



Areas of application of AI in the Historical Archives of the Hungarian State Security

Zoltán Lux, Dániel Havasi-Mészáros, Anna Kulcsár

eArchiving/Historical Archives of the Hungarian State Security Services/*DLM Forum Meeting*
Budapest, 7th November 2024

Topics of the Presentation

- Tasks of the Historical Archives
- Implementing AI-tools
 - Improving the OCR process of degraded scanned documents
 - Recognition of entities (persons, corporate bodies, geographical concepts, events)
 - Retrieval-Augmented Generation
- The status and problems of integrating artificial intelligence tools into the workflow at the Archives of the Hungarian State Security

Tasks of the Historical Archive

- *ordinary archive duties*
- *ensuring person requesting access (insight for the citizen)*
- *providing the possibility of research*



Tasks of the Historical Archive

Finding documents

Researchers have access to database & finding aids.

But sometimes it's not enough...



Tasks of the Historical Archive

We have to find the relevant documents

Based on metadata

Based on free text search

1. assumes good quality ocr text
2. and that all content is ocr-processed

We must provide a suitable search interface for researchers.

The documents to be served must be anonymized.



Tasks of the Historical Archive – AI, OCR and finding information

- We have to find the relevant documents
 - Very good OCR text
 - Improves the result of the free text search
 - Improves metadata quality (NER, RAG – extraction of document content)
 - We hope that by applying RAG, we can add the necessary background and historical knowledge to the searches.
- We must also provide researchers with a suitable search interface
 - NER
 - RAG



Implementing AI tools – improving OCR

- It has been built into the system so that you can revert to the previous OCR at any time.

The screenshot displays a document viewer interface. On the left, a document page is visible with the title "A Magyar Kormányzat Kézléve" and a handwritten number "47/111". The main content of the document is a list of items, with a table of data above it. The table has columns for ID, description, and quantity. The list below the table contains several numbered items, some of which are highlighted in blue. The document text is in Hungarian and discusses the Hungarian Government's actions regarding the trial of Simon György.

| ID | Description | Quantity |
|------------------|---------------|----------|
| 5 V. 9097.001 W | V. 9097.001 W | 200 |
| 6 V. 9097.004 W | V. 9097.004 W | 200 |
| 7 V. 9097.005 W | V. 9097.005 W | 200 |
| 8 V. 9097.007 W | V. 9097.007 W | 200 |
| 9 V. 9097.007 W | V. 9097.007 W | 200 |
| 10 V. 9097.008 W | V. 9097.008 W | 200 |

The list below the table contains several numbered items, some of which are highlighted in blue. The document text is in Hungarian and discusses the Hungarian Government's actions regarding the trial of Simon György.

Implementing AI tools – NER

- Search separately for different entities
- Only a small proportion (about 1/8) of OCR-ed documents have been entity recognised. Searching is more accurate and efficient than in free text.

The screenshot shows a search interface titled "Dosszié/Ügyirat keresés". It features several search filters and a search bar. The filters include:

- Törzsszám (Main number)
- Azonosító (Identifier)
- Ikt./nyilvt. szám (Filing/record number)
- Feldolgozottság (Processing status)
- Tárgy (Subject)
- Ország (Country)
- Megye (County)
- Település (Settlement)
- Nyt./egyéb ikt. sz./Nyitószám (Filing/other filing number/Opening number)
- Irat évköre (Document period)
- Téma évköre (Topic period)
- Átvevő (Receiver)

Below the filters, there are tabs for "Alap" (Basic) and "AI". The "AI" tab is active, showing a search bar with the text "szabó lajos". Other search filters include:

- OCR
- Személy (Person)
- Intézmény (Institution)
- Földrajzi név (Geographical name)
- Egyebek (Others)

The search bar at the bottom contains the text "szabó lajos". The search button is labeled "Keres".

Implementing AI tools – NER

Entities with only name metadata (persons, corporate bodies) should be matched with entities with detailed metadata in the master files. This will not be able to be done fully automatically, but the necessary program development is also quite resource intensive.

The screenshot shows a file management window titled 'Kép felvétel/karbantartás'. It contains a table with columns: 'Oldal', 'Eredeti file neve', 'Cím', 'W', 'Mérete', 'Szél.', 'Mag.', and 'Módosította'. Below the table is a sidebar with four categories: 'Személy', 'Intézmény', 'Földrajzi név', and 'Egyéb'. The 'Személy' category lists names like 'bó László', 'Dr. Iványosi', and 'Dr. Szabó Lajos'. The 'Intézmény' category lists 'Bányagépipari Technikum', 'HM', 'Signum Laudis', etc. The 'Földrajzi név' category lists 'Salgótarján', 'Dózsa Gy. u.', etc. The 'Egyéb' category lists 'Délvidéki Emlékeken', 'ABTL'. A green arrow points from the 'Dr. Szabó Lajos' entry in the 'Személy' list to the search results window on the right.

The screenshot shows a search results window titled 'Személy keresés eredménylista (689)'. It contains a table with columns: 'Neve', 'Célsz. Fedőnév', 'Ügyn. Fedőnév', 'Ü', 'Anyja neve', 'Szül. idő', 'Szül. hely', 'Foglalkozás', and 'Tele...'. The table lists several entries, with 'Szabó Lajos, dr.' highlighted in blue. Below the table is a search bar and a 'Megjegyzés' section containing the text: '1953.11.20. (áv. alhdgy.) a Néphadsereg Pf. 2437. ehárhó osztályra helyezve.'

Implementing AI tools – NER

- GEO entities

The screenshot displays the ABTL online search interface. The main window shows a map of Hungary with 23 geographic locations marked. A pop-up box indicates "23 geographic location". The search results table is as follows:

| Subject | Document's date | Topic's date |
|--|-----------------|--------------|
| Király István beszéde a magyar forradalomról az ENSZ közgyűlésén | 1957 | 1956 |
| Társadalmi jelentés a SZER szociálistársadalmi idején | 1967.05.08 | 1956 |
| Az 1956-os "ellenforradalom" 20. évfordulójával kapcsolatos események | 1976 | |
| Az ellenforradalom 21. évfordulója kapcsán, rendszerváltás idején készített részletes szemleanyagok és iratok gyűjteménye | 1988.01.18 | |
| Az 1956-os magyar forradalom emlékeiről, alakulásáról és utóéletéről című regényes kiadvány | 1988.07.07 | |
| Az ellenforradalom 32. évfordulója megemlékezés emlékkönyve | 1988.10.06 | |
| Az ellenforradalom 32. évfordulója megemlékezés emlékkönyve | 1988.10.10 | |
| Ellenforradalom alatt működő különböző csoportok és fegyveres ellenforradalmi szervezetek, azok tevékenységéről és tagjairól monográfia. Nógrád megyei összeállítás monográfia | 1956 - 1959 | 1956 - 1959 |
| Ellenforradalom alatt működő különböző csoportok és fegyveres ellenforradalmi szervezetek, azok tevékenységéről és tagjairól monográfia. Szeged és környéki monográfia | | |

Implementing AI tools – RAG

The screenshot displays the LIR Online web application interface. The left sidebar contains search filters: 'JELZET' (Reference), 'IRAT ÉVKÖRE (-TÓL)' and 'IRAT ÉVKÖRE (-IG)' (Document date range), 'KIVONAT' (Abstract), and 'KIS- ES NAGYBETŰ MEGKÜLÖNBÖZTETÉSE' (Case sensitivity). The main content area shows a hierarchical tree structure under 'Elhelyezkedés a levéltári struktúrában' (Location in the archival structure). The selected path is: 'Állambiztonsági Szolgálatok Történeli Levéltára' > 'I. levéltár Állambiztonsági iratok levéltára' > '3. szekció Hálózat; operatív és vizsgálati dossziék' > '3.1. A Központi operatív nyilvántartást végző szervezeti egységek által kezelt dossziék'. Below the tree, a table displays document details:

| Jelzet | Levéltári szint |
|----------------------|-----------------|
| HU - ÁBTL - L - 3.1. | Fond |

| Tárgy | Írat évköre | Téma évköre |
|--|-------------|-------------|
| A Központi operatív nyilvántartást végző szervezeti egységek által kezelt dossziék | | |

| Terjedelem (f/m) | Terjedelem leírás |
|------------------|-------------------|
| 2.263,62 | |

Kivonat
Az egykori állampárti titkosszolgálati szerveknek az ÁBTL őrizetében került operatív nyilvántartási rendszere alapvetően két irattípusból áll: kartonból és dossziéből. Ugyan az állambiztonsági szervek dossziéi az egymással összefüggő iratokat őrzik, mégsem tekinthetők őrzési egységeknek, mert elsősorban nem az iratok biztonságos tárolását szolgálták. A dosszié tartalmát ugyanis eleve meghatározta, hogy egy adott ügyben milyen elnevezésű dossziét nyitottak. 1950 után az állambiztonsági szervek ügyvitelének alapegysége a dosszié volt, amely tehát ügyiratként fogható fel. Egy dossziéba 300–350, később 250–300 oldalnyi iratot lehetett elhelyezni. Ezt a mennyiséget meghaladó terjedelmű iratanyag esetén ugyanazon a sorszámon újabb dossziét kellett nyitni. Minden dossziéhoz tartalomjegyzéket és névmutatót kellett csatolni, a lapokat meg kellett számozni és be kellett varrni. A „felderítő munka eredményessége” érdekében, illetve egyéb okokból a különböző állambiztonsági szervek gyakran adták át egymásnak ügyeket, vagyis dossziéikat. Ez minden esetben határozattal történt. Az átadó szerv összefoglaló jelentést készített, amelyben feltüntette a dosszié átadásának okát. A dossziék lezárása szintén határozattal történt, végül az operatív tiszt elhelyezte a zárólapot, amelyre feljegyezte, hány lapot tartalmaz a dosszié. Az operatív dossziékat lezárás után a központi nyilvántartónak küldték meg irattározásra. A megküldött dossziékat a

ed to
ed on research

Implementing AI tools – Anonymization

- We have a significant amount of anonymized digital researcher and citizen documents in pdf format.
 - Anonymized research copies
 - Copies issued for citizen access requests
- It would be nice to be able to anonymise documents that have not yet been anonymised, based on existing anonymisations - as a teaching data set. However, the archives cannot make a mistake here; under current legislation, verification must be built into the process in any case. How much will this facilitate the preparation process?

- The problems of integrating artificial intelligence tools into the workflow at the Archives of the Hungarian State Security
- R&D results should be developed into a stable working system and linked to other working systems.
 - Stable software
 - Supported software
- How and when do we integrate these partial results and solutions into our business process?
 - Make the user's work easier (this is not always the case)
 - Better results from users' work

- The status of integrating artificial intelligence tools into the workflow at the Archives of the Hungarian State Security

- In our 7-year strategic development plan, in 2019, we planned to exploit the information in our data file/digital asset using data mining and text mining tools on a data warehouse basis
- Chat GPT – 2020, AI-hype
- First AI project in 2020-21, objectives: OCR improvement, NER
- Second AI project with Alfred Rényi Institution of Mathematics since 2021
 - **OCR**, NER, Anonymization, RAG

Thank You!



© 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.

Slide xx: [element concerned](#), source: [e.g. Fotolia.com](#); Slide xx: [element concerned](#), source: [e.g. iStock.com](#)

