The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses

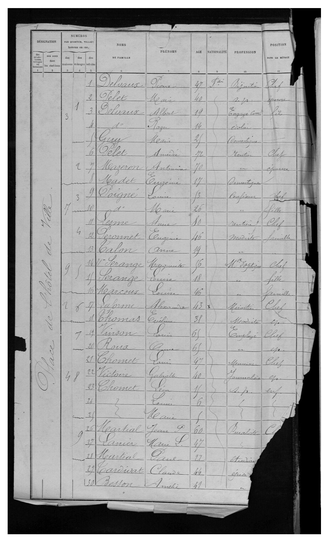
M. Boillet, S. Tarride, M. Blanco, V. Rigal, Y. Schneider, B. Abadie,
L. Kesztenbaum, and C. Kermorvant

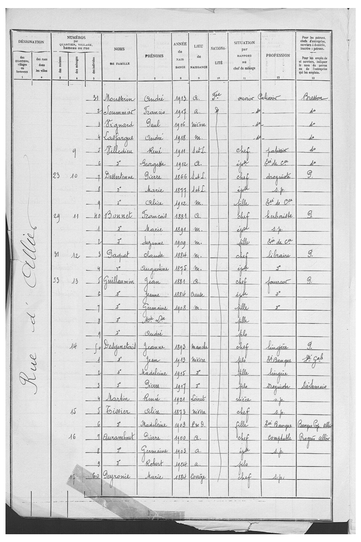
The Socface project

- Gather and process all the french nominal census lists from 1836 to 1936
- Analyze social change over the course of 100 years
- Focus on the pages containing individual information organized by household

Challenges:

- 30 million images to process, dispersed across 100 local archive services
- Columns changed (age vs. year of birth), so did their order in the table
- Quality of preservation varies from year to year





Data collection and information extraction workflow

Data collection

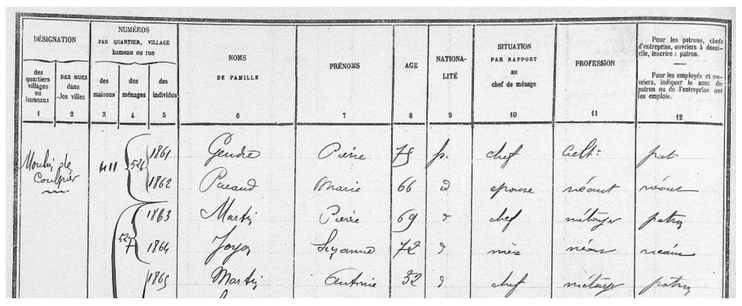
- Images and metadata were presented in a variety of formats and organizational hierarchies
- To facilitate collection, organization and normalization, we developed a web-based platform named Socface-Spider:
  - Import metadata from CSV files
  - Fuzzy identification of place names
  - Image integrity checks via IIIF
  - Export and organization of data

→ Used in more than 50 projects, **successfully validating and organizing more than 9 million images** and their metadata

Ground-truth generation

- We selected 100 single pages from 11 pilot departmental archives for annotation
- Manual transcription tasks carried out on Callico, along with their moderation

→ **Detailed annotation of 33,815 table rows** and their grouping into households



```
<s-h>Gendre <f>Pierre <o>cultivateur <l>chef <e>patron <a>75 <n>française
<s>Paraud <f>Marie <o>néant <l>épouse <e>néant <a>66 <n>idem
<s-h>Martin <f>Pierre <o>métayer <l>chef <e>patron <a>69 <n>idem
<s>Joyoz <f>Suzanne <o>néant <l>mère <e>néant <a>72 <n>idem
...
```

Information extraction

- Page classification using YOLOv8 to filter the pages of nominative lists

→ **99% precision and recall** for the List class

- Full table recognition using DAN

Set	CER	WER
TRAIN	8.94	17.18
VALIDATION	14.30	26.22
TEST	14.47	27.05

Tag	P	R	F1	Support
AGE	0.87	0.87	0.87	1,700
BIRTH_DATE	0.97	0.99	0.98	558
CIVIL_STATUS	0.95	0.93	0.94	1,153
EMPLOYER	0.74	0.76	0.75	237
FIRSTNAME	0.94	0.93	0.94	2,371
LINK	0.85	0.89	0.87	1,838
LOB	0.74	0.76	0.75	788
NATIONALITY	0.67	0.73	0.70	1,287
OBSERVATION	0.37	0.10	0.16	141
OCCUPATION	0.83	0.80	0.81	1,496
SURNAME	0.86	0.82	0.84	1,835
SURNAME_HOUSE.	0.72	0.80	0.76	519
Total	**0.85**	**0.85**	**0.85**	**13,923**

Distributed processing

- 30 million images to process → We relied on public HPC resources
- We enhanced the processing platform Arkindex by integrating the PySlurm library to process the images on the public Jean Zay supercomputer

→ **A first batch of 450,000 images was successfully processed in less than 8 days**

TL;DR

- Socface, a large-scale and collaborative project
- 30 million images to collect and analyze
- Multiple tools and models:
  - Socface-Spider for data collection
  - Callico for data annotation
  - YOLO for image classification
  - DAN for full table recognition
  - Arkindex and Jean Zay supercomputer for processing

Arkindex

DAN

Callico

Paper

Contact information:

Mélodie Boillet
boillet@teklia.com
https://teklia.com/
@_Teklia_