

SOCFACE



Suivre la vie des françaises et des français sur un siècle à partir des archives du recensement: le projet Socface

Christopher Kermorvant
Lionel Kesztenbaum



Colloque Société de Démographie
Historique
11/23/2023

SOC FACE



Collaborative research between:

Historians, economists, demographers

Archivists

Experts in machine learning



Christopher Kermorvant
Lionel Kesztenbaum



Colloque Société de Démographie
Historique
11/23/2023

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).

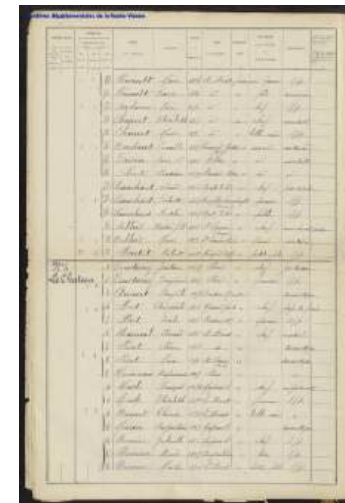
SocFace aims at producing a complete microdatabase of all individuals who lived in France between 1836 and 1936 and at using it to study social change in the long run.



A page from a French census list, showing a grid of handwritten entries. The page is oriented vertically and contains several columns of text, likely representing individual names and their characteristics.



A page from a French census list, showing a grid of handwritten entries. The page is oriented vertically and contains several columns of text, likely representing individual names and their characteristics.



A page from a French census list, showing a grid of handwritten entries. The page is oriented vertically and contains several columns of text, likely representing individual names and their characteristics.

Why Socface? 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**

DÉSIGNATION		NUMÉROS, PAR QUARTIER, VILLAGES, HOMMES EN FEM.			NOMS	PRÉNOMS	ANNÉE	LIEU	NATIONALITÉ	SITUATION	PROFESSION	REMARQUES
des maisons, villages ou hameaux	des maisons des villages	des maisons	des maisons	des maisons	DE FAMILLE		de naissance	de naissance	LITTÉ.	ou chef de ménage		Par les patrons, chefs d'ateliers, ma- nistres & journaliers, in- diquer le nom du patron ou de l'atelier qui les em- ploie.
1	2	3	4	5	6	7	8	9	10	11	12	13
				2	Tribout	Lucie	1886	Reims	F	épouse	ans	
				3	Tribout	Lenni	1909	Paris	"	enfant	"	
				4	Tribout	Alphonse	1912	"	"	"	"	
				5	Tribout	Pierre	1916	Aubervilliers	"	"	"	
				6	Tribout	Audré	1920	"	"	"	"	
				7	Pierre	Henriette	1881	Reims	"	bell. soeur	Magasinier	Bordonneux Aubervilliers
				8	Pierre	Liliane	1912	Bagnollet	"	filie	ans	
				9	Pierre	Lucien	1918	Woisyloté	"	veuve	ans	
				10	Parmenier	Justave	1877	Weschateau	"	Chef	maçon	
				11	Parmenier	Lucie	1867	Lambach	"	épouse	ans	
				12	Parmenier	Maximilien	1885	Paris	"	enfant	Confectionnier	Thulinant
				13	Parmenier	Emma	1900	M. Denis	"	"	empl.	Minist. P. et M.
				14	Parmenier	Ludovic	1902	"	"	"	filie	Alix
				15	Caliez	Jorges	1878	Lill.	"	Chef	Condamner	Magist.
				16	Caliez	Celine	1884	Taché	"	épouse	ans	
				17	Renault	Jean	1869	Combrigny	"	Chef	chef équipe	Magist.

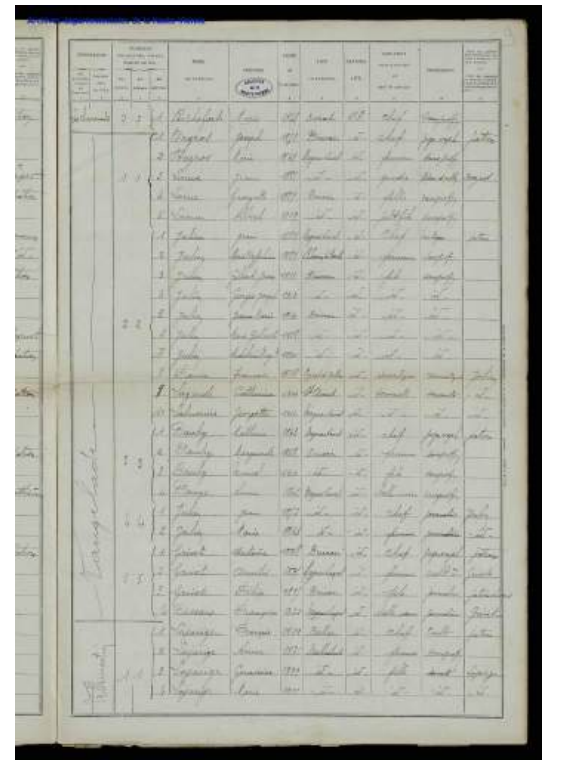
Liste nominative:

Aubervilliers, 1921

Rue
des
Fillettes

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 and NER
 - A source that matters only at a (very) large scale.
 - Individual data at the national level.



Integration of HTR and NER



Sequential approach

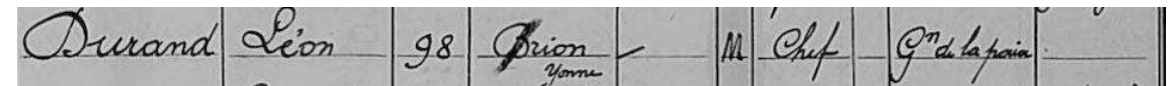
DAN or PyLaia for HTR
SpaCy for NER

End-to-end approach

DAN for HTR+NER
Special tokens are used to tag entities

Integration of HTR and NER

Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
GT + SpaCy	0.0	96.4	Line
PyLaia + SpaCy	17.19	76.3	Line
DAN + SpaCy	8.18	84.0	Page
DAN end-to-end	7.83	85.9	Page



structured text representation of the handwritten document above, with labels for named entities:

surname Durand firstname Léon birth_date 98 lob Brion Yonne civil_status M
link Chef occupation Gardien de la paix

- End-to-end models are competitive
- End-to-end models are easier to train, improve and deploy

Ground-truth preparation

MÉROS DES VILLAGES, OU DE RUE		NOMS	ANNÉE	LIEU	NATIONA-	SITUATION	PROFESSION.
des	des	DE FAMILLE	de	de	LITÉ.	PAR RAPPORT	
maisons.	individus.		NAISSANCE.	NAISSANCE.	LITTE.	au chef de ménage.	
4	5	6	7	8	9	10	11
10		Herouel Jean François	1910	Stand	français	Fils	
11		Cocciac Nicolas	1860	Ormeaux		Domestique	Paul
12		Pérot François	1913	Stand		d°	domestique
13		Pérot Jeanne	1915	d°		d°	d°
14		M ^{lle} Pelleu Amélie	1886	d°		Châtim	Paul
15		Pelleu Amélie	1911	d°		Fille	Costance
16		Kalamine Victorine	1896	d°		Châtim	Paul
17		Kalamine Lucie	1914	d°		Fille	
18		Kalamine Henri	1907	d°		Fils	
19		M ^{lle} Arzel Victoire	1890	d°		Châtim	Repasser
20		Arzel Henri	1915	d°		Fils	Pichon
21		Arzel Yves	1916	d°		Fils	Sans
22		Oulhen Françoise	1854	d°		Châtim	Retraite
23		Oulhen Charles	1855	d°		Trouv	Sans
24		Oulhen Madeleine	1857	d°		Fille	Sans
25		Piquet Yves	1853	d°		Châtim	Retraite
26		Piquet Victoire	1879	d°		Fille	Sans
27		M ^{lle} Pelleu Sophie	1853	d°		Châtim	Sans
28		Oulhen Marie Anne	1866	d°		Châtim	Sans
29		Leanne Jean	1876	Châtim		Châtim	Commercia

Nom
Veuve Arzel

Prénoms
Victoire

Profession
Repasseuse

Situation par rapport au chef de ménage
Chef de ménage

Employeur
Your annotation

Age
Your annotation

Date de naissance
1890

Etat civil (choix parmi : Garçon / Homme / Veuf / Fille / Femme / Veuve)
Veuve

Nationalité
idem

Lieu de naissance
idem

Observations
Your annotation

Skip task Annotate

Not transcribing, but filling a form

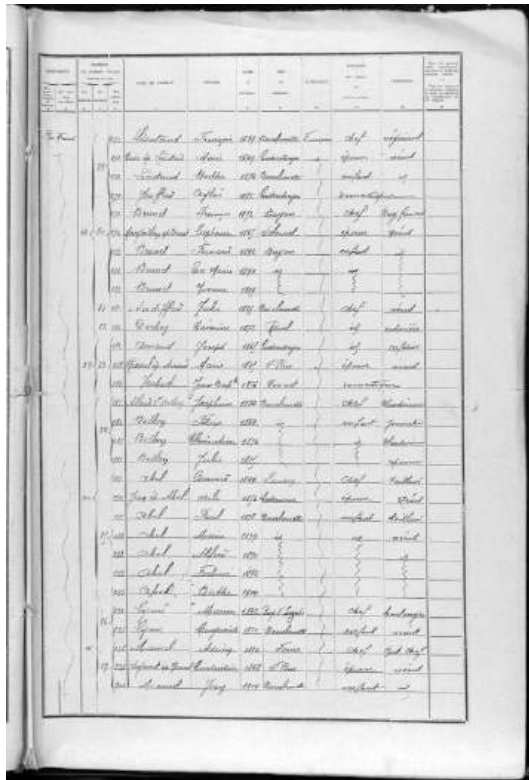
A single form is used for different document documents: all the fields are not always filled

No interpretation of the data: transcribe “idem”

Open-source platform:
<https://gitlab.tekalia.com/callico>

One step beyond : Integrated DLA+HTR+NER

A single model for text line detection, text recognition and typing from full page



```
sumame Léautaud firstnae François occupation négociant link chef birth_date 1839 nationality française lab Barcelonnette
sumame Caire ep Léautaud firstnae Marie occupation néant link épouse birth_date 1849 nationality idem lab Enchartrayes
sumame Léautaud firstnae Berthe occupation idem link enfant birth_date 1876 nationality idem lab Barcelonnette
sumame Jauffred firstnae Aglaé occupation idem link domestique birth_date 1875 nationality idem lab Enchartrayes
sumame Brunet firstnae François occupation brig ferrier link chef birth_date 1852 nationality idem lab Bayons
sumame Margailan ep Brunet firstnae Euphanie occupation néant link épouse birth_date 1865 nationality idem lab Selonnet
sumame Brunet firstnae Fernand occupation idem link enfant birth_date 1892 nationality idem lab Bayons
sumame Brunet firstnae Eva Marie occupation idem link idem birth_date 1894 nationality idem lab idem
sumame Brunet firstnae Yvonne occupation idem link idem birth_date 1899 nationality idem lab idem
sumame Audiffred firstnae Julie occupation néant link chef birth_date 1825 nationality idem lab Barcelonnette
sumame Derbez firstnae Césarine occupation culurière link idem birth_date 1872 nationality idem lab Revel
sumame Arnaud firstnae Joseph occupation cafetier link idem birth_date 1865 nationality idem lab Enchartrayes
sumame Risoulép Arnaud firstnae Marie occupation néant link épouse birth_date 1865 nationality idem lab St Pons
sumame Jaubert firstnae Jean Baptiste link domestique birth_date 1876 nationality idem lab Nvermet
sumame Allard vve Bellon firstnae Joséphine occupation blanchisseuse link chef birth_date 1834 nationality idem lab Barcelonnette
sumame Bellon firstnae Félix occupation journalier link enfant birth_date 1864 nationality idem lab idem
sumame Bellon firstnae Clémentine occupation blanchisseuse link idem birth_date 1876 nationality idem lab idem
sumame Bellon firstnae Julie occupation repasseuse link idem birth_date 1875 nationality idem lab idem
sumame Abel firstnae Edouard occupation tailleur link chef birth_date 1849 nationality idem lab Faucon
sumame Jean ép Abel firstnae Odile occupation néant link épouse birth_date 1856 nationality idem lab Condarnine
sumame Abel firstnae Paul occupation tailleur link enfant birth_date 1878 nationality idem lab Barcelonnette
sumame Abel firstnae Marie occupation néant link idem birth_date 1879 nationality idem lab idem
sumame Abel firstnae Alfred occupation idem link idem birth_date 1890 nationality idem lab idem
sumame Abel firstnae Fortuné occupation idem link idem birth_date 1892 nationality idem lab idem
sumame Abel firstnae Berthe occupation idem link idem birth_date 1900 nationality idem lab idem
sumame Eymé firstnae Marius occupation boulanger link chef birth_date 1842 nationality idem lab Puy de Euzèbe
sumame Lymé firstnae Marguerite occupation néant link enfant birth_date 1880 nationality idem lab Barcelonnette
sumame Maurel firstnae Adrien occupation cant link chef birth_date 1870 nationality idem lab Fours
sumame Signout ep Maurel firstnae Constantine occupation néant link épouse birth_date 1868 nationality idem lab S. Pon
sumame Maurel firstnae Jean occupation idem link enfant birth_date 1904 nationality idem lab Barcelomarle
```

Extension of DAN (Coquenot *et al.* 2023) to NER (Tarride *et al.* 2023)

One step beyond : Integrated DLA+HTR+NER

Evaluation on POP Test Set

Model	LEVEL	CER % ↓	F1 (%) ↑
DAN HTR+NER	Line*	7.8	85.9
DAN DLA+HTR+NER	Page	11.7	85.3

- Full page end-to-end models are competitive
- They are even easier to train, improve and deploy
- Ground truth generation is easier (can be trained from page transcription)

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

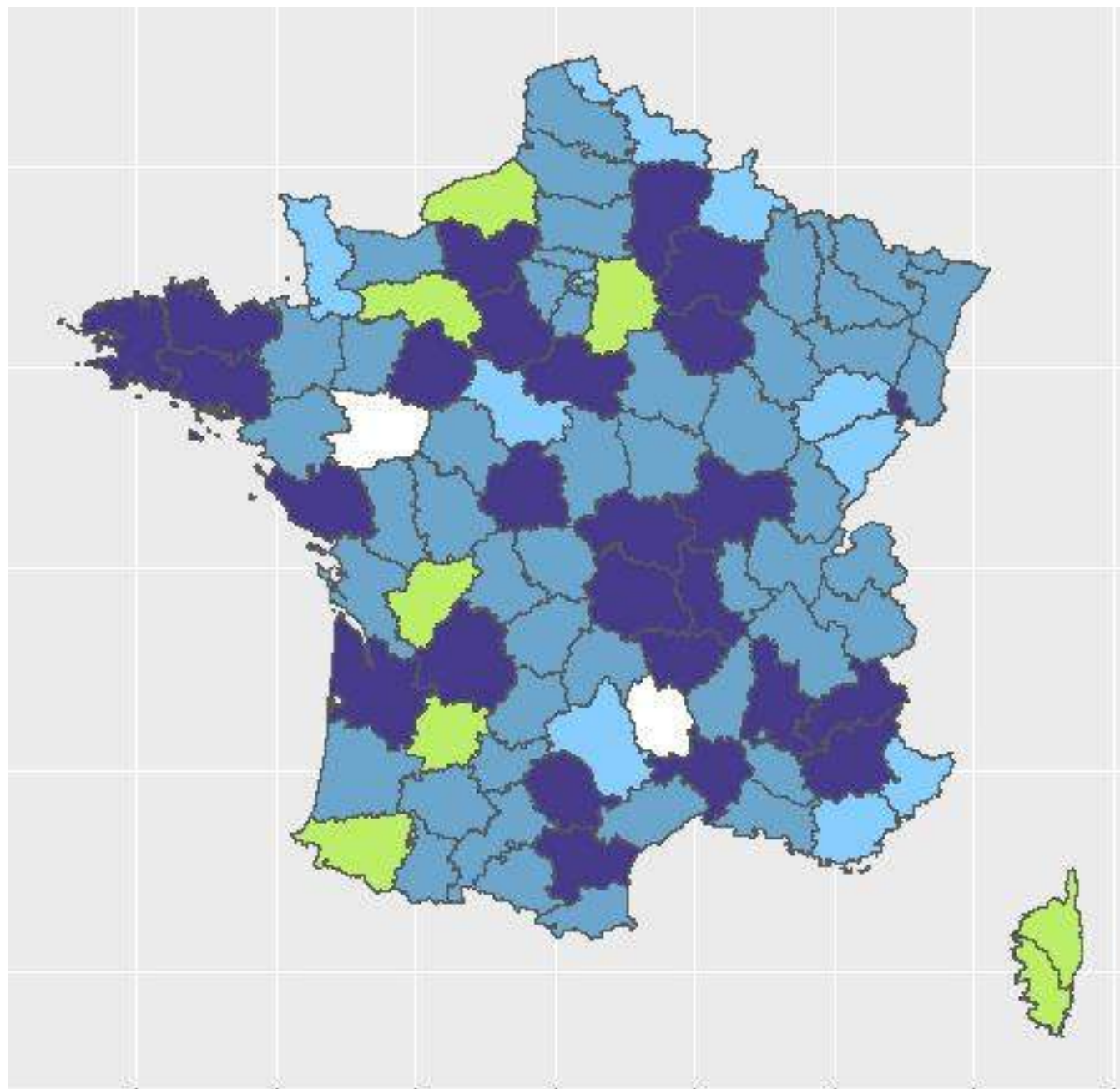
- ❖ Limited socio-economic information: only occupation, rather fragile.
- ❖ Important gaps: no Paris (until 1926), missing areas, etc.

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

Collecting images

State of the project

	Images being collected	10
	Images collected	50
	Not digitalized	2
	Not participating yet	8
	Processing Images	26



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.

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/>

contact@socface.org