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

Mélodie Boillet, Solène Tarride, Yoann Schneider, Bastien Abadie, Lionel Kesztenbaum, Christopher Kermorvant

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

Mélodie Boillet[1][0000−0002−0618−7852], Solène Tarride[1][0000−0001−6174−9865], Yoann Schneider[1], Bastien Abadie[1], Lionel Kesztenbaum[2], and Christopher Kermorvant[1][0000−0002−7508−4080]

[1] TEKLIA, Paris, France
boillet@teklia.com
[2] Institut National d'Etudes Démographiques (INED) and Paris School of Economics (PSE), France
lionel.kesztenbaum@ined.fr

Abstract. This paper presents a complete processing workflow for extracting information from French census lists from 1836 to 1936. These lists contain information about individuals living in France and their households. We aim at extracting all the information contained in these tables using automatic handwritten table recognition. At the end of the Socface project, in which our work is taking place, the extracted information will be redistributed to the departmental archives, and the nominative lists will be freely available to the public, allowing anyone to browse hundreds of millions of records. The extracted data will be used by demographers to analyze social change over time, significantly improving our understanding of French economic and social structures. For this project, we developed a complete processing workflow: large-scale data collection from French departmental archives, collaborative annotation of documents, training of handwritten table text and structure recognition models, and mass processing of millions of images.
We present the tools we have developed to easily collect and process millions of pages. We also show that it is possible to process such a wide variety of tables with a single table recognition model that uses the image of the entire page to recognize information about individuals, categorize them and automatically group them into households. The entire process has been successfully used to process the documents of a departmental archive, representing more than 450,000 images.

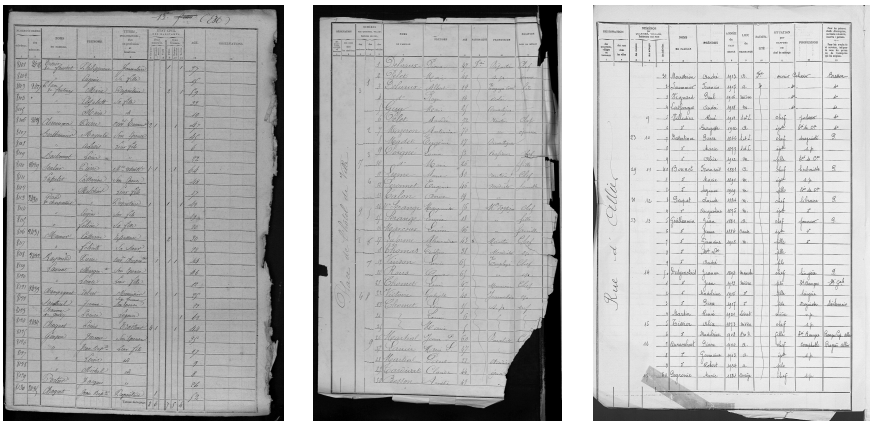
Keywords: Handwritten table recognition · Large-scale data collection · Collaborative annotation.

1 The Socface project

The Socface project[1] involves archivists, demographers, and computer scientists working together to analyze French census documents and extract information

[1] https://socface.site.ined.fr/

on a very large scale. Its objective is to gather and process all the handwritten nominal census lists from 1836 to 1936 using automatic handwriting recognition. Produced every five year, these lists are organized spatially (municipality; wards, hamlets, or streets; houses; households) and summarize the information from the census, listing each individual with some of his or her characteristics, e.g., name, year of birth, or occupation. The project aims at taking advantage of this archival material to produce a database of all individuals who lived in France between 1836 and 1936, which will be used to analyze social change over the course of 100 years. An important impact of Socface will be public access to the nominal lists: they will be made freely available, allowing anyone to browse hundreds of millions of records.



(a) 1836 (b) 1886 (c) 1936

Fig. 1: First page of nominal lists for the commune of Moulins (department of Allier) for three census years. The quality of the pages varies greatly from one year to the next. In addition, the table template evolved over the years: in 1881, civil status was replaced by the column marking the position in the household. In 1906, age is replaced by year of birth, as can be seen on the 1936 example.

As depicted on Figure 1, the data are presented in tabular form. A major challenge in this project and in processing these documents is that the tabular formats have evolved over the 100 years studied. As can be seen in the figure, the columns changed (age vs. year of birth), so did their order on the page. In addition, the quality of preservation varies from year to year and from archival deposit to archival deposit. The very large number of writers makes the task even more complex.

The decentralized nature of the source material has been a significant hurdle for prior attempts at a project of this scale. The images required for the Socface project are dispersed across 100 local archive services throughout France, rather than being housed in a single repository. The project is faced with a

dual layer of variability due to the dispersion of documents, which requires not only the preliminary collection of images, but also dealing with the diversity of the documents themselves and the differing organizational systems and metadata standards employed by each archival service. The collection and systematic analysis of the data is made difficult by its complexity.

The Socface project faces a significant challenge in processing a vast number of documents, with an estimated count of 30 million images. To address this challenge, access to public supercomputing resources is necessary. However, High-Performance Computing (HPC) architectures are not inherently designed to manage such extensive input and output flows. Tailored development efforts are necessary to ensure that images can be efficiently processed by available computing resources, particularly GPUs, and to seamlessly integrate the results into a document management system. This highlights the need for innovative solutions to bridge the gap between traditional HPC capabilities and the demands of large-scale data analysis projects.

In this paper, we describe the methodologies and technological advancements developed in the Socface project, highlighting our contributions to document recognition and historical data analysis. Our work presents a comprehensive approach to processing and analyzing historical census documents on an unprecedented scale. The key contributions of this paper are:

Data collection and normalization: We present a reliable method for collecting, identifying, normalizing, and storing images and metadata from each archival service. This involved developing a standardized protocol for interacting with the various organizational systems found across the 100 local archives, ensuring consistency in the way documents are digitized, classified, and archived. Our approach involves techniques for harmonizing metadata, which facilitates the integration of different data sources into a cohesive dataset.

Deep learning model for handwritten table understanding: A central contribution of our work is the design of a unique deep learning model capable of recognizing and structuring the personal information contained within the handwritten lists, despite the considerable diversity of document formats encountered. This model leverages advancements in full page handwriting recognition to accurately interpret a wide range of handwriting styles and extract structured data from documents whose layouts evolve over time.

High-Performance Computing (HPC) for document processing: A pivotal advancement made in this project is the extension of Arkindex, an open-source platform for automatic document processing, to facilitate communication with High-Performance Computing (HPC) systems via the SLURM workload manager. This extension grants the document processing community the ability to leverage the vast processing capacities inherent to HPC environments.

This paper is structured as follows. Section 2 provides an overview of the main approaches for information extraction from digitized handwritten tables.

Section 3 presents the tool developed during this project to simplify the data collection and normalization from departmental archives. Section 4 presents the census registers and the annotation process. Section 5 describes the training data, presents each step of the proposed information extraction workflow, and discusses the results. The final section 6 describes how we distributed the document processing across a cluster of computers using HPC tools.

2 Related work

Several models are available for detecting tables in document images. However, there are few models that can retrieve both the textual content and structure of tables, especially for historical and handwritten documents. The dominant approach for processing such documents is to first detect the rows of a table and then apply a standard character recognition model at the row level. More recently, models have been proposed to process handwritten tables as a whole by analyzing the image of the entire table. Both approaches will be discussed in the following sections.

Table row processing In the literature, most analysis focus on 2-step pipelines. First, the table rows are extracted using standard text line detection models. These are usually Fully Convolutional Networks (FCN) [9,4], Region-based CNNs (R-CNN) [13,10] or, more recently, Transformer-based [17,3] models. Once the rows have been extracted, standard text recognition using an HTR model is applied, and the columns are often recreated in a post-processing step.

In their work on the POPP dataset, Constum *et al.* [5] addressed the problem using a standard line detection model [1] followed by a line-level text recognition model [7]. As their tables followed the same template from image to image, there was no need to segment the tables into columns, as the information was always presented in the same order. To correctly categorize the retrieved information, they added a `/` symbol in the ground truth to separate the information from the different columns.

In their study, Tarride *et al.* [16] made this method a little more generic by predicting both the text and the category of information recognized. This makes it possible to apply the same model to multiple table templates. To achieve this, the authors transformed the information in the ground truth by adding a token before the start of the text in each column, representing its category. This allowed them to avoid using the `/` symbol, which was no longer useful. The trained model performs very similarly to the model trained by [5], but it is much more general and predicts more information as it categorizes the detected information.

The TableTransformer model [14] goes one step further by extracting both tables and their structure from PDF document images. This means that it can extract not only the rows of tables, but also their columns and cells. This model works very well on printed data and has shown good performance on handwritten

tables [2]. However, as with previous approaches, it cannot recognize the content of the cells directly, so it is necessary to apply a text recognition model afterward.

Full table processing A major disadvantage of processing at the table row level is that, as with conventional text recognition, detection errors have a major impact on the quality of text recognition. Furthermore, if we use a character recognition model that uses context, in particular Transformer models, the context is greatly reduced compared to full page recognition. This is also the case for table processing: when processing at row level only, the context of previous rows is lost, as is the information contained in the table header. Recent advances, particularly in Transformer architectures, make it possible to process and understand the entire page or table without any prior segmentation.

In [16], the authors trained a model on the whole table. The model has been trained to extract the textual content of the table in a structured way: each piece of textual information is extracted with its type, which corresponds to the column label. For this, the authors used the DAN model [6], which combines a convolutional encoder and a transformer decoder, making it possible to process documents at line level, but also at full page level. Due to the small amount of annotated data and the much more complex task, the model performs less well than the table row level model, but is not affected by the quality of the line-detection model, whose impact on text recognition has not been assessed.

Given the immense size and diversity of the documents in the Socface project, it is impractical to set up a complex workflow with sequential models such as in [15] or to use several models tailored to specific document templates. We therefore decided to develop a single, comprehensive model capable of processing entire tables. This model is designed to automatically adapt to the variations inherent in documents, ensuring efficient and accurate recognition and structuring of data without the need for prior segmentation or template-specific adjustments. This approach not only streamlines the processing pipeline, but also overcomes the challenge of handling the project's large and varied dataset with a scalable and flexible solution.

3 Data collection and normalization

A critical component of the Socface project is the comprehensive collection, normalization, and organization of images and metadata from 94 departmental archives services across metropolitan France. The majority of these services have volunteered to participate in the project by providing access to their archival images and associated metadata. In return for their cooperation, they are offered access to all the data automatically extracted from their documents. These archive services use a variety of systems to store their images, including self-hosted solutions, external hosting and the International Image Interoperability Framework (IIIF), as well as different archive management systems. As a result, images and metadata were presented in a variety of formats and organizational hierarchies,

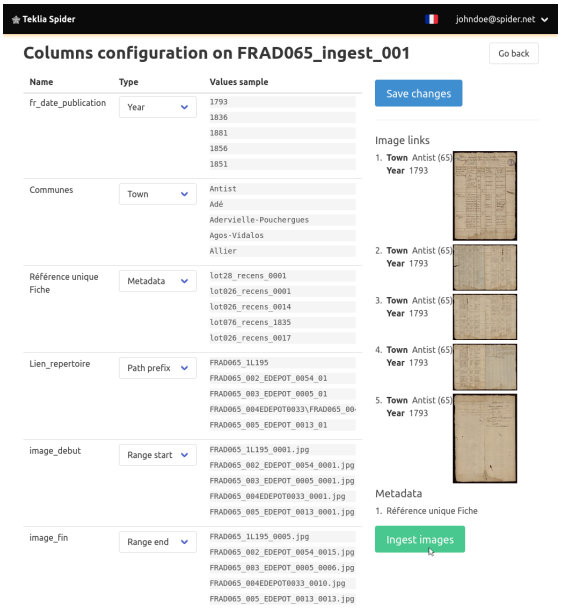


Fig. 2: Configuration interface for retrieving and organizing data from the input CSV file. The "Name" column indicates the fields present in the CSV file. The "Type" column indicates how the CSV fields will be used (whether it corresponds to the year or commune, or if the field should be ignored). If the data displayed in the "Values sample" column is correct, the user will see a preview of the retrieved images with their metadata.

including XML-EAD, CSV, XLSX, XLS and ODT, with no standard naming conventions for cities across the services.

To address the challenge of collecting, organizing and normalizing this diverse dataset, we developed a web-based platform called Socface-Spider. This platform is designed with several key functionalities to facilitate the processing of the collected data:

- Import metadata from CSV files: In response to the diversity of file formats provided by the archive services, all file formats are first converted to CSV. Following conversion, metadata files are imported into Socface-Spider.
- Support for specific CSV formats: Given the variation in the structure of CSV files across different archive services, Socface-Spider includes a feature that allows users to manually select the columns containing the necessary metadata. This selection is facilitated by a user interface, shown in Figure 2, designed to accommodate the specificities of each CSV format. This process ensures that essential data such as the year, city name, archival ID, and image path are accurately identified and normalized for consistency.
- Fuzzy identification of place names: Given the lack of standardized naming of cities across services, the platform uses fuzzy matching techniques to identify

city names within the Cassini index [12]. This index catalogues all official names of communes in France since 1793, facilitating accurate matching of data to specific locations.
- Image integrity checks via IIIF: The platform verifies the presence and integrity of images on the storage server via IIIF access, ensuring that digital artifacts are complete and uncorrupted before further processing.
- Export and organization of validated data: After validation, the platform exports the data to Arkindex, where the images are organized in a standardized manner by census year, municipality, and register. All the metadata collected is linked to the corresponding census registers, creating a structured and accessible dataset.

At the current stage of the project, Socface-Spider has proven its effectiveness and versatility by being used in more than 50 projects, successfully validating and organizing more than 9 million images and their metadata according to the specific requirements of each project.

4 Document organization and content

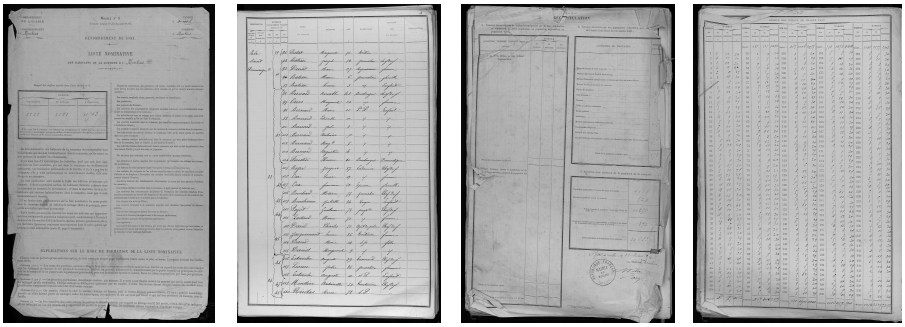
4.1 Description of census registers

The census registers provide a unique window on the demographic fabric of France from the mid-nineteenth to the mid-twentieth century. These nominative lists were systematically compiled every five years from 1836 onwards. Exceptions to this five-year rhythm were due to historical contingencies: the census of 1871 was postponed to the following year due to the occupation of parts of the territory by the Prussian army, and those planned for 1916 and 1941 were cancelled due to war conditions. The censuses were carried out within the municipal framework, systematically listing the inhabitants by household. This organization gave priority to the head of the household, followed by his or her spouse, children, other relatives living in the household, and then any servants or apprentices, among others.

Over time, the content of these communal nominative lists evolved and typically included first and last names, ages or dates of birth, family positions, occupations, nationalities, and occasionally precise addresses. The images obtained from the archival services are systematically organized into registers, each corresponding to a specific census date and commune. The images are mainly scans of double pages and, for certain years and departments, single pages. These include not only the nominative lists, but also title pages, summaries, totals and even blank pages as presented in Figure 3, with most images in black and white, scanned either from the originals or from microfilm, although a few are in color.

At present, our project is concentrating exclusively on the pages containing individual information organized by household. These lists are usually 30 lines long, although variations from 29 to 36 lines have been observed.

The layout of these documents generally begins with columns for street, house and household information, followed by details of individuals such as surname, first name, age (or year of birth) and occupation. At the current stage



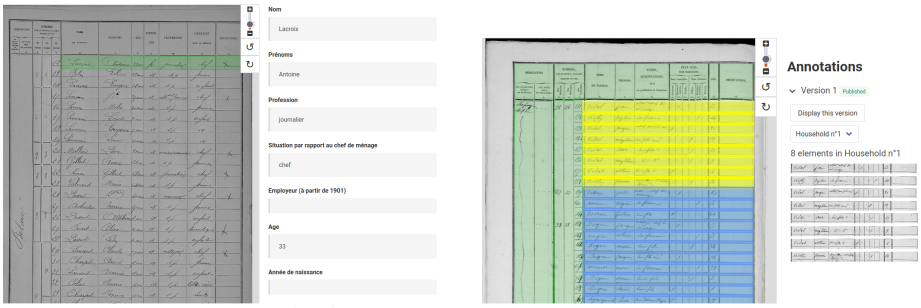
(a) Front page (b) List page (c) Recap page (d) Totals page

Fig. 3: Example of digitized pages from the census of the commune of Moulins (department of Allier) in 1881.

of the project, we are focusing on the recognition and analysis of the individual information contained in these lists.

4.2 Ground-truth generation

Generating ground-truth data is a fundamental step in training deep learning models for automatically extracting individual information from historical census lists. Given the wide variation in documents - including differences in time, format, and scanning conditions - it is imperative to collect and annotate a representative sample that captures the full spectrum of document diversity.



(a) Key-Value mode to annotate the individuals' information.

(b) Grouping mode to group individuals into households.

Fig. 4: Callico interfaces for annotating information on individuals and grouping them into households.

We selected 100 single pages from 11 pilot departmental archives for annotation. This selection was carefully chosen to include all years covered by the

study and to accurately reflect the diversity of page appearance, image quality and table templates present in the archives. These images served as the basis for manual transcription tasks carried out on Callico [11], an open-source document image annotation platform, using two specific annotation modes. First, the KEY-VALUE mode was used to annotate the individual level information. In this mode, the annotator is presented with a full list page, with a highlighted zone corresponding to an individual entry, and is prompted to enter the relevant details into a designated form, as presented on Figure 4a. Second, the GROUPING mode was used to construct the household groupings present in the nominative lists. In this mode, the entire page is displayed with individual zones highlighted, as shown on Figure 4b. Annotators are tasked with selecting zones that belong to the same household, in order to reconstruct of household units from the disjointed individual entries. In addition, we assigned a class to each selected page in order to train the page classification model described in Section 5.1.

Throughout the project, 22 annotation campaigns were conducted - two for each of the 11 selected departmental archives - resulting in an annotated dataset for model training. The first type of campaign, which focused on detailed annotation of individual information, resulted in 33,815 rows of table data. For the household grouping efforts, a total of 532 pages were annotated. Importantly, the majority of these annotations underwent a moderation process where they were either validated or corrected by experts to ensure the highest possible accuracy and reliability of the ground-truth data.

5 Information extraction workflow

This section describes the different models used to extract the personal data. To process only the list pages and extract the individual information, we start the processing by applying an image classification model, as described in Section 5.1. A text recognition model is then applied directly to the pages to extract the individual information. This model and its training parameters and performance are described in detail in Section 5.2.

5.1 Page classification

Since our study focuses on the pages of the nominal lists, and in order to save processing time, we only send images containing list pages to the recognizer. We therefore trained an image classification model with the following classes:

- The FRONT class corresponds to the first page of a register, which contains all the information about the year, the commune, the department and also some instructions for filling in the nominative lists;
- The LIST class corresponds to pages of nominative lists with information on individuals, organized by household and street;
- The RECAP pages contain various tables summarizing information about the population of the parish;

– The Totals pages contain the total number of houses, households, individuals, but also men and women in the commune;
– The Other class contains all other images such as blank pages, black pages or handwritten tables that do not correspond to nominative lists, summaries, or totals.

Training configuration To train a model, we chose to fine-tune the classification model pre-trained on ImageNet [8] available in YOLOv8[2]. We started from the YOLOv8x-cls model and fine-tuned on Socface images during 200 epochs with early stopping and a batch size of 4. The model was trained on square images of size 1024×1024 pixels. The data used is the same as that selected for Callico annotation to train the recognition model, to which we have added pages from classes other than List. In total, we have 1,285 pages, divided into 899 in the training set, 193 for the validation and 193 in the test set.

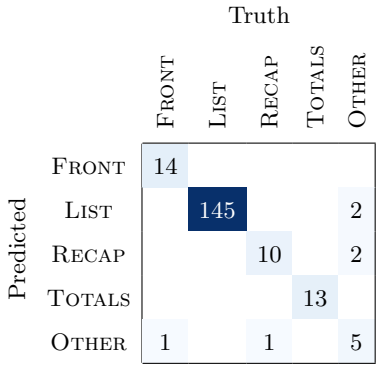
Page classification results The performance results of the model were satisfactory, particularly in terms of minimizing classification ambiguities regarding the List class. The accurate identification of this class is critical, as these pages are subsequently processed by the information extraction model, which requires a high degree of precision and recall to ensure comprehensive data capture. As shown in Table 1a, the model demonstrates exceptional efficiency, achieving precision and recall metrics of at least 99% for the List class.

Table 1: Results obtained by the image classification model.

(a) Results obtained by the image classification model for each set and class. The results on the training set are not shown because the model obtained 100% for the precision, recall and F1-score for all classes.

Class	Validation			Test		
	P	R	F1	P	R	F1
Front	1.0	1.0	1.0	0.93	1.0	0.97
List	1.0	0.99	1.0	1.0	0.99	0.99
Recap	0.92	1.0	0.96	0.91	0.83	0.87
Totals	1.0	0.92	0.96	1.0	1.0	1.0
Other	0.88	1.0	0.93	0.56	0.71	0.63

(b) Confusion matrix of the test set.



Furthermore, analysis of the confusion matrix shown in Table 1b reveals that the model faces more challenges in classifying the Other class, which is

[2] https://docs.ultralytics.com/tasks/classify/

characterized by its considerable diversity. This category combines data that includes tables that are often misidentified as summary pages, as well as pages with printed text that resemble front pages, leading to classification ambiguities and, consequently, reduced performance metrics within this specific class.

Notwithstanding the model's limitations in accurately classifying the OTHER class, its ability to identify list pages remains commendably high, making it sufficiently capable for the purposes of classifying and earmarking pages for subsequent processing by the information extraction model outlined in the following section.

5.2 Handwritten table recognition

The information about the individuals is presented in a table where the individuals are grouped into households. Given the scale of the project and the diversity of the documents, it was not feasible to develop and maintain a processing chain comprising multiple models. We therefore chose the DAN model [6] to perform a full table recognition, which not only extracts the text from the table, but also tags the extracted text in order to categorize the predicted text at the same time [16].

The advantage of this method is that it does not require any segmentation of the page nor the table, as it works directly on the whole page. In addition, some information is marked by vertical lines or ditto labels, so processing the whole page allows better interpretation of these labels compared to, for example, table row-level processing, where the model has much less context to interpret the content of a cell. A second advantage of this method is that there is no need to apply a second model later to label the information, as it is categorized directly at the same time as the text is recognized.

Finally, this model can also be used to predict data in a structured way. In fact, by adding a token indicating the head of the household before the individual information, we are able to directly structure individuals into households without any other model. This structuring involves a post-processing step consisting of going through all the pages of the register, in the correct reading order, to reconstruct households spanning two pages.

Label generation To extract information from individuals and group them into households, we use a unique text recognition model. To train it, we constructed the ground truth transcriptions as described below and shown in Figure 5:

- Each piece of information annotated in the Callico form is preceded by a token indicating the type of information (<s-h>, <s>, <f>, <o>, <l>, <e>, <a>, <n> in the Figure);
- The names of individuals listed as 'heads of household' are preceded by a token (<s-h>) that is different from the other members of the household (<s>), in order to indicate the start of the household;

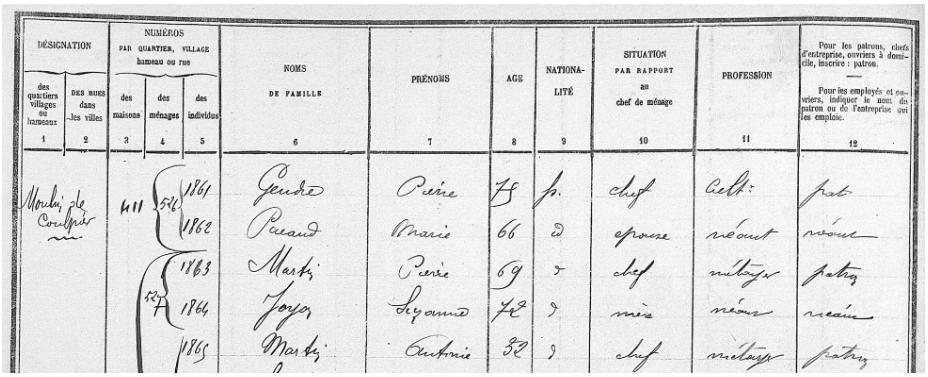


Fig. 5: Table header and first rows of a table from the census of the commune of Neuilly-le-Réal (department of Allier) in 1901. The label used to train the model for this part of the table is:

```
<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
...
```

Note that the order of the entities in the labels is always the same and does not always correspond to the order in which the information appears in the images, as there are multiple templates.

- All information about an individual is concatenated into a single string so that it always follows the same order, even if it is different from the order in the table;
- The transcriptions for each individual are themselves concatenated to represent the whole page in a single string.

Empty cells and rows are not annotated and not present in the transcription.

Model and training configuration DAN [6] is an open-source attention-based Transformer model for handwritten text recognition that can work directly on pages. The encoder is fully convolutional, while the decoder is a Transformer network. It is trained with the cross-entropy loss function. The last layer is a linear layer with a softmax activation function that computes probabilities associated with each vocabulary character. We trained a DAN model on the annotated single pages for 1,000 epochs with early stopping and a batch size of 4. The model was trained on a single GPU A100 with 80Gb. To reduce the memory required for training, the images were resized so that their height was equal to 1900 pixels. Data augmentation was applied during training and the maximum number of tokens to be predicted was set to 2,800 according to the training data.

Table 2: Results obtained by the information extraction model.

(a) Character Error Rate and Word Error Rate obtained by the information extraction model (%).

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

(b) Evaluation of entity recognition on the test set.

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**

Full-page recognition results The performance of the text recognition and household grouping model is shown in Tables 2a and 2b. The CER obtained on the validation and test sets are 14.30% and 14.47% respectively. These values, which may seem rather high, reflect the quality of all the categories of information to be extracted at the level of the whole page. As these metrics are strongly affected by a shift in recognition: an extra word, for example, shifts the entire predicted sequence, they are very difficult to interpret. For this reason, the performance of each entity is presented in Table 2b. From this table, the F1 scores for all the fields, except the "Observation" category, ranged from 70% for nationality to 98% for year of birth. These high scores show that the model is robust and generic enough to handle a large number of documents, image qualities and table templates.

From the table, we can also see that the information contained in the "Observation" columns is very poorly recognized, with an F1 score of 16%. This can be explained by the fact that this category is very poorly represented during training: it appears only 388 times in the manual annotations of the training set, which means that the information is present in about 1% of the table rows.

Finally, although it definitely plays a role, performance in the other categories does not seem to be directly correlated with the number of elements in the training set. Our hypothesis is that the difference in performance between the different categories can be explained by several other factors:

– Some entities are easier to recognize because the possible values are very limited: this is particularly the case for the age and year of birth categories;

– Others entities are more difficult to recognize because they may contain ditto entries, and some annotators have rewritten the text in the corresponding cell rather than annotating it as a ditto.

In order to improve performance and make the results easier to interpret, further standardization of the annotations would be necessary, particularly to reduce the impact of this last factor.

Household extraction Table 2b also shows the performance on the household grouping task, which consists of predicting a different category for the surnames. We can see that 76% of the households were correctly grouped, which seems quite good considering the difficulty of the task. In fact, in some lists, the information is clearly annotated with brackets. But this is not always the case, and sometimes the information is not directly annotated but has to be inferred from the 'link' category, making the task much more complex.

6 Distributed processing on HPC

The Socface project leverages the capabilities of Arkindex, an open source document processing platform that offers a comprehensive suite of functionalities including document organization, visualization, processing, and export. However, we are faced with the monumental task of processing approximately 30 million images. To meet the demanding computational requirements of this huge dataset, we rely on public High-Performance Computing (HPC) resources. However, integration with the HPC infrastructure imposes specific constraints: the compute nodes are isolated from the Internet, requiring pre-staging of data on specialized local storage and orchestration of job submissions through dedicated scheduling systems such as SLURM.

To overcome these limitations, we have developed a three-step strategy to extend Arkindex to take advantage of HPC resources:

– Data preparation and pre-processing: Recognizing the lack of Internet connectivity on HPC computing nodes, the first stage is performed on front-end CPU nodes that do have Internet access. This step involves downloading the required dataset images from the IIIF server, along with essential processing metadata such as image dimensions. These elements are then stored on local storage, making them available to the HPC computing nodes for subsequent processing stages.
– Image processing: With the data pre-positioned on local storage, processing shifts to the HPC's GPU nodes. This stage uses the computing power of the GPUs to efficiently analyze the images. The results of this processing stage are encapsulated in JSON files, providing a structured representation of the results that can be easily transferred and interpreted in subsequent steps.
– Integration and monitoring of results: The final phase moves back to CPU nodes with Internet access. This is where the JSON files containing the

processed data are uploaded to the Arkindex database. This step not only secures the processed data within Arkindex, but also facilitates real-time task status updates. Such updates are critical for monitoring the progress and success of processing tasks, providing insight into operational status, and ensuring that any necessary adjustments or re-processing can be addressed in a timely manner.

To implement these various steps, Arkindex's internal distributed task system has been significantly enhanced by integrating the PySlurm library[3], which enables seamless communication with SLURM. This key development has effectively enabled Arkindex to take advantage of the immense computing resources available in HPC environments, significantly increasing its processing capacity to meet the demands of large-scale projects.

We conducted a processing time evaluation using our distributed processing framework enabled by High Performance Computing (HPC) to manage the extraction of information from a batch of 450,000 images, processed using a distributed architecture that integrated 14 parallel processes on CPU nodes for the initial and final stages of the workflow, and used NVIDIA V100 GPUs to execute the deep learning model responsible for table recognition and entity typing.

The breakdown of processing times for each stage of the workflow is as follows:

- Preprocessing Phase: This initial phase, mainly focused on image download, was completed in an average time of 1.6 seconds per image.
- Table Recognition and Entity Typing: The core processing task of recognizing full-page tables and typing entities within these tables using our deep learning model took an average of 12.5 seconds per image.
- Post-processing stage: The final phase, which included uploading the results to the database along with text position, line-level text recognition and entity tagging, took an average of 7.2 seconds per image.

The entire batch of 450,000 images was processed in less than 8 days, demonstrating the efficiency and scalability of our distributed processing approach using HPC resources.

7 Conclusion

In this paper, we have presented a comprehensive workflow designed to automatically extract information from individual census tables spanning 20 censuses over a century, structured to closely follow the original format of the source documents. This methodology has already proven its effectiveness on thousands of images and will be scaled up to process millions more from numerous French departmental archives by the end of the project.

Our achievement lies in the development of a unified model capable of handling a wide variety of image types, table structures and handwriting styles.

[3] `https://pyslurm.github.io/`

Using a transformer-based architecture, this model allows direct processing of entire tables without the need for prior segmentation, significantly minimizing the potential for errors commonly associated with multi-step processing approaches. Careful label generation ensures comprehensive information extraction across all table variants, covering both the content and the familial arrangements of the listed individuals.

However, the current method has limitations, most notably the inability to process complete registers on a page-by-page basis while retaining the context of previously processed pages. This shortcoming requires additional post-processing to reassemble household units that span multiple pages. Future enhancements will focus on overcoming this challenge by enabling sequential processing of entire registers with the aim of preserving contextual continuity. We also plan to extend the processing to include address recognition, thereby facilitating the reconstruction of household compositions within individual houses, streets, hamlets, and sectors.

8 Acknowledgments

The Socface project is funded by the French National Research Agency (ANR) under the fund ANR-21-CE38-0013. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013446 made by GENCI and was partially funded by the ACADIIE project "Compréhension automatique des documents d'archives pour l'extraction d'informations individuelles" supported by a grant overseen by the French National Research Agency (ANR) as part of the France Relance program.

References

1. Ares Oliveira, S., Seguin, B., Kaplan, F.: dhSegment: A Generic Deep-learning Approach for Document Segmentation. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12 (Aug 2018)
2. Bernard, G., Wall, C., Boillet, M., Coustaty, M., Kermorvant, C., Doucet, A.: Text Line Detection in Historical Index Tables: Evaluations on a New French PArish REcord Survey Dataset (PARES). In: Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration. pp. 59–75. Springer Nature Singapore (Dec 2023). https://doi.org/10.1007/978-981-99-8085-7_6
3. Biswas, S., Banerjee, A., Lladós, J., Pal, U.: DocSegTr: An Instance-Level End-to-End Document Image Segmentation Transformer. In: arXiv preprint arXiv:2201.11438 (2022)
4. Boillet, M., Kermorvant, C., Paquet, T.: Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks. In: 25th International Conference on Pattern Recognition (ICPR). pp. 2134–2141 (Jan 2021)
5. Constum, T., Kempf, N., Paquet, T., Traounez, P., Chatelain, C., Bree, S., Merveille, F.: Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20th Century Paris Census. In: 15th International Workshop on Document Analysis Systems (DAS). p. 143–157 (May 2022). https://doi.org/10.1007/978-3-031-06555-2_10

6. Coquenet, D., Chatelain, C., Paquet, T.: DAN: a segmentation-free document attention network for handwritten document recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1–17. Institute of Electrical and Electronics Engineers (IEEE) (Jan 2023). https://doi.org/10.1109/tpami.2023.3235826
7. Coquenet, D., Chatelain, C., Paquet, T.: End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 508–524 (Jan 2023). https://doi.org/10.1109/TPAMI.2022.3144899
8. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: ImageNet: A Large-scale Hierarchical Image Database. In: IEEE Conference on Computer Vision and Pattern Recognition (ICPR). pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848
9. Grüning, T., Leifert, G., Strauß, T., Labahn, R.: A Two-Stage Method for Text Line Detection in Historical Documents. In: International Journal on Document Analysis and Recognition (IJDAR). pp. 285–302 (Sep 2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (Jun 2016)
11. Kermorvant, C., Bardou, E., Blanco, M., Abadie, B.: Callico: a Versatile Open-Source Document Image Annotation Platform. In: Sumbitted to ICDAR2024 (2024)
12. Motte, C., Vouloir, M.C.: Le site cassini.ehess.fr. Un instrument d'observation pour une analyse du peuplement. Bulletin du Comité français de cartographie **191**, 68–84 (2007)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: 28th International Conference on Neural Information Processing Systems (NIPS). p. 91–99 (Jun 2015)
14. Smock, B., Pesala, R., Abraham, R.: PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4634–4642 (Jun 2022)
15. Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., Vézina, H., Kermorvant, C.: Large-scale genealogical information extraction from handwritten quebec parish records. Int. J. Doc. Anal. Recognit. **26**(3), 255–272 (jan 2023). https://doi.org/10.1007/s10032-023-00427-w, `https://doi.org/10.1007/s10032-023-00427-w`
16. Tarride, S., Boillet, M., Kermorvant, C.: Key-Value Information Extraction from Full Handwritten Pages. In: Document Analysis and Recognition - ICDAR 2023. pp. 185–204. Springer Nature Switzerland (Aug 2023). https://doi.org/10.1007/978-3-031-41679-8_11
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: 31st International Conference on Neural Information Processing Systems (NIPS). p. 6000–6010 (Dec 2017)