Machine Learning for Automated Historical Record Transcription on Edge Devices

Kai Brennan
Walter Scott Jr. College of Engineering
Colorado State University
Fort Collins, Colorado
kaijbren@colostate.edu

Paul Brotelande
Walter Scott Jr. College of Engineering
Colorado State University
Fort Collins, Colorado
brotel@colostate.edu

Abstract*—Record transcription is time-intensive, limiting access to history for casual researchers and archivists. Much of the resources for professional archivists are directed towards "significant" events and disregard the history of common people. This is a significant drawback as it limits the amount of easily accessible knowledge regarding history relevant to the populus. This restricts accessible knowledge of daily life, a historical field too often overlooked in favor of socio-political events or ruling families. To ameliorate this problem, there is significant potential in an transcribed database from vital event indices. This would aid fields reliant on handwritten records, such as genealogy, meteorology, and sociology. In this paper, we modify an existing optical character recognition (OCR) model by transfer learning on a custom-made dataset, making it tailored to delineate and transcribe French and Belgian "ten-year tables" from 1830-1924 (tables décennales, tienjarige tafels; indices of vital events for communes within a ten-year range). Experimental results demonstrate the model's ability to parse entries in various table formats, both printed and handmade, read the common notary hand (most commonly Écriture Anglaise [Ronde]) in antiquated Francophone records, and accurately transcribe names and dates. Our training loss is .3607 and our validation loss at its lowest is 2.26, providing a model useful for creating large searchable databases in the current landscape of Francophonic genealogical data. Anecdotal evidence and comparison also highlight its abilities. This is a novel application, as most OCRs focus on content in paragraphs, not indices. This provides the opportunity to collaborate on a regional level which has not yet been explored, greatly increasing knowledge through the reduction of work needed for transcription. This work sets the stage for expanding regional transcription projects, fostering collaboration among archives, and preserving underrepresented historical records. Additional steps are taken to attempt to deploy this model to a KRIA KV-260 from Xilinx, due to its strengths as a distribution method and a simple management system for archives.

Keywords—*État Civil, Burgerlijke Stand, tables décennales, tienjarige tafels, transcription, genealogy, optical character recognition (OCR), handwriting recognition, transfer learning, cultural heritage*

I. Introduction

In recent years, historical records have become an increasingly valuable resource for sociological and genealogical research, particularly in understanding the lives of everyday people. Countries such as France and Belgium have preserved civil records since their founding (1792 and 1830, respectively), offering a unique opportunity to explore the past through common people's stories. However, the handwritten nature of these records creates significant accessibility barriers.

Unlike in the U.S., where large-scale collaborations have manually transcribed large databases like the 1950 census, transcription efforts in Europe are often hyper-localized. Researchers typically focus on a single commune, limiting collaboration across regions. Municipal archives, such as Beveren in Belgium, rely on labor-intensive searches, even with index tools like "ten-year tables." Within the countries of France and Belgium, historical archivists created indices for each name, date, and event type within the books (sorted alphabetically by last name). However, the usefulness of these depends on the modern researcher to know the precise town or commune the event occurred in. In many cases, this knowledge is lost or otherwise unknown. This specifically creates significant challenges for American genealogists, who often lack specific regional knowledge of their ancestors' origins.

This paper presents AHRT (Automated Historical Record Transcriber), a novel method for transcribing and indexing ten-year table indexes for the stated countries and time periods. AHRT leverages transfer learning to specialize OCR models for regional variations in handwriting, language, and format. Our contributions include:

- Developing reliable text detection for diverse formats, from official government booklets to handmade entries.
- Creating a highly specific OCR model tailored to French and Belgian "ten-year tables" from 1830–1924.
- Introducing a scalable methodology for linguistic specialization within Francophone regions, with examples from Liège (Wallon), Normandie (French), and Nouvelle-Aquitaine (Langue d'Oc).
- Establishing success measures based on accurate name and date transcription, benchmarked against manual transcription rates.
- Beginning implementation of the model on the KRIA KV-260 from Xilinx, an Edge device.

Our goal is to enable the creation of searchable, region-wide databases, significantly reducing transcription labor and making historical records more accessible to researchers worldwide. By developing a user-friendly model that municipal archivists can implement with minimal technical expertise, we aim to reduce transcription burdens, facilitate genealogical research across borders, and preserve an invaluable connection to the social history embedded within these records.

II. Related Work

Optical character recognition has many previous general implementations already, and the fruit of this research is accessible everywhere, from Google Colabs libraries to built-in software in phone cameras. Beyond these all-purpose OCRs, there has been interest in specific applications for historical records. In modern machine learning, there have been numerous projects that seek to solve the problem of time-intensive transcription.

French Word Spotting: First of all, there have been efforts to transcribe antique Écriture Anglaise in literary works [1]. This model targeted a new idea in the expansion of OCR capabilities by "word spotting." Where most models are focused on isolating and identifying characters, this model makes use of the overall shapes of words to transcribe the volumes written by Swiss-French author C.F. Ramuz. This model greatly expanded on existing implementations of OCRs for French transcription in an efficient manner, and provided intrigue into the idea of highly specialized transcription models, as the final evaluation of the model was concerned with the handwriting of a single author. Our model relied on this same concept, due to the fact that we had a significantly limited "word bank." The difficulty in applying this research to historical indices was that this concept was specifically designed for paragraph transcription, rather than having each line separated to be a different entry with a unique topic. This is a completely different skill set for the model to learn, where each word is mostly independent of the surrounding ones. It is essential for a OCR specific to tabulated entries to not be dependent on paragraph-style input, and to exclude unnecessary vocabulary to minimize impact. Our novel application makes use of a highly limited vocabulary and line segmenter to ensure that these fixes are made.

Paris Census Records: More specific to historical record transcription, a large database of entries contained in the Paris census of several diverse years between 1926 and 1936 was used to train models for antique French script [2][3]. This showed the great importance of document segmentation, as the unmodified image uploads they were transcribing were full folios (two pages joined together), with numerous columns and tick marks to indicate information that was the same as above. Additionally, they worked on writer specialization to enhance text prediction accuracy. However, the limits to this research came in with the types of data they were using. The specific years they chose for training were highly intentional, as they had the same table layout. While this research is undeniably useful and served as a significant training dataset for our base model, this type of dependency on an original layout is not feasible for something like the ten-year tables, as these had many different formats depending on the language of the region, the current regime, and the resources of the specific commune. These factors require a more generalized model that knows how to divide between entries even in such edge cases. AHRT circumvents these troubles as we made sure to include not only different scribes and regions, but also to include many layouts in our training data. This allowed for the model to understand that a column is not necessarily always designated to be a field reserved for a specific type of data like surnames or years.

American Genealogical OCRs: Beyond this, there have been numerous American and other national efforts to create machine-learning models to transcribe historical records. Some that have been of great interest to this team are , where historic naval registers from the American Civil War were transcribed through the use of machine learning OCRs [4]. The limits to this come in to their reliance on the specific layout of naval records, which did not experience nearly as much variation during the war as civil records do across numerous regions. Additionally, this record does not span as many regions or dialects as there is not as much linguistic disparity within the Union or the Confederate states of this time period. Our model expands on OCRs such as these by accounting for dialectal versions of various names, dates, and writing styles. This approach lets us cater to the needs of many different archives and overcome the problems originating from hyper-localization.

All of these models create great time-saving advancements that allow researchers to waste less time searching through records. The noteworthy difference between these and our model is that these are focused on the full-length records themselves, while ours is focused on the indices. This difference allows our model to be more efficient for specific use in searchable databases, as the ten-year tables already have the necessary reference data isolated from the rest of the facts of the record. In addition to this, our proverbial word bank is significantly reduced from the other models, as instead of containing a working dictionary of the French language and grammar, it instead focuses on contemporary names, surnames, and date-relevant vocabulary. This specialization means that each of our training examples will be more potent and widely applicable.

As it seems, none of these existing models have been implemented on edge devices, save general-purpose OCRs which are commonly integrated into phones, cameras, and scanners. This is a marked lack that our project has attempted to address, as it would be through easily distributable edge devices that departmental archives would interact with this tool.

III. Data

For the Ocerization of antiquated records, many datasets already exist from independent genealogical and historical research efforts in manual transcription. Save the examples where spelling was corrected manually (independent from the record's actual text) or abbreviations were expanded, these pre-existing datasets are highly usable in their unmodified state. Most of these are focused on paragraph entries, will statements, church registers, and civil records. Additionally, there is the potential to create more data through research projects such as the ones that commonly take place in American genealogical societies, where entire collections are transcribed in their tabulated entries.

Initially, the base model [5] we specialized was trained on a normal handwriting dataset which was then specified for antique French handwriting using two transfer learning stages; one stage was for the language switch, and the second was for the orthographic changes. First, this paragraph-based OCR program provided a significant advantage to us and allowed us to address our focus problem even more closely than we would have been able to without it. The original model was trained using two datasets before we did our transfer learning, and these were a general handwriting dataset of incredible size, and a more specific Parisian census dataset [2][3]. This was one of the datasets that we were originally considering using when we intended to create the OCR from scratch, so it made sense to attempt to improve and apply it as a resource to our specific problem.

Once the preprocessing was taken care of, we began to create a dataset for transfer learning. As the splitter was already successful by this stage, we needed a dataset of single lines from ten-year tables that were good models of handwriting with various character and orthographic variations with regional influences. In order to do this, we created our own dataset specific to the ten-year tables. To maximize our potential success, we included transcriptions from as many different departments within the two countries to account for regional variations. In total, this was realized in 7 regions within France and Belgium. Each of these regions was specifically chosen due to either their adherence to Metropolitan French or different regional dialect. This dataset encompasses regions where the following languages and name traditions were used: Wallon, West Flemish, French, and Occitan.

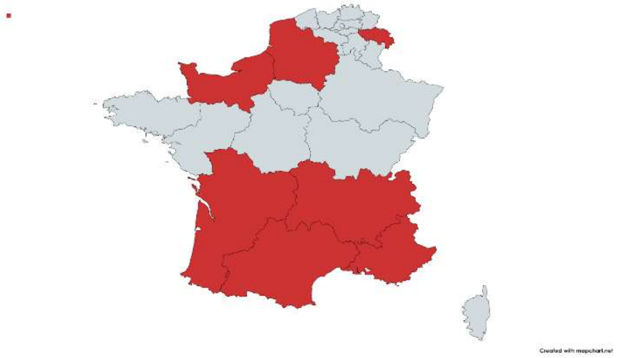


Fig. 1. Map of the Regions of France and Belgium Included in the Custom Dataset.

In the end, the transfer learning was done using this custom-made dataset. This provided us with enough data to learn necessary attributes of these indexes, as the model was already adept at paragraph based handwriting samples of Écriture Anglaise. This is a novel dataset that expands the application of machine learning to index transcription. As previously stated, most existing datasets focus on the content of the records, which is less useful for a low-footprint searchable database. This was our only option to acquire the necessary data for training. Despite significant consideration of other potential ways to get equivalent data, this was the sole opportunity to get data from our target regions, format, and years.

Our new data allowed the model to become highly adept at various styles of entry; in fact, the training data not only collaborates with the line segmenter to work on all styles of index, but it also allows the information in the columns to be written in any order, with three unique training examples provided. Our data-gathering technique has proved to be highly effective, as will be discussed in future sections, and such an adaptable entry format appears to make the model fully independent for the various register styles. This dataset proved effective even though it is smaller than ideal, and it can be very easily augmented with additional data from communes and departmental archives in potential collaborations.

IV. Methods

This novel specification needed to address numerous unique challenges. Beyond the image dividing for indexing, we also had to consider difficulties within the records. Although uniform in their cursive style, they are handwritten, making accurate transcription a challenging task. Second, the structure of the ten-year tables—listing names, dates, and event types—varies somewhat between regions, requiring a model adaptable to minor format variations without compromising precision. To address these challenges, we propose a specialized machine learning-based transcription tool that uses OCR (optical character recognition) and automatic segmentation specialized for the unique characteristics of these records. Through automated text segmentation and region-specific training, our approach allows for a streamlined and robust transcription process, creating a cohesive, region-spanning index of these archival records.

Within the limited regions of France and Belgium, to make our model able to transcribe almost all ten-year tables, we first needed to find a way to isolate each entry line so that we could then identify the names and dates they contained. Within the two countries we are targeting, at certain times and in certain places, the ten-year tables could be either completely hand-drafted or written in official booklets with lines and columns. Since we needed our algorithm to be highly efficient and automated, it was necessary to be able to upload an entire sheet image and automatically segment it, running each line through our OCR.

In order to process these images, we begin by ensuring that the file is completely black and white. This involved converting it to grayscale, before increasing the contrast to make the only shades fully pure (binary; where each pixel is either black or white). This is kept in parallel to the original file so that we can use it to locate text without overwriting the original quality for transcription. We then find the contours of the binary version, which is essentially drawing a line around every dark object in the frame. At this stage, shapes that are insignificant will be ignored, because of being an outlier in its size or placement within another contour line. Several other considerations are taken to ensure we have accurate segmentation, including overlap and padding. We pad each outside contour to ensure that we are not cutting off essential parts of the text entries, and to attempt to find the middle of the rows, even if they are handwritten and irregular heights. We also disregard overlapping rows, so that there are no duplicate entries for the same person just because their words landed on marginally different baselines.

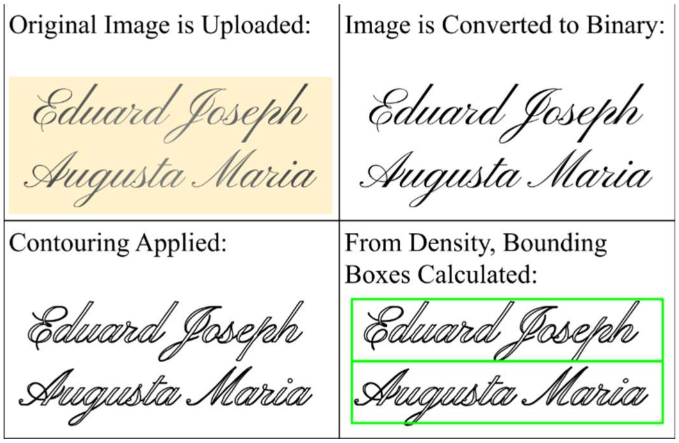
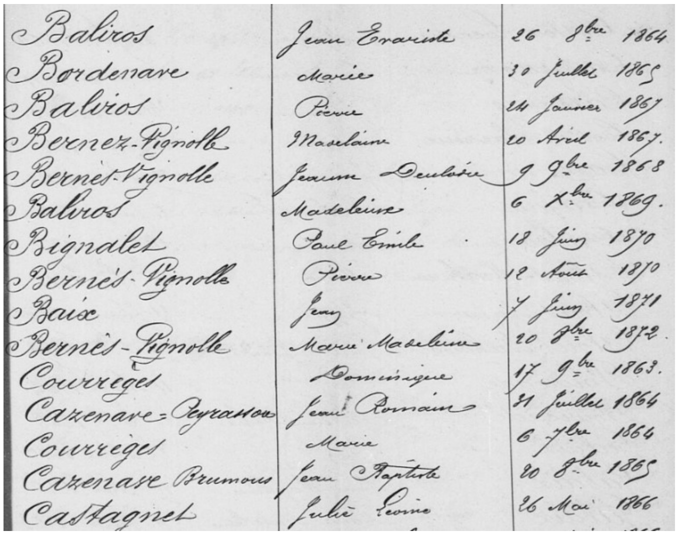


Fig. 2. Illustration of Segmenter Processes.

These processes ensure that each target is actually a unique line of text, rather than being a paragraph or a stray mark. From this, we then have the computer delineate the rows to run them automatically through our transcriber. We do this using the y coordinates of the contours in the binary image to separate the lines within the original image for readability and quality. After this stage, the binary image can safely be disregarded as its purpose has been served. These images retain their original order from the page, so one doesn't have to be concerned about the data being mixed up or losing references.



Baliros	Jean Evariste	26 8bre 1864
Bordenare	Marie	30 juillet 1865
Balirios	Pierre	24 Janvier 1867
Bernez-Pignolle	Madelaine	20 Avril 1867.
Bernes-Vignolle	Jeanne Deulaise	9 9bre 1868
Baliros	Madeleine	6 Xbre 1869.
Bignalet	Paul Emile	18 Juin 1870
Bernès-Pignolle	Pierre	12 Août 1870
Baize	Jean	7 Juin 1871
Bernès-Pignolle	Marie Madeleine	20 8bre 1872.
Courreges	Dominique	17 9bre 1863.
Cazenave-Peyrasson	Jean Romain	31 Juillet 1864
Courreges	Marie	6 7bre 1864
Cazenave Brumous	Jean Baptiste	20 8bre 1865
Castagnet	Julie Léonie	26 Mai 1866

Fig. 3. Example of an original image entry.

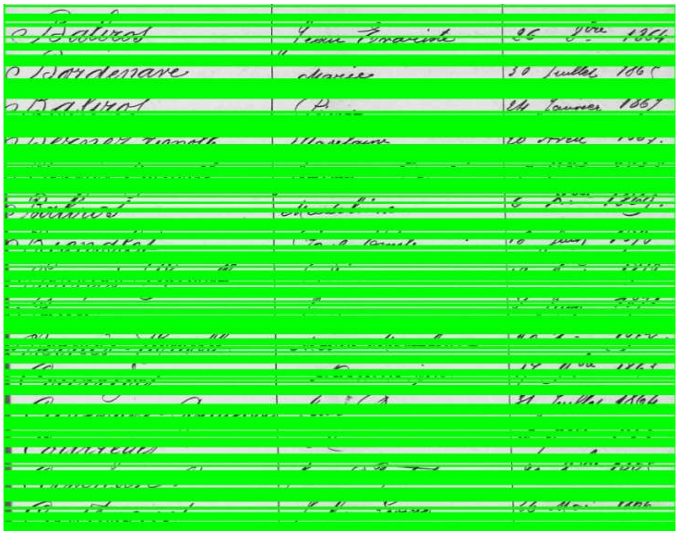
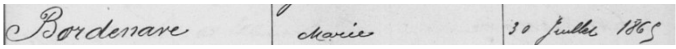


Fig. 4. The image after the preprocessing and contouring are completed and the bounding boxes are drawn superimposed on the original.



Bordenare	Marie	30 juillet 1865

Fig. 5. An example of a properly delineated entry, using the above algorithm (entry #2).

This approach was the only feasible way to adequately respond to the needs imposed by the entry styles of ten-year indices as it made sure errors would be confined within a single line, and that the entries would be properly delineated for their entry into a searchable database. As this was the true end goal of the project, everything had to be formatted around something that would be low-impact and well-organized. The application of this idea greatly informed the two authors on the importance and methodology of input image segmentation, which is applicable to not only scanned records but also to subject identification and other such detection algorithms.

Overall, this segmentation proved to be necessary for the project, even though there were some issues encountered. This happened where it would infrequently divide an entry into multiple parts or cut off ascenders or descenders. This did not happen often. It also did not pose much of an obstacle for the model overall, as long as the majority of the miniscule was visible. Majuscule seemed to be less significant in this regard, and flourishes in the writing were inessential for identification.

In order to successfully complete this project, we adopted a strategy where the training is easily modular to ensure that it is applicable to various regions. This is functional because the standard French handwriting would encompass the entire country, but to create local transcriptions, some region specific names and vocabulary would be necessary. Having such a modular approach should also allow for this model to become language independent, as long as the handwriting style is similar; our strategy tries to assign the skills for dialects and names closer to the end of transfer learning for easy customization.

After deciding upon this route, we began the transfer learning that was required to specialize our model to these types of records. This was essential to the overall success of the project as it took a model that was adept for various word forms in paragraph entries and turned it into an efficient name and date transcriber for single lines. This is a completely different style of reading necessary, but certain skills were transferable. As the model we found was already adept in understanding the graphical forms common in Écriture Anglaise and had been trained on datasets that we were considering using ourselves before we found it, it was the perfect starting point for a specialized ten-year table database creator. Here, we exercised transfer learning to adapt it for our goals. The dataset we utilized for general training was very handy for the ultimate implementation, although smaller than ideal and difficult to make. We struggled for a while to figure out the most efficient method in which to create the data for training and testing as the labeling proved very difficult to create. As previously discussed, we had a large span of regional and time-based records.

At this point, we had to consider the most useful style of output from our program. In order to facilitate easy CSV or searchable database construction, we decided to create the most convenient entry format. This involves simply breaking apart each line of the record image to be its own entry. From here, additional information can be supplemented as desired; the text could be broken up at the space characters to get the entries for each column, and other departmental information could be added. As the regional information is not contained within the handwritten records themselves, this would have to be specified by the project or department that is using it. This CSV format also provides liberty to communes who may have had a different entry order. After the initial transcription, one could also run the output through a specialized language learning model to normalize them. This would enable someone to make a database easy to use without worrying about overwriting the original misspellings or odd characters that were authentic to the manuscript.

After this point came our edge implementation. Unfortunately, by the end of the project, this was not fully realized. In the use of the AMD-specific quantizer and inspector, it seems that regardless of various attempts to reduce the size and to make the model compatible with the specific board we chose, the quantizer brought the accuracy to zero percent every time.

V. Experiments

Overall, this model proved to be very adept within the intended implementation and with standard orthography with some room for regional variations. In this section, we evaluate our model in terms of training and validation loss. Additionally, we compare our model to accessible state-of-the-art OCRs on a sample ten-year table image file.

During the process of transfer learning, our model became definitively more skilled at transcribing the ten-year tables. One can simply compare the loss over time. The points at the start reflect the baseline from the Écriture Anglaise model. As it began to learn from our custom dataset, the loss significantly dropped.

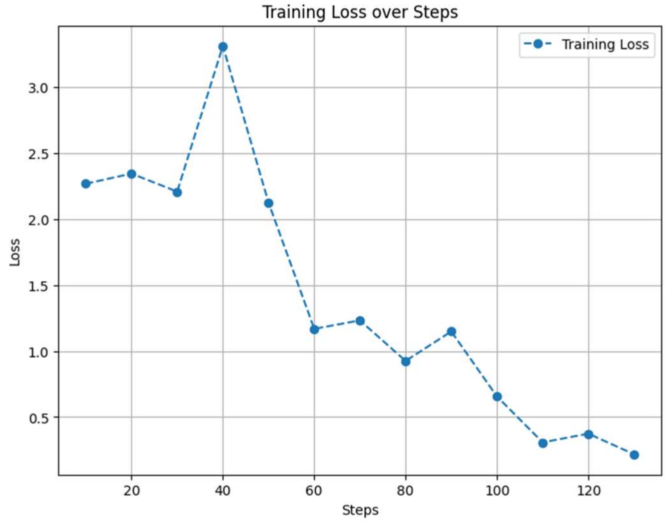


Fig. 6. AHRT Training Loss over Steps for AHRT Transfer Learning.

This steady decline, with marked lack of overfitting, not only serves to demonstrate that our training was effective, but it also shows that a larger dataset should continue to improve accuracy along the same trendline. This same trend can be seen in our evaluation loss.

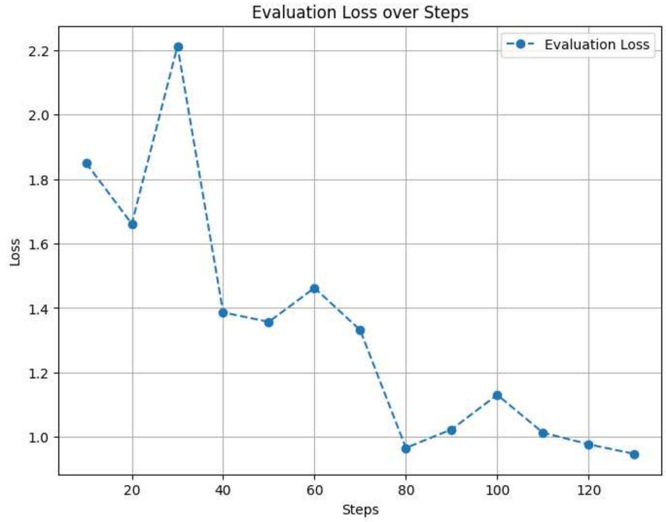


Fig. 7. AHRT Evaluation Loss over Steps for AHRT Transfer Learning.

Interestingly, the evaluation has a spike early in the training because of a difficulty in overwriting no-longer essential knowledge. This model must forget some of the paragraph reading skills and wider vocabulary in favor of a more specialized set that are applicable to Francophonic names and dates. This still retains a few notable words including the titles of vital events as these make infrequent appearances in the data. Thus, after the new vocabulary and formatting have been learned, there is a significant increase in the quality of the guesses. Similarly to the training loss, this trend appears to be cut off rather than bottom out on its own, so more data would be highly advantageous to the model's overall success. This should be promising to archives and other beneficiaries.

To evaluate the contemporary success of our model, and to verify its applicability to the current state of Francophonic historical index transcription, we analyzed our character error rate (CER) against our baseline model [5]. As a supplemental comparison, we compared against a common OCR that is deployed on many devices: Google Lens.

TABLE I. OCR MODEL COMPARISONS

OCR Model	**Transcription Success Metrics**		
	Literal Transcription	***CER***	***% Acc.***
Hand Transcription	Line 8: Bignalet Paul Emile 18 juin 1870 Line 9: Bernès Vignolle Pierre 12 août 1870 Line 10: Bo*r Ie**** ****** *Malfunction with Splitter* Line 11: B******** **** Madeline 20 8bre 1872 *Malfunction in Splitter* Line 12: ********** 9bre 1863 *Malfunction in Splitter* Line 13: Courrèges Dominique 17 9bre	N/A	100
AHRT	Line 8: Bignalet Paule 18 juin 1890 Line 9: Pournès Pierre Pierre 12 aout 1850 Line 10: Bournon Pierre 7 juin 1851 Line 11: Dommèelle Marie Maréline 20 8bre 1872 Line 12: Sémes Louise 17 9bre 1862 Line 13: Courcqes Dominique 17 9bre	18	13.14
Baseline Model (Pre Transfer Learning)	Line 8: Bignalet Paul Emile 18 Sieur 1870 Line 9: Devoirès Pierrelle Pierre 12 Août 1850 Line 10: Baison Seine J Sieur 1871 Line 11: Pourcès Pierrelle Marie Madeleine 20 8bre 1872 Line 12: Saines-Maire 17 Pbre 1863 Line 13: Courrèges Dominique F d°	30	22.4
Google Lens OCR	Line 8: Bignalet Paul Emile 18 Jung 1870 Line 9: Rezerét. Vimolle 12 Hour 18/0 Line 10: Baie Lin 1871 Line 11: 20 Jbre 1872	46	34.33

	Line 12: 17 bre 1863 Line 13: Courreges Dominique =		

In these results, one can see the clear success of our model. Even in the areas where transcription was not entirely accurate, one could easily deduce the reasoning behind this mistake. That is to say, circular letters might be mistaken for other circular letters, and letters with ascenders could be interpreted as other ones with similar forms. This was a notable improvement over other such models. Additionally, as this is a word spotting model, some of these errors could get phased out with more training on the common French names. For now, to see that the characters are being properly distinguished is a very good sign, and error rates as low as we found here prove this project a success.

In this upload stage, due to various factors, like skew or other image artifacts, there were some slight troubles with our image segmentation software. These were notated in the original transcription above, and calculated in the final scores. However, this was a systemic problem that applied to all models tested as the same image inputs were given to each. In spite of this, our word spotting methodology proved highly effective. Even in cases where the segmenter malfunctioned and output half of a name, or only a few characters, the model still had good odds of getting it right if there were only a few letterforms visible.

Out of numerous anecdotal examples, one specific, quantifiable improvement regarded the '5' character. As the Écriture Anglaise 5 is completely different from any recognizable modern 5, written or typed, this character proved to be a repeated stumbling block for many OCRs. However, as there was a good number of examples of this letterform in our custom training data, we found that AHRT had no troubles in distinguishing this.

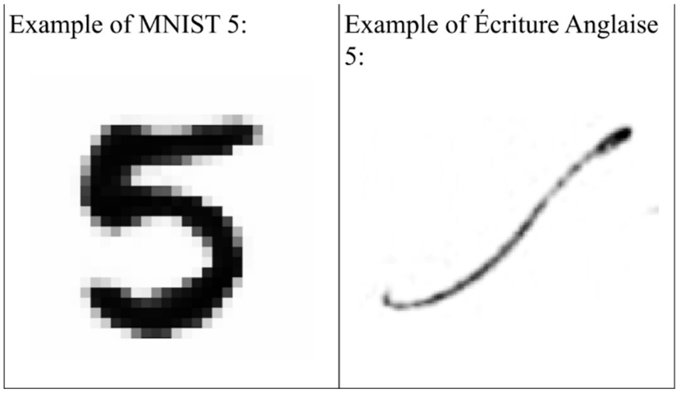


Fig. 8. Comparison of the '5'character in MNIST script and Écriture Anglaise.

This sort of observation is a good sign in the general quality of our transfer learning and model overall as it not only demonstrates the needed specialization for accurate transcription, but it also shows that our dataset can overcome regional differences and quirks.

With some minor changes and expansions, we could have a model that is deployable for archives in Francophone countries. The ones that come first to mind are:

- Expanding our dataset to encompass more regions and names that will commonly appear.
- Adjusting the line segmenter settings to prevent cut offs that affect the OCR as a whole.

Overall, our novel response is an improvement on the original model, and provides the best current implementation of an OCR with historical capabilities such as this. Working with archives or other departmental agencies would continue to revolutionize this software, and would surely create a model that is widely applicable and accurate.

VI. Conclusion

In the early stages of this project, our work was focused on laying the groundwork for creating an image segmentation algorithm that would function regardless of diverse formats. This part proved highly successful, and our results demonstrate a segmenter that splits tabular indices, regardless of the template used. This development makes our model better suited for creating searchable databases. As specified earlier, this allows for a direct pipeline from inputs of scanned half-folios into a CSV or other organized list. Additionally, the processes that we went through to create this functionality taught us a great deal about image modification and sterilization. We learned about all kinds of filters and methods for identifying subjects of interest before ultimately settling on our tool.

After this, we gained significant experience regarding dataset creation. While this was less than ideal, we successfully implemented a system to organize this data and easily import it. This developed a system to label and structure training data for efficient ingestion into the model This phase underscored the critical role of high-quality data. Creating a novel dataset allowed us to appreciate other available information with greater context. This was not easy, although it was essential to the project's success as a whole, and it showed the amount of effort that must go into dataset creation for machine learning models to work well.

In addition, we've deepened our understanding of OCR mechanics to navigate the nuances of various handwriting formats, font weights, and line structures that differ in archival records. This study has guided our approach to a modular design that allows for customization to regional and dialectal variations in names and dates. Our plan necessitated using transfer learning to tailor the OCR to specific languages, regions, and formats. Regarding this transfer learning, we came to understand that it allows for great adaptability of general models. In this project, we continued to build off of a chain of transfer learned models, which continually got more specified for ten-year table indexing. This started with a general handwriting model, was adapted to the French language, then to French text, and finally to the ten-year tables.

Alongside this, we attempted to implement this OCR model on the hardware board that should eventually support the transcription interface. While it was a relatively straightforward part of the setup, resolving the issues with direct image upload was essential because the final tool needed to be accessible to municipal archivists. This integration has been an area of particular focus, ensuring archivists can work with the tool efficiently and without technical complexity. However, it did not become fully realized by the end of the project due to a systemic drop in accuracy to zero. Although the final interface faced challenges with accuracy during deployment, this process underscored the importance of preparation and model quantization for future implementations.

While this project successfully developed a robust OCR model specialized for transcribing ten-year tables, several challenges emerged that highlight areas for improvement. A primary limitation was the dataset size and diversity; despite regional representation, the dataset remains relatively small, reducing the model's capacity to generalize across less-represented handwriting styles and orthographic variations. Expanding the dataset to include more records from underrepresented regions or periods would further enhance the model's adaptability and accuracy. Additionally, the edge device implementation, though conceptually promising, faced technical hurdles during deployment. The quantization process significantly degraded model performance, resulting in a systemic drop in accuracy to zero. Future efforts should prioritize optimizing the quantization pipeline and exploring alternative hardware solutions better suited to handling high-complexity OCR tasks. Lastly, while the model performed well in transcription, it occasionally struggled with ambiguous or degraded entries, which points to the need for advanced post-processing techniques, such as incorporating context-aware language models to correct errors and normalize outputs. Addressing these challenges would not only improve transcription accuracy but also broaden the model's applicability to other archival datasets and languages.

In summary, this project represents a significant advancement in automating the transcription of archival records through a modular and highly adaptable OCR model. From the successful development of a robust image segmentation algorithm to the creation of a specialized dataset and the implementation of transfer learning techniques, each stage brought new insights and challenges. While the hardware integration remains a work in progress, with a notable failure in maintaining accuracy during initial implementation, the lessons learned have positioned us to refine the system for practical use by municipal archivists. By focusing on indexing rather than full record transcription, this project has demonstrated a unique and efficient approach to creating searchable databases, laying the groundwork for further innovation in historical document analysis.

References

[1] Arvanitopoulos, N., Chevassus, G., Maggetti, D., & Süsstrunk, S. (2017). A Handwritten French Dataset for Word Spotting: CFRAMUZ. Proceedings of the 4th International Workshop on Historical Document Imaging and Processing., pp. 25-30. https://doi.org/10.1145/3151509.3151523

[2] Brée, S., Merveille, F., & Paquet, T. (2020). Rapport Scientifique Du Projet Popp: Projet d'océrisation des recensements parisiens. Projet Lauréat CollEx-Persée 2019-2020. https://popp.hypotheses.org/files/2022/10/RAPPORT-SCIENTIFIQUE-DU-PROJET-POPP_diffusable.pdf

[3] Constum, T. et al. (2022). Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20th Century Paris Census. In: Uchida, S., Barney, E., Eglin, V. (eds) Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science, vol 13237. Springer, Cham. https://doi.org/10.1007/978-3-031-06555-2_10

[4] Gleeson, D. T., Shiels, D., Hsieh, W., Harvey, M., & Funk, A. (2024, August 19). Civil War Bluejackets: Citizen science, machine learning, and the US Navy common sailor. Muster. https://musterhistory.com/civil-war-bluejackets-article

[5] Gombert, A., & Beigelman, M. (2023). TrOCR in French: Adapt to French archives [Machine learning model]. Hugging Face. https://huggingface.co/agomberto/trocr-large-handwritten-fr