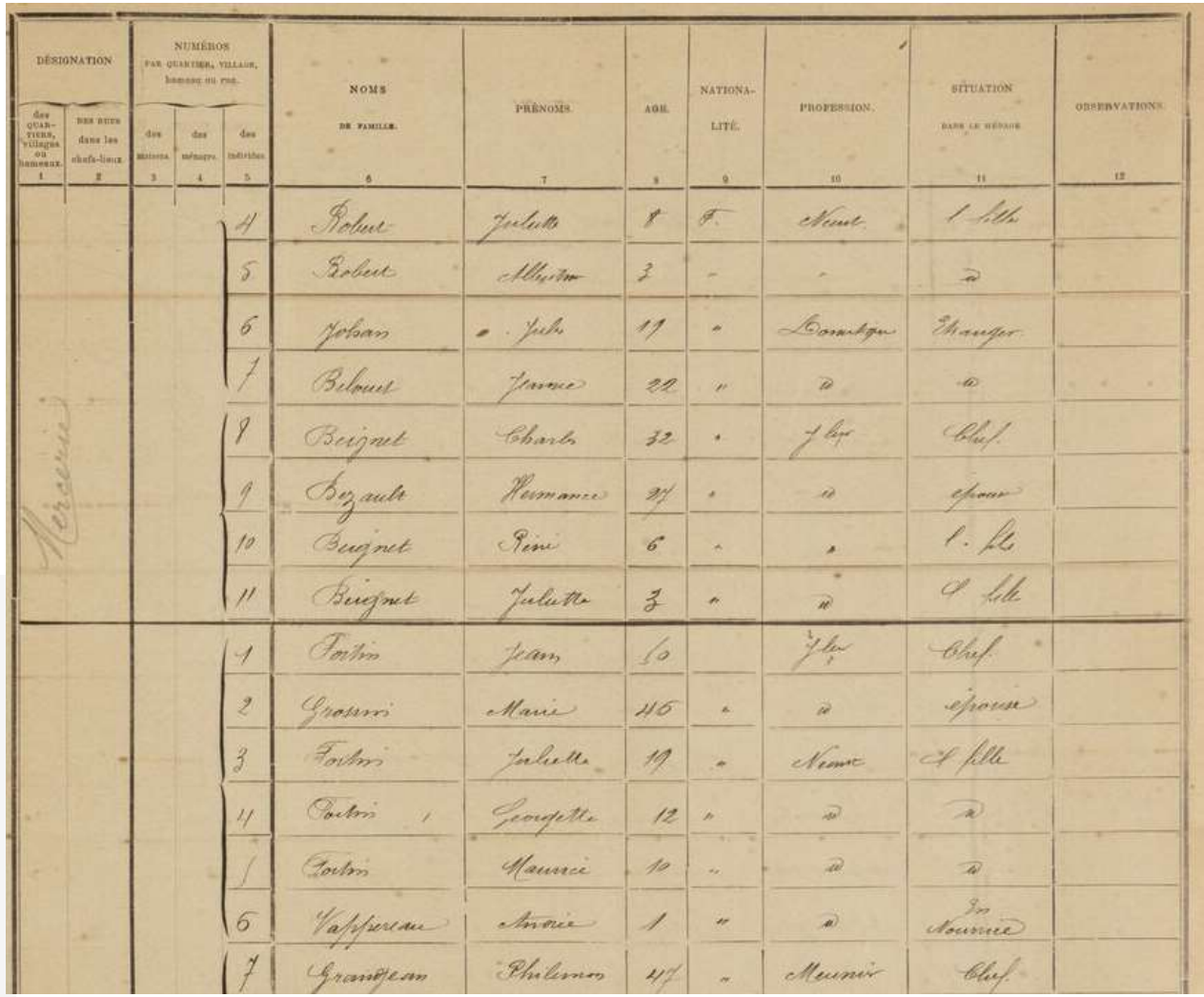
Coherence tests

Encoding information

Aggregated statistics

Example with transcription and named entity recognition

surname Robert firstname Juliette occupation néant link l. bille age 8 nationality française
surname Robert firstname Albertine occupation idem link idem age 3 nationality idem
surname Johan firstname Jules occupation domestique link chauger age 19 nationality idem
surname Belouet firstname Jeanne occupation idem link idem age 22 nationality idem
surname Beignet firstname Charles occupation j les link chef age 32 nationality idem
surname Bezault firstname Hermance occupation idem link épouse age 27 nationality idem
surname Bugnet firstname René occupation idem link l. fils age 6 nationality idem
surname Biignet firstname Juliette occupation idem link l. bille age 3 nationality idem
surname Portin firstname Jean occupation j leer link chef age 50 nationality idem
surname Gronni firstname Marie occupation idem link épouse age 46 nationality idem
surname Fortin firstname Juliette occupation néant link sa fille age 19 nationality idem
surname Portin firstname Georgette occupation idem link idem age 12 nationality idem
surname Portin firstname Maurice occupation idem link idem age 10 nationality idem
surname Vappereau firstname Andrée occupation idem link idem age 1 nationality idem
surname Grangeau firstname Philomèn occupation meunier link chef age 47 nationality idem



Linking individuals: building trajectories to study mobility

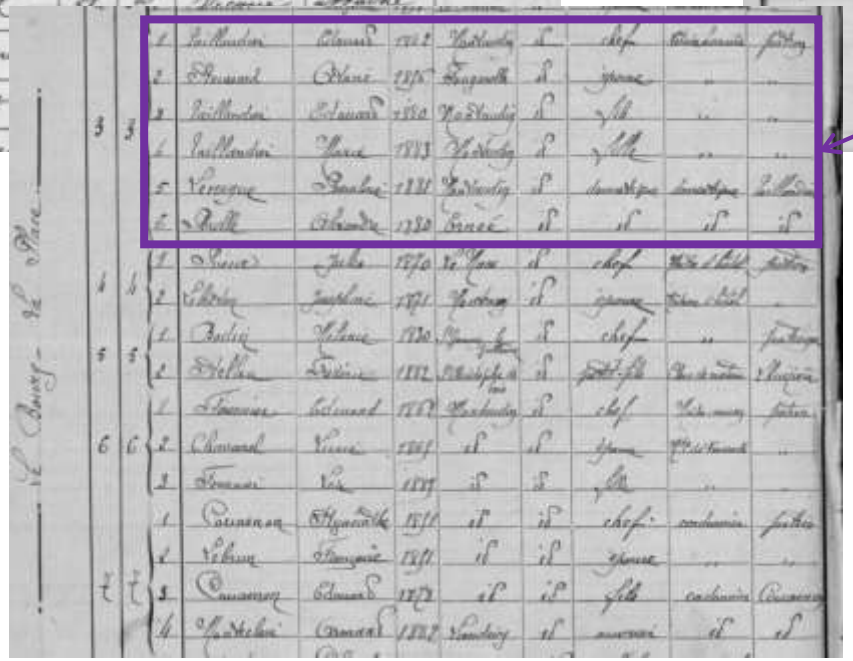
- Life-cycle trajectories from the census
  - Following individuals all their life, through time and space (at least in metropolitan France).
  - As children and in adulthood.
  - Also assessing the transition between the two.
- An essential tool to study mobility, both geographic and social, both within and between generations.
- But also how it changed over time (in 100 years).
- It is also decisive to add value to links with other sources.

Linking individuals between censuses

1886

1896

1906



Taking advantage of microdata at the national level

- Structural change in the long run
  - Transformation of the labor market: spatial variations, gender inequality, …
  - Evolution of transportations: effects on the spatial distribution of the population…
- Shocks
  - Short-, medium- and long-term consequences.
  - E.g., phylloxera crisis; World War One.
- Spatial organization of economic activities (project Landurb)
  - Linking individual information with spatial database.
  - Consequences of the transition from agriculture to industry at the very local level.
- And in the future?
  - The basis for future historical studies as a platform for contemporary quantitative history.
  - Link with other sources (civil registers, military registers, …) and databases (e.g., genealogical records).
  - Connection with the contemporary period.

Dissemination: back to Archival deposits, and beyond

- Raw database to be distributed by the Archives
  - On a national database (*FranceArchives*), with a search engine.
  - Sur les bases des Archives Départementales.
  - Direct links with the image.
- Encoded database to be distributed for research
  - A database where various information are organized and encoded (occupation, place of birth, etc.).
  - A database with probabilistic linkage between individuals.
- Opening on other sources: a model for disseminating French national archives?

Thank you!

https://socface.site.ined.fr/

http://socface.org

contact@socface.org