https://doi.org/10.3311/celisr.40895

From Digitised Manuscript to Database:
Automatic Processing of Civil Registers

Szűcs Kata Ágnes
szucs.kata.agnes@mnl.gov.hu
National Archives of Hungary
Digital Humanist

Vadász Noémi
vadasz.noemi@mnl.gov.hu
National Archives of Hungary
Digital Humanist

Záros Zsolt Béla
zaros.zsolt@mnl.gov.hu
National Archives of Hungary
Senior Developer

Szatucsek Zoltán
szatucsek.zoltan@mnl.gov.hu
National Archives of Hungary
Director

Bánki Zsolt István
bankizsolt@mnl.gov.hu
National Archives of Hungary
Head of Department

Received: 2025.04.14.
Accepted: 2025.08.12.
Published: 2025.09.29.

Cite as: Szűcs, K. A, Vadász, N., Záros, Zs. B., Szatucsek, Z., Bánki, Zs. I. (2025) From Digitised Manuscript to Database: Automatic Processing of Civil Registers, Central European Library and Information Science Review (CELISR), 2(3), p. 309–315.
https://doi.org/10.3311/celisr.40895

This article presents[1] the first milestone in our research on the systematic processing of Hungarian civil registers, encompassing the stages from digitization to data integration into a structured database, culminating in the dissemination of results via an online platform. The initial phase of the project focused on the handwritten birth, marriage, and death certificates of the Abony municipality in Pest County, covering the period from 1895 to 1980. During this pilot initiative we created a comprehensive workflow that facilitated the conversion of digitized images into a structured SQL database, leveraging automated processes and machine learning techniques. The successful completion of this pilot project represents a significant stage, establishing a replicable framework that can be extended to other municipalities, as all components of the workflow are now operational, and can be called sequentially.
handwritten text recognition, HTR, database, civil registry

1. Introduction

1.1 Historical background

The recording of births, marriages, and deaths in Europe was initiated by the Catholic Church following the Council of Trent's decision in 1563. Originally intended for sacramental administration, church registers evolved into essential tools for modern states, which recognized the benefits of comprehensive population records. In Hungary, from 1827, parishes were mandated to submit duplicate copies of their registries to state archives, culminating in state control of the registration process by 1895. Consequently, the population records from the period of 1895 to 1980 contain 37,611 volumes of civil registries from all 3,422 settlements of Hungray.

The digitization of this extensive collection – amounting to 1,908 linear meters or 22 million pages – serves as a vital resource for both amateur genealogists and researchers, offering insights into historical, social, cultural, and economic conditions. Analysing registry data enables various academic disciplines to explore population composition, demographic shifts, social and family structures, migration patterns, cultural practices, and economic activities.

1.2. The project

The project aims at the automated processing of all the civil registry records issued during the period, followed by building the database and providing full-text services. The current paper describes our efforts creating the infrastructure and the completed workflow. We concluded the development phase with the first test runs using the birth, marriage, and death registers from Abony in Pest County, Hungary. The Abony registers total at around 30,000 scanned images of data. The practical testing and fine-tuning of the workflow are still under development.

"The classification of the images includes the structural recognition of their layout."

2. The Workflow

In the case of a single municipality, 25 volumes of birth records, 18 volumes of marriage records and 19 volumes of death records are processed and organized in a database using the automated workflow that was developed for the project (Fig. 1). Before the workflow can start, the digitized images are pre-processed, which prepares them for further processing and publication. The classification of the images includes the structural recognition of their layout. The bounding boxes on the template image file prepare the data to be loaded into the database. After the segmentation of lines, various post-processing, correction and normalization procedures are applied to the output of the handwritten text recognition (HTR). Next, the geographic names are mapped with authority records, and then the name entities in the database are linked together. Finally, the results are published in the database.[2] The following sections will explain the main steps in more detail.

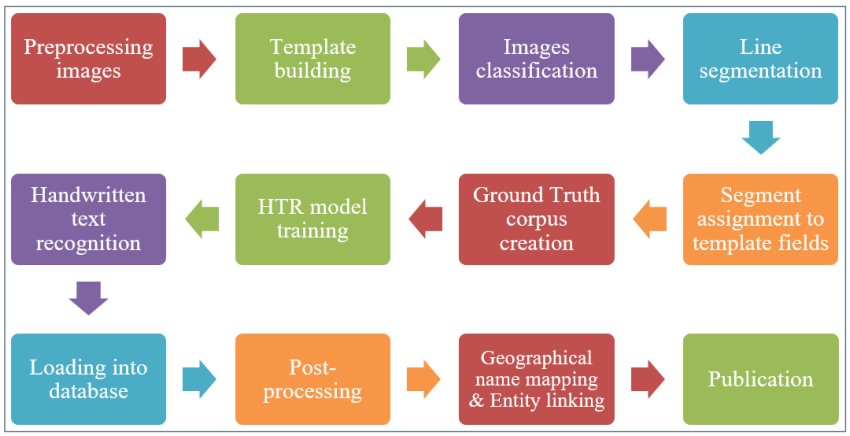

Fig. 1: The Overall Workflow

2.1 Digitized Images and Pre-processing

Considering the scope of the workflow, the digitization process encompasses all civil registries and constitutes an ongoing task, alongside workflow development. To facilitate this process, we employ a combination of manual scanners and three robotic scanner units, which – under expert supervision – significantly accelerate our work. The images are stored in JPEG format, at a scan quality of 300 dpi. This is sufficiently high to enable both machine-based and manual processing, while also helping us meet our storage capacity requirements.[3] Notably, the digitization of the records of Abony presented specific challenges, including the development of a standardized file naming convention, the identification of suitable storage locations, and the definition of an optimal structural framework for the generated images.

https://doi.org/10.3311/celisr.40895

If necessary, the workflow subjects the scanned images to various pre-processing steps: rotation, noise filtering, as well as removing marks, stains and other image defects based on computer vision.[4]

2.2 Template Creation

The Template Builder (Fig. 2) is an in-house developed application that runs in a web environment, through which we categorize the image files by a manually selected reference image – a part of the image containing unique properties. Based on an identical match of the reference image, all images with the same layout will automatically be assigned to a single template during classification. The Civil Registers are either tabular or form-based, according to the legislation in force during the period they had been issued. The document-syntax is usually simple: printed static values in a row or in a header followed by handwritten unique values.[5]

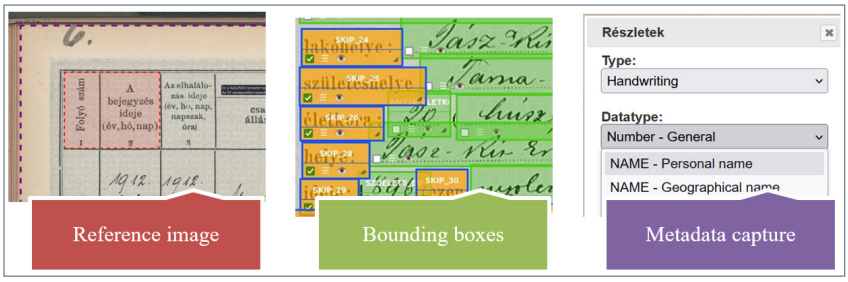

Fig. 2: Template Builder

The other function of the template is to store the recorded metadata from each page and transfer it to the corresponding database entry. Template creation in the web application essentially involves placing bounding boxes on the digitized image, to which additional metadata can be assigned by specifying various attributes. Using the templates, the data can be extracted not just as machine readable text, but also in a structured form (e.g. a certain string must be an equivalent of a mother's name, a religion, a place of birth, etc.). Therefore, it can be linked to a corresponding field in the archival database at the end of the process.

Selecting images upon which templates are created is manual, and assigning the images to templates is an automatic process. On a sample corpus of 250 randomly selected images, the template assignment process is 92.8% accurate. A more challenging task is identifying images that were incorrectly assigned to a template by the script, which occurred in 7.2% of the 250-sample corpus.

At present, we manually identified 42 different civil registry templates for the municipality of Abony from the period under review. This number may increase by including civil records from other communities.

"The Civil Registers are either tabular or form-based, according to the legislation in force during the period they had been issued."

2.3 Handwritten Text Recognition (HTR)

After a template is assigned to an image, the workflow proceeds with the recognition of the handwritten texts. First, the lines visible in the image are segmented into the designed metadata fields on the templates. The segmentation is baseline-driven and performed by the Kraken BLLA module. Kraken is a comprehensive Optical Character Recognition (OCR) system specifically designed to handle historical and non-Latin script materials.[6] Kraken offers fully trainable tools for layout analysis, for determining the reading order of textual elements[7], and for

character recognition, allowing users to customize the system for their specific needs. When we first processed the records from Abony, our workflow utilized the standard BLLA module for line and region detection.[8]

Then, the next step is to produce training data based on the segmented lines that contain the transcribed texts. Through a collaborative effort with volunteers, we transcribed samples of civil records using the Ground Truth Builder interface, generating training data for HTR models. The Ground Truth builder (Fig. 3) is an in-house developed, versatile web-based application that can be used to produce training material for HTR models, to validate and monitor the material produced, and to analyse the texts that have undergone HTR as well.

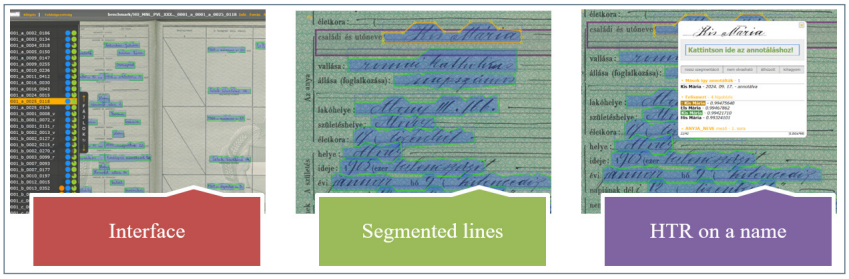

Fig. 3: Ground Truth builder (Annotation Tool)

The training data is used for fine-tuning the TrOCR model[9] in order to produce our own handwriting recognition models. TrOCR is an end-to-end text recognition system that utilizes the Transformer architecture for both image understanding and wordpiece-level text generation, meaning that the model generates text in units, rather than individual characters. (Li et al., 2022)

The model used in the pilot project contained a mixed set of training data: 34,500 segments of civic registry data; 10,000 segments of data from the letters of József Kiss, a 19th century Hungarian poet,[10] and 10,000 segments of additional data, also including writings by several hands. The output of the handwriting recognition is a set of multiple hypotheses per sample with different probabilities (confidence score) assigned to them. The recognised handwriting and the corresponding labels from the templates create the basis of the civil registry database.

During the pilot stage we were developing a benchmark dataset to compare the performance of the models within the project and, in the long run, between projects.[11] The civil registry benchmark corpus consists of 50 instances of each template type mentioned above, giving us a total of 1,991 civil records to evaluate our handwriting recognition models. The benchmark data will be suitable for evaluating the performance of the HTR models and the post-processing and normalisation procedures.

2.4 Loading into Relational Database

The next step is to load the result of the best HTR model into an SQL relational database. The database is managed with SQL Developer and utilizes JSON format for data storage, enabling efficient data management and exchange.[12]

The purpose of the civil registration database is to store the data generated at each processing level in a structured format, and to enable the possibility of identifying and linking all the entries of the same person appearing in different civil registers, thus reconstructing their family ties and indirectly building family trees. Therefore, the database design had to take into account not only the accessibility of each entry at different processing levels but also the need to assign

"The recognised handwriting and the corresponding labels from the templates create the basis of the civil registry database."

 https://doi.org/10.3311/celisr.40895

the individuals in the entries with different roles (e.g. registrar, husband, child, deceased, etc.), and to record their data (e.g. name, place of residence, year of birth/death, occupation, etc.) following a project-appropriate logic.

2.5 Correcting and Normalization

Post-processing of handwriting recognition (HTR) output is essential for correcting model errors and enabling accurate data linkage. We distinguish between the correction of HTR output and the normalization of data.

The structure of numbers and dates in civil records is largely predictable. Given that these records are organized by year, with sequential dates, the birth dates within individual records can be cross verified against the metadata of the volume. This allows for partial verification of the birth date and the age of the recorded individuals.

Civil records, organized by year with sequential dates, allow for partial verification of birth dates and age by comparing them against volume metadata. Corrections are straightforward when fields have a limited set of values. For example, in civil records from 1895 to 1980, the "sex" field is binary – either fiú (boy) or leány (girl) for children, and férfi (man) or nő (woman) for adults –, allowing for accurate corrections even with HTR errors. Similarly, fields such as marital status and religion, which contain 7 and 9 possible values respectively, can be reliably corrected using predefined lists. As the database expands, additional values may be introduced for fields like occupations, cause of death, and first names, which will necessitate more sophisticated correction methods.

Data normalization is applied to specific fields within the data model to standardize and link various entries, despite differences in writing variations. The results of this normalization process are not visible to end-users but ensure accurate search outcomes regardless of the initial HTR output. For instance, gender is normalized to its adult equivalent (e.g., girl is normalized to woman) and religious terms are standardized (e.g., r. kath., római katolikus, and római kathólikus are all normalized to római katolikus to represent Roman Catholic).

Composite data types, such as personal names and addresses, require careful handling due to their distinct structural elements. Names are typically composed of prefixes, surnames, and first names, while addresses include components such as municipality name, street name, and house number. Proper segmentation and entry of these data types are critical for linking records. We fine-tuned the huBERT model (Nemeskey, 2021) for name segmentation by training it on labelled data that identify key elements (such as family name, given name, and maiden name and separate name) for further processing. The fine-tuned model labelled and separated previously unseen names with a high success rate. Its performance was evaluated on 1,000 manually segmented examples, yielding an average F-measure of 0.93, indicating that the fine-tuned huBERT model demonstrated excellent performance in name segmentation, with minor errors, suggesting that the model is effective and requires minimal post-processing corrections.

"Civil records, organized by year with sequential dates, allow for partial verification of birth dates and age by comparing them against volume metadata."

2.6 Personal Record Linkage

Linking records is a longstanding challenge. Since the 1950s, various statistical methods have been used to identify individuals across sources, with machine learning recently surpassing traditional statistical approaches. For civil registers, we draw on prior machine learning experience with data of Hungarian prisoners of

war deported to the Soviet Union.[13] Feature engineering considers both constant and mutable data, using methods ranging from text proximity for names to thesauri-based normalization and geographic distances.

We organise the personal information from registry entries by turning them into individual personal records, which are structured according to each person's role. Despite advances in handwriting recognition and post-processing, data quality for linking personal records remains uncertain. In the late 19th and early 20th centuries, identification relied on familiarity rather than documentation, leading to both inaccuracies in birth details and to record-keeping errors, complicating record linkage.

During the pilot project, we focused on developing a person matching process, with a specific emphasis on generating training data extracted from the workflow for linking individuals across various records. Our objective was to identify and connect multiple references to the same individual (e.g. a person can appear in the role of a mother in birth records and as a bride in marriage record), under a unified identifier in the database. To achieve this, we manually curated a comprehensive training dataset, which will serve as a foundation for further model development and evaluation.

Linking all entries of a person allows for the creation of hypothetical individuals based on record pair linkage confidence, which is never absolute. Users must be able to explore alternative links, and these hypothetical individuals require continuous re-evaluation as the pipeline improves. We then establish relationships between these individuals to create family trees.

"Since the 1950s, various statistical methods have been used to identify individuals across sources, with machine learning recently surpassing traditional statistical approaches."

3. Summary

This project aims to digitize and automate the processing of Hungarian civil records, focusing on birth, marriage, and death registers from 1895 to 1980. The paper introduced the workflow and its main steps starting with records from Abony. We use machine learning and automated processes to convert scanned images into a structured SQL database. A custom-built, modular system for handwriting recognition and template-based data extraction enables the accurate storage of data for future searches and for the linking of entities. Ultimately, this effort will create a detailed, searchable resource for genealogical and historical research, connecting individuals across records to build family trees and to visualize relationships in Hungary's past.

Notes

[1] The paper is based on a conference presentation given at the 18th International Conference on Metadata and Semantics Research in 2024 and reflects that status quo. The participation was founded by the National Research Development and Innovation Office (NRDI) under the project 149512 MEC-R-24.

[2] The public database currently provides volume-level metadata and images, while detailed results remain forthcoming as research is ongoing. The database is available at: https://adatbazisokonline.mnl.gov.hu/adatbazis/allami-anyakonyvek. (Accessed: 2025.08.06.)

[3] For Abony, the 25 649 TIFF files occupy 1.8 terabyte compared to the 429 GB storage space of the JPEG files.

https://doi.org/10.3311/celisr.40895

[4] It is not possible to describe every single step in detail, as this would exceed the scope of this study. Pre-processing images is particularly important in cases where scanned images come from external sources and do not meet our standards.

[5] However, in many cases, we encounter postscript comments written in the margins, ignoring the original structure.

[6] The documentation for Kraken is available at: https://github.com/mittagessen/kraken/blob/main/kraken/blla.mlmodel (2024) and at https://kraken.re/4.3.0/api_docs.html (2024).

[7] Reading order is the sequence in which Kraken arranges lines of text to reflect how a document is meant to be read, particularly in multi-column layouts or those containing non-linear elements such as notes on the margins or the side of the document.

[8] For the source code of the BLLA module available at: https://github.com/mittagessen/kraken/blob/9a218ce8/kraken/blla.py (2025).

[9] The documentation for TrOCR is available at: https://github.com/microsoft/unilm/tree/master/trocr (2024).

[10] The correspondence dataset was provided by the Hungarian National Museum Public Collection Centre (HNMPCC) National Széchényi Library and the Petőfi Literary Museum. For more details: https://dhupla.hu/collection/kiss-jozsef-levelezes (2025).

[11] The HTR benchmark is under development, and its future availability is uncertain. Given its specialized nature, the corpus is likely to be of limited utility for projects outside of similar, registry-related domains but its potential for public release has not yet been determined.

[12] The database size was expanded since the pilot phase of the project. However, for reference, the database schema, including indexes, currently requires a storage space of 7,07 GB. The schema comprises 50 tables, containing a total of 23,254,720 records. Notably, the data presently covers 4 settlements, and the composite fields have not yet been fully processed to produce a final output.

[13] For reference see: https://aol.mnl.gov.hu/gyujtemeny/szovjetunioba-elhurcoltak

Sources/Literature

Li, M. et al. (2022) *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models* (No. arXiv:2109.10282). arXiv. https://doi.org/10.48550/arXiv.2109.10282

Nemeskey, D. M. (2021) *Introducing huBERT*. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. p. 3–14. ISBN 978-963-306-781-9. Available at: https://acta.bibl.u-szeged.hu/73353/ (Accessed: 2025.08.06.)