Recognizing the Structure and Content of Hungarian Civil Registers

Kata Ágnes Szűcs **Noémi Vadász** **Zsolt Béla Záros**
National Archives of Hungary
Budapest
szucs.kata.agnes@mnl.gov.hu zaros.zsolt@mnl.gov.hu
vadasz.noemi@mnl.gov.hu

Abstract

The study[1] evaluates key steps in a system for processing data from digitized Hungarian state register records (1895-1980) into an SQL database. It examines how template selection and post-processing impact data accessibility and integration. The research details the compiled datasets, annotation processes, and evaluation functions used to measure processing quality, emphasizing template selection and post-processing to improve the overall workflow and the accuracy of the published data. An evaluation method for publishing structured data provides a model for similar projects.

1 Introduction

In 1895, Hungary replaced ecclesiastical procedures with a unified civil registry. The project processes registers (birth, marriage, death) from 1895 to 1980,[2] transferring data from digitized images to a SQL database using automated and machine learning tools. In the pilot project, we have chosen a single municipality (Abony in Pest County) for testing before implementing the long-term plans to reach national coverage. The paper outlines evaluation methods developed in conjunction with the current workflow and provides a conceptual workflow overview.

Abony produced 25 birth, 18 marriage, and 19 death register volumes. These used two layouts: form-based and columnar. The former averaged 600 pages/volume; the latter 300. Consequently, for a single municipality, more than 30,000 scanned pages of data are processed and organized into a database through an automated workflow that integrates various open-source tools.

The process starts with image scanning, then layout classification via templates. Text is segmented, and handwriting is recognized. Handwritten text recognition (HTR) output is structured into a database, followed by post-processing: correction, normalization, standardization which are essential to make sense of the data. Geographic names are mapped to namespaces, enabling entity linking within the database. The final phase is structured data publication to an online platform.

This study focuses on two major, but not consecutive workflow steps: template selection and post-processing (PPR). We explore how their evaluation can refine early-stage results.

2 State of the Art

Recent advancements in handwritten text recognition and post-processing reflect a shift toward hybrid systems combining neural architectures with rule-based or knowledge-driven corrections. These address challenges in historical documents, such as spelling variation, degradation, and domain-specific patterns. Transformer models like TrOCR (Li et al., 2021) and ViT (Dosovitskiy et al.) yield strong results, yet template matching remains vital for semi-structured data like civil registers due to its interpretability and accuracy (Hashemi et al., 2016; Brunelli, 2009).

Such hybrid methods also mitigate the scarcity of annotated data in low-resource domains (AlKendi et al., 2024; Peng et al.). In Hungary, huBERT (Nemeskey, 2021) supports historical name segmentation. Remaining issues include inconsistent formatting and orthography. Legal and archival systems are adapting to digitization (Hohol), while models like HTR-JAND (Hamdan et al., 2024) apply attention and distillation to boost scalability and precision.

[1] The paper was founded by the National Research Development and Innovation Office” (NRDI) under the project 149512 MEC-R-24.

[2] Decree-Law No 17 of 1982 formally abolished the civil registration districts, as the specialized apparatus had already been merged into the administrative organization of the municipal council. (Hohol)

Proceedings of Recent Advances in Natural Language Processing,pages 1215–1223
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-140

3 Templates

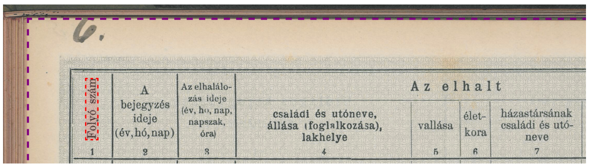
In the fields of digital humanities, computer vision techniques have been increasingly applied to analyze and interpret large collections of visual and textual data. For example, in art history, computational approaches have transformed how we interpret images, enabling analysis of visual styles across large datasets. (Lang and Ommer, 2021). These methods also help improve the accessibility of digitized cultural archives by allowing new ways to explore and interpret visual content. (Jaillant, 2024).

Template matching is a fundamental technique in computer vision that enables machines to identify specific patterns within images by comparing them to a predefined reference template (Hashemi et al., 2016).

In this study, the term ”template” can be interpreted in two distinct contexts: one refers to the computational method that locates reference images based on similarity scores, while – in the context of the Template Editor[3] – the other describes a standardized layout model used to interpret register pages.

3.1 Template Matching

Unlike machine learning-based approaches in computer vision,[4] template matching relies on mathematical similarity measures to locate objects or patterns in visual data making it useful in scenarios requiring precise and rule-based pattern recognition,[5] even though more advanced AI-driven techniques are now available. (Brunelli, 2009).



6.

Folyó szám	A bejegyzés ideje (év,hó,nap)	Az elhalálozás ideje (év, hó, nap, napszak, óra)	Az elhalt			
			családi és utóneve, állása (foglalkozása), lakhelye	vallása	életkora	házastársának családi és utóneve
1	2	3	4	5	6	7

Figure 1: Reference image (in red) on a template image [detail]

In the project, reference images were manually selected from digitized records in the municipality of Abony, based on features like layout, structure, and fonts (see Figure 1). This allowed us to classify record types across birth, marriage, and death registers into consistent categories. While the registers are standardized at a national level, differences in design may appear over time or between municipalities.

At the current stage of the project, 42 template types have been identified manually to classify the registers from Abony, covering both form-based and tabular layouts (see Table 1). However, the total number of templates may evolve as classification continues, since the diversity in form and structure means a fixed number cannot yet be determined.[6]

Nr. of templates	Layout	Register Type
6	form-based	birth
5	columnar	birth
6	form-based	marriage
11	columnar	marriage
8	form-based	death
6	columnar	death

Table 1: Identified templates

The template matching procedure in the workflow is currently defined by three key parameters: similarity score, pixel distance, and Kraken binarization[7] with an adaptive threshold.

In the beginning, the process employs a similarity score threshold of 0.66, which was determined based on preliminary experiments and has demonstrated good results. Setting this threshold to a lower value would raise the likelihood of confusing the templates based on the reference image, while increasing the threshold would impose overly strict criteria, potentially leading to the exclusion of templates that belong to the appropriate category. Additionally, the allowed pixel distance between the reference box and its detected location in the examined image is set at 500 pixels, which was established to accommodate variations in image alignment and potential distortions, thereby enhancing the robustness of the matching process.

The search terminates upon the first successful match, and the results are subsequently categorized into ”hit” and ”missed” folders for further analysis. In case of no initial match, a second pass attempts

[3] An in-house developed web application.

[4] https://opencv.org/about/

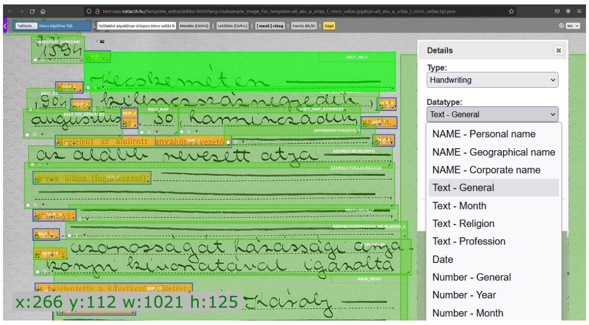
[5] e.g. Object detection, optical character recognition, medical imaging.

[6] While we have outlined the identified templates in Table 1, the absence of standard evaluation metrics reflects the exploratory nature of this research phase. As we progress, we will refine our methodology and update our findings, ensuring that the classification process remains robust and adaptable to the complexities of the data, thereby addressing potential concerns regarding the standardization of template quantities in future evaluations.

[7] https://kraken.io/

broader detection, followed by a merging step to update classifications. If some templates still remain unrecognized, a final round of matching is performed using a lower threshold and adaptive binarization techniques with Kraken. This increases processing time but improves accuracy, especially with unclear or faded documents.

3.2 Template Editor



Details

Type: Handwriting

Datatype: Text - General

Figure 2: The interface of Template Builder

The Template Editor serves two key purposes: it stores the reference image used for classification and captures metadata through a structured annotation interface. The reference image act as visual marker defining specific register subtypes, while the metadata is collected by drawing bounding boxes around fields of interest and assigning attributes via drop-down menus. (see Figure 2).

Given the structured format of the data, the expected content of the captured fields is largely predictable. This predictability enables the template creation process to systematically extract metadata from the handwritten (variable) and pre-printed (static) content. Templates not only structure the visual layout but also help standardize data extraction, ensuring compatibility with the underlying database model.

By leveraging the inherent structure of the source material, template preparation optimizes automation, enhancing efficiency while reducing manual effort. Additionally, this function of the Template Editor identifies and structures individual fields in preparation for database integration, ensuring compatibility with predefined data attributes.

The database aims to provide users with a searchable platform for genealogical and historical research. It also stores data from every processing phase in a structured manner, providing the possibility of identifying and linking identical persons (entities) in the registers, thus enabling the possibility of indirectly building family trees. To support this, the database architecture had to manage both different roles individuals might hold (e.g. registrar, parent, deceased) and the various data fields associated with them (e.g. names, places, dates, and occupations).

3.3 Evaluation and Results

The evaluation of template matching is based on a curated test set, originally built from 250 files and later expanded. All files were automatically assigned to a template and manually reviewed, annotated, and corrected to ensure accurate benchmarking. The result was a dataset of 448 files representing all but two known template types that were subsequently identified misassignments. Since the automated template assignment process did not always yield examples of each template type, in cases where no instances were found, manual search and supplementation were not feasible, therefore, no additional sampling was conducted to supplement the dataset with examples from these template types.

The results were evaluated as a labeling task, we calculated precision, recall and F-1 score for each template type. The averaged results are as follows in Table 2:

Precision	Recall	F-1 Score
0.9093	0.8991	0.8932

Table 2: The average template matching results

On average, the system achieved high accuracy, with 85% of the templates are classified correctly (36 templates), meaning that the F-1 score is at least 0.8. In 9.5% of cases this value is below 0.8% (4 templates) indicating a worse performance of the template selection. In two other instances, no result was measured at all, as no such template was detected in the database. These under-performing columnar templates will require further refinement in future work.

The average performance shows that most of the automated templates work well, some exceptions could be improved. In terms of the efficiency of the overall processing workflow, it is important to achieve the best possible result, as further errors may be caused by a template matching error in the segmentation and PPR. Templates with low precision and recall will be reviewed, closely examined, and fine-tuned by adjusting the parameters.

The template assignment method has demonstrated sufficient robustness and effectiveness to justify its continued use as a foundation for further research and experimentation. This decision allowed us to optimize resources by focusing on workflow development rather than further refining template matching at that stage. Its performance has been deemed adequate to support the next stages of the project, with ongoing evaluation and refinement expected to yield further improvements.

With the pipeline now complete and resources allocated, we have begun experimenting with improvements. The existing template matching process is computationally time-intensive, and evaluation has revealed potential for precision enhancement, particularly in tabular registers. We are currently working on a rule-based error correction and have also started to experiment with the introduction of transformer-based solutions such as the ViT model.[8] Initial experiments have yielded promising results, demonstrating that better performance can be achieved with minimal training data.

To evaluate the VIT model, we selected three highly similar tabular templates from marriage registers. This selection was based on the confusion matrix generated during the assessment of previous results, which indicated a frequent misclassifications between these templates. Within the municipality of Abony, a total of 472 images were manually classified for the training set and 143 as the benchmark corpus. For details see Table 3 below. In the case of the *szul_ketsor* template, the benchmark was determined to be larger, as many of the images were selected sequentially from only two folders, thereby contributing limited value to the experiment. The collection of images for the training set was conducted manually, but given the large size of the corpus and the limited time available, it was practically infeasible to manually curate the required number of images.

Template name	Train Set	Bench	Error
all_akv_b_tablazat	152	50	0
all_akv_b_szul_ketsor	160	158	0
all_akv_b_szul_ketsor_2	160	35	0

Table 3: Initial VIT results

The training lasted 100 epochs and the results obtained are notably promising, with no errors identified across the benchmark dataset. However, these findings should not be regarded as conclusive. Further training sets evaluations are necessary, including the development of larger training and benchmark datasets, as well as testing on additional templates to comprehensively assess the efficacy of the procedure.

4 Post-processing and Normalization

After the template selection several steps are undertaken in the workflow prior to the implementation of the PPR. These steps include line segmentation, HTR and loading of data into the database. Consequently, a text corpus is automatically generated and systematically organized within a SQL database according to the previously established structures. The output of handwriting recognition requires post-processing to correct errors from the HTR and normalize data for database searches and entity linking. Such post-processing workflows have been increasingly studied in historical records contexts (AlKendi et al., 2024).

It is crucial to distinguish between *correction*, which aims to accurately reflect the content of the digitized document (including typos, misspellings and hyphens), and *normalization*, which seeks to link identical individuals in the database through standardized data that may have different spellings. Consequently, post-processing involves verifying the content of civil registry fields from previous processing steps, correcting HTR model errors, and normalizing specific data fields. Thus, PPR serves three primary purposes: 1) to restore the text as visible in the original document, akin to a diplomatic edition with the highest degree of accuracy; 2) to support the identification and linking of individuals in the database; and 3) to facilitate database searches.

As for some fields PPR procedures are still under development that will not be further discussed, the next section describes data correction and normalization in more detail.

4.1 Simple Cases

The most straightforward data to correct are those for which a finite set of expected values can be predefined. In the case of death and birth registers between 1895 and 1980, the sex field for a child or deceased individual contains only two possible values: *fiú* (boy) or *leány* (girl) and *férfi* (man) or *nő* (woman) respectively. Since these string pairs dif-

[8] https://huggingface.co/docs/transformers/model_doc/vit

fer in both length and character composition, even in the presence of HTR errors, the original value can be identified with a high degree of certainty.

Similar fields such as marital status and religion, despite including multiple values, remain easily correctable due to their enumerable nature. This is consistent with Hamdan's findings (Hamdan et al., 2024), which highlight the advantages of using simple algorithms and predefined lists for post-processing historical text data. Currently, the database contains 7 categories of marital status and 9 religious affiliations. The list of religions may expand as other municipalities may add newly found religious denominations to our list.

Moreover, data normalization is frequently applied to sex, religion, and family status as well. For instance, religious confessions, such as *r. kath.*, *római katolikus*, *római kathólikus* are standardized to *római katolikus* meaning roman catholic. For gender, all data are normalized to their adult equivalent (e.g., girl → woman) facilitating linking people in the database.

4.2 Correcting and Normalizing Numerals and Dates

Several date-related and numerical correction and normalization techniques are planned for future incorporation, but at the moment, although the possibility exists, they are neither implemented nor evaluated (see Limitations for further details).

For numeric data, PPR checks whether the values contain the expected number of digits and fall within the defined range specified for the type of entry (e.g., day values are at most two digits and range from 1 to 31). In civil records, the dates fall within the project's scope (1895-1980), though birth years may extend earlier. Each date field is assigned a default start and end date – typically January 1, 1895 to December 31, 1980. If a year, month, or day is successfully identified by the HTR, the corresponding value is replaced in the database. When all date fragments are precisely identified, the start and end dates become identical, creating a final date version. If the exact date cannot be identified, only parts of the date, then the date range is narrowed down to the found year, month, or day. When searching the database, this allows us to obtain relevant results even based on dates that may not be fully identified.

4.3 Correcting with Lists

For corrections involving lists, the following algorithm is employed: after preprocessing removes punctuation, whitespace, converts the string to lowercase, and potentially eliminates accents, the algorithm checks for the presence of the data in the list. If the piece of data is not found, a Damerau-Levenshtein distance is calculated between the list elements and the target string, selecting the element with the shortest distance if it falls below a specified threshold. Thresholds may vary by data type; shorter lists (e.g., religion, marital status) and brief strings allow for more permissive distance metrics, whereas longer lists with closely related strings necessitate stricter thresholds.

When we cannot predict the content or length of the list in advance, the post-processing procedure becomes more complex. The fields of *occupation*, *cause of death*, and *first name* belong to this category, among which first names are of particular importance, since the correct breaking down of names is essential from a research point of view, as users will mainly search by surnames and first names, and it is also one of the strongest data for linking people in the database.

For these fields, historical period-specific lists are required. While registrable given names are available, period-specific data and frequency are key. Although reliable data on frequency before the 20th century is limited, the most common given names are available by century[9] and, from the 20th century onward by decade[10]. These sources offer insights into naming trends over time. Our list of female names consists of 293 names, while the list of male names consists of 234 names.

For occupations, a manually curated list of nearly 1,400 occupations was compiled from the registries of Abony and expanded to 1,747 elements to account for broader regional variations. Future expansion may include additional occupations from urban centers, reflecting a wider occupational spectrum. We hypothesize that in large cities we may find additional occupations from *ács* (carpenter) to *zsákfoltozó* (back patcher or sack mender) that occur more frequently.

Although cause of death is less critical for link-

[9] https://hu.wikipedia.org/wiki/Keresztnevek_gyakoris%C3%A1ga_Magyarorsz%C3%A1gon_a_19._sz%C3%A1zadig

[10] https://hu.wikipedia.org/wiki/Keresztnevek_gyakoris%C3%A1ga_Magyarorsz%C3%A1gon_a_20._sz%C3%A1zadban

ing individuals, its standardization can enhance searching in the database. To facilitate their resolution, the Hungarian Society for Family History Research[11] has published a list of the most common causes of death, including their Latin and modern Hungarian equivalent, based on Wikipedia sources. The list of causes of death used in the PPR process contains 818 items.

Both occupation and cause-of-death are characterized by a significant number of variations in terminology. In order to successfully link the personal names that appear in the civil registers, it would be useful to standardize the names of occupations that represent the same but appear in different variations. However, further refinement requires a deeper understanding on historical context to unravel the true meaning of historical occupations. In the future, it might be worth experimenting with other methods, such as semantic classification using word vectors to enhance categorization and interpretation.

4.4 Evaluation and Results

When evaluating the post-processing routines (PPR), it is essential to recognize that their effectiveness depends on the performance of preceding steps. If template classification or segmentation is inaccurate, or if the HTR output is of poor quality, the PPR cannot yield high-quality data.

This study focuses on fields that have been fully developed and evaluated, as the PPR procedures for various database fields are still under development and subject to frequent revisions. A test set of 124 files was created, containing examples of each of the 42 templates mentioned earlier. The PPR must manage the contents of 60 different fields, with solutions currently available for 48 of them. Due to the complexity of preparing and annotating the test data, we can evaluate 36 fields at present. Among these, 17 fields contain two data types (e.g., string and digit, or two dates within a time interval), both of which are assessed by the current procedure. In total, 3,402 data fields required manual annotation. As the separation of composite fields within the PPR has not yet been completed for all field types and the HTR model is still under development, the data in the test set may continue to evolve.

In this sample of 3402 data fields, we manually performed PPR on all data, with manual annotation based on the current hypotheses generated by the HTR model, keeping the rules used in PPR in mind. Scanned images were not utilized during annotation to eliminate the influence of the HTR model's performance on the evaluation of the PPR procedures.

	nr./fields	**TN**	**ACC**	**lowest ACC**
text	1,603/19	143	0.95	0.74
digit	1,309/26	103	0.96	0.67
date	510/3	0	0.76	0.5

Table 4: The first column shows the number of occurrences for a given data type, and next to it the number of fields in which these occurrences were found. The second column (TN) shows the true negative hits, the third (ACC) the accuracy value for all fields and the last (lowest ACC) the worst accuracy value.

During the evaluation, the manually annotated results were compared to the PPR output on a field-by-field basis, yielding separate results for each data type. The accuracy measure reflects the ratio of matches between the manual and automatic solutions. False negatives were recorded when the automatic PPR failed to provide a solution that the manual process did, while false positives were noted when the automatic PPR produced a solution that the manual process did not. In some instances, it was impossible to manually enter the expected result from the PPR due to segmentation errors, which could lead to unexpected values, such as an alphabetic entry in a field designated for digits or a completely empty field where a value is required. In these cases, we do not expect the PPR to return a result. Although these true negative values can be misleading when aggregating accuracies, they were still included in the evaluation. Therefore, it is crucial to consider the true negative ratio when interpreting the results.

The post-processing procedures encompass various tasks, including classifying the HTR output with the highest confidence score according to the assigned data type for each civil registration field and inserting it into the corresponding database field. These data fields may return strings, digits, dates, or constructed data types, such as surname and first name, town name and address. Additionally, some fields require multiple data types simultaneously (e.g., a string and a digit). Consequently, the performance of the PPR functions was evaluated based on data type. Table 4 presents the results for three data types (text, digit, and date).

Based on the results, further work is needed to correct the date fields, and the evaluation should be

[11] https://macse.hu/society/en/kezdolap.php

extended to other fields not yet addressed. Additionally, it would be beneficial to enhance the test sample coverage by incorporating data from other municipalities.

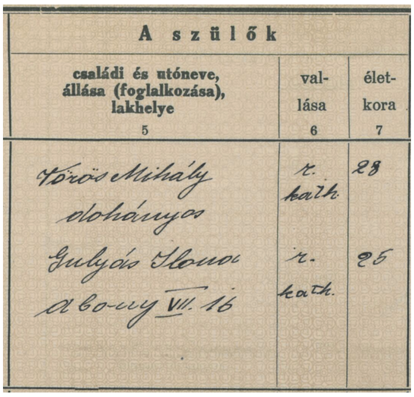
The evaluation of name and address fields, which are critical for entity linking and searchability, has not yet been conducted. Due to their complexity and significance, the necessary correction and normalization procedures are still under development, and their current status will be briefly discussed in the next section of the paper.

4.5 Correcting Structural Data

Post-processing the output of handwriting recognition involves complex procedures for certain data types, particularly composite fields found in historical paper-based records. In such field multiple independent pieces of information are stored in a single cell. In the state register project, composite data is expected for *personal name* and *address*, which have distinct structure. Names are composed of prefix(es), surname(s)[12] and first name(s), and the address is composed of the name of the municipality and other parts of the address, such as street name, house number, address number, etc.[13]

For instance, birth registers contain a cell with the child's first name, sex, religion, and another contains both parents' names, occupations, and residences, all in one box (see Figure 3). This necessitates breaking down the cell's content into individual data elements for separate post-processing to correct potential handwriting recognition errors and normalize the data. The content can be semi-structured, often following the header's order but lacking strict formatting, leading to inconsistencies across cells or registrars[14].

In addressing the place of residence, we separate the fields with additional information (e.g., street name) from the municipality name. This decision is based on the irregular format of additional address components and the municipality name's importance for linking individuals. In post-processing the address, after tokenization we classify the first token as the name of the municipality with remaining tokens as additional address components. However, historical spelling variations complicate this process. While modern Hungarian rules do not allow two-word settlements, older records may present names differently (e.g., Püspök Ladány instead of Püspökladány). Address post-processing requires further refinement, especially in standardizing historical spellings.



A szülők		
családi és utóneve, állása (foglalkozása), lakhelye 5	vallása 6	életkora 7
Török Mihály dohányos	r. kath.	28
Gulyás Ilona Abony VII. 16	r. kath.	25

Figure 3: First column: **Family and given name, occupation and address of the parents**, Second column: **Religion**, Third column: **Age**.

Accurate recording of personal names in civil registry records is crucial for linking individuals across records which is a challenge widely explored in family history record linking (Peng et al.). Identifying and cross-referencing the same person relies on correctly representing family and given names. The first step in post-processing names is segmentation, defining name component boundaries and categorizing them (e.g., family name, given name, maiden name, and descriptors like widow or deceased).

Name segmentation is treated as a tagging task using the Hungarian BERT model (huBERT) (Nemeskey, 2021). For fine-tuning the model, a training dataset was created from unprocessed surnames extracted from Abony, containing 1,000 examples of various name structures, including surnames, given names, and their variants with prefixes indicating maiden names or statuses like widow. Each element is labeled according to the Hungarian equivalents of the following set of labels: widow, family name, maiden name, born as, given name, deceased, allowing the model to classify previously unseen names, facilitating post-processing by distinguishing family names, given names, and complements. The model was trained for 2 epochs.

The training data consists of manually labeled names from birth records, with each label occurring in various combinations. Table 6 presents a labeled

[12] In Hungarian the surname comes first in the order of names.

[13] The official address register has evolved over time, with entries sometimes limited to parcel numbers, while at other times including street names, house numbers, postal codes, and other details.

[14] E.g., the address and the occupation data are often not recorded separately for the mother

example from the training data.

token	label
néhai	néhai (deceased)
Matusanka	vezetéknév (family name)
Jánosné	asszonynév (maiden name)
született	született (born as)
Grász	vezetéknév (family name)
Rozália	keresztnév (given name)

Table 5: A sample from the name segmentation training data.

Currently, name correction focuses on first names using a list-based approach, with plans to integrate additional contextual factors in the future. To improve accuracy, external resources like name lists, dictionaries, and historical frequency data could refine handwriting recognition outputs. Contextual clues from birth certificates, such as the child's sex or the father's name, may also aid validation. Future developments will explore these factors, particularly surname frequency and geographic distribution.

4.5.1 Evaluation and Results

As mentioned before, name segmentation was considered as a tagging task with evaluation metrics including precision, recall and F-measure calculated per tag, and averaged across tag. The performance was assessed on 100 randomly selected data points, produced using the same procedure as the training data for fine-tuning huBERT. To avoid errors from prior steps affecting the results, cases without names (e.g., due to segmentation errors) were excluded from the test data. The results are shown in Table 6.

	nr	Prec.	Rec.	F-1
widow	3	0.667	1	0.8
family name	110	0.953	1	0.976
given name	89	0.955	1	0.977
maiden name	18	0.783	1	0.878
born as	15	1	1	1
avg		0.872	1	0.926

Table 6: Precision, recall and F-measure results for name segmentation and the nr. of the occurrences of each labels in the evaluation set.

The name segmentation performed well, particularly in recall. Future evaluations on a larger, more diverse dataset, including names from other municipalities, would be advantageous. Given that the model was trained for only two epochs without parameter fine-tuning, further optimization could enhance performance. Additionally, expanding the training set may improve the model's generalization. As more municipal data is processed in later project stages, new name components may arise, necessitating further improvements to the segmentation approach.

4.6 Limitations

The name segmentation and address separation approaches are promising but have limitations. Not all post-processing steps for names and addresses, essential for linking civil register records, have been fully developed or evaluated. Current methods rely on list-based corrections that overlook historical variations and formatting inconsistencies. Further refinement is needed, especially as we expand to more municipalities and larger datasets.

For numerical and date fields, we plan to examine volume metadata and entries on the same page, as they follow chronological order. Contextual data can help determine individuals' birth dates, and comparing numerical and textual date representations may enhance accuracy. However, due to time constraints, these improvements could not be implemented at this stage.

5 Conclusions

This ongoing project continues to yield evolving results, marking a sufficient stage to conclude the pilot phase. The workflow is sequential, where each step influences the next, yet its modular design allows for independent analysis and reporting of components.[15]

The study evaluated key steps in the state register workflow, focusing on template matching and post-processing procedures like name segmentation, address, and date normalization. While initial results are promising, further optimization, expanded training sets, and refined methodologies are necessary. Future work should integrate historical context for occupations and names, enhance date normalization, and improve entity linking accuracy, supporting more reliable database integration.

Template matching, while effective, highlighted areas for improvement, especially in terms of accuracy for tabular registers and overall promptness. Inspired by successful outcomes of template match-

[15] Currently, we cannot share any code, but we aim to make it publicly available in the future.

ing strategies outlined by (Lee, 2019), we plan to implement similar hybrid methodologies into our workflow. The study demonstrated the effectiveness of template-based approaches in historical document processing, which align with our ongoing efforts to improve performance in processing of the Hungarian state registers.

As civil register processing continues, the results provide valuable insights. Developing HTR models and refining the PPR method will increase accuracy and facilitate entity linking across civil records, supporting broader data integration. Civil registers will be publicly accessible online per legal guidelines.[16].

References

Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. Advancements and challenges in handwritten text recognition: A comprehensive survey. 10(1):18.

Roberto Brunelli. 2009. *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale.

Mohammed Hamdan, Abderrahmane Rahiche, and Mohamed Cheriet. 2024. HTR-JAND: Handwritten text recognition with joint attention network and knowledge distillation.

Nazanin Sadat Hashemi, Roya Babaie Aghdam, Atieh Sadat Bayat Ghiasi, and Parastoo Fatemi. 2016. Template matching advances and applications in image analysis.

Réka Hohol. Az elektronikus anyakönyvi rendszer előtti és utáni időszak. (2):64–76.

Lise Jaillant. 2024. Introduction to the special issue: Using visual ai applied to digital archives. *Digital Humanities Quarterly*, 18 (2).

Sabine Lang and Björn Ommer. 2021. Transforming information into knowledge: How computational methods reshape art history. *Digital Humanities Quarterly*, 15 (3).

Benjamin Charles Germain Lee. 2019. Machine learning, template matching, and the international tracing service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans. *Digital Scholarship in the Humanities*, 34.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *CoRR*, abs/2109.10282.

Dávid Márk Nemeskey. 2021. Introducing `huBERT`. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, pages 3–14, Szeged.

Xujun Peng, Huaigu Cao, and Krishna Subramanian. Information extraction from historical semi-structured handwritten documents.

[16] Death registers after 30 years, marriage registers after 75 years and birth registers after 100 years. See: `https://net.jogtar.hu/jogszabaly?docid=a1000001.tv`