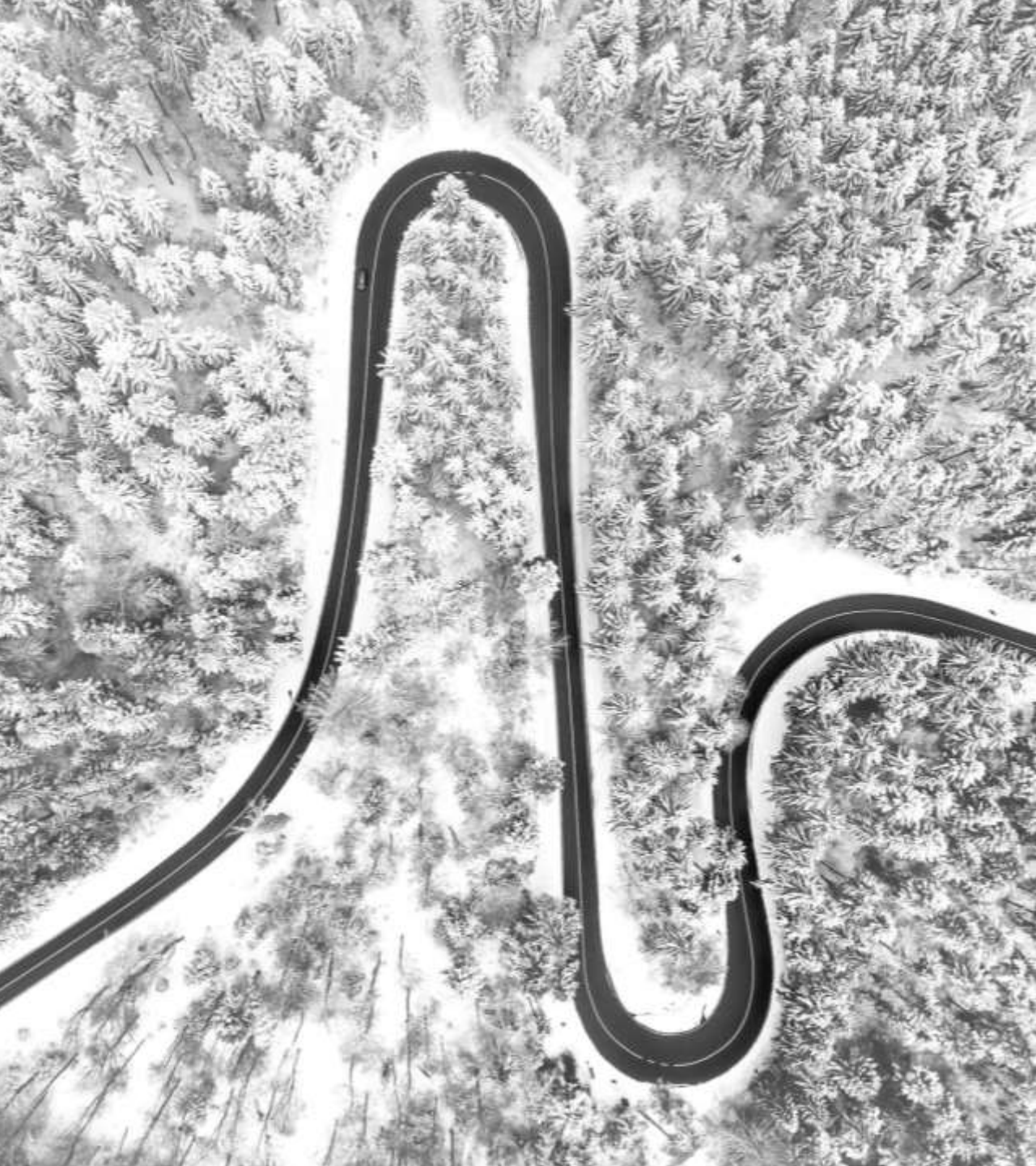




The Use Case of Artificial Intelligence in the National Archives of Hungary



Roadmap

- Relevant Projects in the NAH
 - Tax Census of 1828
 - Hungarian Prisoners in Soviet Camps
 - Population Exchange - Building a typing model using crowdsourcing
 - MIREL – Relation Maker by AI
 - Data Enrichment with Geographical Namespace
- Automatic Processing of Civil Registers

Relevant Projects in the National Archives of Hungary

Tax Census of 1828

- European Digital Treasures HTR (hand-written text recognition) project
 - Searchability function in hand-written archival documents
 - The first HTR project of the National Archives of Hungary (2021)
 - Participating countries: Malta, Norway, Portugal, Spain
- Content
 - 170.000 images
 - 53 administrative units (counties) and free royal towns
 - Personal names and economic data in tabular form (14 different columns about each person/family)
 - No uniform spelling rules and guidelines, observation of the peculiarities of the Hungarian language

Tax Census of 1828

- The workflow
 - NAH: training the software by transcribing 400 pages (8000 names) (Transkribus) (2 archivists)
 - Transkriptorium: Created the handwriting recognition model and ran it on a subset of the census (30%)
 - NAH: Verification phase, the work of the software was checked, corrected or approved by 70 volunteers
 - Transkriptorium: Improvement of the model and process of the entire census (100%)
- Ground Truth (GT)
 - The aim of the validation process is to find the Ground Truth (GT) among the suggested words
 - GT: a term which indicates the final, accepted transcription. It has a 100 % probability value.
 - Words are surrounded by text boxes, they mark the area on the page to which the transcription belongs. Text boxes vary in color, different colors represent different transcription probabilities.

Tax Census of 1828

Browser tabs: Beérkező levelek (3), Google Naptár - 202..., Magyar Nemzeti Lev..., Általános validáló_..., kaposvári_buli.docx, Az 1828. évi ország..., Kézírás felismerés m...

Address bar: adatbazisononline.mnl.gov.hu/hwview/1828/1A1D1BC185C64D7BC8B90D34FDA07420/?search=Radványi&Alconfidence=50

Search bar: Radványi

AOL • Magyarország (Hungaria)/Szabad királyi városok/Vetero -Zolium (69) - 19

Konfidencia 50%

Nominum	Contribuentium	N u m e r o											a quibus Censu solvitur.								
		Præsumptio	Honorariæ	Civis	Coloni	Inquilini	Subinquilini	Fratres	Filii	Filiæ	Servi	Avellani	Opifices	Mercatores	Quantores	Nro.	H.R.	sr.			
160. Daniel Gräß	2	2													770	118	29	28270	5002	97	129826
161. Vidua Radványi	2	2													1	9	12	272	26	46	97
162. Paulus Radványi	2	2													1	9	1	139	24	22	5
163. Paulus Szikora	2	2													1	9	40	88	8	40	
164. Georg Magna	2	2													1	9	40	128	12	28	
165. Vidua Guberda	1																				
166. Math Blavico	2	2													1	9	40	24	2	24	
167. Joana Chovan	1	1													1	9	40	40	1		97
168. Sam Makovitzky	2	2													1	9	12	126	12	26	
169. Georg Schmälik	1	1													1	9	1	126	12	26	7
170. Samuel Hlavaty	2	2													1	9	40	48	1	40	7
171. Paulus Bajnoczy															Præsum			22	2	12	
172. Georg Rozenberger	2	2													1	9	40	144	14	24	
173. Georg Bucsan Senior	2	2													1	9	12	224	22	24	177
174. Paulus Zsubritzky	2	2													1	9	24	220	22		5

- Daniel Gräß
- Vidua Joannes **Radványi**
- Paulus Radványi
- Paulus Szikora **Radványi (100.0%)**
- Georgius Magna
- Vidua Guberda
- Mathias Plavec
- Joannes Chovan
- Samuel Makovitzky
- Georgius Schmälik
- Samuel Hlavaty
- Paulus Bajnoczy
- Georgius Rozenberger
- Georgius Bucsan Senior
- Paulus Zsubritzky
- Nobilis Juliana Janyits
- Andreas Digut
- Joannes Szokan
- Georgius Fabry



Hungarian Prisoners in Soviet Camps

- In 2019, Hungary received from Russia the scanned images of the files containing the basic data of 682,000 Hungarian prisoners of war and civilian abductees, and the database created from them.
- Automatic transliteration was done in the Hungarian Research Centre for Linguistics.
- The data tables include the original Russian text in addition to the Hungarian transcription/translation, and the register card itself.
- The database search engine not only searches in the primary name lists, but also examines other name variants, since in many cases it is impossible to decide unambiguously which is the correct version.

Hungarian Prisoners in Soviet Camps

Dani Lajos - őrmester | Magyarország, Fehér megye, Székesfehérvár járás, Pátka, 1920

HU MNL OLX 10874

1920

Azonosító	231797
Név	Dani Lajos
Vezetéknév (gépi átírás)	Dani [1.00], Danyi [0.88], Dányi [0.68], Dann [0.65], Dáni [0.58], Deme [0.58], Dany [0.46], Dein [0.46], Dene [0.46], Gyányi [0.43], Doni [0.40], Gyáni [0.36], Gyene [0.30], Dame [0.27], Gyemi [0.18], Domé [0.13], Dáme [0.13], Gyámi [0.12] - Дани
Utónév (gépi átírás)	Lajos - Люш
Apai utónév (gépi átírás)	Mihály - Мигай
Nemzetiség	magyar - венгр
Rendfokozat	őrmester - сержант
Születési hely (gépi átírás)	Magyarország, Fehér megye, Székesfehérvár járás, Pátka - Pátka - с. Падка, р-н Сейкеш-Фешрвар, Венгрия, о. Фегир, Вен.
Születési év	1920
Fogságba esés helye (gépi átírás)	Csehszlovákia, Malomárok - Malomárok - с. Мала-Фракл, обл. Пототранска-Жука, Чехослов.
Fogságba esés időpontja	1945.01.23
Távozás dátuma	1947.08.19
Fogolytábor	242/9. sz. tábor - лагерь № 242/9
Nyilvántartási szám	0784106
Távozás oka	átadták a 36. sz. táborba - передан в лагерь № 36
Kapcsolódó táborok	> 242 sz. Gorlovka hadifogolytábor, Sztálingrádi régió, Oroszországi SZSZSZK / L-242 > 36 sz. hadifogolytábor / FPPL-36

Képek



005/0600-
0699/00667/00002921.JPG



005/0600-
0699/00667/00002922.JPG

Population Exchange

Building a typing model using crowdsourcing

The image displays two examples of a crowdsourcing interface for text transcription. Both examples feature a document snippet with a highlighted line of text and a corresponding pop-up window.

Left Screenshot: The document snippet shows the text "ányya: / további 3 tulajdonos. Tarnóc-i". The highlighted text is "t további 3 tulajdonos. Tarnóc-i". The pop-up window contains the following elements:

- Text: "t további 3 tulajdonos. Tarnóc-i"
- Call to action: "Kattintson ide az annotáláshoz!"
- Status buttons: "rossz szegmentáció", "nem olvasható", "áthúzott", "kihagyom"
- Feedback: "Mások így annotálták - 0", "Felismert - 0 hipotézis"
- Field information: "TARSTULAJDONOSOK mező - 2. sora"
- Progress: "8/12" and ID: "|dd8a79a2|"

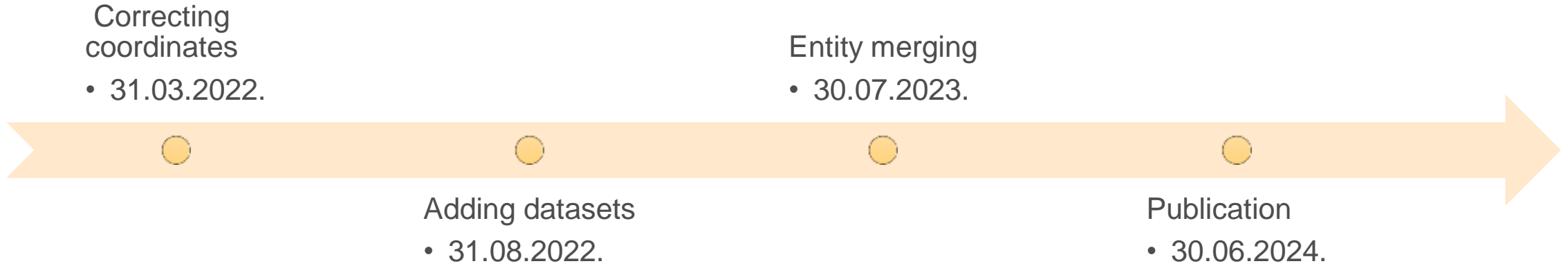
Right Screenshot: The document snippet shows the text "telek nagys. /:...../értéke/:.....". The highlighted text is "2/12 = 500.- Kcs". The pop-up window contains the following elements:

- Text: "2/12 = 500.- Kcs"
- Call to action: "2/12 = 500.- Kcs"
- Status buttons: "rossz szegmentáció", "nem olvasható", "áthúzott", "kihagyom"
- Feedback: "Mások így annotálták - 0", "Felismert - 0 hipotézis"
- Field information: "ERTÉKE_0 mező - 1. sora"
- Progress: "8/11" and ID: "|fe1d9045|"

MIREL – Relation Maker by AI

- MIREL 1.0 - Data Enrichment by Personal Names (for Hungarian prisoners in Soviet camps)
- MIREL 2.0 - Data Enrichment by Linking Geographical Entities (e.g. Police Records)
- MIREL 3.0 - General Validation Application (for Civil Registers)

Data Enrichment with Geographical Namespace



	Geographical names	Entities	Name variants
MNL GEO	75.276	70.854	4.422
GeoNames	19.172.396	12.237.573	6.934.823
Geotaurusz	135.414	109.047	26.367
The Historical Atlas of Medieval Hungary	56.084	24.148	31.936

Data Enrichment with Geographical Namespace

- Basic principles and steps for building a database of entities and data enrichment by linking entities
 - **Rule based connection:** identical strings of names and roles, same or almost same geographical coordinates.
 - **Model based connection** using AI: Similarity of strings (using various metrics e.g. soundex, jaccard), acceptable distance between geographical locations defined by coordinates.
 - **Predefined datasets** for teaching AI and for human validation (size and choosing of elements of these datasets is crucial)
 - An AI application creates its own **rules for prediction**
 - According to the results of validation, the model of AI can be developed, refined.

Data Visualization - Police Records

The screenshot displays a web application interface for searching police records. The browser's address bar shows the URL: `test.zala.natarch.hu/adatbazisok/adatbazis/magyar-keralyi-belugyministerium-allamrendeszeti-kartotekrendszere-csendorkartonok/kereses`. The page title is "Magyar Királyi Belügyminisztérium államrendészeti kartotékrendszere – Csendőrkartonok".

Találatok szűrése (Search Filters):

- Születési dátum (Date of Birth):** A bar chart shows the distribution of records from 1875 to 2023. The x-axis is labeled "Év" (Year) and the y-axis represents the number of records.
- Születési hely (Place of Birth):**
 - Budapest (6582)
 - Debrecen (416)
 - Szeged (388)
 - Szolnok (383)
 - Újpest (352)
- Lákhely (Residence):**
 - Budapest (9030)
 - Újvidék (764)
 - Bécs (749)
 - Újpest (483)
 - Miskolc (482)
- Valás (Religion):**
 - római katolikus (23954)
 - nem beazonosítható (11043)
 - izraelita (9003)

Search Results:

- Search input: "Írd be a keresőszót!"
- Search button: "Keres"
- Results summary: "60933 dokumentumban / 60933 érintett oldal / 53792 földrajzi koordinátával (0:53 másodperc)"
- Warning: "Túl sok találat, 5000 került megjelenítésre"

Map Visualization:

- Map showing geographical distribution of records across Europe and parts of Asia.
- Map controls: "Ikon nézet", "Tíca nézet", "Táblázat", "Térkép".
- Map markers: Colored circles (green, yellow, orange, red) with numbers indicating the count of records for each location.

Automatic Processing of Civil Registers



1	2	3	4
Atony
...
...
...



Handwritten text in a cursive script, possibly a list or a letter, with some lines underlined.



Handwritten text in a cursive script, continuing the list or letter from the previous page.

Handwritten text in a cursive script, appearing to be a list of names and dates, possibly a family register or a list of births.

199	Béla
198
197
196
195
194
193
192
191

Handwritten text in a cursive script, possibly a list of names and dates, continuing the family register or list of births.

Handwritten text in a cursive script, possibly a list of names and dates, continuing the family register or list of births.

Handwritten text in a cursive script, possibly a list of names and dates, continuing the family register or list of births.

A gyermek utóneve, neve, vallása	A szülők családi és utóneve, állása (foglalkozása), lakhelye
4	5
Flóra leány. r. kath	Urbán István földművelő. Cs. Farga Rozália Házasság 1921.

Aims of the Project

- Processing and publishing civil registers
- Publishing digitised images and extracted data together
- Extract structured data content and load it into a database, in parallel with automatic handwriting recognition
- Long-term plans: development of family tree building and visualisation tools

Civil Registers 1895-1980

- 3.422 settlements
- 37.611 volumes
- 1.908 linear metres
- 22 million pages

Civil Registers of Abony

- 63 folders (cca. 600 pages per folder) -> almost 40.000 pages
 - Birth registers: 25 folders
 - Marriage registers: 19 folders
 - Death registers: 19 folders

The Workflow of Processing Civil Registers

I.

preparing the
images for
processing

III.

segmentation

V.

loading into
database

VII.

connecting
data to
namespace

VIII.

linking
persons

II.

image
classification

IV.

HTR

VI.

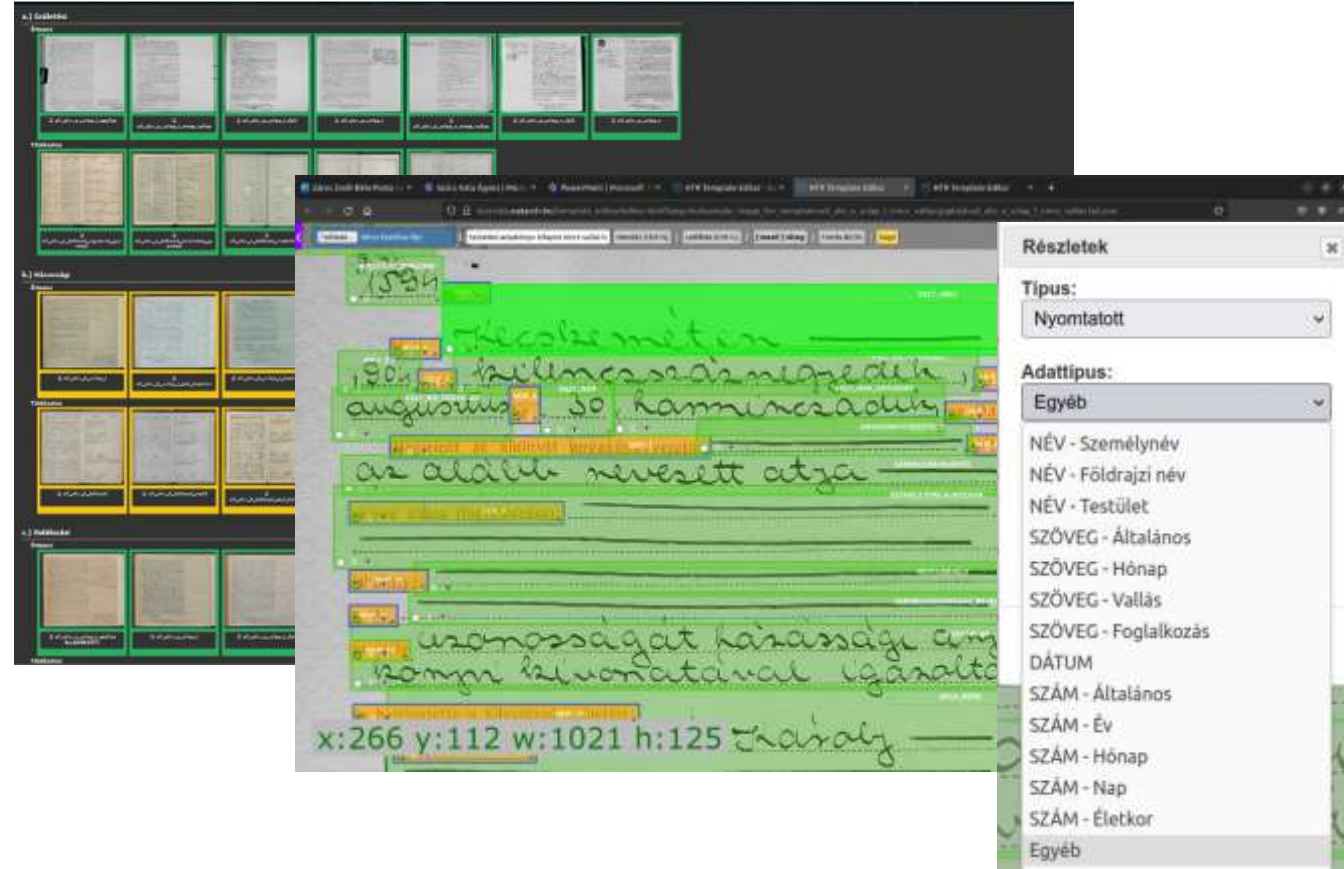
data
normalization
and correcting

IX.

publication

Image Classification: Templates

- Annotation application, web service (Template Builder)
- Creating fields, recording their metadata
- Tagging labels on a form and record their metadata, in particular their static values
- Assigning images to a template based on the template set



Template Builder

The image shows a digital document with a template overlay. The document is a handwritten form with fields for birth records. The template overlay consists of colored boxes with labels like 'REGYÉES-SZÁMSZÁM', 'KÉLT EV', 'KÉLT NO SZÜNET', 'LAKÓHELY', 'A szülő', 'A születés', and 'A gyermek'. A black pen icon is positioned over the template, pointing to the document text.

Document fields (handwritten):

- REGYÉES-SZÁMSZÁM: 376
- KÉLT EV: 1904
- KÉLT NO SZÜNET: szeptember 28
- LAKÓHELY: Pászapáti
- A szülő: evangélikus református
- A születés: evangélikus református
- A gyermek: evangélikus református

Template labels:

- REGYÉES-SZÁMSZÁM
- KÉLT EV
- KÉLT NO SZÜNET
- LAKÓHELY
- A szülő
- A születés
- A gyermek

Handwritten Text Recognition

- Annotation application, web service
- Training data for HTR:
 - Tax census of 1828 (100 pages)
 - Registers of Abony (cca. 1.300 pages)
 - Cardfiles of the Comprehensive Dictionary of Hungarian (almost 6.000 cardfiles)
 - Correspondence of József Kiss (cca. 900 pages)
- Evaluating data with the same method
- Training TrOCR model (still under development)
- Best result on handwriting: 6 % CER, 15–18% WER

Creating Training and Evaluating Data for HTR

The image shows a software interface for processing handwritten documents. On the left, a table lists document IDs. The main area displays a scanned document with handwritten text and colored bounding boxes. A pop-up window is open over the text 'Abádszalók', showing a list of similar words and their confidence scores.

Document ID	Color
004811733_00423_R	Blue
004811733_00513_L	Green
004811733_00513_R	Yellow
004811744_00176_R	Orange
004811751_00319_R	Blue
004811757_00140_L	Green
004811757_00140_R	Yellow
004811758_00502_R	Blue
004811806_00666_R	Green
004811808_00157_R	Yellow
004811813_00216_L	Blue
004811813_00216_R	Green
004811814_00373_L	Yellow
004811814_00373_R	Blue
004811816_00434_L	Green
004811816_00434_R	Yellow
004811816_00443_L	Blue
004811816_00443_R	Green
004811825_00579_L	Yellow
004811825_00579_R	Blue
004811826_00120_L	Green
004811826_00120_R	Yellow
004811827_00113_L	Blue
004811827_00113_R	Green
004811838_00180_L	Yellow
004811838_00180_R	Blue
004811841_00677_L	Green
004811841_00677_R	Yellow
004811850_00267_L	Blue
004811850_00267_R	Green
004811850_00396_L	Yellow
004811850_00396_R	Blue
004811850_00448_L	Green
004811850_00448_R	Yellow
004811852_00530_L	Blue
004811852_00530_R	Green
004811853_00341_R	Yellow
004811853_00357_L	Blue

Document text (handwritten):

a kinek állása (foglalkozása): *feladatok elvégzése*

lakóhelye: *Abádszalók*

személy *aranyos*

és bejelentette a következő

családi és utóneve

vallása: *római*

állása (foglalkozás)

lakóhelye: *Abádszalók*

születéshelye: *Abádszalók*

életkora: *25*

családi és utóneve *Timre*

vallása: *római katolikus*

Pop-up window content:

Abádszalók

Kattintson ide az annotáláshoz!

rossz szegmentáció nem olvasható áthúzott kihagyom

Mások így annotálták - 3

- Abádszalók - 2023. 10. 27. - annotálva
- Abádszalók - 2023. 11. 14. - annotálva
- Abádszalók - 2023. 11. 14. - annotálva

Felismert - 9 hipotézis

- Abádszalók - 0.9999
- Abádszalók K - 0.9999
- Abádszalók k - 0.9999
- Abádnalók - 0.9999
- Aládszalók - 0.9997
- Abádszatók - 0.9997
- Abádszaló h - 0.9997
- Abádszaló - 0.9768
- Abádszalott - 0.9297

HELY.LAKHELY mező - 1. sora

[862, 483, 1140, 60]

handwritten

name_geo

hu

11/39

Loading Data into Database



SKIP_24	lakóhelye: Jász-Kis-ér	ANYJA_LAKHELYE
SKIP_25	születéshelye: Tama-Évös	ANYJA_SZULETESHELYE
SKIP_26	életkora: 20 éves	ANYJA_ELETKORA_SZOVEGES
SKIP_27	születési helye: Jász-Kis-ér	SZULETESHELYE
SKIP_28	születési év: 1946	SZULETES_EV_SZOVEGES
SKIP_29	születési nap: 20. szeptember	SZULETES_NAP_SZOVEGES
SKIP_30	születési hónap: szeptember	SZULETES_HO_SZOVEGES
SKIP_31	születési napszám: 20	SZULETES_N

- SQL database
 - Fields created on templates and their recorded metadata
 - Text recognised by the HTR
 - Corrections and normalised forms made during post-processing

Postprocessing Data

Aims

- Making data searchable
- Helping to link persons

Tasks

1. Treatment of hypotheses: now only the hypothesis with the highest confidence
2. Managing overloaded fields
3. Correction, normalization, standardization
4. Extracting further data

A gyermek utóneve, neme, vallása	A szülők családi és utóneve, állása (foglalkozása), lakhelye
Flóra.	Urban István földművelés.
leány.	Cs. Fuvga Rozália
v. káth	Abony. VI. 252/2

The Output of PPR

Elhunyt - lakhely

Abony 🏡

Abony [0.9998]

1x ELHUNYT_LAKHELYE | PPR_ID: 47014612 | JAVITVA:1 | TELEPÜLÉS: Abony | NEVTER: 12436576 | DEBUG: Abony

Elhunyt - foglalkozás

cipész

: cipész [0.0507]

1x ELHUNYT_FOGLALKOZASA | PPR_ID: 47014612 | JAVITVA:1 | TXT: cipész

Elhunyt - vallás

római katolikus

: római katolikus [0.9768]

1x ELHUNYT_VALLASA | PPR_ID: 47014612 | JAVITVA:1 | TXT: római katolikus

Elhunyt életkora szöveges

nyolcvanhat (86)

86 (nyolcvanhat [0.731])

1x ELETKORA_SZOVEGES | PPR_ID: 47014612 | JAVITVA:1 | TXT: nyolcvanhat | NUM: 86

Bejelentő - név

Halicia György

Halicia György [0.779] Holicza György [0.538] Alalicia György [0.3802] Halica György [0.3607] Halicza György [0.2898]

1x BEJELENTO_NEVE | PPR_ID: 47014612 | JAVITVA:1 | TXT: Halicia György | CSALÁD: Halicia | UTÓ: György | DEBUG: Halicia György

Bejelentő lakhelye

Abony 🏡

Abony [0.8839]

1x BEJELENTO_LAKHELYE | PPR_ID: 47014612 | JAVITVA:1 | TELEPÜLÉS: Abony | NEVTER: 12436576 | DEBUG: Abony

Bejelentő foglalkozása

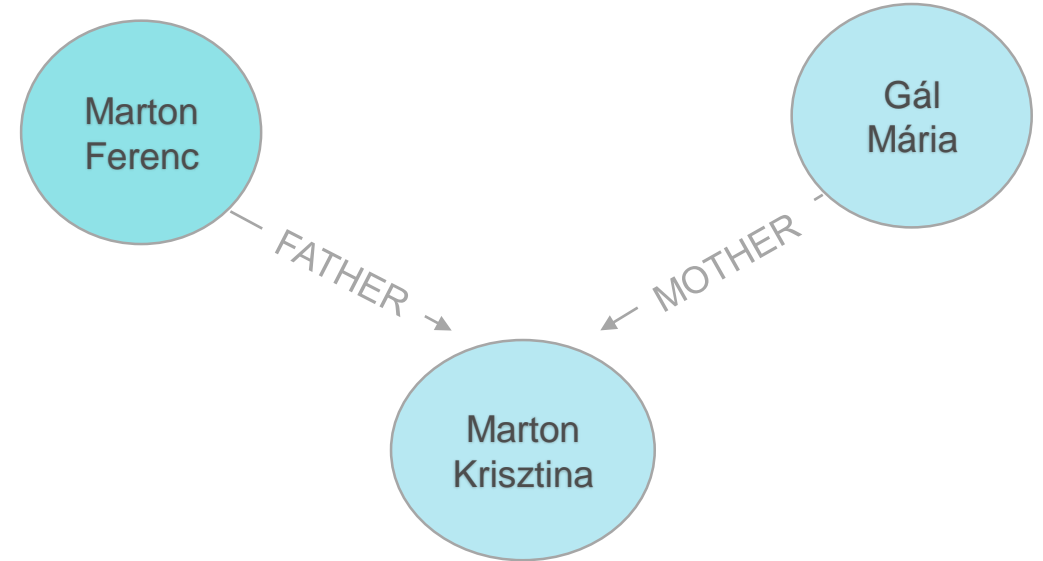
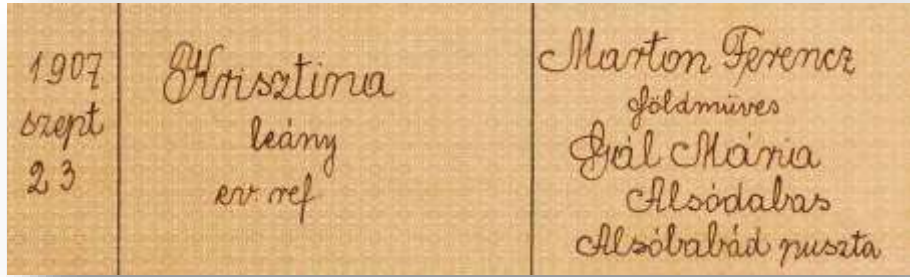
kórházi gondozó

: kórházi gondoró [0.3916]

a [0.732]

2x BEJELENTO_FOGLALKOZASA | PPR_ID: 47014612 | JAVITVA:1 | TXT: kórházi gondozó

Linking Persons



name	relation	birth place	birth date
Marton Ferenc	father	Alsódabas	1883
Gál Mária	mother	Alsódabas	1886
Marton Krisztina	child	Alsódabas	1907

Thank you

<https://adatbazisokonline.mnl.gov.hu/>



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.