





Belfort Birth Records Digitization: Preprocessing, and Structured Data Generation

Wissam AlKendi¹^a, Franck Gechter^{1,4}^b, Laurent Heyberger²^c and Christophe Guyeux³^d

¹CIAD, UMR 7533, UTBM, F-90010 Belfort, France

²FEMTO-ST Institute/RECITS, UMR 6174 CNRS, UTBM, F-90010 Belfort, France

³FEMTO-ST Institute/DISC, UMR 6174 CNRS, Université de Franche-Comté, F-90016 Belfort, France

⁴LORIA, UMR 7503, SIMBIOT Team, F-54506 Vandoeuvre-lès-Nancy, France

{wissam.al-kendi, franck.gechter, laurent.heyberger}@utbm.fr, christophe.guyeux@univ-fcomte.fr

Keywords: Belfort civil registers, segmentation, handwritten text recognition, preprocessing, text skew.

Abstract: Historical documents are invaluable windows into the past. They play a critical role in shaping our perception of the world and its rich tapestry of stories. This paper presents techniques to facilitate the digitization and transcription of the French Belfort Civil Registers of Births, which are valuable historical resources spanning from 1807 to 1919. The methodology focuses on preprocessing steps such as binarization, skew correction, and text line segmentation, tailored to address the challenges posed by these documents. They contain various text styles, marginal annotations, and a hybrid mix of printed and handwritten text. The paper also introduces this archive as a new database by developing a structured strategy for the components of the documents using XML tags, ensuring accurate formatting and alignment of transcriptions with image components at both the paragraph and text line levels for further enhancements to handwritten text recognition models. The results of the preprocessing phase show an accuracy rate of 96%, facilitating the preservation and study of this rich cultural heritage.

1 INTRODUCTION


Handwritten text recognition (HTR) has recently emerged as one of the most significant areas of the pattern recognition field. Numerous researchers have developed methods to transcribe a variety of documents, such as historical archives (Philips and Tabrizi, 2020), books, letters, and general forms using spatial (offline) (Wang et al., 2021) or temporal (online) (Gan et al., 2020) processes. Transcription is the process of automatically translating handwritten text within a digital image into a machine text representation, often resulting in a more accessible format. This process includes applying various approaches, such as preprocessing methods, where the image is prepared for analysis (e.g., noise reduction, normalization), as well as deep learning techniques for recognizing handwriting text. Finally, post-processing methods are used to fine-tune the output, correct errors, and improve


readability (Chacko and P., 2010). Figure 1 illustrates the essential preprocessing tasks, crucial for preparing the image for analysis before employing deep learning techniques for text recognition.


Nowadays, systems can analyze document layouts (Diem et al., 2011) and recognize text at several levels, including characters, text lines, paragraphs, and entire documents. Furthermore, these systems can distinguish distinct handwriting styles written in a variety of languages (Carbune et al., 2020).


Despite significant advancements in recognizing modern handwritten text, there are still many challenges that need to be addressed in transcribing historical documents. These documents often feature unique characteristics such as angular and spiky letters, ornate flourishes, and overlapping words and text lines in various handwriting styles (Bugeja et al., 2020). Additionally, the reproduction quality of these documents significantly increases the time required for the preprocessing phase (Nikolaidou et al., 2022).

The historical significance of Belfort's records stems from the city's distinctive demographic history in the nineteenth century. Following France's defeat to Prussia in 1871 and Germany's absorption of

^a <https://orcid.org/0000-0003-4239-9964>

^b <https://orcid.org/0000-0002-1172-603X>

^c <https://orcid.org/0000-0002-3434-5395>

^d <https://orcid.org/0000-0003-0195-4378>

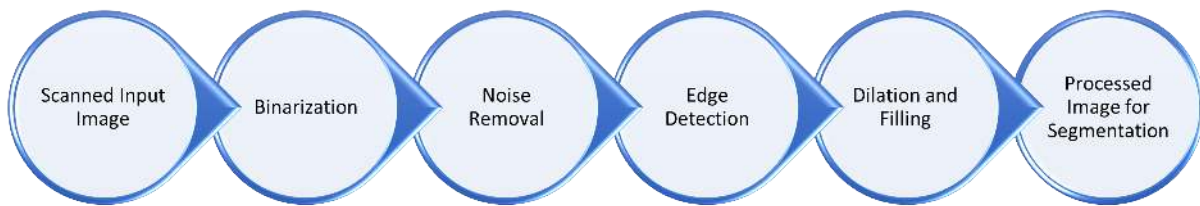


Figure 1: Preprocessing components: the essential preprocessing steps involved in the transcription of handwritten text within digital images.

Alsace-Lorraine, Belfort witnessed a rapid population increase. This growth was primarily driven by the entry of Alsatians, particularly those from Mulhouse, who decided to remain in French territory or relocated to Belfort to work in the expanding textile and mechanical industries established thereafter 1871 (Delsalle, 2009). This metropolis consequently serves as a unique observatory for three fundamental changes: urban expansion, migration, and the sexualization of social relations. Figure 2 shows a visual representation of a sample page from these civil registers of birth.

The study of birth records provides us with a wealth of historical insights on a wide range of topics. These include the duties and practices of midwives, changes in cohabitation and the age of parents when they have their first child, instances of out-of-wedlock births, child recognition patterns, and the selection of witnesses for official declarations. These records also revealed information about child naming conventions and the prevalence of home versus hospital births. Each of these components offers a distinct perspective on the societal and cultural dynamics of the time. To successfully examine and address these historical concerns, it is critical to create a comprehensive knowledge database that will allow for their study and resolution. This procedure entails transcribing archive contents using two methods. The first method is human-powered transcribing. However, this method is costly and time-consuming. The second is automatic transcribing, which employs computer science techniques.

This paper aims to address the preprocessing phase challenges associated with transcribing the Belfort birth records, with a focus on improving data quality, segmenting documents, and developing structured data representations to improve the preservation and accessibility of historical birth records. Handwritten text recognition of such documents poses a significant challenge, particularly in the context of mixed handwritten and printed text. Despite significant advancements in Optical Character Recognition (OCR) techniques in recent years, these techniques have not adequately addressed the challenges

associated with transcribing these documents. The distinctive characteristics of these records necessitate a unique approach for accurate transcription. Hence, there is a strong need to develop new methods to achieve satisfactory results in transcribing these records. This entails exploring novel strategies and methodologies to overcome the impediments and ensure the accuracy and efficacy of the transcription process.

The remainder of the paper is organized as follows: Section 2 presents a summary of recent achievements in the field. In Section 3, we detail the characteristics and challenges associated with the Belfort Civil Registers of Births. Section 4 is devoted to providing a summary of the experimental results. Finally, in Section 5, the conclusion is drawn with suggestions for future research directions.

2 A STATE-OF-THE-ART OVERVIEW

Several studies have been published to improve the image processing field and address the challenges presented by the preprocessing phase, including binarization/thresholding, noise removal, and contrast enhancement methods. This phase is frequently employed in many important areas, as demonstrated in (Hussain et al., 2023; Al-Khalidi et al., 2019)).

In a study by (Philips and Tabrizi, 2020), the authors surveyed this primary phase in the historical document transcription process, which includes steps such as binarization, skew correction, and segmentation processes. Their findings emphasized the vital importance of transcription accuracy as a requirement for meaningful information retrieval in archival texts. Furthermore, in (Binmakhshen and Mahmoud, 2019), authors conducted a rigorous examination of several document layout analysis (DLA) techniques to detect and annotate the physical structure of documents. The investigation focused on the many stages of DLA algorithms, such as preprocessing, layout analysis approaches, postprocessing, and performance evaluation. This research lays the groundwork

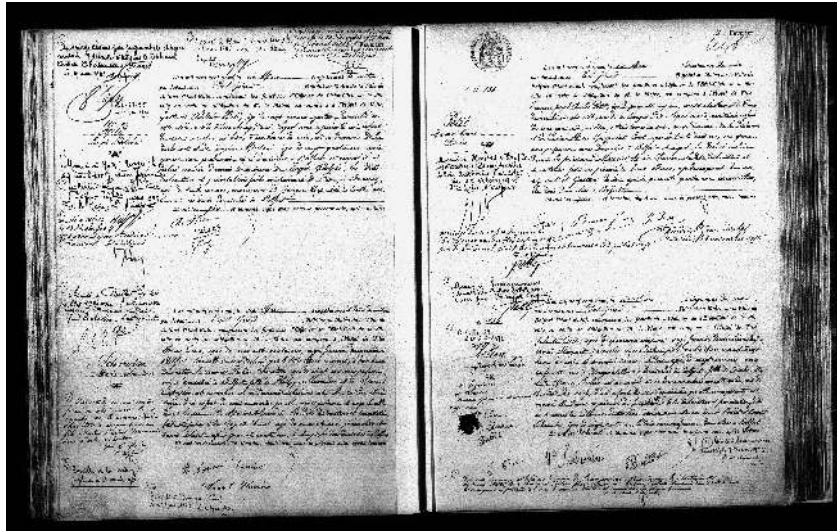


Figure 2: Sample of Belfort Civil Registers of Births

for the development of a universal algorithm capable of handling all types of document layouts.

Several key steps of the preprocessing phase are outlined below.

2.1 Binarization

This step entails transforming digital photos into binary images consisting of two collections of pixels in black and white (0 and 1) (Mustafa and Kader, 2018). Binarization is useful for dividing an image into foreground text and background.

(Tensmeyer and Martinez, 2020) focused on historical document image binarization, establishing a standard benchmark that played a critical role in driving research on the subject. The work covers a variety of additional approaches and techniques, including statistical models, pixel categorization with learning algorithms, and parameter tuning. Additionally, it discusses evaluation metrics and datasets.

2.2 Noise Removal

This step involves removing unnecessary pixels from the digital image that could impact the original information. The noise might arise from the image sensor and electronic components of a scanner or digital camera.

(Ganchimeg, 2015) presented a hybrid binarization approach for improving the quality of historical and ancient documents. They combined global and local thresholding techniques while keeping minimal computational and temporal costs. Furthermore, the study demonstrates that identifying the characteristics

of noise patterns is critical for choosing appropriate methods for their removal. Similarly, (Chakraborty and Blumenstein, 2016) presented a survey on removing marginal noise from historical handwritten document images, such as Australian Archives, which poses unique layout complexities. The survey is intended to assist researchers in determining the best methods based on the type of marginal noise in their datasets, as well as to highlight the limitations of existing approaches, paving the way for the development of more general and robust solutions.

2.3 Edges Detection

This step entails recognizing the edges or boundaries of the text within the image by connecting all continuous points that have the same color or intensity, utilizing techniques such as Laplacian, Sobel, Canny, and Prewitt edge detection (Ahmed, 2018).

2.4 Skew Detection and Correction

Skew describes the misalignment of text within a digital image. In other words, it specifies the number of rotations required to align the text horizontally or vertically. To achieve this goal, several skew detection and correction approaches have been proposed, including Hough transformations and clustering.

They are detailed in (Biswas et al., 2023), in which the authors discussed various methods for detecting skew angles in the text images, including Hough Transform, Projection Profile-based methods (PP), Nearest Neighbor clustering (NN), and Cross Correlation.

2.5 Segmentation

This stage involves breaking the handwritten text image into letters, words, lines, and paragraphs, frequently using the image’s pixel properties. The segmentation process has been carried out using a variety of methodologies, including threshold methods, edge-based methods, region-based methods, watershed-based methods, and clustering methods. The segmentation stage is regarded as one of the most important procedures that can considerably increase the accuracy of HTR models.

(Singh et al., 2022) describe different classical and learning-based line segmentation approaches for Handwritten Text Recognition (HTR) in historical manuscripts. The study suggests that connected components and graph-based techniques are beneficial solutions for situations such as overlapping lines and touching characters in historical manuscripts. Additionally, in (Pach and Bilski, 2016), the authors applied a binarization algorithm based on Gaussian filtering to address the non-uniform luminance within selected pages from the Latin Theologica Miscellanea documents. Additionally, they utilized a Hough Transform Mapping technique to better handle the diverse conditions of handwritten manuscripts, including cases with overlapping characters. Furthermore, (Plateau-Holleville et al., 2021) employed a modular analytic pipeline that utilizes cutting-edge image processing and machine learning algorithms to automate data extraction from the proposed French vital records. This pipeline includes document preprocessing and text segmentation, utilizing histograms and the EAST method to locate text boxes, with the objective of aiding in the manual annotation of the training dataset. The results reveal that the EAST approach is effective for identifying text boxes in such documents.

On the other hand, (Saabni et al., 2014) computed an Energy Map for both binary and grayscale images to minimize the non-text areas by extracting connected components along the text lines. Additionally, they employed a seam carving technique to find seams (paths of least energy) in the computed energy map, which aids in identifying the text lines. The study also introduced a new benchmark dataset comprising historical documents in various languages. Similarly, (Louloudis et al., 2009; Papavassiliou et al., 2010; Alaei et al., 2011) proposed valuable methodologies on text line segmentation. While (Chakraborty and Blumenstein, 2016) acknowledges that the literature on page segmentation for historical handwritten documents is limited compared to non-historical documents, suggesting a

research gap in the field.

3 METHODOLOGY

3.1 Belfort Civil Registers of Births

The birth records in the Belfort commune’s civil registries include 39,627 records written in French, each scanned at 300 dpi. These records were chosen for their consistency, as they include Gregorian birth dates beginning in 1807 and are accessible until 1919 due to legal limits. Originally, these registers were handwritten entries but were changed to a partially printed version with certain areas left vacant for recording precise data about the newborn’s statement. The timing of the changeover to the hybrid preprinted/manuscript format varied between communes.

In Belfort, the shift happened in 1885, affecting around 57.5% of the 39,627 documented declarations. The archive can be accessed online up to 1902 using the following link: <https://archives.belfort.fr/search/form/e5a0c07e-9607-42b0-9772-f19d7bfa180e> (accessed on January 20, 2024). Furthermore, we have acquired permission from the municipal archives to examine data up to 1919.

3.2 Records Structure

These records provide crucial information such as the child’s name, parents’ names, and witnesses, among other details. Figure 3 depicts a sample record from civil registers. Table 1 outlines the structure and content of a record within the archive.

3.3 Automatic Transcription Challenges

The transcription of records in the Belfort archives presents a variety of challenges, as outlined below.

3.3.1 Document Layout

The Belfort birth documents use two unique document layouts. The first type has double pages, each containing a single complete record, whereas the second type has double pages accommodating two complete records per page. Nonetheless, certain pages may contain entries that start on one page and continue to the next.

3.3.2 Reading Order

It is critical to understand the order in which text areas are to be read, including both the primary text and any

Table 1: The structure and contents of a record in the Belfort Civil Registers of Births as presented in (AlKendi et al., 2024).

Structure	Content
Head margin	Registration number. First and last name of the person born.
Main text	Time and date of declaration. Surname, first name and position of the official registering. Surname, first name, age, profession and address of declarant. Sex of the newborn. Time and date of birth. First and last name of the father (if different of the declarant). Surname, first name, status (married or other), profession (sometimes) and address (sometimes) of the mother. Surnames of the newborn. Surnames, first names, ages, professions and addresses (city) of the 2 witnesses. Mention of absence of signature or illiteracy of the declarant (very rarely).
Margins (annotations)	Mention of official recognition of paternity/maternity (by father or/and mother): surname, name of the declarant, date of recognition (by marriage or declaration). Mention of marriage: date of marriage, wedding location, surname and name of spouse. Mention of divorce: date of divorce, divorce location. Mention of death: date and place of death, date of the declaration of death.

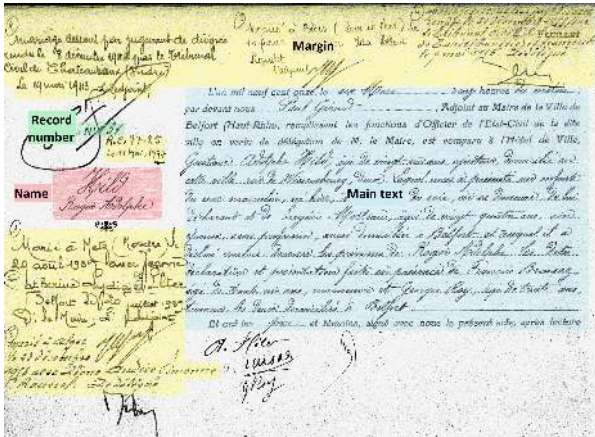


Figure 3: A record from the Belfort Civil Registers of Birth with annotations indicating different sections. The main body of the text, which contains detailed information, is highlighted in blue. Annotations in the margins are highlighted in yellow. The record number, a unique identifier for the record, is highlighted in green. The record name is highlighted in red. Both the record number and name represent the header margin of the record.

marginal annotations. The record should be read in the following order: the record number first, followed by the name, the main content, and any marginal notes last. This arrangement requires reading from left to right and top to bottom.

3.3.3 Hybrid Format

Some registers incorporate entries with both printed and handwritten texts, as depicted in Figure 2. These handwritten portions might vary greatly in style, legibility, and ink density, necessitating a smooth transition between OCR and handwriting recognition modes during transcription. Furthermore, the spatial arrangement of printed and handwritten text can be complex, with handwritten comments appearing in margins, between printed lines, or even superimposed on the printed content.

3.3.4 Marginal Annotations

These mentions are about the individual’s birth but inserted thereafter, often in different handwriting styles and using different scriptural instruments. They are also positioned differently in comparison to the main text of the declaration.

3.3.5 Text Styles

The registers show a variety of handwritten techniques, including angular and spiky letters, varying character sizes, and intricate flourishes, resulting in instances where words and text lines overlap within the script.

3.3.6 Skewness

Skewness is the misalignment of handwritten text caused by the way it was written by writers. Many lines within the main paragraphs and margins show text skew anomalies, including vertical text (rotated 90 degrees). Efficient methods are required for adjusting image skewness, regardless of the degree of rotation.

3.3.7 Deterioration

The images show text deterioration caused by fading ink and page smearing, including ink spots and yellowing of the pages.

3.4 Preprocessing Methods

Several methods and filters have been applied to the images to remove noise and enhance the visibility of the text for effective edge detection and automatic extraction. The suggested methodology segments text images into two levels: text paragraphs and lines. To achieve this, a manual document layout analysis methodology is employed to identify the major components of the records (record number, name, primary content, margins) on the pages. The process involves determining the coordinates of these major components using a custom interface tool, while the coordinates of the text lines within them are determined automatically. The key steps of the preprocessing phase are outlined below. Figure 4 shows the general pipeline of methods applied to the images.

- **Grayscale Conversion:** The original images are converted to grayscale, which simplifies subsequent processing and lowers computer complexity.
- **Gaussian Blur:** A Gaussian blur (GB) is applied to the grayscale images to minimize noise and improve feature recognition. Equation 1 represents the Gaussian blur operation applied to the grayscale images.

$$GB(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) * \text{gray}(x,y). \quad (1)$$

where σ_x and σ_y represent the kernel sizes, and σ is the standard deviation.

- **Adaptive Thresholding:** Adaptive thresholding is utilized to generate a binary image that highlights the edges and features of interest. Equation 2 depicts the adaptive thresholding process performed,

yielding binary images. The Gaussian approach is used with a fixed block size to generate the adaptive threshold value $T(x,y)$ based on the image's local characteristics, eliminating any overlapping between lines if it exists.

$$\text{thresh}(x,y) = \begin{cases} 255, & \text{if } GB(x,y) \leq T(x,y) \\ 0, & \text{if } GB(x,y) > T(x,y) \end{cases}$$

$$T(x,y) = \text{mean}(x,y) - 2 \times \text{stddev}(x,y). \quad (2)$$

where $\text{mean}(x,y)$ is the mean pixel value of the local neighborhood centered at (x,y) , and $\text{stddev}(x,y)$ is the standard deviation of pixel values in the same local neighborhood.

- **Morphological Operations:** Dilation operations are conducted to refine the detected edges by expanding the boundaries of the text, making bright regions brighter and dark regions darker. These processes enhance and manipulate the structure of the edges observed in the preceding stage, increasing their clarity and significance. Equation 3 represents the dilation operation applied to the images using a horizontal kernel of size 1×205 . This operation is employed to identify the core of the text line. It computes the maximum value among neighboring pixels along the horizontal direction.

$$\text{Image dilation}(x,y) = \max_{i=0}^{204} (\text{thresh}(x,y+i)). \quad (3)$$

where (x,y) is the position of the pixel within the images, and i is the horizontal offset within the kernel.

- **Contour Detection:** This technique identifies and extracts contours from the filtered images. The parameters have been configured to obtain only the external contours or outlines of the text lines. It also simplifies and compresses contours by excluding unnecessary points, hence saving memory. This process returns a list of contours as well as a hierarchical representation of those contours. These contours are then examined and filtered using thresholds that represent the width and height of the text lines in the paragraphs.
- **Skew Correction:** A skew correction process is applied to the text line images in various steps, including the utilization of filters, the Canny edge detection algorithm, and the Hough Line Transform. First, we used the Numpy function 'np.arctan2' to determine the slopes of the

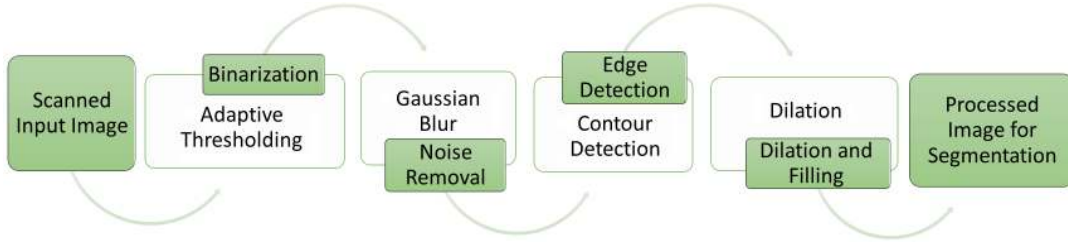


Figure 4: Preprocessing components: the essential preprocessing steps involved in the transcription of process of Belfort Civil Registers of Births.

detected text lines. These slopes provide information about the lines' orientation, allowing us to identify the main horizontal line. We next computed the median angle using the acquired slopes and converted it to degrees. Secondly, we determined the center of the text line image and generated a rotation matrix using `cv2.getRotationMatrix2D` as depicted in Equation 6. This matrix is crucial for the subsequent rotation operation. Additionally, we identified the most frequent color in the text line image's border region to fill empty spaces in the background after skew correction. This is done by counting the occurrences of each unique integer value in the pixel values and selecting the maximum count to effectively find the most common color. Finally, this process applies an affine transformation to the text line image using the calculated matrix M . The flag parameters (`cv2.INTER_CUBIC` and `cv2.BORDER_CONSTANT`) are used to specify the interpolation method for cubic interpolation and the border value parameter for a smooth rotation of the image around its center. This results in a more accurate representation of text line images.

$$\text{Angle} = \text{median}(\arctan 2(\Delta y, \Delta x)) \times \frac{180}{\pi} \quad (4)$$

where Δx and Δy represent differences in the y and x coordinates of line endpoints, respectively.

$$\text{Center} = \left(\frac{N_x}{2}, \frac{N_y}{2} \right) \quad (5)$$

where N_x and N_y are the dimensions of the image.

$$M = \text{cv2.getRotationMatrix2D}(\text{Center}, \text{Angle}, 1). \quad (6)$$

where 1 indicates that there is no scaling applied in the transformation.

Figure 5 illustrates example of the skew correction step applied to text line image.

The parameter and threshold settings are automatically adjusted based on the characteristics of the input images. Figure 6 shows examples of the preprocessing phase.

3.5 Structured Data Generation

Based on our previous research (AlKendi et al., 2024), there is a strong need for a new deep learning technique specifically designed for transcribing the French Belfort Civil Registers of Births, given the numerous challenges it presents. This phase involves manually transcribing a portion of the archive pages to create a training dataset for the deep learning model. Additionally, it entails the development of an accurate methodology for linking the identified image regains to the transcribed text. This will significantly improve the labeling process for future development attempts.

3.5.1 Manual Transcription

The records were manually transcribed through two different procedures. First, our co-authors and their students graciously contributed their effort free of charge. Second, there were costs associated with some circumstances. The transcribed records feature a diverse range of writing styles and span various historical periods. During the transcription process, several tags were employed to ensure the proper structuring and identification of the record components. Table 2 presents the various types of tags utilized in the manual transcription process. So far, a total of 207 .txt files representing 637 records have been successfully transcribed. Each .txt file may consist of 4 or fewer records. Additionally, the files contain 709 margin texts. The overall number of text lines in all the record components is 13,913, and the total number of words is 120,131 words and 745,582 characters. Figure 7 provides an example of transcriptions with tags.

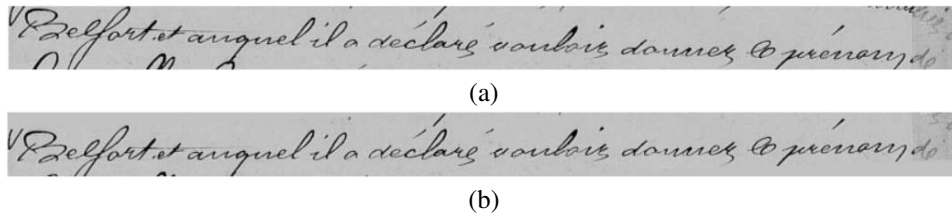


Figure 5: Examples of the skew correction steps applied to a text line image from Belfort Civil Registers of Births: (a) Original text line image. (b) Result After skew correction.

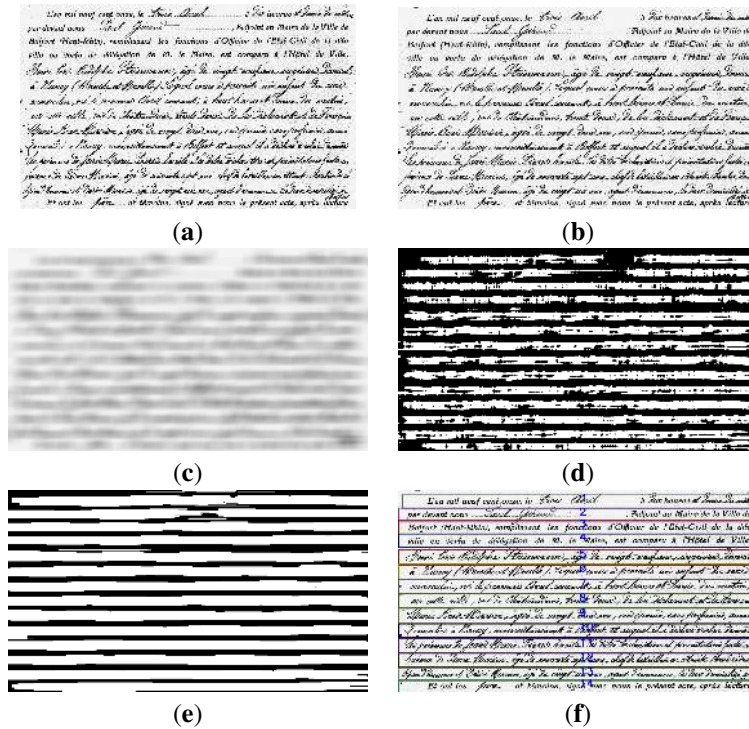


Figure 6: Examples of the preprocessing steps applied to a record from Belfort Civil Registers of Births: (a) Original record. (b) After grayscale conversion. (c) With Gaussian blur. (d) After adaptive thresholding. (e) Following morphological operations. (f) Result after contour detection for text line extraction.

3.5.2 XML File Generator

An XML file generator tool has been developed to prepare the transcribed records for input into the deep learning model. This is achieved by linking each component within the images to its corresponding transcription at both the paragraph and text line levels based on the added tags. Each .xml file contains information for double pages, encompassing four records. The XML files have been designed to include the following properties:

- **Image Information:** This section includes details such as the image name, type (single or double page), image height, and image width.
- **Reading Order:** As illustrated in Section 3.3.2, this section identifies the reading order and pro-

vides a unique index number for each record, specifying the type of components within the reading order (title number, title name, paragraph, margin).

- **Record Title and Name:** This part of the XML file contains information about the record's title and name, including their coordinates within the image and the corresponding transcribed text.
- **Paragraphs and Margins:** This section contains details about the paragraphs and margins within the record. It includes their coordinates within the image and the corresponding transcribed text. Additionally, information about the text lines within the paragraphs and margins is provided, including a unique ID, coordinates within the image, and the corresponding transcribed text.

Table 2: The set of tags utilized in the manual transcription process.

Tag	Description
<code><begin></code>	Begin of the record.
<code><text>...</text></code>	Main text of the record.
<code><margin>...</margin></code>	Margins text.
<code><ptext>...</ptext></code>	Printed text.
<code><striped>...</striped></code>	Striped text.
<code><unreadable>...</unreadable></code>	Unreadable text.
<code><added above>...</added above></code>	Small text added above the text line.
<code><added below>...</added below></code>	Small text added below the text line.
<code><page></code>	Start new page.

```

<begin>
N° 199
Steinmann.
Jean Marie Pierre Emile
<text>
<striped>L'an mil neuf cent onze, le <striped>trois Avril<ptext>, à<ptext> dix <ptext>heures <ptext>et demie du matin<ptext>,
par devant nous <ptext> Paul Giroud<ptext>, Adjoint au Maire de la Ville de
Belfort (Haut-Rhin), remplissant les fonctions d'Officier de l'Etat civil de la dite
ville en vertu de délégation de M. le Maire, est comparu à l'hôtel de ville,<ptext>
Henri Léon Rudolphe Steinmann, âgé de vingt neuf ans, ingénieur, domicilié
à Nancy (Meurthe et Moselle). lequel nous a présenté un enfant du sexe
masculin, né le premier Avril courant, à huit heures et demie du matin,
en cette ville, rue de Châteaudun, trente deux, de lui déclarant et de Françoise
Marie Rose Marion, âgée de vingt deux ans, son épouse, sans profession, aussi
domiciliée à Nancy, momentanément à Belfort et auquel il a déclaré vouloir donner
les prénoms de Jean Marie Pierre Emile. Les dites déclaration et présentation faites en
présence de Pierre Marion, âgé de soixante sept ans, chef de bataillon en retraite, Chevalier de la
légion d'honneur et Désiré Marion, âgé de vingt six ans, agent d'assurances, les deux domiciliés à <added below>Belfort</added below>.
<ptext>Et ont les<ptext> père <ptext>et témoins, signé avec nous le présent acte, après lecture<ptext>
<striped>
<margin>
Décédé à PETRA (Jordanie)
le 8 avril 1963.
Le 20 janvier 1965
</margin>

```

Figure 7: Examples of the tags utilized in the manual transcription process of Belfort Civil Registers of Births.

4 RESULTS

4.1 Text Line Segmentation

A two-phase automatic segmentation process is designed to segment the components of an image into text lines. The first phase involves applying a segmentation tool to the main paragraphs and margins, achieving an accuracy of 92% for the main paragraphs and 77% for the margins respectively. However, in certain cases where text lines heavily overlap, such as in cursive writing styles, the tool incorrectly detects multiple lines as a single text line. To address this issue, a second phase has been introduced to refine the results of the first phase. This refinement examines the height of segmented lines and, if it exceeds a predetermined threshold, applies a secondary phase to split the lines horizontally, increasing the accuracy to achieve 96% for the main paragraphs and 84% for the margins respectively.

To assess the accuracy of our tool, we employ the Detection Rate as an evaluation metric. This metric quantifies the tool's ability to accurately identify and segment individual lines and compares the number of detected lines with the number of lines in the ground truth. Additionally, manual visual verification is conducted to ensure the quality of results.

The Detection Rate is defined as follows:

$$\text{Line Detection Accuracy (LDA)} = \frac{TP}{GT} \quad (7)$$

where TP represents True Positives, indicating the correctly detected lines by our tool, and GT represents the Total Number of Lines in the Ground Truth dataset.

To determine optimal parameters for our specific dataset, automatic determination has been performed based on the total number of text lines in the transcription. Table 3 presents the final parameters values that enhance the tool's effectiveness across different writing styles. Moreover, Table 4 shows the parameters values used in the text lines skew correction step.

Additionally, a 5-pixel padding has been added above and below the height of segmented text lines when establishing their coordinates. This padding improves the visibility and accuracy of segmentation, particularly for various writing styles.

Figure 8 illustrates the accuracy achieved through the text line segmentation process applied to the Belfort Civil Registers of Births.

An accuracy comparison was conducted to evaluate the effectiveness of the suggested segmentation tool compared to online artificial intelligence commercial systems such as DOCSUMO, Ocelus, and Transkribus, which offer free document recognition

Table 3: Parameters values used in the text lines segmentation step.

Step	Parameter	Value
Gaussian Blur	Kernel Size	(101, 51)
Gaussian Blur	SigmaX (Standard Deviation in X)	61
Adaptive Threshold	Maximum Value	255
Adaptive Threshold	Adaptive Method	ADAPTIVE_THRESH_GAUSSIAN_C
Adaptive Threshold	Threshold Type	THRESH_BINARY_INV
Adaptive Threshold	Block Size	71
Adaptive Threshold	Constant from Mean	2
Dilation	Kernel Size	(1, 205)
Dilation	Iterations	1

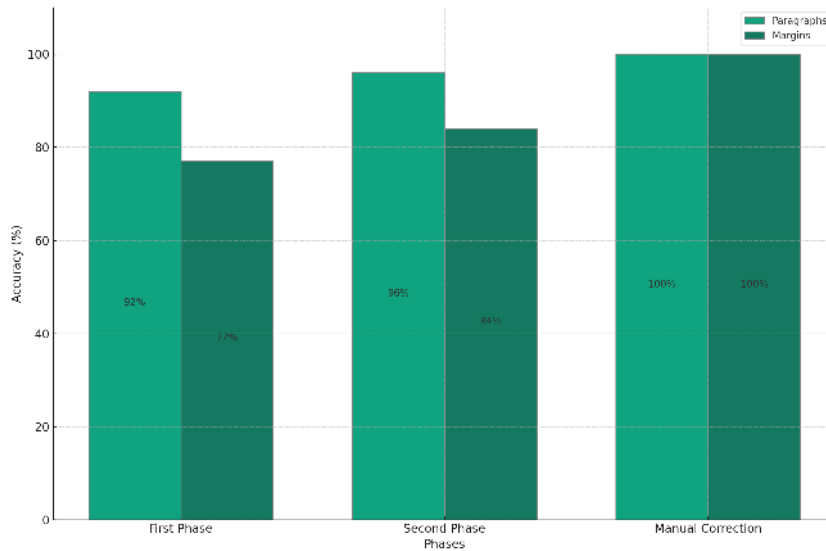


Figure 8: Accuracy percentages of text line segmentation process across the various phases.

Table 4: Parameters values used in the text lines skew correction step.

Parameter	Value
Gaussian Blur Kernel Size	(5, 5)
Canny Edge Threshold1	50
Canny Edge Threshold2	150
HoughLinesP Threshold	100
HoughLinesP MinLineLength	100
HoughLinesP MaxLineGap	55

trials. The experiment was conducted on the paragraph image displayed in Figure 3 in terms of text line segmentation. Table 5 depicts the accuracy rates of the aforementioned systems.

4.2 Results of XML File Generation

An automatic verification tool has been developed to verify and correct the formatting of our manual transcriptions. This tool ensures that the tags are em-

Table 5: Accuracy comparison of Belfort Civil Registers of Birth segmentation process with commercial systems. The links of these systems were accessed on February 3, 2024.

System	Accuracy (%)
DOCSUMO (URL)	97
Ocelus (URL)	99
Transkribus (URL)	97
Our method	96

ployed correctly within the transcriptions, maintaining accurate alignment between the image components and the corresponding transcriptions when we generate the .xml files.

We have processed .txt files that encompass complete transcriptions, resulting in the creation of 90 .xml files. Each .xml file contains complete transcripts (four records), totaling 360 records. An example of one such XML file is presented in Figure 9. It is important to note that our manual transcription process is still ongoing as we work towards completing

the remaining paragraphs within the .txt files. This ongoing effort will enable us to generate additional .xml files in the future.

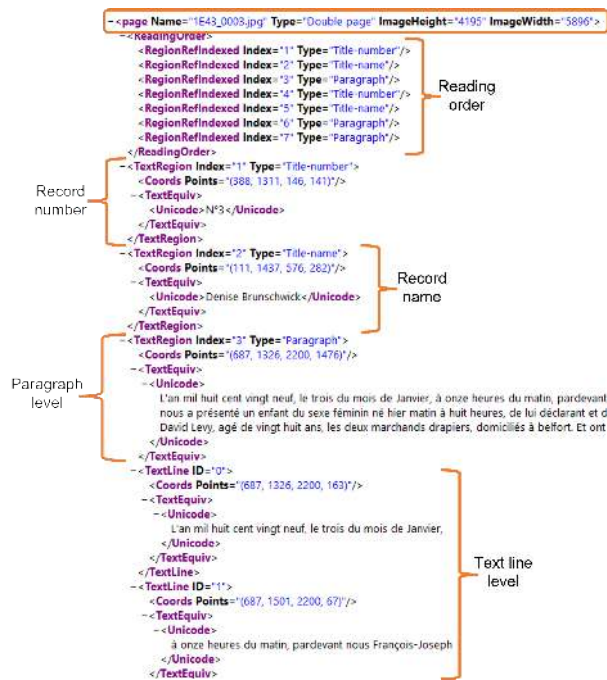


Figure 9: Examples of the Structured Data Files for Belfort Civil Registers of Births

5 CONCLUSIONS

This paper has presented a comprehensive approach to digitizing and transcribing the Belfort Civil Registers of Births. The methodology involves a variety of preprocessing steps, including binarization, skew correction, and segmentation. Despite the numerous impediments posed by these historical documents, such as different text styles, marginal annotations, and the hybrid nature of the texts (printed and handwritten text), we have developed distinctive solutions that successfully carried out the segmentation process. The creation of an automatic verification tool and the XML file generator ensures that transcriptions are properly formatted and aligned with their corresponding image components. Our results show a high level of accuracy in text line segmentation, which is critical for the effective structuring of these valuable historical documents

In conclusion, the work presented in this paper makes a significant contribution to the field of handwritten text recognition, particularly in the context of historical documents, by introducing a new valu-

able structured dataset. However, employing artificial intelligence techniques in the preprocessing phase could further refine accuracy, especially in the segmentation process. Future research will focus on developing an automatic document layout analysis tool and a deep learning model to transcribe the remaining records automatically, with the ultimate goal of facilitating the recognition and study of this rich cultural heritage.

REFERENCES

- Ahmed, A. S. (2018). Comparative study among sobel, pre-witt and canny edge detection operators used in image processing. *J. Theor. Appl. Inf. Technol.*, 96(19):6517–6525.
- Al-Khalidi, F. Q., Alkindy, B., and Abbas, T. (2019). Extract the breast cancer in mammogram images. *Technology*, 10(02):96–105.
- Alaei, A., Pal, U., and Nagabhushan, P. (2011). A new scheme for unconstrained handwritten text-line segmentation. *Pattern Recognition*, 44(4):917–928.
- AlKendi, W., Gechter, F., Heyberger, L., and Guyeux, C. (2024). Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging*, 10(1):18.
- Binmakhashen, G. M. and Mahmoud, S. A. (2019). Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.
- Biswas, B., Bhattacharya, U., and Chaudhuri, B. B. (2023). Document image skew detection and correction: A survey.
- Bugeja, M., Dingli, A., and Seychell, D. (2020). An overview of handwritten character recognition systems for historical documents. *Rediscovering Heritage Through Technology: A Collection of Innovative Research Case Studies That Are Reworking The Way We Experience Heritage*, pages 3–23.
- Carbune, V., Gonnet, P., Deselaers, T., Rowley, H. A., Daryin, A., Calvo, M., Wang, L.-L., Keyser, D., Feuz, S., and Gervais, P. (2020). Fast multi-language lstm-based online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(2):89–102.
- Chacko, B. P. and P., B. A. (2010). Pre and post processing approaches in edge detection for character recognition. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 676–681.
- Chakraborty, A. and Blumenstein, M. (2016). Marginal noise reduction in historical handwritten documents—a survey. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 323–328. IEEE.
- Delsalle, P. (2009). *Family History: Parish and Civil Status Registers, from the Middle Ages to the Present Day: Demography and Genealogy*. Presses universitaires de Franche-Comté, Besançon.

- Diem, M., Kleber, F., and Sablatnig, R. (2011). Text classification and document layout analysis of paper fragments. In *2011 International Conference on Document Analysis and Recognition*, pages 854–858.
- Gan, J., Wang, W., and Lu, K. (2020). In-air handwritten chinese text recognition with temporal convolutional recurrent network. *Pattern Recognition*, 97:107025.
- Ganchimeg, G. (2015). History document image background noise and removal methods. *International Journal of Knowledge Content Development & Technology*, 5(2):11–24.
- Hussain, S. A.-K., Al-Nayyef, H., Al Kindy, B., and Qassir, S. A. (2023). Human earprint detection based on ant colony algorithm. *International Journal of Intelligent Systems and Applications in Engineering*, 11(2):513–517.
- Louloudis, G., Gatos, B., Pratikakis, I., and Halatsis, C. (2009). Text line and word segmentation of handwritten documents. *Pattern recognition*, 42(12):3169–3183.
- Mustafa, W. A. and Kader, M. M. M. A. (2018). Binarization of document images: A comprehensive review. In *Journal of Physics: Conference Series*, volume 1019, page 012023. IOP Publishing.
- Nikolaidou, K., Seuret, M., Mokayed, H., and Liwicki, M. (2022). A survey of historical document image datasets. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(4):305–338.
- Pach, J. L. and Bilski, P. (2016). A robust binarization and text line detection in historical handwritten documents analysis. *International Journal of Computing*, 15(3):154–161.
- Papavassiliou, V., Stafylakis, T., Katsouros, V., and Carayannis, G. (2010). Handwritten document image segmentation into text lines and words. *Pattern recognition*, 43(1):369–377.
- Philips, J. and Tabrizi, N. (2020). Historical document processing: Historical document processing: A survey of techniques, tools, and trends. *ArXiv*, abs/2002.06300.
- Plateau-Holleville, C., Bonnot, E., Gechter, F., and Heyberger, L. (2021). French vital records data gathering and analysis through image processing and machine learning algorithms. *Journal of Data Mining and Digital Humanities*, 2021.
- Saabni, R., Asi, A., and El-Sana, J. (2014). Text line extraction for historical document images. *Pattern Recognition Letters*, 35:23–33.
- Singh, H., Kaur, N., and Kaur, H. (2022). Analysis of line segmentation methods for historical manuscripts. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 2043–2045. IEEE.
- Tensmeyer, C. and Martinez, T. (2020). Historical document image binarization: A review. *SN Computer Science*, 1(3):173.
- Wang, Y., Xiao, W., and Li, S. (2021). Offline handwritten text recognition using deep learning: A review. *Journal of Physics: Conference Series*, 1848(1):012015.