

SOC FACE



## 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 (Bourdelaïs & 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

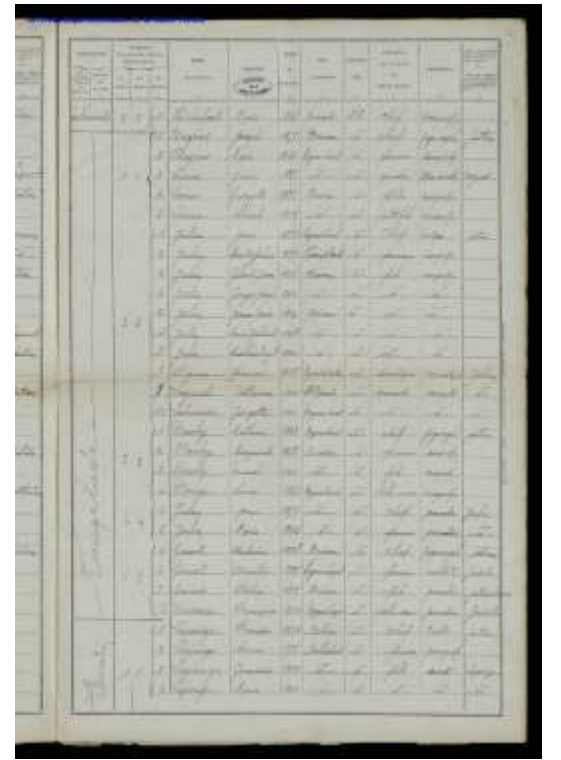
NOM	PRENOM	AGE	SEX	PROFESSION	SITUATION	FAMILIARITÉ	
						PRENOM	AGE
1	André	Renard	27	M	1 <sup>er</sup>	libre	-
2	André	Leblond	27	M	2	1 <sup>er</sup>	-
3	André	Leblond	26	M	3	1 <sup>er</sup>	-
4	André	Leblond	26	M	4	1 <sup>er</sup>	-
5	André	Leblond	26	M	5	1 <sup>er</sup>	-
6	André	Leblond	26	M	6	1 <sup>er</sup>	-
7	André	Leblond	26	M	7	1 <sup>er</sup>	-
8	André	Leblond	26	M	8	1 <sup>er</sup>	-
9	André	Leblond	26	M	9	1 <sup>er</sup>	-
10	André	Leblond	26	M	10	1 <sup>er</sup>	-
11	André	Leblond	26	M	11	1 <sup>er</sup>	-
12	André	Leblond	26	M	12	1 <sup>er</sup>	-
13	André	Leblond	26	M	13	1 <sup>er</sup>	-
14	André	Leblond	26	M	14	1 <sup>er</sup>	-
15	André	Leblond	26	M	15	1 <sup>er</sup>	-
16	André	Leblond	26	M	16	1 <sup>er</sup>	-
17	André	Leblond	26	M	17	1 <sup>er</sup>	-
18	André	Leblond	26	M	18	1 <sup>er</sup>	-
19	André	Leblond	26	M	19	1 <sup>er</sup>	-
20	André	Leblond	26	M	20	1 <sup>er</sup>	-
21	André	Leblond	26	M	21	1 <sup>er</sup>	-
22	André	Leblond	26	M	22	1 <sup>er</sup>	-
23	André	Leblond	26	M	23	1 <sup>er</sup>	-
24	André	Leblond	26	M	24	1 <sup>er</sup>	-
25	André	Leblond	26	M	25	1 <sup>er</sup>	-
26	André	Leblond	26	M	26	1 <sup>er</sup>	-
27	André	Leblond	26	M	27	1 <sup>er</sup>	-
28	André	Leblond	26	M	28	1 <sup>er</sup>	-
29	André	Leblond	26	M	29	1 <sup>er</sup>	-
30	André	Leblond	26	M	30	1 <sup>er</sup>	-
31	André	Leblond	26	M	31	1 <sup>er</sup>	-
32	André	Leblond	26	M	32	1 <sup>er</sup>	-
33	André	Leblond	26	M	33	1 <sup>er</sup>	-
34	André	Leblond	26	M	34	1 <sup>er</sup>	-
35	André	Leblond	26	M	35	1 <sup>er</sup>	-
36	André	Leblond	26	M	36	1 <sup>er</sup>	-
37	André	Leblond	26	M	37	1 <sup>er</sup>	-
38	André	Leblond	26	M	38	1 <sup>er</sup>	-
39	André	Leblond	26	M	39	1 <sup>er</sup>	-
40	André	Leblond	26	M	40	1 <sup>er</sup>	-
41	André	Leblond	26	M	41	1 <sup>er</sup>	-
42	André	Leblond	26	M	42	1 <sup>er</sup>	-
43	André	Leblond	26	M	43	1 <sup>er</sup>	-
44	André	Leblond	26	M	44	1 <sup>er</sup>	-
45	André	Leblond	26	M	45	1 <sup>er</sup>	-
46	André	Leblond	26	M	46	1 <sup>er</sup>	-
47	André	Leblond	26	M	47	1 <sup>er</sup>	-
48	André	Leblond	26	M	48	1 <sup>er</sup>	-
49	André	Leblond	26	M	49	1 <sup>er</sup>	-
50	André	Leblond	26	M	50	1 <sup>er</sup>	-
51	André	Leblond	26	M	51	1 <sup>er</sup>	-
52	André	Leblond	26	M	52	1 <sup>er</sup>	-
53	André	Leblond	26	M	53	1 <sup>er</sup>	-
54	André	Leblond	26	M	54	1 <sup>er</sup>	-
55	André	Leblond	26	M	55	1 <sup>er</sup>	-
56	André	Leblond	26	M	56	1 <sup>er</sup>	-
57	André	Leblond	26	M	57	1 <sup>er</sup>	-
58	André	Leblond	26	M	58	1 <sup>er</sup>	-
59	André	Leblond	26	M	59	1 <sup>er</sup>	-
60	André	Leblond	26	M	60	1 <sup>er</sup>	-
61	André	Leblond	26	M	61	1 <sup>er</sup>	-
62	André	Leblond	26	M	62	1 <sup>er</sup>	-
63	André	Leblond	26	M	63	1 <sup>er</sup>	-
64	André	Leblond	26	M	64	1 <sup>er</sup>	-
65	André	Leblond	26	M	65	1 <sup>er</sup>	-
66	André	Leblond	26	M	66	1 <sup>er</sup>	-
67	André	Leblond	26	M	67	1 <sup>er</sup>	-
68	André	Leblond	26	M	68	1 <sup>er</sup>	-
69	André	Leblond	26	M	69	1 <sup>er</sup>	-
70	André	Leblond	26	M	70	1 <sup>er</sup>	-
71	André	Leblond	26	M	71	1 <sup>er</sup>	-
72	André	Leblond	26	M	72	1 <sup>er</sup>	-
73	André	Leblond	26	M	73	1 <sup>er</sup>	-
74	André	Leblond	26	M	74	1 <sup>er</sup>	-
75	André	Leblond	26	M	75	1 <sup>er</sup>	-
76	André	Leblond	26	M	76	1 <sup>er</sup>	-
77	André	Leblond	26	M	77	1 <sup>er</sup>	-
78	André	Leblond	26	M	78	1 <sup>er</sup>	-
79	André	Leblond	26	M	79	1 <sup>er</sup>	-
80	André	Leblond	26	M	80	1 <sup>er</sup>	-
81	André	Leblond	26	M	81	1 <sup>er</sup>	-
82	André	Leblond	26	M	82	1 <sup>er</sup>	-
83	André	Leblond	26	M	83	1 <sup>er</sup>	-
84	André	Leblond	26	M	84	1 <sup>er</sup>	-
85	André	Leblond	26	M	85	1 <sup>er</sup>	-
86	André	Leblond	26	M	86	1 <sup>er</sup>	-
87	André	Leblond	26	M	87	1 <sup>er</sup>	-
88	André	Leblond	26	M	88	1 <sup>er</sup>	-
89	André	Leblond	26	M	89	1 <sup>er</sup>	-
90	André	Leblond	26	M	90	1 <sup>er</sup>	-
91	André	Leblond	26	M	91	1 <sup>er</sup>	-
92	André	Leblond	26	M	92	1 <sup>er</sup>	-
93	André	Leblond	26	M	93	1 <sup>er</sup>	-
94	André	Leblond	26	M	94	1 <sup>er</sup>	-
95	André	Leblond	26	M	95	1 <sup>er</sup>	-
96	André	Leblond	26	M	96	1 <sup>er</sup>	-
97	André	Leblond	26	M	97	1 <sup>er</sup>	-
98	André	Leblond	26	M	98	1 <sup>er</sup>	-
99	André	Leblond	26	M	99	1 <sup>er</sup>	-
100	André	Leblond	26	M	100	1 <sup>er</sup>	-

Liste nominative:  
La-Ferté-Saint-Aubin  
Loiret, 1896

DÉSIGNATION		NUMÉROS PAR QUARTIER, VILLAGE, BOULEVARD OU RUE.			NOMS	PRÉNOMS	AGE	NATIONA- LITÉ.	PROFESSION.	SITUATION	OBSERVATIONS.
des QUAR- TIERS, VILLAGES OU BOULEVARD.	DES RUES dans les chefs-lieux	des maisons	des cabanons	des indivisibles	DE FAMILLE.					DANS LE MÉRIAGE	
1	2	3	4	5	6	7	8	9	10	11	12
				4	Robut	Juliette	8	F.	ouv.	1 fille	
				5	Robut	Allyette	3	"	"	"	
				6	Jehan	Jeh.	17	"	Compteur	Marier	
				7	Bilout	Jeanne	22	"	"	"	
				8	Beignet	Charles	32	"	J. l'gr	Chef.	
				9	Syzault	Hermanne	24	"	"	épouse	
				10	Beignet	Rini	6	"	"	1. fil	
				11	Beignet	Juliette	3	"	"	1. fil	
				1	Fotin	Jean	50	"	J. l'gr	Chef.	
				2	Grasseri	Marie	45	"	"	épouse	
				3	Fotin	Juliette	19	"	Nouv.	1. fille	
				4	Fotin	Georgette	12	"	"	"	
				5	Fotin	Maurice	10	"	"	"	
				6	Vassureau	Stroie	1	"	"	3 <sup>es</sup> Nouv.	
				7	Grandjean	Philémon	47	"	Muniir	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.

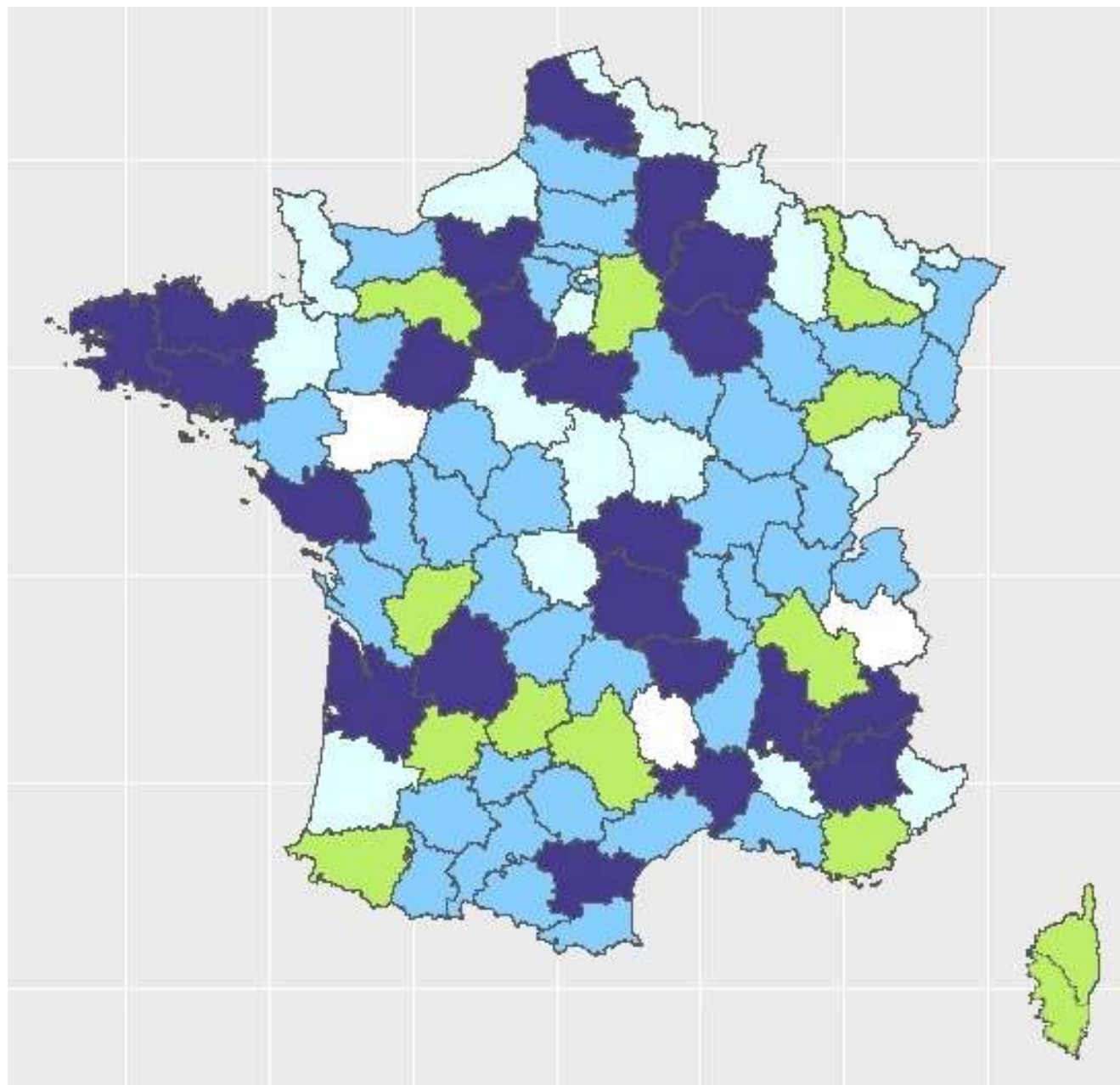
➔ A common challenge: **size**, millions of images, hundreds of millions of records...



# 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).

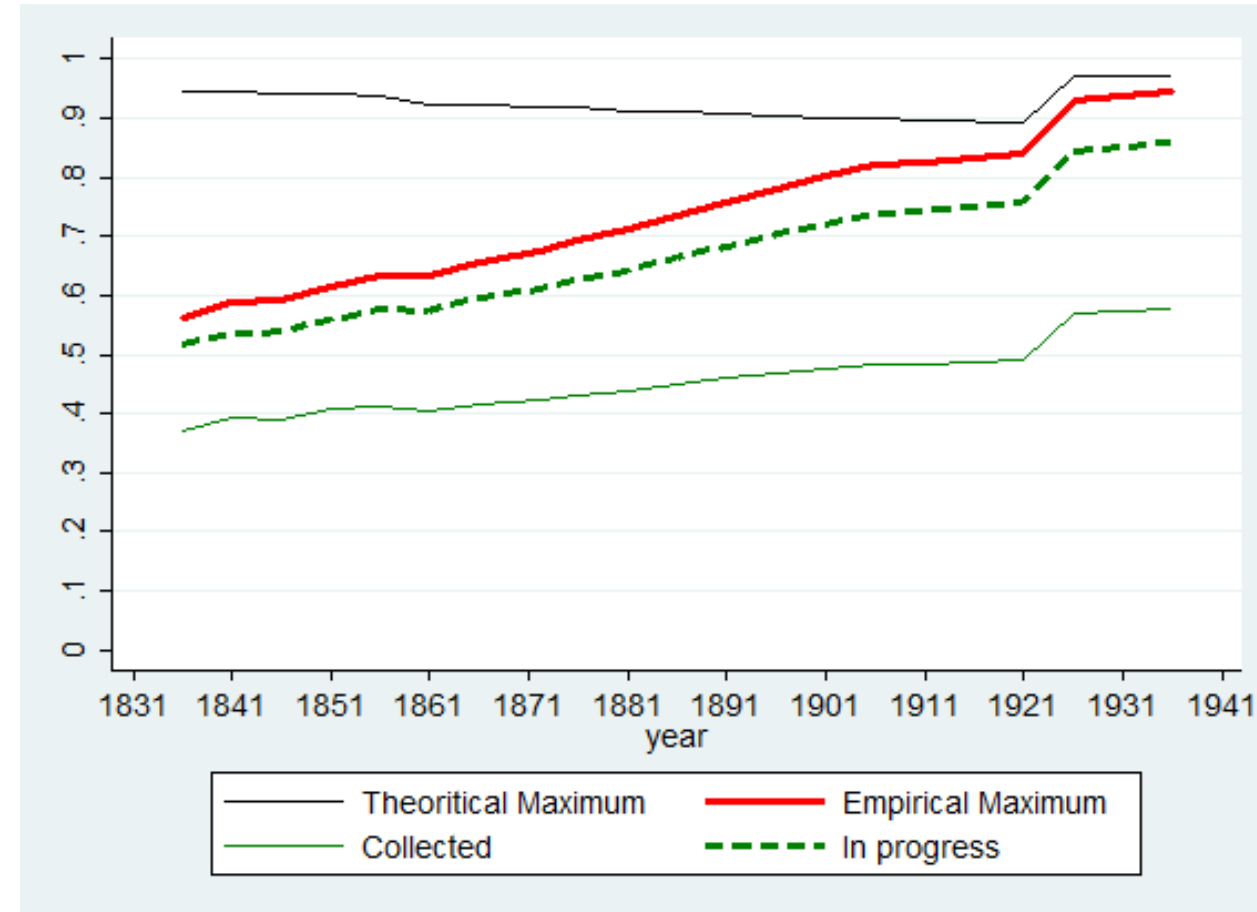
➤ Reality (as of today):

- ❖ 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.

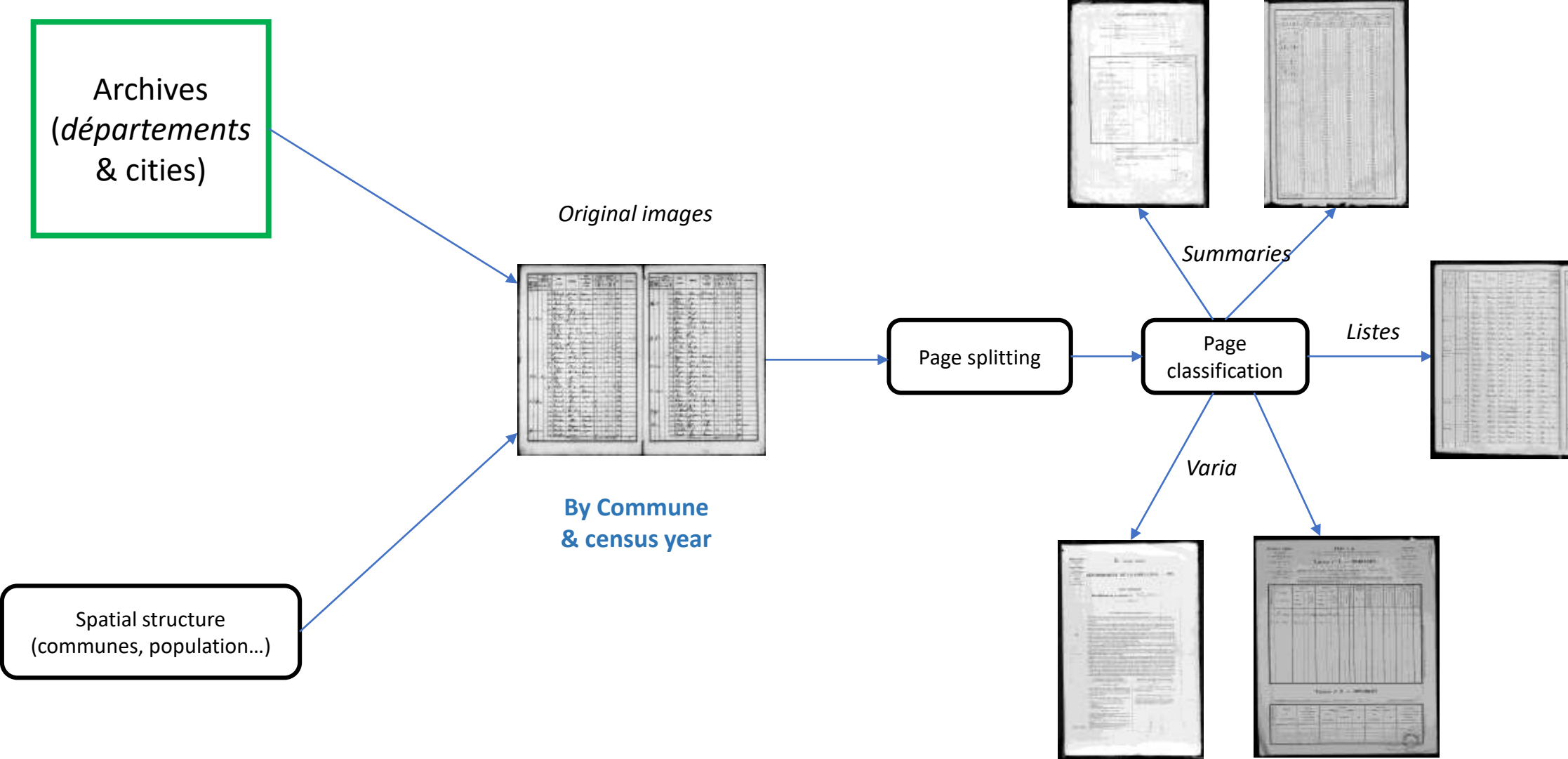
## ➤ Linking individuals across time and space

- ❖ 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.

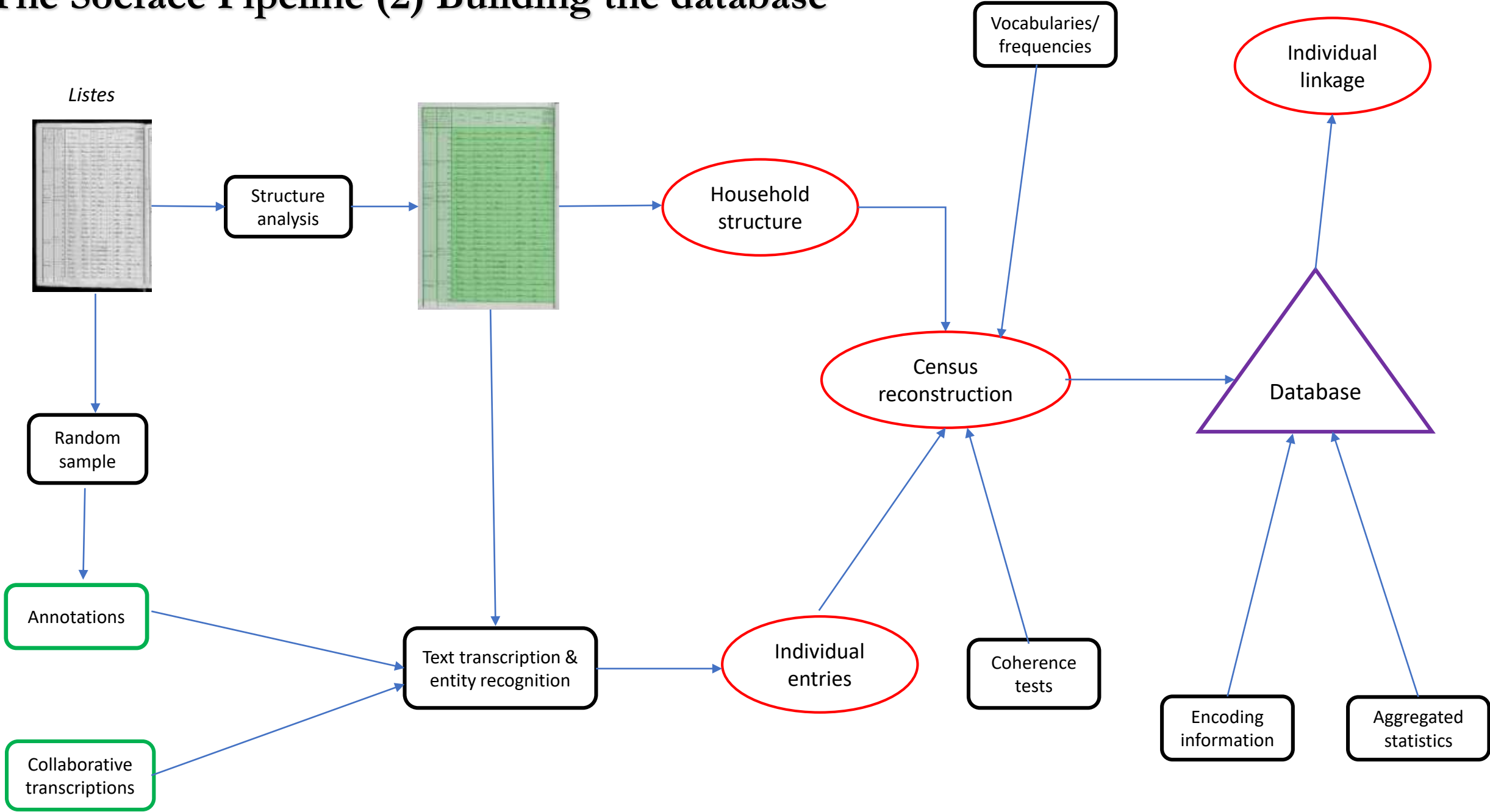
## ➤ Social Science

- ❖ 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



# The Socface Pipeline (2) Building the database



# Example with transcription and named entity recognition

DESIGNATION	NUMÉROS PAR QUARTIER, VILLAGE, BOULEVARD OU RUE.			NOMS DE FAMILLE	PRÉNOMS	AGE	NATIONALITÉ	PROFESSION	SITUATION DANS LE MÉNAGE	OBSERVATIONS
	des QUARTIERS, VILLAGES ou BOULEVARDS	DES RUES, BOULEVARDS ou PASSAGES	des maisons							
1	2	3	4	5	6	7	8	9	10	11
Maraîchers			4		Robert	Juliette	8	F.	néant	l. fille
			5		Robert	Albertine	3	-	-	»
			6		Johan	Jules	19	»	domestique	chauger
			7		Belouet	Jeanne	22	»	»	»
			8		Beignet	Charles	32	»	»	chef
			9		Bezault	Hermance	27	»	»	épouse
			10		Beignet	René	6	»	»	l. fils
			11		Beignet	Juliette	3	»	»	l. fille
			1		Portin	Jean	50	»	»	chef
			2		Gronni	Marie	46	»	»	épouse
			3		Fortin	Juliette	19	»	»	l. fille
		11		Portin	Georgette	12	»	»	»	
		1		Portin	Maurice	10	»	»	»	
		6		Vappereau	Andrée	1	»	»	l. épouse	
		7		Grangeau	Philomèn	47	»	meunier	chef	

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

La Place.

31	36	92	Baillandier	Edmond	44	♂	Noblesse	chef de ménage
		93	Houssard	Octavie	30	♀	propriétaire	sa femme
		94	Baillandier	Edmond	5	♂	"	son fils
		95	Baillandier	Maurice	3	♂	"	son fils
		96	Boussard	Welfépine	72	♀	propriétaire	sa femme
		97	Bouzelet	Emile	20	♂	ouvrier	ouvrier
32	36	98	Nelamoy	Amable	40	♂	Commissaire	chef de ménage
		99	Beaudou	Eugénie	32	♀	M <sup>re</sup> au commerce	sa femme
33	37	100	Chapin	Julien	43	♂	rentier	chef de ménage
		101	Ferat	Mari Anne	60	♀	rentière	sa femme
		102	Julie	Joseph	42	♂	ouvrier	chef de ménage
34	38	103	Grandin	Nicolas	22	♂	Noblesse	ouvrier
39		104	Bichard	Jeann				
40		105	Briguel	Marie				

1896

S. Bourg - La Place

1	Julie	Joseph	72	♂	rentier	chef
2	Renard	Marie	42	♀	rentière	
3	Buchard	Jeann	47	♀	propriétaire	
4	Chodose	Marie Charles	40	♀	propriétaire	chef
5	Nelamoy	Amable	50	♂	propriétaire	époux
6	Grandin	Eugénie	43	♀	"	épouse
7	Baillandier	Edmond	53	♂	rentier	époux
8	Houssard	Octavie	40	♀	propriétaire	épouse
9	Kobus	Marie	37	♀	propriétaire	
10	Vary	Julie	35	♀	rentière	épouse
11	Mile	Marie	36	♀	"	épouse
12	Vary	Julie	11	♀	"	fils
13	Vary	Marie	7	♀	"	id
14	Mile	Eugénie	27	♀	propriétaire	sa femme

1906

S. Bourg - La Place

1	Baillandier	Edmond	1882	44	♂	chef	commissaire	époux
2	Boussard	Marie	1885	30	♀	épouse		
3	Baillandier	Edmond	1880	5	♂	"		
4	Baillandier	Maurice	1883	3	♂	"		
5	Boussard	Welfépine	1882	72	♀	propriétaire	sa femme	
6	Bouzelet	Emile	1880	20	♂	ouvrier	ouvrier	
7	Nelamoy	Amable	1870	40	♂	Commissaire	chef de ménage	
8	Beaudou	Eugénie	1881	32	♀	M <sup>re</sup> au commerce	sa femme	
9	Chapin	Julien	1883	43	♂	rentier	chef de ménage	
10	Ferat	Mari Anne	1880	60	♀	rentière	sa femme	
11	Julie	Joseph	1882	42	♂	ouvrier	chef de ménage	
12	Grandin	Nicolas	1882	22	♂	Noblesse	ouvrier	
13	Bichard	Jeann						
14	Briguel	Marie						

# 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\*\*](mailto:contact@socface.org)