

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 Al-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

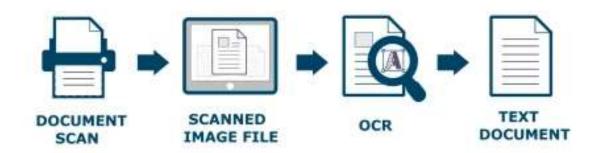
Beased on free text search

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
 - Impoves 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

Alfréd Rényi Institute of Mathematics



• NER





Implementing AI tools – improving OCR

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

D hught- 4 4 6 407 17 1266/1945/11.82	0 5 V 90997_003 w	V_90997_003.W	N.:	2070022	1476 2296 Revoluterpands	292		5 V_90997_003 w	V_90897_003.W	- N.	3379022	1476	2096 Restoregants	202
12/14	6 V_90937_004.W	V_90937_004 #	.14	3282454	1452 2258 filendue gaztie	202		6 V_90937_004.W	V_90987_004 M	- 14	3282454	1452	2258 fileschate pactile	200
* Reguer Lostdrasmig Revolvent 17 (21.)	T V 86937_005.W	V_30897_005.9	54	3456118	1584 2302 Rendsargapida	202		7 V 80837 101 a	V_90197_005.9	- 54	3456118	1504	2302 Rendstergateda	202
and the second se	A Line works through the set of the set o	(V_W293_D01 W		SILTIN	1431 2250 Freedompattie	1212	2	ELA MART TOTAL	N_10292_000 M		30,2210.2	1431	228. Perpisingette	1244
A Meptinded.gok Drazdiges Taracca an insettes surtation missis missis	0 9 V_103837_007-6	V_90997_067 w			1540 2335 Revolutergatide	- 203		9 V 100907 002-6	V_90907_007.W	. M.	753397/6	10.000	2335 Revolvegande	- 200
J z p p Slach Sykogy ellent binigyet, -maijten a lebrecent Adult deag Th 17/144/S. st. 1668 4tt Fabruar ht 7 mep au listerst houst, a sept-	10 V_90997_008.w	V_96897_068.W	N .	3461862	1520 2275 Renduergazda	202		10 V_90997_008.w	V_90897_088.W	- N -	3461862	1520	2275 Renduergante	202
	840					le l	11/43							
a ved sizement a 2p. 363.5.1.s. 40 1.v. jotilje slapjan a binbarag bagelis-	2.11						4							-
pinase mints of a Linking to any single any link an englishes regard any intert any	2													-
a relationsmet & 20.2005.1.1e. do 1.v.perity simpler a birthering and its rilies minet de a 11.0.0.00 for single regulations from the signification mileight commune forther meghanese a birease for the birtherin birght is non- vinguist diverts de unglisses a bireased	 A Hépberőságok Orez 	got Tanàcsa				1.00		oninceagor. Orsebacs Trevel	CER.:					
	7 -1 ' '70							PV: 1269/1949/1142						
	ACIT N. 1250/1946/Eaz 1	£					A We	gyne Köclérseség Nevébern						
# Peptirjeagok Grazagos Tandous a népbirjung itéletenek a fibintenest														
Elested réards a II. Bours. 20. 5 - 0 3. poptys elegipte contraining a sectorization a veglatt fördnitetdedt 1 /egy/ 44. bir Sour adradkit.	A Matyor Job revealed to Law		-						Folder from offshire cerego an e					
		os Takácsa az izgetés hiintette miat vár							nt, enelyber a debrecen répbin					
A nivigying samaisigi persesat ugyasinida a oddelar semilagi pers unat a fantiak a tulaani résiden ér, illetve visarsatasti a		bünügyst-mésikben 0 debreceni népbi tabruár hő 7 nepián téletet tuzott - e nép							o 77 rispan illefertit bozott - a napil	Wertstee # na	CROH LENSING O	onnoentee	ment a bip this is a line only	
		inčeltes miet a Bp 365',\$ i b porte el m					et expe		porte clepito e binosso e bol					
		 e. és-J c.pontis eleptin'a biobsés 							dentoff invitvance to kebby their throw					
I mil ok ul Å s s									Chemical subscription of a second research for Ch	Contraction of the second s	THE OLD WHEN		out in a raveleners	
		a alamian and bline when the interest of	1010040042-0				opening and a second se	hat.						
	Pittee miet és e il tin 28	 a alapján enyhités véget bejelemet sé otadot mévéres (Hebby-tel) térmelée 					Victoria de la composición de	tut:	a a application deleterarie a What	and a state of				
At the same tankes as staffren bindag ditts bintanig dayjad wer- distort tärydisisen hisronanger, indelse samt i trifelleriseget "agr heigtalse tärybelt täretkerteiset ann tahalt	P #8ea miet és ell tin 28 in leégi penesza folytés m	glanot nyivaxos fitiabbvisi i targyalas						net. Isbininingoli Orszeigon Teneros	a a néphérody réletensé a mbur	Ref Col				
As Grendges Tankes as eleffektu birdeng áltel e binterég derjad usz- 4.115100t tészáltászas hláryozságot, Amájos ég el, 1976 ilevinséget agy hegytalas tényhelt vövekbeztetési am talált	Pittee miet és e il tin 28	glanot nyivaxos fitiabbvisi i targyalas					Hungah	het: obkrunkgok Orszegon Tenecs bis reszerie A Nonov 2011e 2 p	ere- maintenine wilden antro-	NAME OF COLUMN				
Au Orenágos Tanáco az eleffeku byrdeág últal a bintonég Mayjaul ung- 4.1151005 tégyáltászon hláryoraágot, Jonahlyos myst, Joriellevenseget negy hejtalas tégyélt kevelketetésés men tanátt legy az elefrozi téálett tényáltást téálterdag elepjél elfágette	P žises miet és e il IIn 28 is leégi penesza folytés m v iszgálat alé vete és meg	glanot nyivaxos fitiabbvisi i targyalas					Averative Averation of the Average Ave	het obkroningolk Orszingols Tinnakos bio renzerti w Nickow 2011 – e 2 p mitobarrhetelek 1. (legy) ere bio	sonta alapján mejsemmisti - en tirre mésséki					
As Commigne Tarmace as electrica birdang ditel a bintonig displaul ung- dilative takopilakane historianakat, inakino eng at, leviellevane get negr heighalan displati takomikativasi ana tahaki ingo az aletrozu italati tanjatistat italteradas alapidal alitagette Tarvingt adratt a mahor forg as a binterada singidal alitagette Tarvingt adratt a mahor forg as a binterada singidal alitagette Tarvingt adratt a mahor forg as a binterada singidal alitagette Tarvingt adratt a mahor forg as a binterada singidal alitagette	P täsa miet és a il fin 20 in leigi penesza folytán m viszgálat alá vete és meg	glanot nyivaxos fitiabbvisi i targyalas					Annesk Vitako 7 retai	het obkraningoli, Orsziegon Tinnikos bia renzerini ili Nonov 2011 – e 2 p m Obvarteteneti 1. (egy) (evi bia logyika a veneti ologi peneti sati	sorita alapján mejsertivisti - es tsrea méndéli Pagytsztrán a védelem nemmele					
As ferminger Tarnkor as wirdfrikt byndang filml a bintonnig dispjaud une- 6. Holinoit täöykiläkken hiäryornkapt, hemäinen kui. 1000 tilevenuset magy heiphalas täykkelt tähykiläst ibäirennan sinkit igg as sinäfördu itäisti tähykiläst ibäirennan sinpikul alläypette Tärvönyt adrumti a näyöttönän äs a täänttöö tiusekajanit a 15 mm ja 5. Aben fegisti anyagi samakada, okot salaitoita magamet valysalanon 5. Aben fegisti anyagi samakada, okot salaitoita magamet valysalanon	P bisse mint és e li 11n.20 in leégi penessa folytés m viargábát díá vette ás meg e 7 kéleset * (glantof nyiñviðson i fálabbvítei i tárgyalás norða á krivefkierð	n-				Ausent Vitako 7 retui Autot a	het obkroningolk Orszingols Tinnakos bio renzerti w Nickow 2011 – e 2 p mitobarrhetelek 1. (legy) ere bio	sorita alapján mejsertivisti - es tsrea méndéli Pagytsztrán a védelem nemmele					
As foremigne Tankon az eleffeku birleng iltel e bintonég Meyjewi wez- dilutitott tégyilikénen hikryornagok, jaméh ne my i jeriellevenueget negy néghelmen tényésit kevekketéses an tankit igy az elefrozi téleti tényétiket tédirerdes eleptési elfégedte Terrényi séritt a néghtrőnég és a bintetés tiszhésénel a (1 m) terien 5 dén fordall a mysdi mensékéi men terien tisz és a men telégi	P tisse mint és e li 11 n.20 in leigi penessa folytés m viargélet sik vete és meç ex 7 kelenet * A Béphinpségok Országo	glanot nyivaxos fitiabbvisi i targyalas	er-				Annex Visito 7 vita Avita e Ladar	het bistrangok Orszegon Tendes bis rozsten i Norsov (0.1 – e.2) artoburielenet 1. (ogy/ en bor log/wisz a reminiségi bieneszt introlek intérnerő részben i 2 sk 61 és 3.	sorta alegión mesorravial - en tirro ménéki Nagyaaméné - édelem anarosik Siék-e-casandésit ja	b Dute	nyinfinctions (faircess	iger, havilyes short	
4. Creatigne Tarnich az eteffektu berdaig által a bintonég Marjaul ung- áltistnött tésziltésen hiéryorasgot, Annélo es mat kiér ag négralas tényisti tészetkettésen ann kiéki igy az eleffozi téslett tényittése téslernéss eleptési elfégette Torvenyt ag zett a négbiréseg és a bintetés tiszbiaségette a fisk ze 5. dem foglati annyal samakége koks tiszbiaségete a fisk ze 5. dem foglati annyal samakége koks talattés eleg met a tisziséget metingette a bintésegt kövültésegéte a ugy a váda test migsigeru tors désel antésége.	P tisse mint és e li 11 n.20 in leigi penessa folytés m viargélet sik vete és meç ex 7 kelenet * A Béphinpségok Országo	glence nyivaxos (tietove) i tegyelés nada a következő T enácsa a népbiróság féletének a lób 55 3 -e: 29 ortja alegján megsemmini	er-				Annank Vitako 7 nita Antar Actor	het bistrangok Orszegon Tendes bis rozsten i Norsov (0.1 – e.2) artoburielenet 1. (ogy/ en bor log/wisz a reminiségi bieneszt introlek intérnerő részben i 2 sk 61 és 3.	sorita alapján mejsertivisti - es tsrea méndéli Pagytsztrán a védelem nemmele	b Dute	ryŵdiastener	hileyoosi	igar, hanilyas sigar.	
A: Aveniges Tarmics an eleficitu bridaig ditai s bintoneg Misyjaul ung- dilativat sésydiláten hisrochagot, Annalo as get, isticieltarana et agrikejtalat sénydiláten hisrochagot, Annalo as get, isticieltarana et agrikejtalat sénydel tentte ténydiláta töltkeráng elegyikel elfőgette Türvényt sérint a népöltőség és a büntetés timskégénig a filósizat S dem fegitelt anyad samakégé vört talástotta egyikel elfőgette S dem fegitelt anyad samakégé vört talástotta egyikel elfőgette S dem fegitelt anyad samakégé vört talástotta egyikel elfőgette S dem fegitelt anyad samakégé vört talástotta egyiket elfőgette meringető el titokegét avallenyetes se igy s vádlatása migelgoru binte alasal ajtokla.	P blace met és el IBn 28 is lefej penssa folytén m v szyblet dik vete és meg ** 7 keleset * (A Bépkipségok Onstégo k iszábó részére 8-Bábor vedict fővűrelesel 1 / égy	glence nyivaxos (tietove) i tegyelés nada a következő T enácsa a népbiróság féletének a lób 55 3 -e: 29 ortja alegján megsemmini	n- Kreskul				Antesta Vitako 7 reita Antesta Az Orritaria	het bitmingole, Orszegole Tankiss bitmingole, Orszegole Tankiss bitmingole, Annova 20, 1 – 2 / stroburntenbelt 1 / orgy/ kalo bot logoksz s knownere felszben kö k fől kel 3 széggel Tankiss az knobbek a	sertig alaption menorewold - en Steen onlineitii Huggestricht a de geen commente Methon of constitutiity Method a betraceitig alaptinal a	b Dute	ny white the set	hileyana	itur, haniliyas sigat	
A: Greenigos Tanzos az eleffeku birdaig illel s bintosoig kisjaki usy- 6. Holtott töyüllikisen hidrigornágot, Jonah ovak ok. Hortellevenseget neg helghalas tönyösil körnökatása inginál előgedte Türvényt adristi s négobirősár és s büntetés tinuságainal előgedte Türvényt adristi s négobirősár és s bűntetés tinuságainal előgedte Türvényt adristi s négobirősár és s bűntetés tinuságainal előgedte hortageles s bindeségi könüllenyektes és gyr válós ést ujudgöru histotasásal aujtotága. Mihás es öbő sz öteságár Tanisz s héppirósás taltatásása a felettéségi tarvénégi könéségi könöségi tanisz s héppirósása az és előtettéségi s felettéségi s könöségi könöségi a tarvénégi könöségi a tarvénégi könöségi tarvásásá az előségi a tarvénégi könöségi a tarvásásása a tarvénégi könöségi könöségi a tarvásásása könöségi a tarvásásásása a tarvénégi könöségi könöségi a tarvásásásásásásásásásásása a tarvénégi könöségi a tarvásásásásásásásásásásásásásásásásásásás	P bleannet és al (16-20) is lategi panasza folytás m viszgálat dík-veltés és meg *** A blebent * t A blebinpságok Onstago k kobbo részéra II-Bibrov vedichtfölüntetését I (eg) A hapugyást sesemiség p	glance njevaco s tšebovali i tegvalao tota a kovarkato T anácsa o něpbiroság fili letimek a töb Cs \$ -a 20 ortin alapjím megareminiá - dvo botním meseká	n- Kreskul				Kristek Vitelio 7 nito Kristen Kristelio Vitelio Vitelio	het biblin nigolik, Orszengon, Tinnics biblin nigolik, Orszengon, Tinnics biblin nigolik (Tinnick) hittiblik, nikowa (Tinnick)	sontja ningstvo mesonovovali – na Gree ordinalski Dragostatolon a védijilam namratski Slobov vrastanikasit ja otolog obel a čestatoving elapitel a del nest taloži	b Dute	v, blinsten	hileyooni	itur, henilyas sigat	
 Ar Gressigne Tarmos ar eteffrön biplang illel a bintorsig kerjaul ung- biltittet tädysiltäeten hiärporsigni, jonalu on ang ol, jorisiltersauget ang reistalas tädyselt törnökente son an talait. Igr ar elefronu itälatt tädysiltäet itälkandas alagraal elöspette Turvinyt adrintit a näyderösäy är e täntetäs tiusekisänäl elöspette Turvinyt adrintit a näyderösäy är e täntetäs tiusekisänäl elöspette son fudlalt enymel samasadagi värat talaitoitte se gaart biltelasion, miningette elä suitekset talaitoitets es uja välatitet uja tänte turvinyt adrintita. Alter turvinyt adrintitas Turvinyt elängitteta elä sittä alasta instalasta elänteta. Alter turvinyt elängitteta eläntetä eläntetteta eläntetteta eläntetteta eläntettäänä elänteta. 	P bleannet és al (16-20) is lategi panasza folytás m viszgálat dík-veltés és meg *** A blebent * t A blebinpságok Onstago k kobbo részéra II-Bibrov vedichtfölüntetését I (eg) A hapugyást sesemiség p	glantot nyk-ákos i tilebitukai i tergvalás note a koverkező T enécsa a népbiróság félélének a töb Ita 5 -a 20ortja alegján megsemnisé - révu börtöne méssikő	n- Kreskul				Australia Vitalia Tretta Agi On Vitalia Vitalia Vitalia Vitalia	Inter- tion interpols, Orszangon, Tamaras to reactor all Norsov (2011 – 92, arritopuntationed 1. (egg), and opo- ingular a summaling planness of taking and the second planness of taking and taking and taking and taking and taking and taking and menologist maliptalism langupati Acquirely and metodologist maliptalism langupati Acquirely and exclusions	sontja ningstvo mesonovovali – na Gree ordinalski Dragostatolon a védijilam namratski Slobov vrastanikasit ja otolog obel a čestatoving elapitel a del nest taloži	b Dute	syðfinstem (hileyaan	itust, hankiyos ségut	
An Oremaigne Tarmice an eleffektu birdanig ditel a bintosnig displaud uns- dintatuota takoyallaksen hidroyanakot, junkijuotang displaud uns- magr heighnalse fanyaksi kiraktentisei ana tainit. Ingr an eleffort itäleti tänyällänä itälternän atappial elfägette Tärvänyt adrintit a näyötrösnig än e bäntetäs tiuseksinil a 15 M-28 5 Mone fagilalt arynyd semakangi okut valaitoita anginert belyteleny- mistagette a bintänän kiraktenyt kuulangintakset ei yötä ota tuiseksinit tuiseisen taistata provei kuulangintatoita ei tuiseksinit a 15 M-28 5 Mone fagilalt arynyd semakangi okut valaitoita anginert belyteleny- mistagette a bintänän kiraktenyt ei aja väänä tuiseksinit a fagilation tuiseisen autotas. Albit es okoi as Branagee Tamics e näpötränas tääntösette a fagulatione atamistagette väänän proveitariat muonamistation ohe endettettet	P blaca met és a til för 20 in leitig ponessa folyten m viszgálat alle velte és meg *** * 1 A Béptinpségok Országo k ksztöb részér a 1- Btorov veldot förüvisetéset 1 (eg A nápagyét szemisség) azet a femlék a kümend ré ladóksála A zőrszágo Tanica az él	glantot nyk-ákos i tilebitukai i tergvalás note a koverkező T enécsa a népbiróság félélének a töb Ita 5 -a 20ortja alegján megsemnisé - révu börtöne méssikő	n- Lanashad B B panas" menga				Kinimk viteRo 7 nitta kontri e Agiton kinimke kinimke kinimke Kinimke	For the constant of the second sec	soritja nispjalo responsovati - na Store melaniki Urogostavlon a odoblani sumovan Slebon otazanikatil ja stanigoškel na bantarali ja etanigoškel na bantarali ja det nem tuliki miseo niegiske vitogedta	an Bayan nyegi ya In Essana	ny de Banstrom I	hileyean	igur, hendyrn ságat	

the OPE OPER STRATEGIES.







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.

Törzsszám	~	Azonosító 0	
lkt./nyilyt. szám		Feldolgozottság	
T árgy 🛛			~
Ország			¥
Megye			¥
Település			
Nyt./egyéb ikt. sz./ Nyitószám			
lrat évköre			
éma évköre[
Átvevő			
lap Al			
OCR			٨
	szabó lajos		٨
Szemely			Λ
Szemely Intézmény			-
Intézmény	4		^



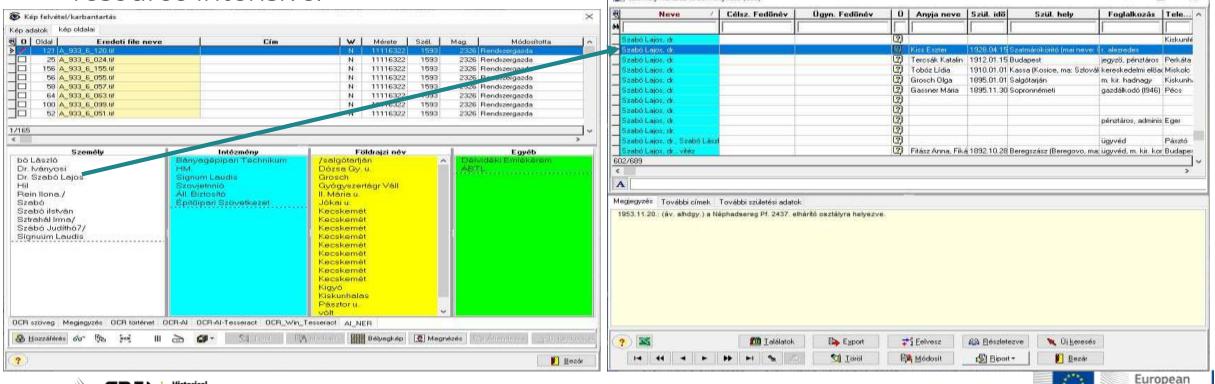


Implementing AI tools – NER

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.

×

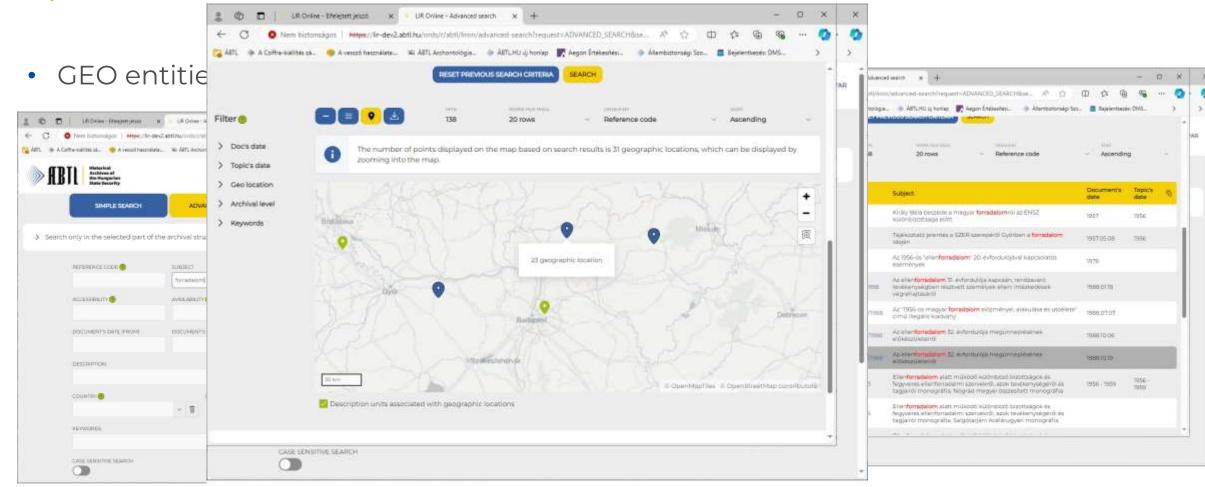
Commission







Implementing AI tools – NER

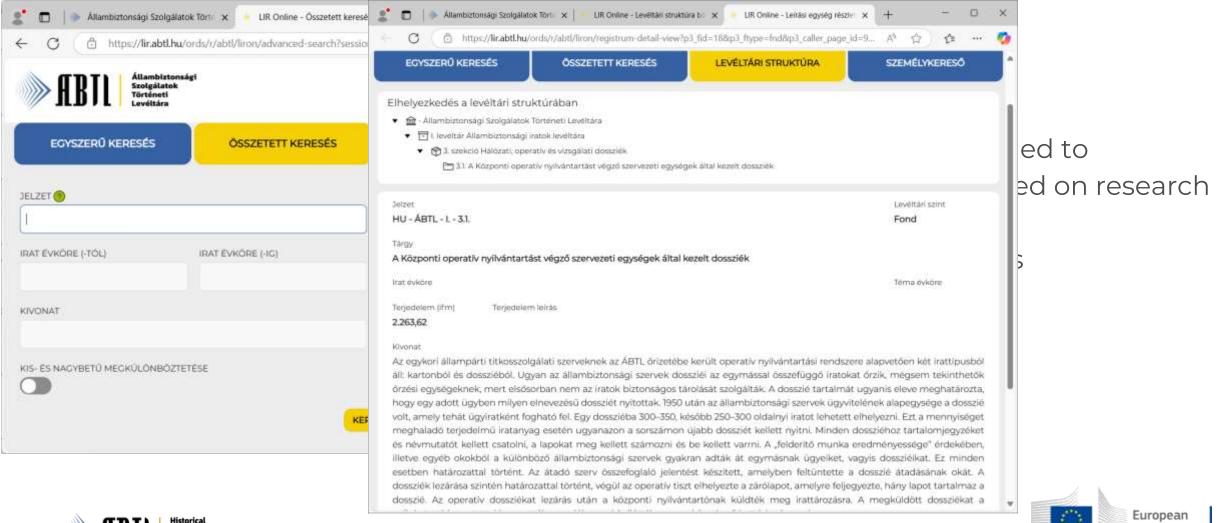








Implementing AI tools – RAG



Commission

Archives of the Hungarian State Security



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 <u>CC BY 4.0</u> 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