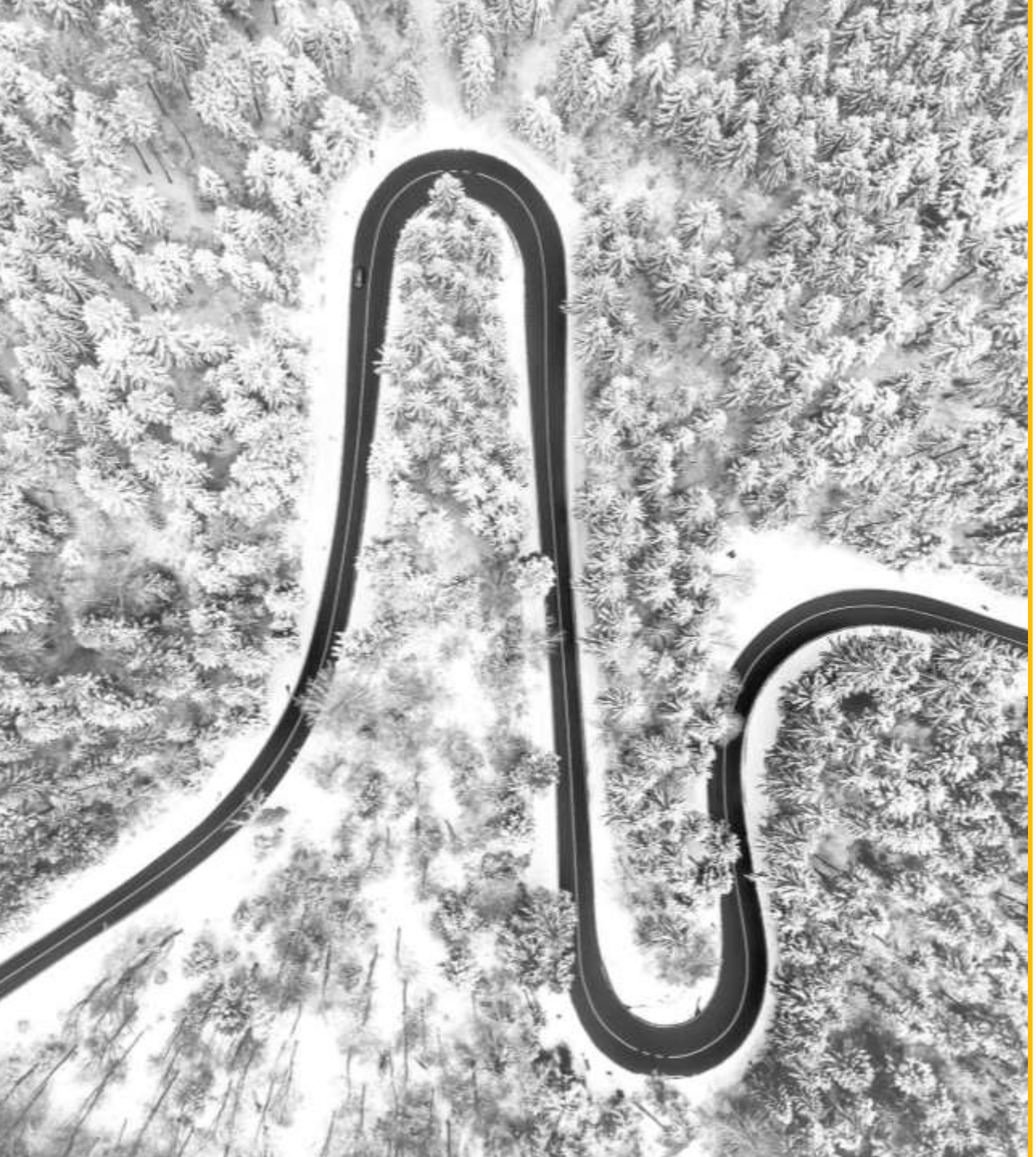
The Use Case of Artificial Intelligence in the National Archives of Hungary

Noémi Vadász, Zsolt Bánki



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

European Commission

Relevant Projects in the National Archives of Hungary

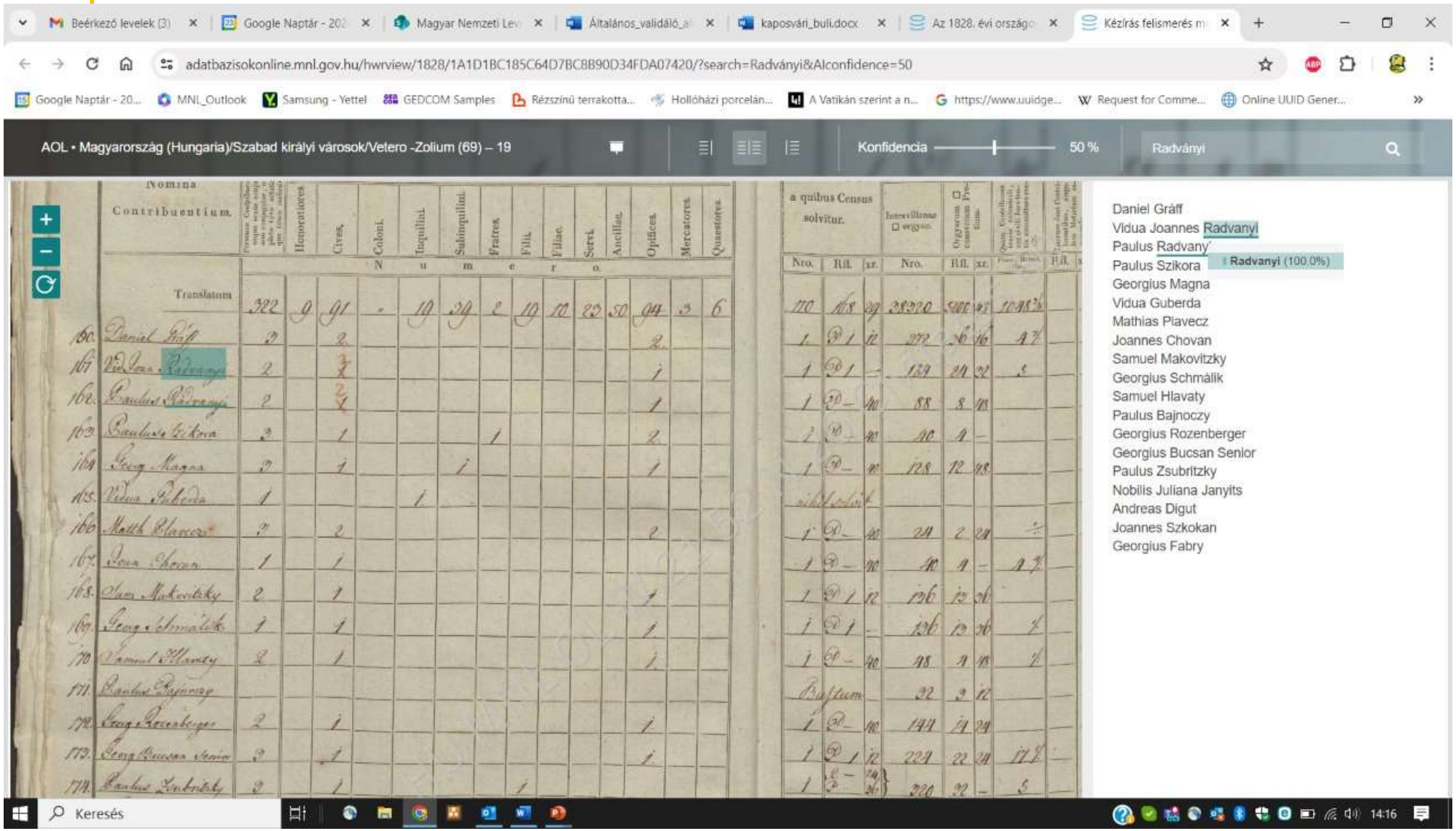
Tax Census of 1828

- 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 sofware 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



European Commission

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 OL X 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 - Лоюш
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	

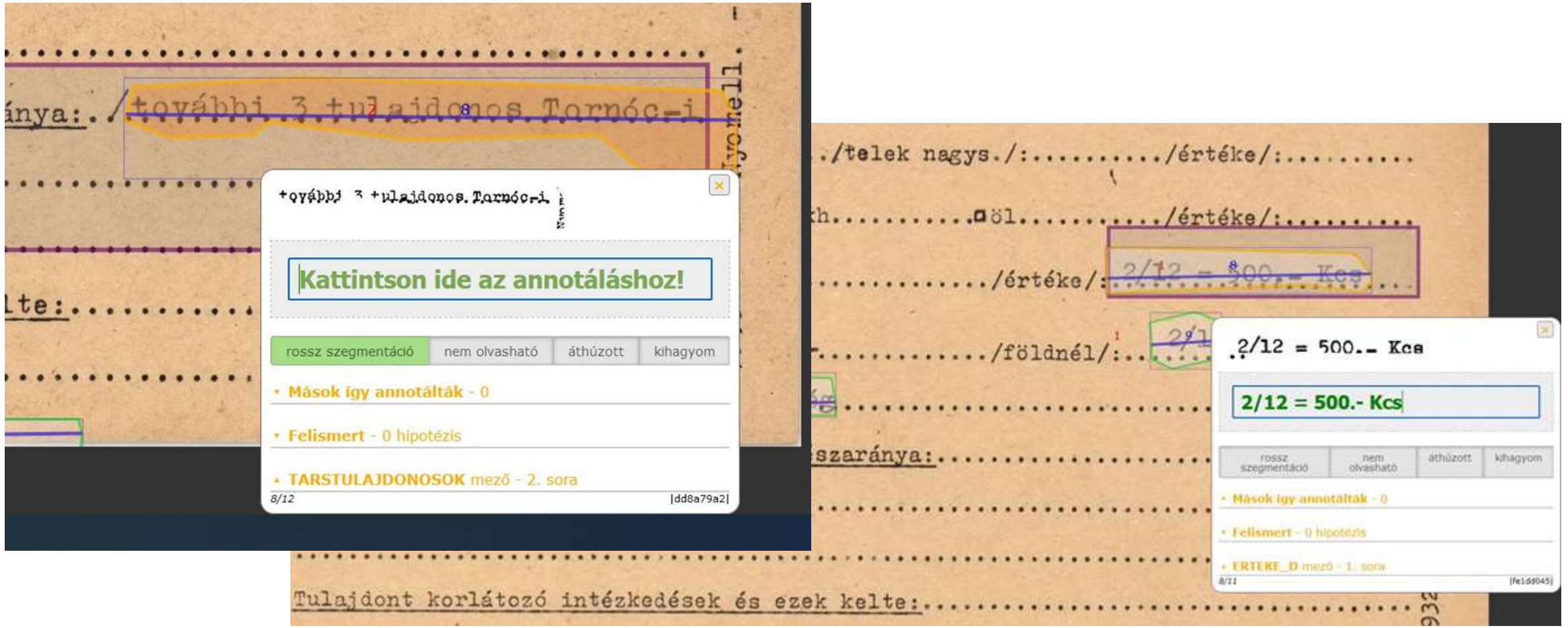
005/0600-0699/00667/00002921.JPG

005/0600-0699/00667/00002922.JPG

European Commission

Population Exchange

Building a typing model using crowdsourcing

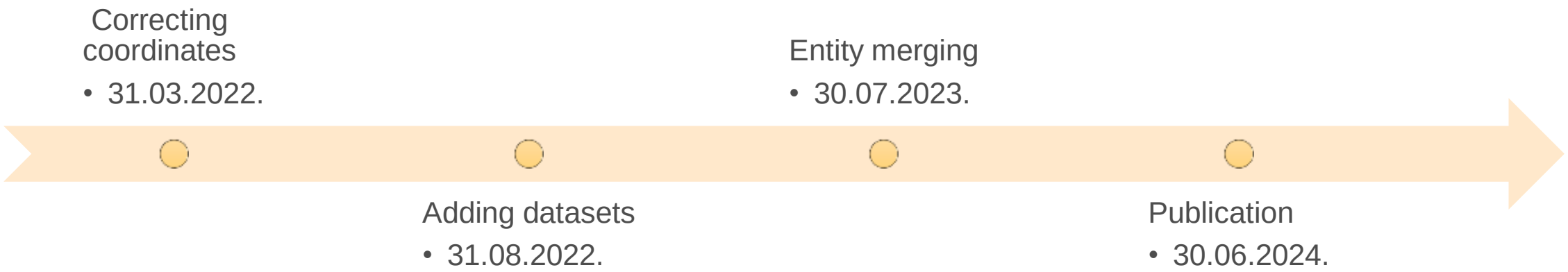


European Commission

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 Geogaphical Entities (e.g. Police Records)
- MIREL 3.0 - General Validation Application (for Civil Registers)

Data Enrichment with Geographical Namespace



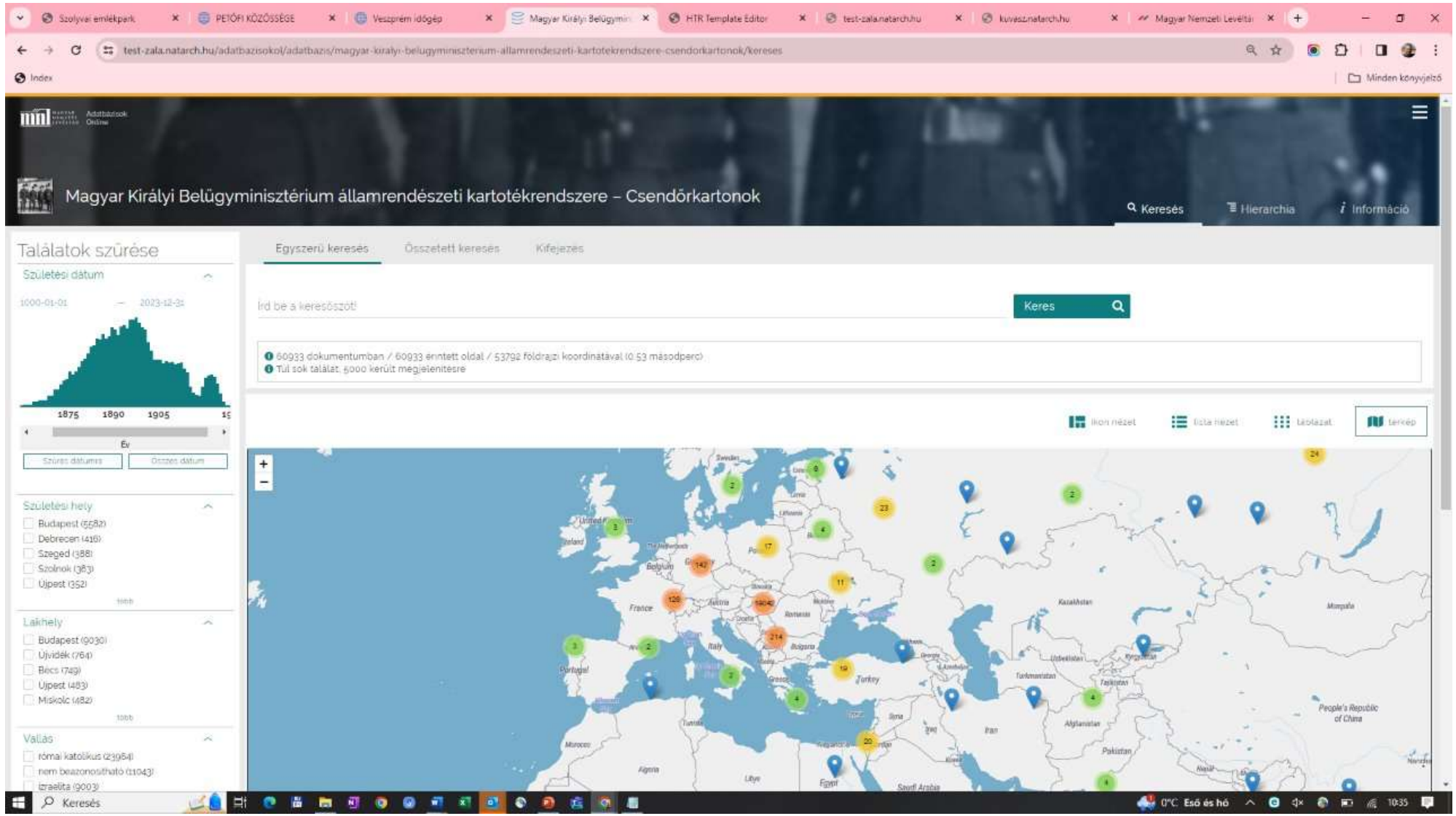
	Geographical names	Entities	Name variants
MNL GEO	75.276	70.854	4.422
GeoNames	19.172.396	12.237.573	6.934.823
Geotaurusz	135.414	109.047	26.367
The Historical Atlas of Medieval Hungary	56.084	24.148	31.936

European Commission

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



European Commission

Automatic Processing of Civil Registers

European Commission

A gyermek utóneve, neme, vallása	A szülők családi és utóneve, állása (foglalkozása), lakhelye
4	5
Ilona. leány. r. kath	Urbán Istóné földmunkás. Cs. Farga Rozália

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

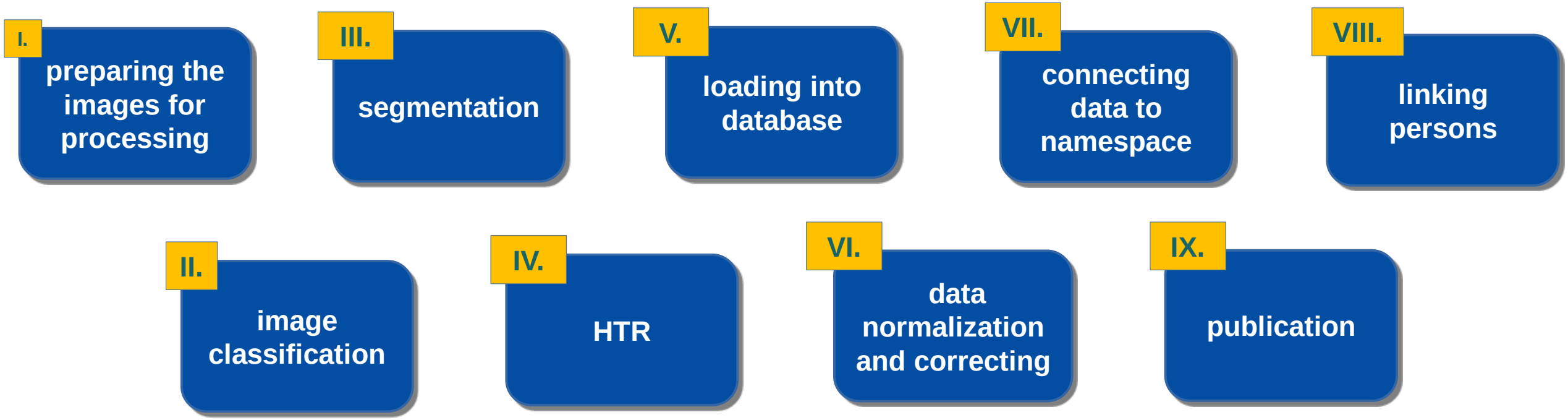
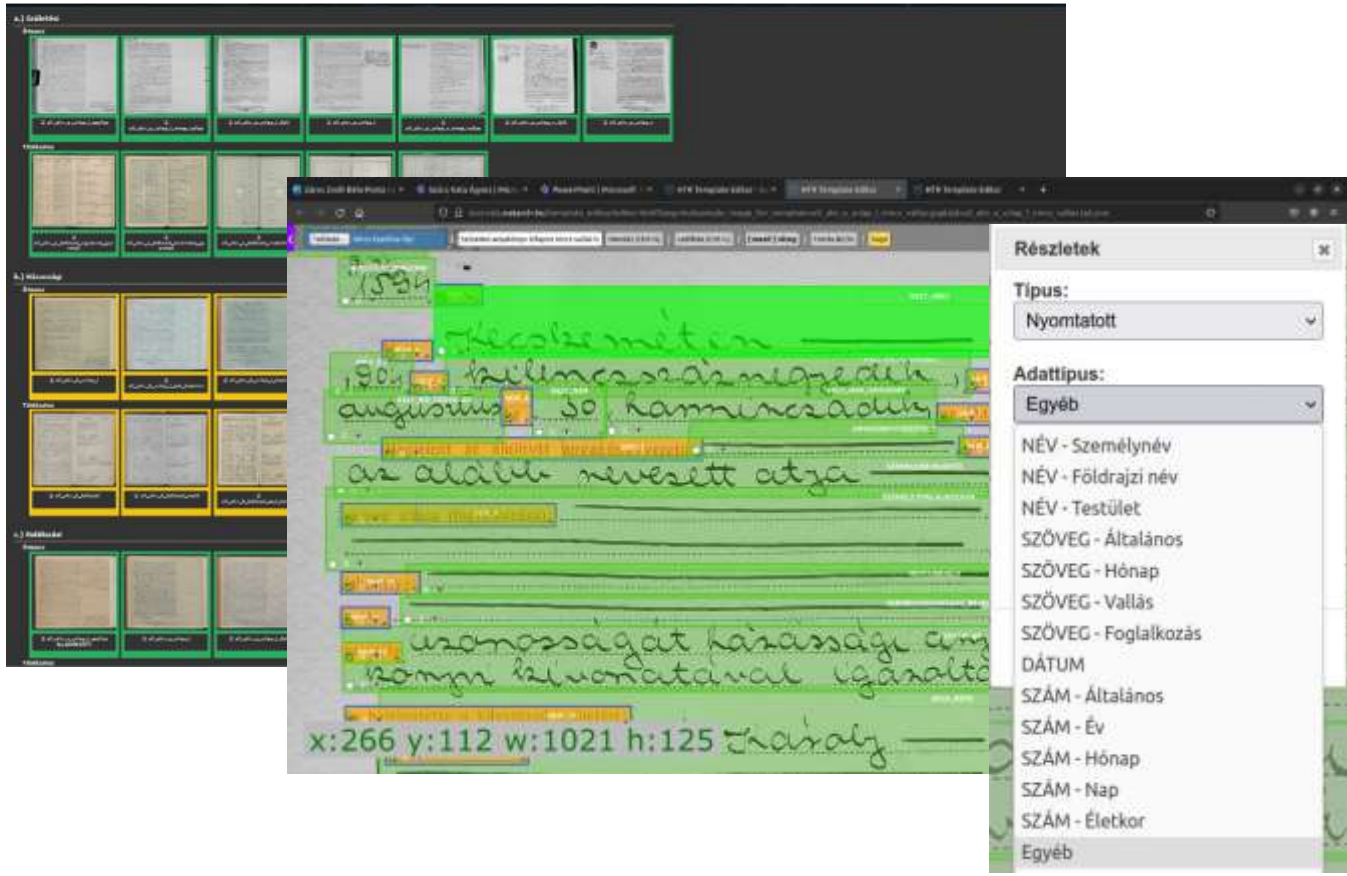


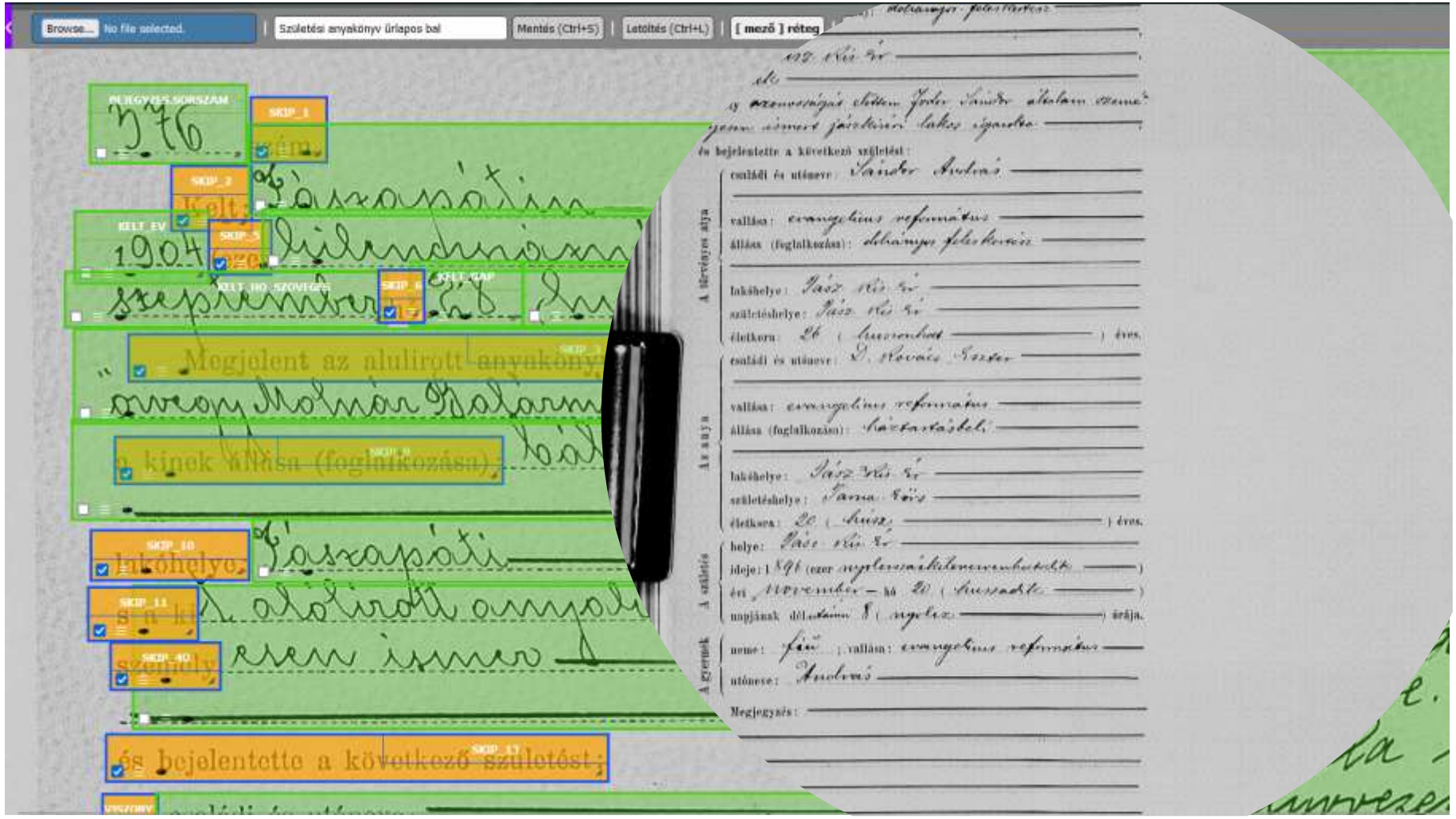
Image Classification: Templates

- 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



European Commission

Template Builder

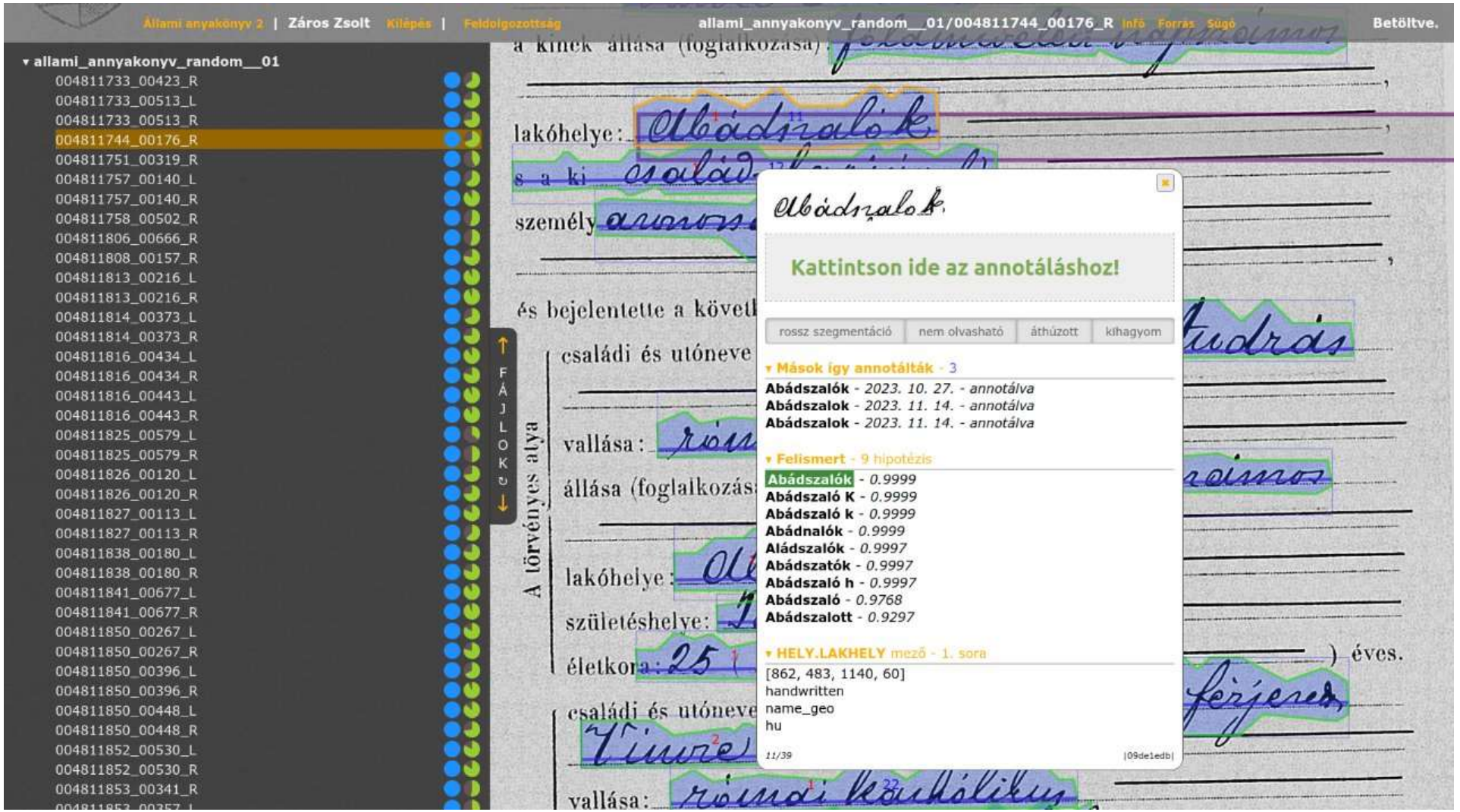


European Commission

Handwritten Text Recognition

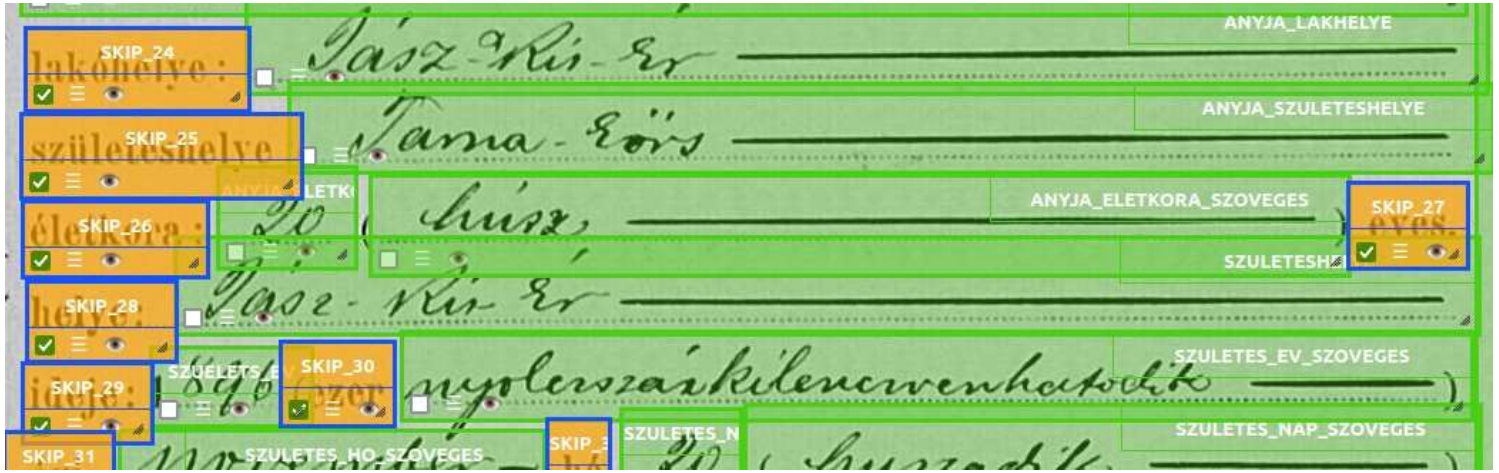
- Annotation application, web service
- Training data for HTR:
  - Tax census of 1828 (100 pages)
  - Registers of Abony (cca. 1.300 pages)
  - Cardfiles of the Comprehensive Dictionary of Hungarian (almost 6.000 cardfiles)
  - Correspondence of József Kiss (cca. 900 pages)
- Evaluating data with the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

Creating Training and Evaluating Data for HTR



European Commission

Loading Data into Database



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

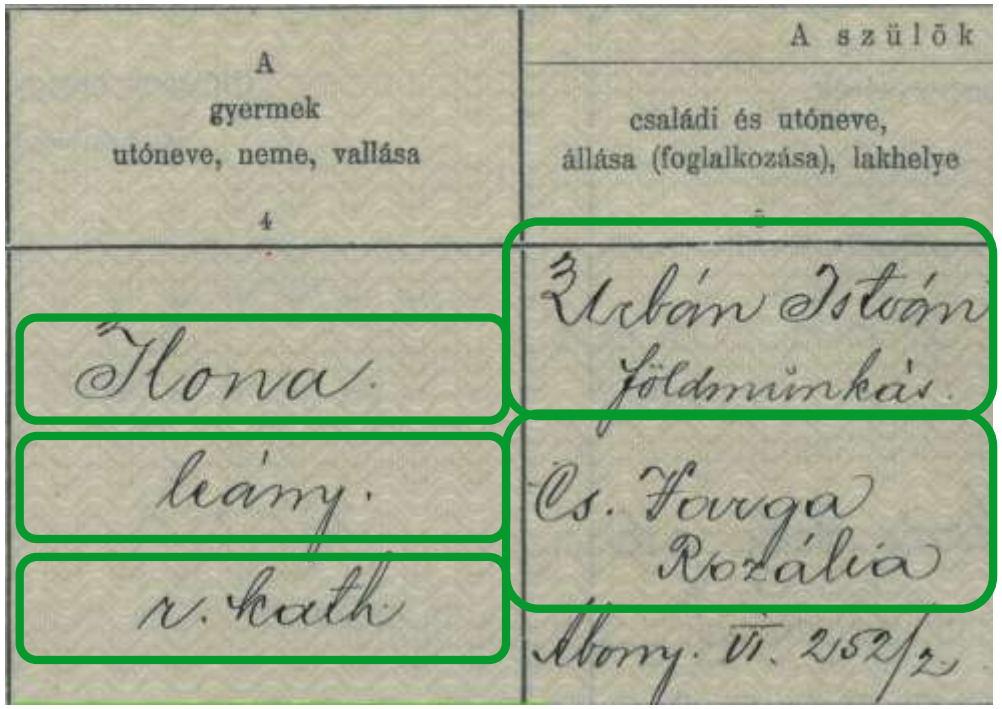
Postprocessing Data

Aims

- Making data searchable
- Helping to link persons

Tasks

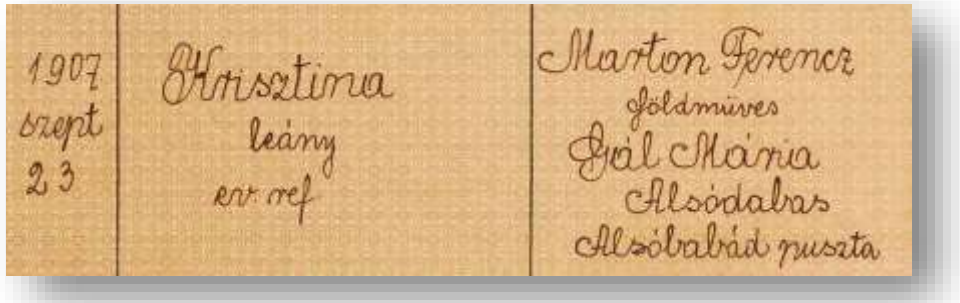
1. Treatment of hypotheses: now only the hypothesis with the highest confidence
2. Managing overloaded fields
3. Correction, normalization, standardization
4. Extracting further data

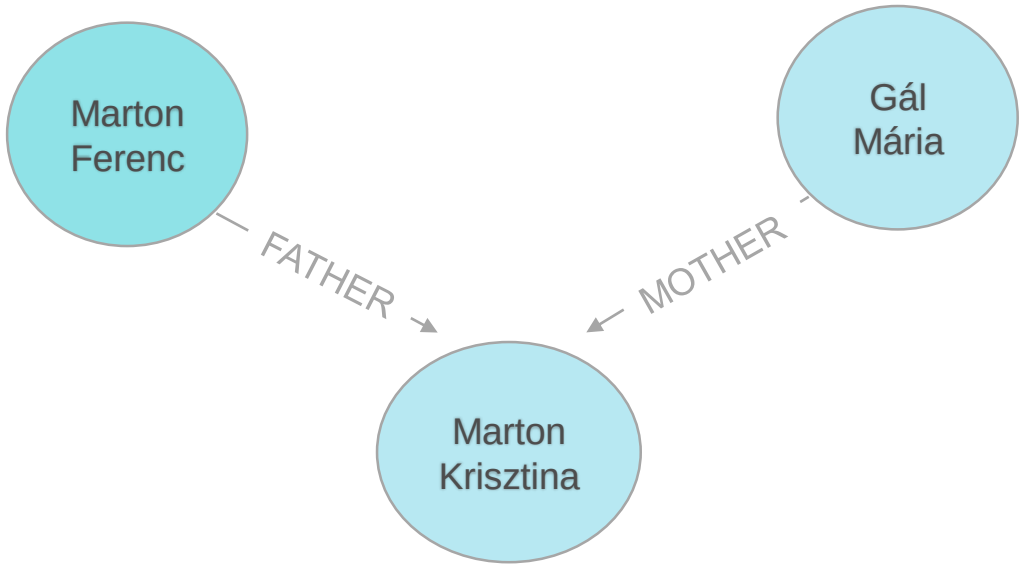


The Output of PPR

Elhunyt - lakhely	**Abony** Abony [0.9998] *1x ELHUNYT_LAKHELYE | PPR_ID: 47014612 | JAVÍTVA:1 | TELEPÜLÉS: Abony | NEVTER: 12436576 | DEBUG: Abony*
Elhunyt - foglalkozás	**cipész** : czipész [0.0507] *1x ELHUNYT_FOGLALKOZASA | PPR_ID: 47014612 | JAVÍTVA:1 | TXT: cipész*
Elhunyt - vallás	**római katolikus** : római kathólikus [0.9768] *1x ELHUNYT_VALLASA | PPR_ID: 47014612 | JAVÍTVA:1 | TXT: római katolikus*
Elhunyt életkora szöveges	**nyolcvanhat (86)** 86 (nyolczvanhat [0.731] *1x ELETKORA_SZOVEGES | PPR_ID: 47014612 | JAVÍTVA: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 | JAVÍTVA:1 | TXT: Halicia György | CSALÁD: Halicia | UTÓ: György | DEBUG: Halicia György*
Bejelentő lakhelye	**Abony** Abony [0.8839] *1x BEJELENTO_LAKHELYE | PPR_ID: 47014612 | JAVÍTVA:1 | TELEPÜLÉS: Abony | NEVTER: 12436576 | DEBUG: Abony*
Bejelentő foglalkozása	**kórházi gondozó** : kórházi gondoró [0.3916] a [0.732] *2x BEJELENTO_FOGLALKOZASA | PPR_ID: 47014612 | JAVÍTVA: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/