International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-023-00428-9

ORIGINAL PAPER

An end-to-end pipeline for historical censuses processing

Rémi Petitpierre[1] · Marion Kramer[1] · Lucas Rappo[1]

Received: 23 August 2022 / Revised: 10 January 2023 / Accepted: 20 February 2023

Abstract

Censuses are structured documents of great value for social and demographic history, which became widespread from the nineteenth century on. However, the plurality of formats and the natural variability of historical data make their extraction arduous and often lead to ungeneric recognition algorithms. We propose an end-to-end processing pipeline, based on optimization, in an attempt to reduce the number of free parameters. The layout analysis is based on semantic segmentation using neural networks for a generic recognition of the explicit column structure. The implicit row structure is deduced directly from the position of the text segments. The handwritten text detection is complemented by an intelligent framing method which significantly improves the quality of the HTR. In the end, we propose to combine several post-correction approaches, neural networks, and language models, to further improve the performance. Ultimately, our flexible methods make it possible to accurately detect more than 98% of the columns and 88% of the rows, despite the lack of graphical separator and the diversity of formats. Thanks to various reframing and post-correction strategies, HTR results reach the excellent performance of 3.44% character error rate on these noisy nineteenth century data. In total, more than 18,831 pages were extracted in 72 censuses over a century. This large historical dataset, as well as training data, is made open-access and released along with this article.

Keywords Historical document processing · Layout analysis · Tabular document understanding · Handwritten text recognition · OCR post-correction

1 Introduction

Historical census records are dense, structured and massive sources of demographic data, whose extraction has been regarded as a goal of major importance for the humanities and social sciences [1,2]. Indeed, from the early eighteenth century, with the rise of the administration, this kind of document multiplied in many areas. Population censuses in Europe and North America are typically structured in columns indicating the names of the inhabitants of each household, their age, their occupation, and sometimes their religious affiliation. If this high-level structure facilitates historical analyses, the extraction of tabular documents implies high requirements in terms of quality so as not to bias subsequent analyses.

In point of fact, while one might think that the structure of tables, generally repeated from page to page, makes them a relatively generic document, the physical reality is often quite different. The columns are full of additions, remarks, abbreviations, arrows and various brackets, erasures and corrections. In addition, the writing was usually the work of many different censors, each with their own way of interpreting and reporting the data, often with ill-defined instructions. In Lausanne, for instance, until 1898, censuses were taken twice a year by district commissioners, who also performed numerous tasks to preserve public order. It was not uncommon to forget a section of street or a few apartments, and the administrative documents of the time cast a critical eye on the diligence of these commissioners [3].

While attempts to automatically extract censuses have been presented in the past, research has naturally focused on the more regular sources of the early twentieth century, benefiting from a regular structure. These projects include the study of French genealogy [4] and US Census Records from the 1930 s and 1940 s [5,6]. In both of these cases, the tabular documents were structured in a uniform manner, with uniformly sized cells and a constant number of cells on each

✉ Rémi Petitpierre
remi.petitpierre@epfl.ch

Marion Kramer
marion.kramer@epfl.ch

Lucas Rappo
lucas.rappo@epfl.ch

[1] Digital Humanities Institute, EPFL, Lausanne, Switzerland

Published online: 17 March 2023

page. The text extraction problem therefore simply comes down to a cell extraction issue. Another strategy to reduce the complexity of the extraction problem is to focus on a single column, thus avoiding offset issues. In this way, Pedersen has achieved the extraction of the occupational codes from nineteenth–twentieth-century Norwegian Historical Population Registers [7]. In that case, the values were corresponding to a limited number of solutions. However, to manage the natural variability of historical tabular documents, some studies have also suggested the adoption of a fully probabilistic perspective [8,9]. For instance, [9] use character n-gram embedding to address keyword spotting problem. While a probabilistic approach is certainly the most relevant for this particular task, it does not permit to produce deterministic extraction results for research in historical demography.

Regardless of the method used, processing historical tabular documents can be decomposed into several successive steps, including document layout analysis (DLA), handwritten text detection (HTD), and recognition (HTR). In this article, HTR post-correction will also be discussed.

Text detection itself is a complex problem, especially in the case of tabular historical documents, since the text can easily overflow into another column and the text lines in the table are almost systematically interrupted by long empty spaces, contrary to a continuous text. Moreover, in the case of registers, the text lines extend over two different pages, which is not the case for classical texts. The primary challenge is therefore to reconstruct these lines. For these reasons, the flexibility required for HTD favors the use of methods based on neural networks. These technologies have already proven their performance on HTD problems in the past, for example in the context of ICDAR2017 competition [10]. This potential has also been demonstrated in the processing of tabular documents, such as the cBAD track b corpus [11]. More recently, specific approaches have been developed for better managing the detection of the text segments overlapping with vertical separators and even use the latter as a lever [12,13].

The analysis of the structure of the document, such as the separation of the document into columns and rows, is the object of layout analysis. While some methods rely on horizontal and vertical rulings and their intersections to detect cells, many historical documents, such as those to be discussed here, do not contain a graphical separation of rows. In this case, row detection must be based on regularity patterns and logical rules [14]. Therefore, row detection in itself is a particularly difficult problem. Indeed, unlike a continuous line of text, the row in a table can contain several empty cells. Moreover, a single cell may actually span several successive rows, in which the content is continued. To solve these graphical and hierarchical issues, several approaches combine graphs and neural networks [15–17]. Other methods include clustering techniques [18].

To detect columns in tabular documents, traditional DLA methods rely on successive image filtering, transforms, binarization methods, and/or machine learning classifiers [19,20]. More recently, several approaches have relied on semantic segmentation methods [21–23], or object detection [24] based on convolutional neural networks (CNNs). Tools based on these technologies are generally more generic and easily adaptable. This is relevant in the case of historical sources that may express different typologies and variability, which requires the use of a robust algorithm.

In terms of HTR, several methods have recently achieved very respectable results on various benchmarks. However, it is necessary to differentiate between the raw performance of these algorithms and their performance after applying a language model or other spelling correction. For the present study, we have pre-selected three competitive models in terms of raw performance for which an open-source implementation is available [25–27]. Some studies suggest that transfer learning, which consists in bootstrapping the learning by first pre-training the model on a generic dataset, before fine-tuning it on a more specific one, might be key in the context of HTR. If this is confirmed on more complex and irregular datasets, it could lead to a significant reduction of the number of annotations needed. According to [28], pretraining with a large and standard database can decrease the character error rate (CER) on small datasets by more than half for the same number of annotations. In that study, CER ranged from 3 to 9% depending on the parameters. Alternatively to standard HTR, Clawson has proposed a system of glyph library specifically designed for the extraction of censuses, in which HTR is treated as a glyph classification problem. They achieved very high accuracy on repetitive entries (such as gender or birthplace) [6].

However, as good as HTR algorithms can be, reading ancient-digitized documents remains challenging and the output is usually noisy. For this reason, OCR post-correction is a highly regarded topic. In particular, it was the subject of a dedicated competition at ICDAR 2017 and 2019 [29]. The simplest approaches rely on computing a Levenshtein's distance [30] between the predicted tokens and a dictionary of candidate types [31–33]. However, this method only works in the case where a rich and near-exhaustive dictionary is available. Moreover, it does not take into account the context in which the word is found. As this can be problematic for some data types, many studies [34–36] prefer to use language models instead. These language models can be derived from statistical models [37] and/or neural networks [38,39]. The attention scale can vary from the token to the character. However, with the exception of character-based models, word embedding-based algorithms are not directly relevant for the specific case of censuses, as they rely on the concept of contextual co-occurrence and meaning. Several studies also include morphological features, derived from the longest

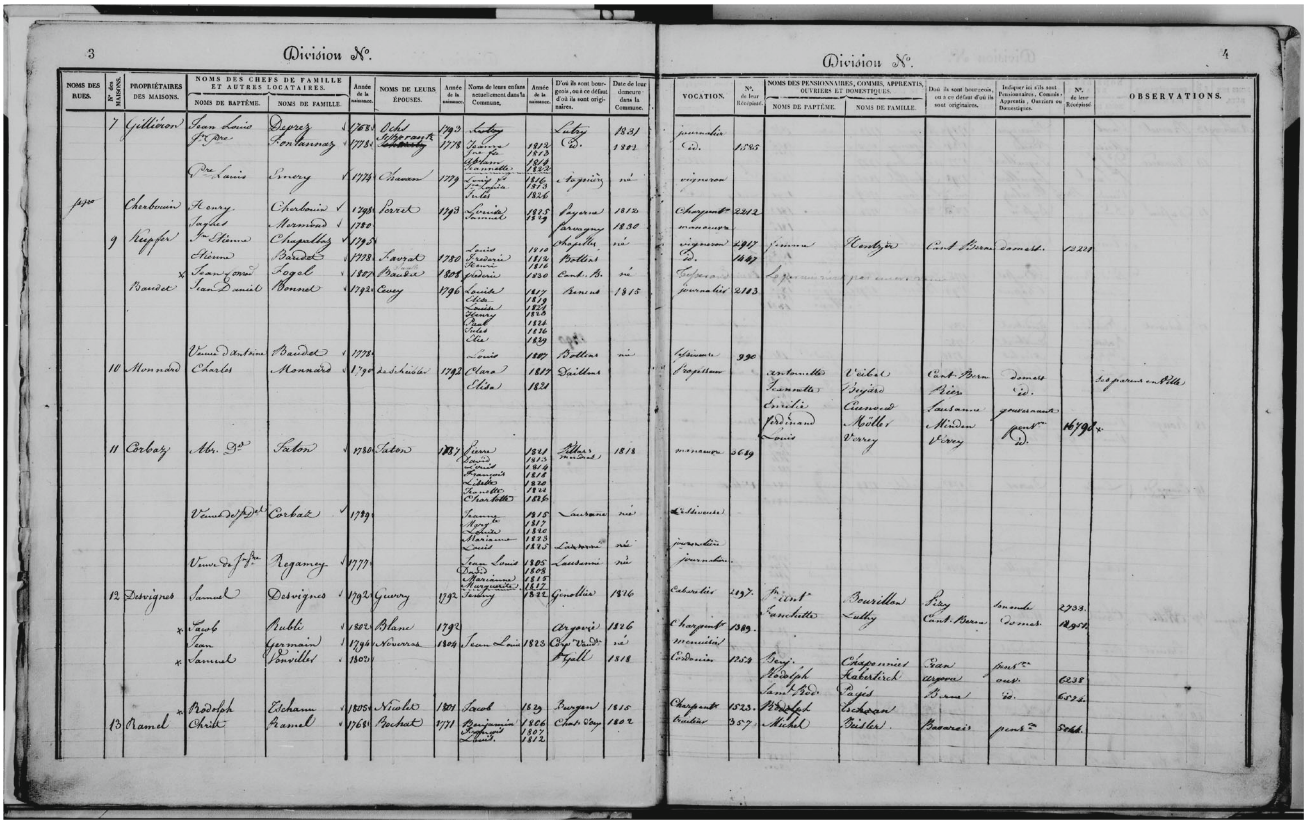


Fig. 1 Second page of the 1832 census of Lausanne

common subsequence ratio (LCSR), or n-grams characters, at certain stages of the process [40]. In general, the main limitation of language models is the need for large and clean corpora for training. Like dictionary-based models, they do not handle well the presence of new words. In a different perspective, Cao has presented a density-based approach for the post-correction of surnames in recent census data, based on a hidden Markov model [41].

In this study, we propose a flexible and generic processing pipeline to handle the entire processing chain of tabular historical documents. Although we recognize that a combination of deep learning with traditional computer vision methods is often necessary for the processing of complex documents, such as censuses, the introduction of conventional algorithms is often accompanied with logical criteria and free parameters, which are specific to the document. In this research, on the contrary, we seek to reinforce the generic character of deep learning by post-processing the prediction with flexible computer vision algorithms, optimized automatically to the specificity of the corpus.

To develop our algorithm, we rely on a naturally protean corpus. Our corpus consists of 72 censuses of the city of Lausanne, drawn up between 1805 and 1898 (see Fig. 1). This corpus has the advantage of changing typology and structures several times over the years, which makes it possible to verify the robustness and genericity of the extraction pipeline. In total, more than 18,831 pages were extracted in 72 censuses.

2 Methods

The document processing pipeline can be divided into 5 phases. First, the structure of the pages, and more specifically the columns, is extracted using semantic segmentation and computer vision. Second, the text segments are detected using a generic baseline detection algorithm and a novel intelligent framing method. Third, the text segments are transcribed after selecting an efficient HTR algorithm. Fourth, the results are structured in a tabular database, after detecting the rows. Finally, the HTR results are post-corrected using statistical models and neural network-based approaches.

2.1 Layout analysis

The structure of the censuses changed nine times between 1805 and 1898, with columns appearing or disappearing, and changing format, width, and position, over time. Among the main changes, one can note the introduction of the name, surname, and origin of the residents in 1813, the removal of the date of arrival of the family in the commune in 1849, the addition of a column corresponding to the year of birth of the residents in 1859, and the first name of the wife in 1883. In 1886, a more important reorganization took place, with the appearance of the days and months of birth, as well as the residence permit numbers for the members of the household, and finally the occupation of the children. Following these reorganizations, the corpus could be roughly separated into 6 main types that display some structural proximity: 1805 (St. Laurent section), 1805-1810, 1813-1848, 1849-1858, 1859-1885, and 1886-1898. The number of columns varies between 18 and 22.

2.1.1 Projection profiles

Projection profiles are used as a baseline method, for comparison. In this classical approach, the images are thresholded with Otsu algorithm [42]. The page is slightly tilted to correct rotation when necessary, by maximizing the ratio between maximum and minimum values on the vertical projection profile, so that the column separators become vertical. The vertical projection profile is then computed on each page. An adaptive threshold is used to detect the column separators, until the number of separators reaches the expected number of column separators, plus the sides of the page and the central binding, i.e., all main vertical lines.

2.1.2 Semantic segmentation

The structure extraction is mainly based on a CNN-based semantic segmentation using a torch implementation of dhSegment [22]. In a second step, a flexible post-processing is applied to improve the extraction.

First, the images are downsized to a vertical size of 1000 px. A proportional sample, with respect to the 6 structural types mentioned above, totaling 135 pages, is annotated on CVAT (cvat.org) software. The annotation ontology has 3 classes: header, even columns, and odd columns. A UNet [43] CNN using ResNet101 as encoder, is pretrained on ImageNet. Training is then performed using Adam optimizer [44], with a minibatch of 8, and a learning rate of 5×10^{-5}. The data are randomly augmented with several filters (blur 3×3, rotate ±3°, shear 5%, contrast-limited adaptive histogram equalization (CLAHE, [45]), elastic transform, random shadow), with an independent probability of 20% each time. A 12% validation subset is manually selected and the remaining 88% are used to train the CNN during 400 epochs, with early-stopping patience of 50 epochs. In a second step, the weights of this second network are used to train more specific networks for each of the 6 structural types.

The output probability map is used for post-processing. The page rotation is corrected when necessary. The angle is estimated as the tilt of the minimum area bounding rectangle corresponding to the 'columns' semantic class. The column contours are extracted with a Canny filter applied to the probability map restricted to background and column classes. Linear vertical separations between columns are then identified by Hough transform. The closest lines are partially clustered together by KMeans, where K is based on the number $n \in [19, 23]$ of expected columns in the considered structure ($K = n+5$). The $n+3$ widest columns with regard to the separators are then extracted, i.e., the n columns, plus the empty spaces to the left and right of the document, and at the middle binding. In a second step, the median widths between the columns separation are calculated for each year. For each page, the extracted column widths are automatically compared with the expected median width over the year, to check the accuracy of the extraction. When the difference between the expected position of a column separation (measured from the previous separation) and the position of the closest separation actually found is greater than δ_{max} pixels (here $\delta_{max} = 20px$), the position is considered to be imprecise or incorrect. In that case, it is automatically corrected, based on the median width for the year, for this particular column. Indeed, the width of the columns must be relatively constant for the same year since the tables are printed. The width of the central binding, which is highly variable, is ignored. The position of the header is simply established where the probability of this semantic class is maximal.

2.2 Extraction of text segments

An approach based on baseline identification is adopted [12]. In a second stage, an intelligent framing method is used to delineate the text segment boundaries (Fig. 2). Two areas are defined for each text segment: the patch area and the core area. The patch area is created by applying a padding x_1, x_2, y_1, y_2 on the baseline in the 4 directions (up, down, left, right). The core area is defined similarly by a padding x_3, x_4, y_3, y_4, where $x_3 < x_1$, $x_4 < x_2$, $y_3 < y_1$, and $y_4 < y_2$.

The image patch area is then binarized using Otsu's method [42]. When the patch is vertically overlapping with the core area of a neighboring segment, it is cropped accordingly, to avoid any overlap. An overlap is detected only if the neighboring core area is horizontally encroaching by a proportion greater than t_1, in order to avoid diagonal or side overlaps to be taken into consideration. Finally, the empty (white) areas of the patch are removed and the patch is tightened around the text pixels. Any visible column separator, detectable when the proportion of vertically aligned black pixels in a patch exceeds a threshold t_2, is also removed to restrict the area to the text pixels. Finally, a slight uniform padding z is applied in all 4 directions. The above 11 parameters are entirely optimized using Tree-structured Parzen Estimator (TPE, [46]) on 34,888 manually annotated text segments (see next section). The optimization problem

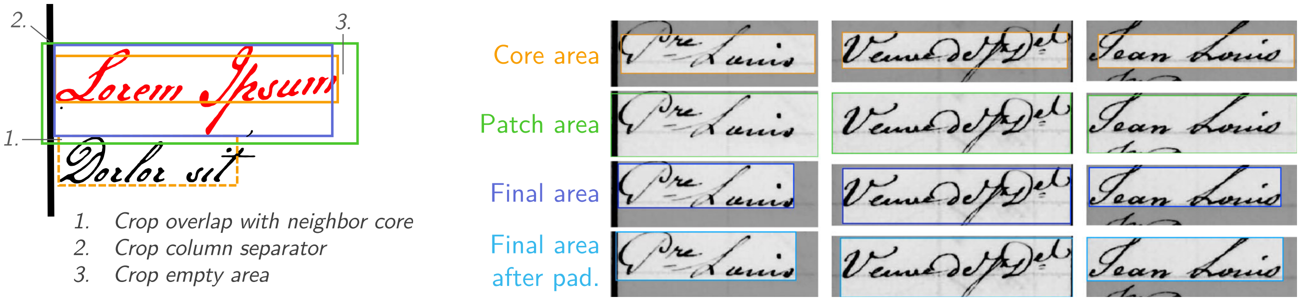


Fig. 2 Visual summary of the intelligent framing method

is $max\,(J\,(X, Y))$, where $J\,(X, Y) = \left|X \cap Y\right| / \left|X \cup Y\right|$ is the Jaccard index, X is the area covered by the predicted text segments on the page, and Y is the area covered by the manually annotated text segments.

2.3 Handwritten text recognition

A total of 34,888 text segments, corresponding to a sample of 83 pages, from which at least one per census year, were manually annotated by an expert using the open-source software CVAT [47]. Text segments were delineated and transcribed literally during this process. The transcription includes punctuation, hyphens, and abbreviations. Diacritical signs (arrows, brackets) were however not considered.

The intelligent framing method described in the previous section was applied to standardize the format of the detected text segments. For handwritten text recognition, the transcription labels are reduced to 46 unaccented characters, including the unaccented lowercase letters of the French alphabet, the numbers 0-9, and 9 symbols, plus the space. This *Lausanne Historical Censuses Dataset 35k* [48] is made open-access and released along with this article. The dataset is split into 85% for training, 10% for validation, and 5% for testing. The Bentham dataset [49] is used for pretraining the network. The network is then trained for 1000 epochs, with an early stopping clause fixed at 25 epochs, a batch size of 16 and a degressive learning rate starting from 5×10^{-4}, then decreasing by a factor $\frac{1}{2}$ at each plateau.

After comparison with two other architectures [25,27], the model described by Puigcerver [26] is finally selected. The model is trained separately for the period 1832-98, which is overperforming compared to the entire period 1813-98.

2.4 Structuration in a tabular database

To structure the registers in a tabular database, one of the crucial issues is the separation into rows. In the present corpus, for instance, the data are organized by household, and the baseline theoretically always contains the name and/or surname of the head of the household, and several other fields, such as occupation, origin, etc. When the household includes several children or other occupants, the first one is usually also on the baseline, next to the head of the household, and the others follow below, one per row. The height of the rows is irregular, as is the number of households per page.

To begin with, the relative vertical position of each text segment on the page is calculated with respect to the vertical position of the table header, detected by semantic segmentation during the layout analysis step. More precisely, the header position is approximated by a sliding median, computed over a window of a few pixels, from the position of the lower bound of the header. Text segments that are above this bound are ignored. This can typically comprise the header titles themselves, or the page numbers, as well as various side annotations. For the following steps, this relative vertical position, with respect to the header, is used.

The vertical position of the baselines (i.e., of the household heads) is estimated according to the relative position of the text segments with regard to the column header. The position of each baseline is computed iteratively and relatively, with respect to the position of the previous baseline. A minimum threshold t_3, corresponding to 50% of the average height of the rows in the corpus, is applied. The position estimation is thus flexible and relative.

Once the baseline position is recovered, the remaining segments on the page are assigned, column by column, starting from the column containing the household head. The vertical position of the baseline is slightly readjusted for each new column, so as to take into account the effective tilt of the row. When a column contains a list of segments (e.g., list of children), they are separated horizontally, on the same principle as the baselines.

In a post-correction step, the numeric columns are cleaned from the alphabetic elements, and vice versa. However, when the 'intruder' string is long enough $l_1 \geq 3$, the segment is not necessarily considered as an error, but if coherent with the alphanumeric type of the neighboring column, the element is considered to have undergone a shift and is therefore replaced in that column. The laterality of the shift is determined with regard to the position (beginning or end) of the 'intruder' segment in the string.

2.5 HTR Post-correction

The HTR post-correction methods are trained (when necessary) and evaluated on 14,571 text segments divided into three separate subsets (70% train, 20% validation, 10% test), taken from the corpus annotated for HTR training. The segments selected are those belonging to the columns first names, surnames, origins, and occupations, on which the post-correction step is developed and deployed. As the predictions on the HTR training data are naturally better than those that would be obtained on independent data, a random noise is applied to increase the CER from 3.07% to 5.14%, similar to the error observed on test set (see Table 2). Thus, using actual HTR predictions as a base permits to create more plausible mistakes in the data, while adding additional noise stimulates the sensitivity of post-correction models.

2.5.1 Dictionary realignment

First, the possibility of correcting the data against specialized dictionaries or lexicons is assessed. This is a common baseline approach for correcting noisy and topical OCR data, which consists in mapping the tokens from the noisy extraction on the closest type from the lexicon. This experiment both aims at providing a baseline for comparison with the other methods deployed and assessing the relevance of such approach when working with local historical demographic data. In this perspective, the Register of Swiss Surnames from the Historical Dictionary of Switzerland [50] was used to build a database of 44,664 surnames and 593 places of origin, including a list of the historical villages of the canton of Vaud, as well as all the Swiss cantons and the nearby historical European states. Moreover, a list of 11,414 historical occupations in French from the database of the International Institute of Social History is retrieved [51]. We also include the dataset of the National Institute of Statistics and Economic Studies (Insee, [52]) of France to retrieve some 5,410 first names attributed in France until 1939, supplemented manually with 625 additional first names observed in the censuses.

The realignment of the text segments is performed column-wise. A maximum Levenshtein distance l_2 is automatically fitted on the training subset, so as to maximize the true-positive corrections and minimize the false-positive, when matching a word with a correction-candidate.

2.5.2 Statistical models and neural network-based approaches

In a first step, the OCR predictions are subjected to a simple precleaning, by normalizing spaces (e.g., double spaces) and removing most punctuation marks. Indeed, punctuation can be very useful for formatting dates for instance, but is carrying a negligible amount of information for the columns concerned by the precleaning (first names, surnames, origins, occupations). In fact, they multiply the number of types in a an undesirable way, when it comes to post-correction. In the case of historical censuses such as this one, the spellings ’jean-samuel’ and ’jean samuel’ do not carry different historical or linguistic information; they are simply the result or an arbitrary choice made by the censor.

In a second step, five popular post-correction strategies, three based on statistical models [53–55], and two based on neural networks [56,57] are tested, compared and combined with precleaning step described above. In a first set of experiments, the effectiveness of these methods in terms of CER reduction is assessed individually. In a second step, the total performance of the post-correction, when successively applying the precleaning, and the external method from the literature is evaluated. In the last experiment, the HTR precleaning is applied first, followed by a majority vote of the three best algorithms in the literature. That is, at least two methods must agree on the post-correction for the intermediate token to be effectively modified in its final post-corrected form. An open-access implementation is used to train the five methods from the literature [58].

3 Results

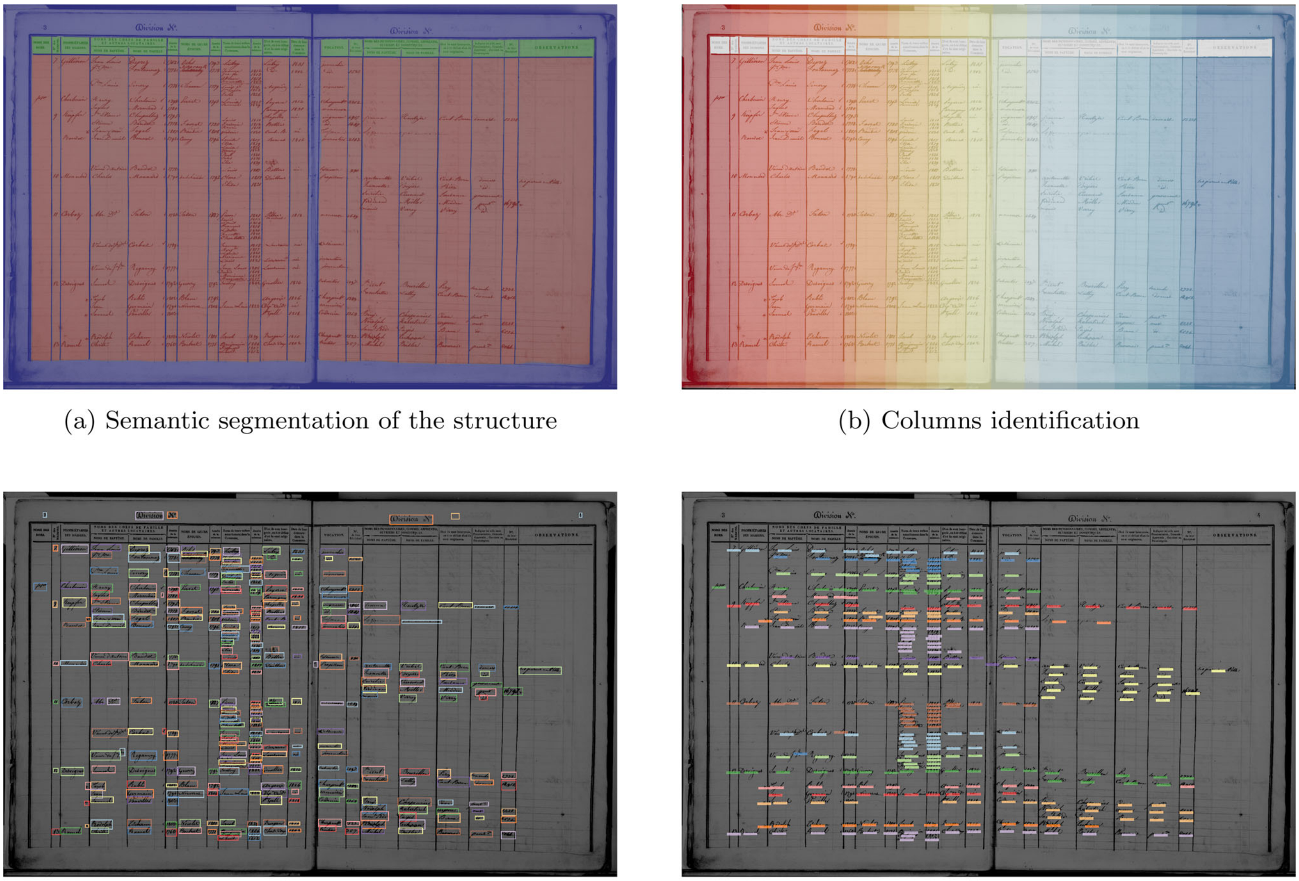
3.1 Layout analysis

First, the method based on projection profile is evaluated manually, on a sample of 0.7% of the pages, drawn randomly, i.e., 125 pages. In the end, this baseline approach permits to successfully detect 87.5% of the column separators, against 98.0% for the method based on semantic segmentation. Besides, the percentage of pages that are extracted without any error is 16.8% when using projection profile, against 92.2% with semantic segmentation. The performance of the semantic segmentation approach and the typology of errors encountered is detailed hereafter.

The segmentation CNN (Fig. 3a) optimized on the whole training data shows a high performance, with a mean intersection over union (mIoU) of 94.3%. The fine-tuned models trained specifically on the corresponding structural subcorpora perform better, with a mIoU of 97.3% for 1805-13,

Table 1 Column separator detection assessment

Typology of error	Prevalence
LE1 - false positive	0.17 %
LE2 - false negative	0.23 %
LE3 - offset	1.59%



(a) Semantic segmentation of the structure

(b) Columns identification

(c) Detection of the text segments

(d) Household attribution

Fig. 3 Results from the main steps of the extraction pipeline on the second page of the 1832 census

97.1% for 1832-1848, 98.3% for 1849-58, 97.8% for 1859-85, and 98.0% for 1886-98.

For assessing the quality of column detection in more details (Fig. 3b), circa 2% of the pages are randomly selected and validated manually. Three typologies of errors are encountered. The first one (LE1) is the column separator false positive, meaning that the algorithm has detected too many column separators. The second one (LE2) is the column separator false negative, i.e., one or more columns are not detected by the algorithm. The third error typology (LE3) is the column separator offset, where the correct number of columns is detected, but an offset of more than 20 pixels between the predicted separation and the ground truth is observed. These errors are not mutually exclusive and can be encountered on the same page, for different columns.

The most frequent errors are offsets (LE3). Most of the time, LE1 and LE2 occur on the same page, compensating each other. A frequently identified source of mistake is the central binding of the document, which can hide the whole or a portion of a column (LE2). Sometimes, a dark and sharp central binding itself can be confused with a column separator (LE1). Other times, slight pencil lines are found in addition to the printed ones (e.g., to separate the dates into DD/MM/YY format), and this can lead to false positive (LE2) and offsets (LE3). However, this occurs mostly in the late censuses and does not account for the majority of LE2/LE3 errors. It should be noted, however, that the error rate is not relative to the number of pages that contain misdetected columns, since it is much more common for there to be multiple errors on one page. Of the 409 manually assessed pages, there are 32 pages with incorrectly detected columns, which represents 7.82% of all pages.

3.2 Extraction of text segments

In order to validate the extraction of text segments (Fig. 3c), the hyperparameters are optimized on 80% of the annotated pages (66 pages, c.a. 27,800 text segments). Then, the segment-wise mIoU is computed on the remaining 20% (17 pages, c.a. 7,100 text segments). Our approach shows a mIoU of 66.6% (median 72.8%), which is considered rather high. In comparison, the software Transkribus [59], also using Grüning's algorithm [12] to detect baselines, obtains a mIoU of only 15.9% (median 10.7%).

Most importantly, the better fit induced by the intelligent framing method results in a clear increase in HTR inference performance (Table 2). For the period 1832-98, the error for the Puigcerver model is reduced by 14% (from 5.74 to 4.93 CER) and by 21% for Flor (from 6.73 to 5.34 CER). For the Bluche model however, the error stagnates at a high level, with even a slight increase.

3.3 Handwritten text recognition

It can be noted in Table 2 that incorporating the first years of the censuses lowers the performance of the optical character recognition (OCR). For this reason, the years 1832-98 were extracted with a specific model. The raw HTR/OCR performance reaches a character error rate (CER) of 4.93%, in the best case, using the Puigcerver architecture [26].

A milestone output of the project is also the release of the *Lausanne Historical Censuses Dataset 35k* [48]. This quality dataset can foster research on historical registers and more globally on the transcription of French historical sources. In comparison, the IAM dataset has only 13k training examples, and the Bentham-R1 dataset only 12k, although the text segments are longer.

3.4 Structuration in a tabular database

The quality evaluation of the structure, and primarily the quality of row detection (Fig. 3d), relies on manual data verification. One random page is assessed for each slice of 200 pages in a yearly census. The errors are sorted in three typologies. The first typology of structural error (SE1) is related to missing words and induced 'inner' offsets. More specifically, it accounts for offset errors due to an undetected or missing word and offsets within a household. This often concerns the residents columns, or the children columns, for instance when the name of a child is matched with the wrong birth year. The second type (SE2) is linked to structural inconsistencies within a household. In this category, we consider any error that is directly due to inconsistency in the source. This can be caused by the commissioner using an unexpected and uncommon format, such as brackets, arrows, or other forms of indication to correct or change the assignment of a group of information to another household. Concretely, the children's names may for example be written on multiples rows, at the same height as another (unrelated) household, and a bracket or another diacritical signs is used to show where the names actually belong. As the brackets are not detected, this leads to mismatched items. The third typology of error (SE3) represents the offsets between rows, i.e., some part of a household is attributed to another household. This kind of error can happen, among other things, because of the central binding, when both pages are not well aligned.

Table 3 summarizes the results of the manual assessment. One can notice that SE3 is almost twice as frequent as the other two types of errors. The structural incoherences (SE2), which corresponds to human errors, concern 1 in 30 households. This is about the same proportion as SE1. The low prevalence of type SE2 errors is explained as they are only observed in some years of the censuses, probably due to poorly trained commissioners.

Table 2 HTR performance with various network architectures, measured with the character error rate (CER %) and word error rate (WER %) on test set, given with a 95% confidence interval (CI). The datasets marked with a star (*) did benefit from the intelligent framing method

Data	Arch.	CER % ±CI	WER % ±CI
1805-98*	Flor	6.02 ± 0.65	21.33 ± 1.92
	Puigcer.	5.36 ± 0.66	17.84 ± 1.79
	Bluche	11.60 ± 0.91	37.30 ± 2.28
1832-98*	Flor	5.34 ± 0.65	20.52 ± 2.05
	Puigcer.	4.93 ± 0.65	17.79 ± 1.94
	Bluche	10.80 ± 0.90	36.76 ± 2.47
1832-98	Flor	6.73 ± 0.70	24.84 ±2.07
	Puigcer.	5.74 ± 0.66	20.83 ±1.97
	Bluche	10.55 ± 0.88	35.71 ±2.27
Bentham	Flor	9.58 ± 0.86	38.65 ± 1.89
	Puigcer.	10.12 ± 0.94	39.15 ± 1.87
	Bluche	15.64 ± 1.20	47.29 ± 1.95

Table 3 Table layout assessment

Typology of error	Prevalence
SE1 - Words fn / inner offset	3.47 %
SE2 - Rows structural incoherences	3.22 %
SE3 - Rows offset	6.09 %

3.5 HTR Post-correction

The character error rates (CER) resulting from the various post-correction strategies are gathered in Table 4. Despite the relatively low CER after HTR, the realignment on specialized lexicon, used as a baseline approach for comparison, is ineffective at improving the CER further, no matter how low the threshold on the Levenshtein distance is set.

In raw performance, with a CER ranging from 3.89% to 4.18%, the standard algorithms perform well. On average, 15.3% of the errors are corrected and up to 18.6% for Luong. However, one can notice that the results of the different algorithms are systematically better when they are applied after HTR precleaning (21.8% on average, and up to 25.3% for the best algorithm). The relative performance of the standard algorithms is quite similar when applied directly on the raw

Table 4 CER of the different post-correction algorithms applied on the validation and test sets, given with a 95% confidence interval (CI)

	CER % ±CI	Δ w. HTR
HTR	4.78 ± 0.45	–
Preclean (Pc)	4.47 ± 0.43	−6.5%
Garbe [53]	4.18 ± 0.47	−12.6%
Norvig [54]	3.99 ± 0.46	−16.5%
Stefanovic et al. [55]	4.01 ± 0.45	−16.1%
Bahdanau et al. [56]	4.17 ± 0.46	−12.8%
Luong et al. [57]	3.89 ± 0.45	−18.6%
Pc + Garbe	3.88 ± 0.45	−18.8%
Pc + Norvig	3.68 ± 0.44	−23.0%
Pc + Stefanovic	3.71 ± 0.44	−22.4%
Pc + Bahdanau	3.85 ± 0.44	−19.5%
Pc + Luong	3.57 ± 0.43	−25.3%
Pc + Luong ∩ Norvig ∩ Stefanovic	**3.44 ± 0.42**	**−28.0%**

OCR (−15.3%), or after precleaning (−14.5%), so the two steps add up. Both statistical and neural network-based methods show a substantially similar performance on our dataset, despite the relatively repetitive nature of the data.

Finally, the last experiment is the one that results in the best performance. Applying precleaning followed by a majority vote of several algorithms combined achieves a CER of 3.44%, which is a reduction of more than a quarter of the error rate.

3.6 Output of the extraction pipeline

As an example, the result of the extraction of the first households and columns of Fig. 1 is presented in Table 5. Overall, the output is readable and structured. The qualitative analysis of this excerpt shows that, in situ, errors are not necessarily surprising and also reflect the expected limitations of automatic approaches.

Looking more precisely at the mistakes found in Table 5, we notice for example that two entries are found in row 2, under ‘Name of the Spouse‘: ‘sprant‘ and ‘schmotz.‘ In the corresponding image, there are indeed two text segments at this place. However, the one below, which might in fact be read as ‘schwarz‘ or ‘schuatz,‘ is heavily crossed out. It also hinders the readability of the upper one, probably ‘schrantz.‘ Another example is the particular inscription under the street names, detected as ‘fose‘ and which we interpret rather as ‘fsno‘ or ‘sno‘ (probably an abbreviation of ‘sans numero,‘ with means without number). A typical example of a SE3 row detection error is the presence of a child named ‘louis‘ in the Chapallaz household (l. 6). This child is actually listed on the bottom household (Baudel-Favrat), but is located clearly above the baseline, although it does not really overlap with the upper cell either (which would be counted as an SE2 error). In the Bonnet-Cevey household (l. 9) a segment of text was not detected in the column of children’s birth years (SE1). The algorithm identifies this gap and introduces a midpoint (·), to indicate an empty field. However, the gap is not detected at the correct position, and the third child (the second Louise) receives the birth year 1819, while it is actually the birth year of her sister (Elise). The attribution is then compensated for the following children in that household.

4 Discussion

Simple and inflexible methods like projection profiles fail to accurately render the tabular layout and detect column separators. However, our results show that an excellent performance can be achieved using a common semantic segmentation model, trained on few annotation examples. In fact, our results demonstrate that a semantic segmentation model can be efficiently fine-tuned with less than a dozen samples, as long as it has already learned a general idea of the structure. For instance, the mIoU reaches 97.3% when post-training on the 1805-13 data based on a subset of only 7 training samples. An adequate data augmentation is probably one of the reasons for this good result. In general, the results of the layout analysis show that column detection is not a problem when it can be based on clear separators. The visual separators are both solid markers for the computer vision methods, but they also reinforce the intrinsic structure of the document. Indeed, the columns are in the vast majority of cases respected by the commissioners, although some overruns have been spotted in the corpus. As stated in the results section, the most frequent errors (1.59%) concerning the detection of column separators are by far imprecisions, or offsets (LE3). This is expected as the algorithm includes the expected number of columns for each year, therefore limiting cases of failure LE1 and LE2 (0.4%). The balance is explained by the presence of important graphical challenges, such as the partial disappearance of columns in the central binding (see Fig. 4b).

In a global way, the intelligent framing method leads to a very noticeable increase in similarity between the automatically detected text segments and the manual annotations. This step appeared to be essential for the generalization of the HTR performance, when validating the inference capacity on the test set. Adequate framing of the text segments plays an important role in the quality of the HTR. For documents such as censuses, the risk of several text segments encroaching on each other, or with other elements, especially column separators, is important. Moreover, some letters, in particular capital letters in cursive script, may be poorly detected due to disparity in height, or width (see Fig. 4). Overall, with a decrease of 14% to 21% CER for the best HTR models,

Table 5 Result of the end-to-end extraction pipeline on the first rows and columns of the second page of the 1832 census (see Fig. 1). In this visualization, the years were additionally post-corrected using a string distance

Street	Nr.	Owner	First name	Surname	Birth	Name of the spouse	Birth	Names of the children	Birth	Origin or bourgeoisie	Arrival	Occupation
.	7	gillieron	jean louis	deprez	1768	ochs	1793	lutry	.	lutry	1831	journalier
.	.	.	jn pre	fontannaz	1778	sprant schmotz	1778	jeanne	1812	lutry	1802	journalier
								jne fse	1813			
								aplam	1814			
								jeannette	1822			
.	.	.	pre louis	ernery	1774	chavan	1779	ig fs	1816	nagniere	ne	vigneron
								jn lauise	1813			
								jules	1826			
fose	.	cherbonnin	henry	cherbouin	1798	perret	1793	louise	181	payerne	1812	charpent
								samuel	1829			
.	.	.	jaques	mermoud	1780	.	.	.	.	farvagny	1830	manoeuvre
				mermoud								
.	9	rupfer	jn etienne	chapallaz	1795	.	.	louis	1810	chapelles	ne	vigneron
.	.	.	etienne	baudel	1778	favrat	1780	frederic	1812	bottens	.	vigneron
								pudevre	1816			
.	.	.	jean fonrod	fogel	1807	baudel	1806	frederic	1830	cont s	ne	tesserand
.	.	baudet	jean daniel	bonnet	1792	cevey	1796	louise	1817	renens	1815	journalier
								elise	.			
								louise	1819			
								henry	1826			
								paul	1824			
								jules	1826			
								elie	1829			
.	.	.	veuve d antoine	baudet	1778	.	.	louis	1807	bottens	nee	tessiveuse
.	10	monnard	charles	monnard	1790	de scheibler	1792	clara	1817	daillens	.	professeur
								elisa	1821			
.	11	corbaz	abr	jaton	1780	jaton	1887	pierre	1821	villars	1818	manoeuvre
								david	1813	emadiez	nee	lessioeuse
								louis	1814			
								francois	1818			
								lisette	1820			
								jeanette	1822			
								charlotte	1826			
.	.	.	veuve de f del	corbaz	1789	.	.	jeanne	1815	lausanne	nee	journaliere
								margot	1817	lausanne		
								louise	1820			
								marianne	1823			
								louis	1825			

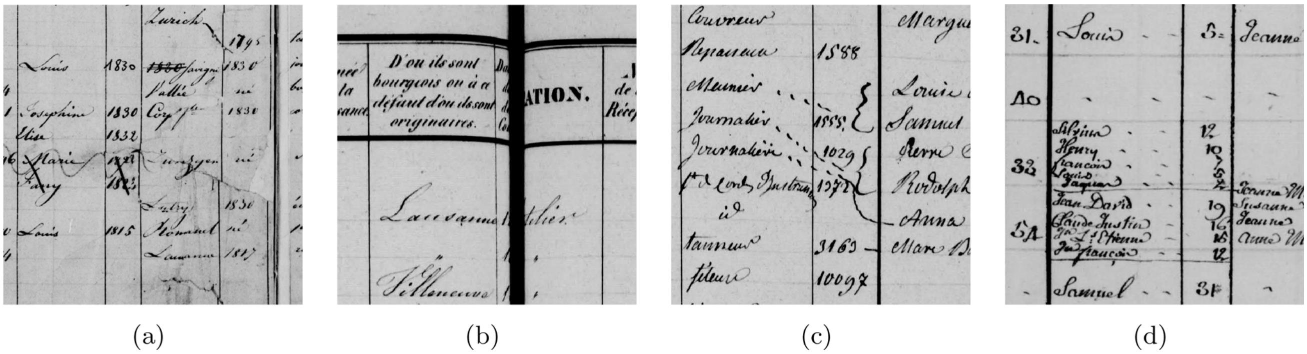


Fig. 4 Examples illustrating some of the difficulties of the extraction task. **a** Material damages, tears, fading ink, chaotic erasures. **b** Partial disappearance of the columns in the central binding. **c** Spatial discontinuity of the rows, reassignment braces. **d** Overlapping text, inconsistent line spacing and text size, challenging handwriting

it represents a clear improvement over [12]. Pragmatically, knowing that a large part of the remaining text recognition errors can probably be avoided only by improving the HTR models themselves, we believe that the effect of applying the intelligent framing heuristics is substantial. Besides, the performance of text segments extraction also impacts the detection of rows and households, which is probably the most challenging task of the pipeline.

In the end, the HTR achieves 4.93% CER, which seems excellent, especially in comparison with the Bentham reference dataset, whose complexity is similar. The repetition of words in the censuses, however, as well as the large size of our training dataset probably contributes to this performance. Moreover, the limited number of classes (characters) also plays a role. Obviously, the quality of the HTR varies according to the regularity of the censor's handwriting and its representation or not in the random sample used for training. While some parts of the document are easily readable, others are more noisy (see Fig. 4a, 4d).

The structuration in a tabular database, which involves the detection of rows and households, is the most challenging step. The difficulty in solving this problem is mainly due to the variability of the formats. The way in which children and residents are listed was identified as a frequent source of mistakes. On the one hand, the text in lists is sometimes written quite smaller, which can cause issues with row detection (see Fig. 4d). On the other hand, the row spacing with the next household is not always respected, which can obviously lead to allocation errors since the name on the list is physically already overlapping with the next household. Additionally, household detection is sensitive to the presence of diacritical signs. For instance, the meaning of quotation marks, added by the commissioner, is not always clear: they are sometimes used to indicate a repetition (ibidem), sometimes on the contrary to indicate an empty datapoint. Moreover, as the censuses were historically used as a support for the administrative management of the city, the pages are covered with numerous diacritical signs, like crosses, arrows, reassignment braces, or erasures, related to administrative changes, or notes (see Fig. 4a, c). These marks are detected as text segments, and their presence can complicate the detection of the segments relevant to locate the household, for instance the head of the household or the list of children.

Finally, the HTR post-correction step is conclusive. If the results indicate that the use of dictionaries, even topical ones, seems not to be relevant for historical data of this type, other strategies can be effective to reduce the HTR errors. The ineffectiveness of approaches based on string distance seems to be mostly induced by the difficulty to find relevant and fairly exhaustive dictionaries, as this strategy is found to be effective in other contexts [60]. Beyond the historical and cultural differences, the intrinsic variability of historical documents is also an issue. Consequently, for every topic analyzed here (first names, surnames, places, professions), the risk of distorting the data even exceeds the probability of correcting it. Statistical or neural network-based methods, however, show promising results. In the end, we notice that the combination of several methods, through a conservative majority vote, a strategy that limits false positives, is the most effective setting.

In general, the pipeline is designed to limit the cascading impact of errors on other steps of the process. However, many dependencies remain. For instance, the detection of columns is independent from the detection of text segments and HTR. It may, however, slightly impact the detection of rows and households, although the latter was designed to be robust to missing fields, natively present in the document. Conversely, the detection of rows heavily relies on the detection of text segments. This double dependency may explain the lower performance of the row detection since previous errors are accumulated in this step. The importance of optimizing text segment detection for HTR was emphasized above. Similarly, the performance of post-correction seems proportionally better when the initial CER is lower, as was illustrated by measuring the impact of preprocessing. This is due to a higher type-token ratio and a lower variance in the HTR data, which improve the statistical separability of the variants.

The overall qualitative analysis of the output shows the nature of the errors encountered in the final results. The quality of the database is satisfying, as it is readable and exploitable for research in historical demography, with an adequate research protocol and methodology. The data are also readily usable for ordinary historical research or genealogy. The legibility is probably above that of the original documents, for people who are not trained in paleography, nor familiar with local patronyms and toponyms.

5 Conclusion

The various validation steps highlighted the main challenges of an automatic extraction. Among these, the natural variability of the raw historical data, due to the length of the period covered by the corpus, but also differences in layout between commissioners, changes in the table structure, as well as corrections and notes added a posteriori. Digitization flaws, such as columns disappearing in the central binding, also contribute to the complexity of the problem.

To conclude, the pipeline allowed the extraction of over 6.2 million datapoints from 18,831 pages of historical censuses using a robust algorithm, leveraging optimization strategies to dynamically adapt to structural changes. Our method can handle both explicit (e.g., printed separators) and implicit (e.g., cells consisting of multiple lines of free text without separators) layouts. HTR, combined with an efficient pre-processing of text segments, and complemented with a combination of post-correction strategies, reaches the excellent character error rate of 3.44%. These various improvements allow us to obtain quality data that can be exploited for research in demography or social history.

Funding Open access funding provided by EPFL Lausanne. This work was supported by the public fund for Collaborative Research on Science and Society (CROSS 2021) “Names of Lausanne: the evolution of family names in administration records 1803–1900”, and the College of Humanities at EPFL.

Data availability statement The images of the digitized censuses used in this research are entirely available on the website of the Archives of the City of Lausanne [61]. All training data and the resulting historical dataset are made open-access and released along with this article [48, 62].

Declarations

Conflict of interest The authors declare they have no competing financial or non-financial interests.

References

1. Ruggles, S.: (2014) Big Microdata for Population Research. Demography 51(1):287–297. https://www.jstor.org/stable/42919999
2. Williams, L., Godfrey, B.: Bringing the prisoner into view: english and Welsh census data and the Victorian prison population. Australian Historical Stud. **47**(3), 398–413 (2016). https://doi.org/10.1080/1031461X.2016.1208258
3. Municipalité, de Lausanne AVL RB 14-023. Procès-verbaux de la Municipalité de Lausanne. p 376 (1828)
4. Sibade, C., Retornaz, T., Nion, T., et al.: Automatic indexing of French handwritten census registers for probate geneaology. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. Association for Computing Machinery, New York, NY, USA, HIP ’11, pp 51–58, (2011) https://doi.org/10.1145/2037342.2037352
5. Nion, T., Menasri, F., Louradour, J., et al.: Handwritten information extraction from historical census documents. In: 2013 12th International Conference on Document Analysis and Recognition, pp 822–826, (2013) https://doi.org/10.1109/ICDAR.2013.168
6. Clawson, R., Bauer, K., Chidester, G., et al.: Automated recognition and extraction of tabular fields for the indexing of census records. In: Zanibbi R, Coaũsnon B (eds) Document Recognition and Retrieval XX, International Society for Optics and Photonics, vol 8658. SPIE, pp 170 – 180, (2013) https://doi.org/10.1117/12.2004788
7. Pedersen, B.R., Holsbø, Andersen, T., et al Lessons learned developing and using a machine learning model to automatically transcribe 2.3 million handwritten occupation codes. Histor Life Course Stud 12:87 (2022) https://doi.org/10.51964/hlcs11331
8. Andrés Moreno, J.: Search and information extraction in handwritten tables. Universitat Politècnica de València, Master thesis (2021)
9. Lang, E., Puigcerver, J., Toselli, A.H., et al.: Probabilistic indexing and search for information extraction on handwritten german parish records. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, pp 44–49, (2018) https://doi.org/10.1109/ICFHR-2018.2018.00017
10. Simistira, F., Bouillon, M., Seuret, M., et al.: ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 1361–1370, (2017) https://doi.org/10.1109/ICDAR.2017.223, ISSN: 2379-2140
11. Diem, M., Kleber, F., Fiel, S., et al.: cBAD: ICDAR2017 Competition on baseline detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 1355–1360, (2017) https://doi.org/10.1109/ICDAR.2017.222, ISSN: 2379-2140

12. Grüning, T., Leifert, G., StrauSS, T., et al.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019). https://doi.org/10.1007/s10032-019-00332-1
13. Guerry, C., Coüasnon, B., Lemaitre, A.: combination of deep learning and syntactical approaches for the interpretation of interactions between text-lines and tabular structures in handwritten documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp 858–863, (2019). https://doi.org/10.1109/ICDAR.2019.00142, ISSN: 2379-2140
14. Coüasnon, B., Lemaitre, A.: Recognition of tables and forms. In: Doermann D, Tombre K (eds) Handbook of Document Image Processing and Recognition. Springer, London, p 647–677, (2014). https://doi.org/10.1007/978-0-85729-859-1_20
15. Clinchant S., Déjean, H., Meunier, J.L., et al.: Comparing machine learning approaches for table recognition in historical register books. arxiv, (2019). https://arxiv.org/abs/1906.11901
16. Déjean, H., Meunier, J.L.: Table rows segmentation. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp 461–466, (2019) https://doi.org/10.1109/ICDAR.2019.00080, ISSN: 2379-2140
17. Schreiber, S., Agne, S., Wolf, I., et al.: DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 1162–1167, (2017) https://doi.org/10.1109/ICDAR.2017.192, ISSN: 2379-2140
18. Zucker, A., Belkada, Y., Vu, H., et al.: ClusTi: clustering method for table structure recognition in scanned images. Mobile Netw. Appl. (2021). https://doi.org/10.1007/s11036-021-01759-9
19. Liang, X., Cheddad, A., Hall, J.: Comparative study of layout analysis of tabulated historical documents. Big Data Res. **24**(100), 195 (2021). https://doi.org/10.1016/j.bdr.2021.100195
20. Breuel, T.M.: The OCRopus open source OCR system. In: Document Recognition and Retrieval XV, vol 6815. International Society for Optics and Photonics, p 68150F, (2008) https://doi.org/10.1117/12.783598
21. Shen, Z., Zhang, R., Dell, M., et al.: LayoutParser: a unified toolkit for deep learning based document image analysis. (2021) arXiv, https://arxiv.org/abs/2103.15348
22. Oliveira, S.A., Seguin, B., Kaplan, F.: dhSegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp 7–12, (2018) https://doi.org/10.1109/ICFHR-2018.2018.00011
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440, (2015). https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html
24. Li, M., Cui, L., Huang, S., et al.: TableBank: table benchmark for image-based table detection and recognition. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1918–1925, (2020). https://www.aclweb.org/anthology/2020.lrec-1.236
25. de Sousa Neto, AF., Bezerra, BLD., Toselli, AH., et al.: HTR-Flor++: a handwritten text recognition system based on a pipeline of optical and language models. In: Proceedings of the ACM Symposium on Document Engineering 2020. Association for Computing Machinery, New York, NY, USA, DocEng '20, pp 1–4, (2020). https://doi.org/10.1145/3395027.3419603
26. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 67–72, (2017).https://doi.org/10.1109/ICDAR.2017.20, ISSN: 2379-2140
27. Bluche, T., Messina, R.: Gated convolutional recurrent neural networks for multilingual handwriting recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 646–651, (2017). https://doi.org/10.1109/ICDAR.2017.111, ISSN: 2379-2140
28. Aradillas Jaramillo, JC., Murillo-Fuentes, JJ., M. Olmos, P.: Boosting handwriting text recognition in small databases with transfer learning. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp 429–434, (2018). https://doi.org/10.1109/ICFHR-2018.2018.00081
29. Rigaud, C., Doucet, A., Coustaty, M., et al.: ICDAR 2019 competition on post-OCR text correction. In: 15th International Conference on Document Analysis and Recognition, Sydney, Australia, pp 1588–1593, (2019) . https://hal.archives-ouvertes.fr/hal-02304334
30. Levenshtein, VI., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady, Soviet Union, pp 707–710 (1966)
31. Bell, S., Marlow, T., Wombacher, K., et al.: Automated data extraction from historical city directories: the rise and fall of mid-century gas stations in providence, ri. PLOS One **15**, 8 (2020). https://doi.org/10.1371/journal.pone.0220219
32. Haldar, R., Mukhopadhyay, D.: Levenshtein distance technique in dictionary lookup methods: a improved approach. (2011)http://arxiv.org/abs/1101.1232
33. Berenbaum, D., Deighan, D., Marlow, T., et al.: Mining spatio-temporal data on industrialization from historical registries. J. Environ. Inf. **34**(1), 28–34 (2019). https://doi.org/10.3808/jei.201700381
34. Häläinen, M., Hengchen, S.: From the Paft to the Fiiture: a Fully automatic NMT and word embeddings method for OCR post-correction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). INCOMA Ltd., Varna, Bulgaria, pp 431–436, (2019) .https://doi.org/10.26615/978-954-452-056-4_051
35. Hakala, K., Vesanto, A., Miekka, N., et al.: Leveraging text repetitions and denoising autoencoders in ocr post-correction. (2019) arXiv, https://arxiv.org/abs/1906.10907
36. de Sousa Neto, A.F., Bezerra, B.L.D.: Toselli AH Towards the natural language processing as spelling correction for offline handwritten text recognition systems. Appl. Sci. **10**, 21 (2020). https://doi.org/10.3390/app10217711
37. Kissos I, Dershowitz N.: OCR Error Correction Using Character Correction and Feature-Based Word Classification. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp 198–203, (2016)https://doi.org/10.1109/DAS.2016.44
38. Mikolov, T., Chen, K., Corrado, G., et al. Efficient estimation of word representations in vector space. arXiv, (2013) https://arxiv.org/abs/1301.3781
39. Devlin, J., Chang, M., Lee, K., et al. BERT: pre-training of deep bidirectional transformers for language understanding. (2018) http://arxiv.org/abs/8100.4805
40. Roy, A., Ghosh, S., Ghosh, K., et al. An unsupervised normalization algorithm for noisy text: a case study for information retrieval and stance detection. (2021) http://arxiv.org/abs/2101.03303
41. Cao, H., Rawls, S., Natarajan, P. 1990 us census form recognition using ctc network, wfst language model, and surname correction. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp 977–982, (2017) https://doi.org/10.1109/ICDAR.2017.163
42. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst., Man, Cybernet. **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
43. Ronneberger, O., Fischer, P., Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241 (2015)

44. Kingma, D.P., Ba, J. Adam: a method for stochastic optimization. http://arxiv.org/abs/1412.6980 (2014)
45. Zuiderveld, K.: Contrast limited adaptive histogram equalization. Gr. Gems **4**, 474–485 (1994)
46. Bergstra, J., Bardenet, R., Bengio, Y., et al.: Algorithms for hyper-parameter optimization. Adv. Neural Inf. Process. Syst. **24**, 87 (2011)
47. (2020) Computer vision annotation tool. https://cvat.org/
48. Rappo, L., Petitpierre, R., Kramer, M.: Lausanne Historical Censuses Dataset HTR 35k (2023). https://doi.org/10.5281/zenodo.7711178
49. Sánchez JA (2016) Bentham Dataset R0. https://doi.org/10.5281/zenodo.44519
50. (2021) Register of Swiss Surnames. https://hls-dhs-dss.ch/famn/?lg=e
51. (2021) History of Work. https://historyofwork.iisg.nl/search.php
52. (2021) Fichier des prénoms Etat civil. https://www.insee.fr/fr/statistiques/2540004?sommaire=4767262
53. Garbe, W.: 1000x faster spelling correction algorithm. (2012) https://towardsdatascience.com/symspellcompound-10ec8f467c9b
54. Norvig, P.: How to write a spelling corrector. (2007) http://norvig.com/spell-correct.html
55. Stefanovič, P., Kurasova, O., Štrimaitis, R.: The n-grams based text similarity detection approach using self-organizing maps and similarity measures. Appl. Sci. 9(9):1870 (2019)
56. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. http://arxiv.org/abs/1409.0473
57. Luong, M.T., Pham, H., Manning, CD.: Effective approaches to attention-based neural machine translation. (2015) http://arxiv.org/abs/1508.04025
58. de Sousa Neto, AF. arthurflor23/spelling-correction. (2020) https://github.com/arthurflor23/spelling-correction
59. Colutto, S., Kahle, P., Guenter, H., et al.: Transkribus. a platform for automated text recognition and searching of historical documents. In: 2019 15th International Conference on eScience (eScience), pp 463–466, (2019) https://doi.org/10.1109/eScience.2019.00060
60. Priambada, S., Widyantoro, DH.: Levensthein distance as a post-process to improve the performance of OCR in written road signs. In: 2017 Second International Conference on Informatics and Computing (ICIC), pp 1–6, (2017) https://doi.org/10.1109/IAC.2017.8280534
61. Corps de police (1898) Recensements communaux pour 1804-1813 et 1832-1898. https://vidy-archives.lausanne.ch/adm-c1-rc-106
62. Petitpierre, R., Kramer, M., Rappo, L., et al.: 1805-1898 Census Records of Lausanne: a Long Digital Dataset for Demographic History (2023). https://doi.org/10.5281/zenodo.7711640

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.