## Zoltán Szatucsek Unlocking the potential of Handwritten Text Recognition: European Collaboration in the National Archives of Hungary

ELTE – Chung Ang Artificial Intelligence Conference July 17-July 18

## Short introduction



- intro
- 1723 established (happy birthday
- 21 archives at 72 locations
- 667 employees
- 300 kilometers of records
- 59TB of digital
- 36.917.310 files
- OAIS digital archives system for LTP

Oldest record in NAH is the renewal of the founding charter of the Veszprémvölgy from 1109 no dince nouveration mit even perpiente officialitimo par colutiono penovaries 7 bos prustegui de monas apa prenerita montalia Carla ane penourront bechan Chi wei printegui fei P tephana de onitous princiens end nond teri crashinto cas afterness effer multime comanabile accellicate guarida laciarconi funter anor cu sulto melgane rane of cervinana he weers pulear over an nout here of ca commerce one monsterry illud ne ani anerougilia illa empore obennue ande ofte pullene pequi el penouse fehet fune es que muentale e usual delis fine one contradie nonelimente quing ettaj i que tully regil et confentu bezprementil architactie pler mante & uertar a manifuno revi columbana reference villa zadberin cultin mentionibi fedita ad iplan unta priment e conun hui cille habitator ib: Inla zamon meta danubin dedie ny lef Arphan cu quanta manhomb; sepostea dedit in dedie tione cete mouna manhond meade ulla Se eria meade ulla vi pulatore. Tora que plus ullo dominin abbante e Tore ien forreiden ulle ernaula port en fore nauderil monaleris it deche ena milita milito er pe uplan milita deche polin u fild ad op partin fed an ef comunit tem into hora fild haber preserione Augioni. Tercis uilla et muna day: annea kindla keva crom in the un allin opens condur kindla crom nor off ga commit nee lice dicar un que darda adailque Sorana unla eff forlout bar commung e en popto quecherfu epi Verana una epidera n qua nulla habe fili podrac comune na ci plo abbande Jona nulla e genecar cui cia comune Promite utili ualti non haber abbanita l'generaticat nin totimo do Les comunitores esperiariosas ce dual nationel totropie et uni corna core ce uni lagenarii cubose Ceii oni fe nel unitore di sacore. Ce mulla polotinic haber abbanita initi unnore ci mnea fus her bee dech for por in portu modaca, promone buloni adop monului fer mare formentin ver aut hultenti un ra lingua ale tori monalteri grece lenoni de adnoranim ur se concordia neterif anoui agasfeeretter cer ratido uer ratif. lea uvra mueltigacione in detiente pi sponia une quela mnouo-que non muenume macres. and los active accrementar in polla april. a metic, and dould de hil que in hos builters continent officiers mauere maledie tron min mediel di sele marie avoiil aplois avoinin fidelini che he mineceri puilegio ienorai e higtmean where pean de pustione grotie de plana

## Why do we are so enthusiastic about AI?

Archival objects are unique and need context



## European Digital Treasures - Crowdsourcing



- Norway National Population Register, 1960-
- Portugal General Register of Charges of King John VI, 1792-1826
- Hungary National Census of 1828
- Malta Immigration records, Customs Department, 1905-1966
- Spain Passport Issuance Register of the Consulate General of Spain in Buenos Aires between 1936 and 1939





#### Találatok szűrése Egyszerű keresés Összetett keresés Kifejezés MI magabiztosság küszöb 50% Q Keres Ország ~ Magyarország (177508) 194649 dokumentumban / 194649 érintett oldal (1,13 másodperc) Horvátország (16886) 1 Túl sok találat, 5000 került megjelenítésre Megye ~ Szabad királyi városok (12883) Bihar vármegye (8158) lista nézet ikon nézet Pest-Pilis-Solt vármegye (7517) Nyitra vármegye (6812) Zala vármegye (6766) 19ha több Média todon The state of the second of the Kép (194649 db) Névtér használata Ne használja a névteret Személy HU-MNL-OL-W-21-6-N-26-Vas varmegye-465 HU-MNL-OL-W-21-6-N-26-Vas varmegye-465 HU-MNL-OL-W-21-6-N-26-Vas varmegye-465 HU-MNL-OL-W-21-6-N-26-Vas varmegye-466 HU-MNL-OL-W-21-6-N-26-Vas varmegye-466 🕓 🛢 Az 1828. évi országos összeirás 🖝 Magyar 🕓 🛢 Az 1828. évi országos összeirás 🖝 Magyar 🕓 🛢 Az 1828. évi országos összeirás 🖝 Magyar Testületi 🕓 🛢 Az 1828. évi országos összeirás 🖝 Magyar 🕓 🛢 Az 1828. évi országos összeírás 👉 Magyar Földrajzi

#### Rendezés

Relevancia szerint

Sorrend

#### Megjelenítés

Hierarchia

**Q** Keresés

Konfidencia 🗕

+							Reference in								ないであったい	4										
Q	Nomen Juris	dic	dict	ion	nis	100	zi	an		ist.	ifer	rei		Nor	nen	Lo	ci d	cons	crip	oti	Re	") ín			tin the first struct	
	1-	2			8			N. S.	Contraction of the second	)***				10-		7		Contraction of	11-	199207		2	13	5	14*	
	AND DON	um utri- tine, aut e als ex-		v	7 i n	e a e	•			naria, uneta			P e	c o		r	a.	Pe	c u d	e 5.	S y I	lvae.				
	Nomina Contribuentium.	ruene Contribuenti usque sexus cònjug non conjugatae, ind	Capaces Metret. Poson.	Lucrum dant,sin- gulam Metretam	Posoniensem as- sumendo. Remirunt Fos-	sorres.	Frocreant post unum Fossorem Urnas.	Modium currandre- tium Procreationia post unum Posso- rem obvenium.	Falcastra.	Laerum daut, singu- li Paloustri exten-	Boves jugales.	Vaccae foetae, nut matres.	Faccae steriles.	uvenci et Juven- cae 5. amos su- perantes.	uvenci et Juven- cae 2. annos su-	el curritori el ar- malater 5, alboa	pi curriferi et ar- sentales 2. anaos uperantes.	res unum annum compleates.	tigeri anam an- nam complentes.	ompleates.	une occupant ter- um capacem.	Inferunt, annoannum pensante , e Glan- d'ustione et Quae- stu Lignorum.	Kdurotham	5	Observationes.	
		Te	Nro.	R.O.	Xr.	Nro.		Rfl.  xr.	Nro.	Rn.	xr.	N	u	m	e-	-	1	0	3-		~	Rit.	x-   Rfl	. <u>xr</u> .		
;	Ofif Janie 7							•							in the	-			-	1		-		1		
12 177 196	Steger Mallyar	2	-	195								1		1	-	- 11	1	-	-0	5	-	1.0	-	1		
Section Production	deserge, Martin	S STATE	+							1		-	•		1	1	- 1		<u>`</u> 0'	-		1		-		1.1
4	Bruther bring	e		1		-	-	• •	-	•	2	-	-	-		-	-	$\underline{\mathscr{O}}$	-	-	-	-		-	A share a share was the	
5	Chif Mattane of	1 2				-	•		-		. 2			-	4	1		-	-	-	-	-				
6	Reman Maria			-				•	-		-	1	-	1		6	-	1	-	-	-	-				
7	Acger Mihal	-		•	-			•			1	1	1	i	2-1	1-	-	-	-	-	-	-				1 Carlos
8	Sochen en Marga	14	-	-	•	-	-	4	-	12	. 6	10.000		Ð)	-	I-	-	-	-	-	-	-				
9	Brukner Markar	20 157 CH					1 2 2		A COLOR	-	2	T				-	-	-	-	-	-	-		-		
10	Row anor it	0010303							TEL	-		E	1	1	1	1		-	-	-	-	-				D. C.E.M
it	Heid James on	15 1 2 2							-	A STATE	4	-		-	-	1.	-	-	-	-		-		11		E E
12	leger Mihal d	-	150	The state					(	3		-	-	-	2			1	100	-	1		1			
A CONTRACTOR OF	ang Mihay			7.19						2	-	-			1	1	-	-	-	-	-	-		1		The level
14	Lang Janvene		1	-				1						-	-	1	-									1 4

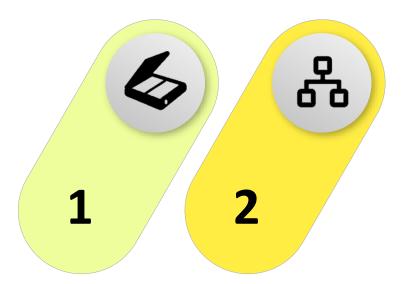
Pfif János öreg Szeger Mátyás öreg Reperger Márton öreg Brukker György Pfif Mátyásné öreg Czerman Mátyás Szeger Mihály Hovsretter Mátyás Brukner Mátyás öreg Roz János öreg Neid János öreg Steger Mihály öreg Lang Mihály Lang Jánosné Arthofer Mihály öreg Arthofer Márton öreg Czáfel György Czáfel Mátyásné öreg Schaden György Kapel Mátyás

## How to continue?

- Recognized the importance of volunteers
- Learnt the basic concepts and some not so basic (KWS, hypos)
- What is easy and what is difficult
- Instead of a PoC project service
- Contractors are not 100% reliable
- There are open source components available
- Transkribus is expensive
- -Archives' use case is different from DH



## Classification



Classification

• classifier

• template builder

## Scanning

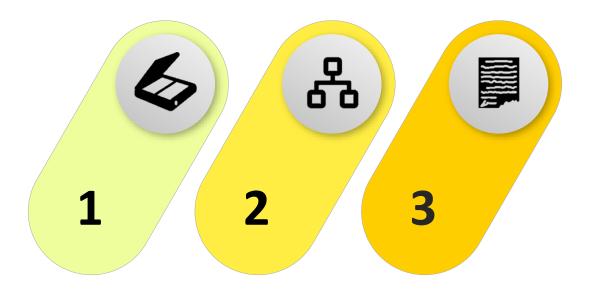
- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning
- archiving

15 szám. Helt Abomyban 1895. Kelt: Abomybanlever my oler száz kilener ven 1895 (ezerstadik) evi sktober ho 13 oktober-10 13 (-) napján. /tirenharmadik mapyan Hegjelent az alulirott anyakönyvezető -- A lexaras és ar unes . Goncrol Palhelyek kitöllése utan helyek kitöllése utan ele az alairois előtt-kitient, hogy a veimek likiselye: Aborn/ Goldeshmied tanya/ kiuroa khal nem lettek sa sit alolulott anyakomyoverető lentő megjegyzi hogy-T, a bejelentist vei maler veir kilomer -1, a begelentest und is bejelentette a következő születést: nyalozorazkelenczven és bejelentette a következő születést: ölinik mi aktaber (esaládi és utónere: Gönczöl Pálstodik evi aktaber tiven harmadikan tor. tent .-vallasa: romai katholikusállása (toglalkozása): gazdaságbom éves - II, as aty a husson hirseled lence eves\_ lakéhelye: Abony Boldschmied tampa III., as anya kuoronszületéshelye: Aborny. életkora: 29 (\_\_\_\_\_ ofer eves. my , a sailetes excompolar ( esaladi és ntóneve: Molnar Julianna saar kilener ven stodik in oktober tiren stode vallasa: vomai Eatholikus. ken tortent del elit állása (foglalkozása) : \_\_\_\_\_ let negyed tiren beto rakor lakohelye: Abony születéshelye: Ale Img crol Palv. K életkora : 28 (mathe Freedy. helye: Abom\_gahomyvaerato" Az elso" podanynyal vali megegera's hiteliie, Vielt Aborybar 1895. ok., tobr hi 13-a. Mathichiad ideje: 1.895(ezer ---ivi aktober no 10 (napjának délelőtt %/3-) órája. neme: fill ; vallåsa: vomai katholikus Machily utóneve: Pal augalling vorgets" 11.13. 79/1920 Megjegyzés :-Felolvastatván -----Az első példánynyal való megegyezés hiteléül helybenhagyatott és alairatott. Kelt Abompban 1895 évi Oktober no 13 napjan. mothe forsefuk Mathegoicef Goncrol Palls anyaköny vezető. bejelentö.

395 - nyolerszar kilenervene oktober - - 1.- ( első ----gardalkodo loldberto; gardalkodo/fold birtor 44 (negyren negy 44 (negyren negy Landsing & Gizella izraelita-26 husronhat\_ 895 my alezozár kilenezven oto September 24 churron negy saga a felmutatott harassagi bi stober 10 1.

Részletek x Típus: Kézírás  $\mathbf{v}$ Adattípus: NÉV - Földrajzi név v Nyelv: Magyar  $\sim$ Statikus érték: Ok . . . alivent

## Segmentation



#### Scanning

- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning
- archiving

### Classification

• template builder

• classifier

- Segmentation
  - layout analysis
  - polygonization
  - baseline detection
  - word segmentation
  - skipping

II, a srilete's werne our internet in the second of the se		
and my alice so as interview 188 and the second of the sec		15 szam.
and my alice so as interview 188 and the second of the sec	10 11 1	
Attended for abilitalis high of the for the former of the second of the	Helt Abomyban 1895.	Kell, Allomyban
Attended in program inter a line and a paraground in the left of the series of a since interest intere	otor my oler stat hilener ven	1840 (eet
At leveniss is an ince "Gimeric Reil helyek kitölleve etiin helyek kitölleve etiin hikis a alaina selött hitient harge vainut hikis a legen arain tette sai 1 alailient angakimpereete sai 1 alailient anga sai 1 alai	Hirenharmadik maria	verouer-h 10 14- e napján
Ar obs polamnyal vals melograds Ar obs polamnyal vals melograds	- A lexaras és ar irres	Gineral Az autorit anjakonyrrazari - Ediati
hitenst hvorg a vanne better proverkend men letter hitenst hvorg av vanne better hitenst hvorg av vanne better hitenst hvorg he hort hvor -5, a befelentest ner -5, a befeletest ner	nelyer kitollese utan	a kinek állása (foglalkozisa): gardaságban éves
travakkhal 'nem letter inner talanti angen anniget inner talanti ter inner talanti angen inner i	- at alacrastate	cseled
tento Universitati legi universitati legi universitati universitati della legi universitati universitati universitati universitati universitati universitati universitati della legi universitati d	szavakkal nem lettek	Inkohelye Aborry Goldstonnied turnyal
ny oleven vie kielemenen in state in st	Ruton. Hours arest bei	
nyaler vier kelemer en under interes i	tento megjegyzi hogy-	stenely carero tomec
Ar ore his optiminger all treen harmadikain ter tenet tenet II. ar aty a historn her II. ar aty a historn her II. ar anya historn my olae abas II., a saileted evenny ole historn 199 interes was histore histore Albony interes was interes wa	molor wink .	da hajalaatatta a Liinathani antilottat
tienen harmaaikain tir tenet. II., ar atya tusaan hu tener word II., ar anya tusarn myolae abes II., a saiteese atusarn one II., a saiteese atusarn II., a saiteese atusarn one traiteese abes II., a saiteese atusarn isteen 29 II., a saiteese atusarn one state hilener wen ätöde ten törtent ale elekt tet negyed turen keto alese 23 mathe forese b saite forese b isteen 19 Mathe forese b isteen 19 Nele Story ban 195 ok. inne fell allan voinai katholikus tine 13 m Uele forese a tine 23 m Mathe forese a tine 13 m Nele Story ban 195 ok. tine 2 Pall Magegris. Ar elos palsannyngel valo myokan. Mathe 19 mas angegreds Myokan. My tenes palsannyngel valo My tenes atus atus atus My tenes atus atus atus My tenes atus atus My tenes atus atus My tenes atus atus My tenes atus My t	sodik evi aktaber	esaládi és utóneve: Gönczöl Pal
II. ar aty a tusan he terner even III. ar anya tusan he itilia lotaliaisii gardasiagbau even itilia arga tusan he itilia lotaliaisii gardasiagbau even myolar abas II. a saiitetei uremoore tulian itilian itilian itilian myolar abas II. a saiitetei uremoore tulian itilian itilian itilian a itilian tortakee abas itilian tortakee abas tilian tortakee abas itilian tortakee angada curem keto valadi is nimeri lathalikee miathie goraed a curem keto valadi is nimeri lathalikee miathie goraed a curemoore tulian tortakee itilian tortakee valadi is nimeri lathalikee itilian tortakee valadi is nimeri lathalikee miathie goraed a curemoore tulian valadi is nimeri lathalikee miathie goraed a curemoore tulian tortakee waan aharingoreesto valadi is nimeri pula alla itilian tortakee itilian tortakee	tiren harmadikan tot	, , , , , , , , , , , , , , , , , , , ,
tenner even III. ar anya kusern nyolar abes III. ar anya kusern nyolar abes III. a sriileteis wernoore teilail in iliane Allony in oktober teren ötöele kin törtent del elött til miane Vonai Valhalikud hin törtent del elött til miane Za mathi förned i Vale del pened i valla vonai Valhalikud til mathi förned i valla vonai Valhalikud til miane Za mathi förned i til an in törtent del elött til miano Za mathi förned i valla vonai Valhalikud til an in törtent del elött til an in tortent del elött til an in tortent del elött til an inter vonai Valhalikud til an inter vonai Valhalikud magna hängen vala i til an inter vonai katholikud til an inter vonai		
III. at anya kuston nyolor abes III. a stanya kuston in oli abes III. a stanya kuston in oli abes in oktober teren stoar in other stoar	- the are aty a kusson hi-	
II, a srilete's werne our internet in the second of the se		E lakoholy Abom / Goldschmind tama
II, a snillete's werning over talking in the interes shall be undered Molnor Sulianna invite kilener ven ötörde to ter oktoles teren stord ter oktoles teren stord tet megyed usen keto """"""""""""""""""""""""""""""""""""	they any a kuoron	- szülatéshelye Aleman
Ar oktober tren stode to in oktober tren stode to him tottat ext ext a total tet negyed usen keto mathe forest is allow mathe forest is any aktograde of the delay of the total Mar delay of t		Aletkorn 29 - Coves
Ar elso palsannyal valo megogrados hieles. Ar elso palsannyal valo megogrados hieles. Hieles. Hieles. Hieles. Hieles. Hieles. Hieles. Hieles. Hie	In a souletes exerned	esaládi és utóneve: Molnar Julianna
hen tortent del elitti in ilian (logializzan) tet negyed uren keto ale or mathe from il falle angehengenesto de elso pellangnyet velo tet 20 pellangnyet velo helde from ilian del interest tet allas: vomai hatholikis angelity volute tet allas: vomai hatholikis angelity volute tet allas: vomai hatholikis angelity volute tet allas: vomai hatholikis angelity volute heldestatavan	eve oktober tires itig	alle romail totlalit
Az elso palannnyal valo magogrado hade at a series Az elso palannnyal valo magogrado hadeal. Az elso palannyal valo magogrado	teen tortont d'a l'11	
Jonce il Palle mathie gover il Palle any akingonai li Pallengogel oele Ma else pallangogel oele magemai hielelie : in Patte Story ben 1995. ok. in Majaak al Vatte II allas: romai kathalikes angelige Story ben 1995. ok. in Majaak al Vatte II allas: romai kathalikes angelige Story ben 1995. in Majaak al Vatte Story ben 1995. in Story ben 1995. in Majaak al Vatte Story ben 1995. in Story ben 1	rale negyed tiren keto-	
Miathie govers is alline intervent in the line is all any or all intervent in the line is all any or all intervent in the line is all and all little is a line intervent in the line is a line intervent is alline intervent is all any intervent intervent is all any int	(1) 人人工具有效主要工作的需求工作的主要工作的需要工作的制造的主要工作。	A CONTRACTOR OF A CONTRACTOR O
Az elso pellannynyel valo magegyado Az elso pellannynyel valo magegyado hiekel. Az elso pellannynyel valo magegyado	Joncrof Pally	
Az első pellennynyel vels megegyezés hibéki. Az első pellennynyel vels megegyezés	mathe forsef v.k	
Az elso pellannynyel való magogyados htelefel.	the also moral and	Tileje : 1.8 96 (ezer
Hilde Story bar 195 ok. tote hi 13- in Machi ford angelety wight i Maging 201 - Allin : romai katholikus angelety wight i Haging 201 - Allin : romai katholikus Az elso palannynyal valo magogyado hisilia.	mogeneros hitoloist	
Az elso pelitanynyal való-megegyazós hitelési.	Welt Aboryban 1895. ok.	The second statement in the second statement of the se
Az elso pelitanynyal való-megegyazós hitelési.	tober his 13-an Marchie Goined	neme: fill ; allása: vomai katholik.
Az elso pelitanynyal való megegyozós hitelési.	( A A augality variato.	atóneve: Pat
Az elso pellanynyal való megogyezős htelölül	· = · /1/3.79/1920	Megjegyzés :
Az elso peldenynyal való megegyezős hitelédi.	Cine Stall	
Az elso peldenynyal való megegyezős hitelédi.	Carry Contraction of the Contrac	
Az elso peldenynyal való megegyezős hitelédi.		
Az elso peldenynyal való megegyezős hitelédi.		
hitelöll. relybenhagyatott és_alawatott		Felolvastatván
relybennagyatou es autovalou		
Kett atbomyban		elybenhagyatott es_acacratott
al gip in gktober og D majan.		and the second
( ) Matheforer Mathe Forsefut Goneral Dallas	( Thatlegoief	Mathe Forsefuk Goneral Poils
ayakinyverea. bejelenik.	( = E =) ( prestations to	payakonyvvezeto.
Com see	THEY KENT	





Helt Abompban 1895. 200 exer my oler saar kilener ven otadik) evi sktober ho 13 twenhormadik mapjan - A lexarces es ar irres 600 helyek kitottese után de ax alairas elatt. hitient, hvory a kamok stavakkal nem lettek kutva. Ugganaxert beje lento megjegyzi, hogy-- I, a bejelentist eret myolcz szaz helenczven 1200 vodik evi aktaber toren har sils

Flett Abomyban 1895. ever my oler stat kilenerven stadik evo aktober ho 13 (toxenhormadik) mapjour A lexaras es ar unes 000 helyek kitollese utan - de sax alairas elotthitunt, hogy a ramok stavakkal nem lettek kutva. Ugyanarert beje lento megjegyzi hogy--I, a bejelentest erei myolex wax helenexven 1200 vodik eve oktober tiren harmadikan tot. tent\_

## Preprocessing

# よ

#### Scanning

- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning
- archiving

## Classification

• classifier

- template builder
  - polygonization
    - baseline detection

Segmentation

• layout analysis

- word segmentation
- skipping

#### Preprocessing

- binarization
- deslant correction
- noise removal
- tation thinning
  - blurring

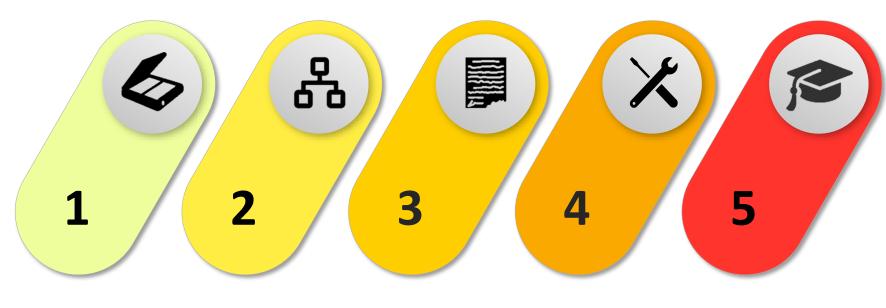
Helt Abampban 1895. ever my alex stax kilence ve stadile en aktalo hol twentrownadik Inapja A lexards es ar une elyck kitaltese uta ar alawaselott. hitient, hogy a dram navakhal men letter autra Ugyamaxert bas ento miegjegyti hogy lefelentest eres yoles was helener Frick eve aktabes ween harmordikan to ent II, an atra husson hilence eves and anya kuoron. myolve aver sculeted exerned oktober trendtod en tostant del cliff at negyed tixen keta

Helt Abanyban 1895. lever my oler szaz kilence ven otadik) evi aktaber ho 13 [tirenharmadik] mapjan - A lexards és ar irres helyek kitöltése után - de ax alairas el att\_ kitient, hogy a skamok szavakkal nem lettek kutva. Ugyanaxist bejo lento megjegyzi hogy--I, a bejelentist eret myolez vráskilener ven ötödik évé oktobes tiven harmadikan tor. tent .\_\_\_ -II., an atya turon hi-lence eves. -III. an anya kuoron myolar eves. II, a skilletes exerny of skak kilener ven ötodi k ive oktober tiren stode ken tortant del cliff tet negyed tiren keto Mathe Jorsef v. k any akomy versto Uz elso" pelsamy ny al valo" "Melt Abory bar 1895. ok. tober ki 13-a Mache Spieg " augating voordet." 11 13. 79/1920

Gönczöl Palls?

Gancrol Palls

## Annotation



#### Scanning

- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning

#### • archiving

#### Classification

• classifier

- template builder
  - polygonization
    - baseline detection

Segmentation

• layout analysis

- word segmentation
- skipping

## Preprocessing

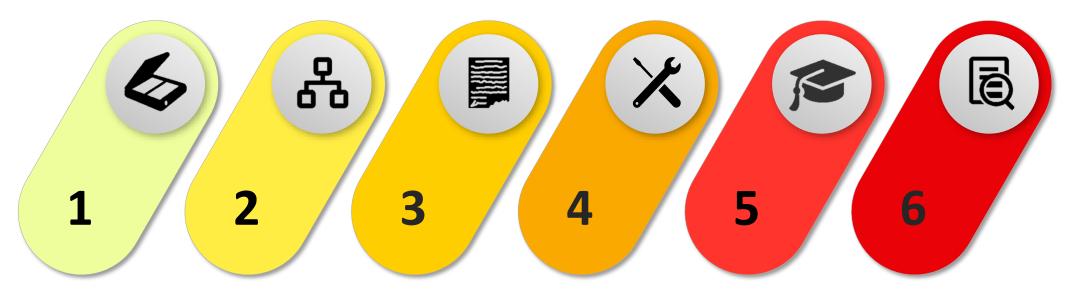
- binarization
- deslant correction
- noise removal
- ation thinning
  - blurring

#### **Training AI**

- randomization
- annotation app
- crowdsourcing

Felhasználó: 1 anyakonyv/HU\_MNL\_PML\_XXXIII\_0001\_a\_0001\_a\_0001\_0015 info Forrás Betöltve. Megjelent az alulirett anyakönyvvezett oglalkozisa): gardon cre lakóhelye: Abo 1. chraits at attantain s a ki7 12 ATTA ATTACK OT személy l' lakohelye: Aborny Goldstonied tanya ! és bejelentet Kattintson ide az annotáláshoz! esaládi rossz szegmentáció nem olvasható áthúzott kihagyom atya vallása : 27/ Mások így annotálták - 0 Felismert - 0 hipotézis állása A HELY.LAKHELY mező - 1. sora Sou Man - 21 [63beb0f8] 15/82 n 34/1unn willetishelve 1 , 361 életkora 38/19 lar ( applide de ntin ,42 Falling. vomai & Ast bali

# Building and applying model



#### Scanning

- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning
- archiving

#### Classification

• classifier

- template builder
  - polygonization
    - baseline detection

Segmentation

• layout analysis

- word segmentation
- skipping

- Preprocessing
- binarization
- deslant correction
- noise removal
- ation thinning
  - blurring

#### **Training AI**

- randomization
- annotation app
- crowdsourcing

#### Handwritten TR

- optical model
- language model
- page reconstruction

# Building models

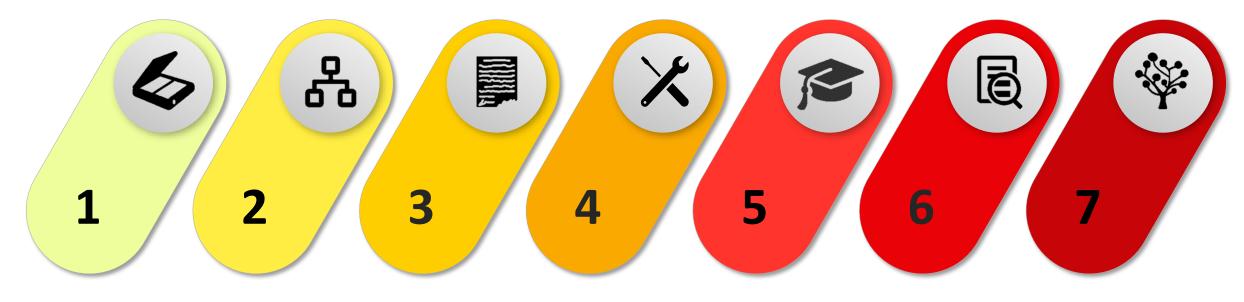
## • TROCR

- Pylaia
- Tesseract
- Kraken



"pageNr" : 2, "transliteration" : "none", "spots" : [ {	
spors . [ ]	"txt" : "Balus",
	"rp" : 0.000644671,
	"pos" : [ 355, 3885 ],
	"size" : [ 203, 163 ]
}, {	
	"txt" : "Galus",
	"rp" : 0.000662552,
	"pos" : [ 355, 3885 ],
	"size" : [ 203, 163 ]
}, {	
	"txt" : "István",
	"rp" : 0.999855,
	"pos" : [ 607, 3886 ],
	"size" : [ 284, 163 ]
}, {	
	"txt" : "Paius",
	"rp" : 0.000715723,
	"pos" : [ 355, 3885 ],
	"size" : [ 203, 163 ]
}, {	
	"txt" : "Pal",
	"rp" : 0.000368085,
	"pos" : [ 319, 3885 ],
	"size" : [ 131, 163 ]
}, {	
	"txt" : "Pala",
	"rp" : 0.00019393,
	"pos" : [ 347, 3885 ],
	"size" : [ 187, 163 ]
}, {	
,, , , , , , , , , , , , , , , , , , ,	"txt" : "Palas",
	"rp" : 0.00129407,
	"pos" : [ 355, 3885 ],
	"size" : [ 203, 163 ]
}, {	
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	"txt" : "Palcs",
	"rp" : 0.000450743,
	"pos" : [ 355, 3885 ],
	"size" : [ 203, 163 ]
}, {	
	"txt" : "Pales",
	"rp" : 0.00978169,
	"pos" : [ 355, 3885 ],
	,

# Linking records



#### Scanning

- surveying
- planning
- item description
- equipment
- preprocessing
- logistics
- scanning
- archiving

#### Classification

• classifier

- template builder
  - polygonization
    - baseline detection

Segmentation

• layout analysis

- word segmentation
- skipping

- Preprocessing
- binarization
- deslant correction
- noise removal
- tion thinning
  - blurring

#### Training AI

- randomization
- annotation app
- crowdsourcing

#### Handwritten TR

- optical model
- language model
- page reconstruciton

### Publication

- database
- postprocessing
- record linking
- indexing
- search GUI

István Q	80%	Ţ								
+ 7	Lace marton	2			1	-				Szaberentz Vis Mihály öreg
G	Lani Anvaro	2	-		1			1		Horváth Irinyes
a la companya da companya d	Verinte Latving	9		1	1	1.	-		100	Horváth György Mihály Tar David
	<b>István</b> (82.1%)	• 0		Part of	1				1	Bata Mihály Márton Laz Lazo
1. 10	Astván (15.8%) 0 Ostván (7.9%)	14	21.5.8					1		István Paius Ferintz István
11.	<b>Stván</b> (4.1%)	7		1		H ANA		1		Janes Tóth Paus Balus
12.	Acha Gall	h		1	_1_		1	×	-	Páll Roka István Horváth
1.2	Himrich Krow	2	-	-	1		-	1	1	Paus István Vöröserencz Ferencz
14	Pate Synan	2	-		1					Vörös Ferentz Lugosi
15	Voris Ference	9			1		T.			István Nemeth Tolvaj György
the second se	Juandi Scrento	0			1			-		Györgyörgel örgel Vörös János
	Semeth Moring	9		-	1	-	-			Oros Pos Mihály
18	Tologi Current inges	2			T		1	a fair	i	
ta	Vin Inna	4.			1			t		

## HTR pipeline demo

#### National Archives of Hungary handwritten text recognition demo

#### Detailed information of the models

#### Image testing Write now and test

A szülök törvenyes haras Megjegyzés : felmutatott harassagi bisaga a ranyitiannyal igarolog lott.

and south to many him to a plant when the south to be a south to be the south to be a south to be the south to be a southet to be a south to b Enable and set the preprocessing of an image sent for recognition

# Pages: 1 2 3 Enable and set the preprocessing of an image sent for recognition Segmentation Kraken / Bila model Model

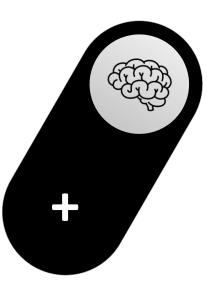
TrOCR / MNL anyakönyv 2023.02.11. / 39. checkpoint. CER=3,53 WER=11,96

#### Recognized te

Megjegyzés: A szülők törvényes házas- [0.9958] Megjegyzés: A szülők törvényes házas- [0.9957] Megjegyzés: A szülők törvényes házas- [0.9957] Megjegyzés: A szülők törvényes házas - [0.9957] Megjegyzés: A szülők törvényes házas - [0.9956] -sága a felmutatott házassági bi- [0.9999] -sága a felmutatott házasági bi- [0.6163] -sága a felmutatott házasági ki- [0.3684] -zonyitvánnyal igazolva lett. [0.9545] szonyitvánnyal igazolva lett. [0.9509] -zonyitvánnyal igazolva lett. [0.9486] -zonyitvánnyal igazolva lett. [0.9486] <

#### <u>Click on me</u>

## Text processing



#### Postprocessing

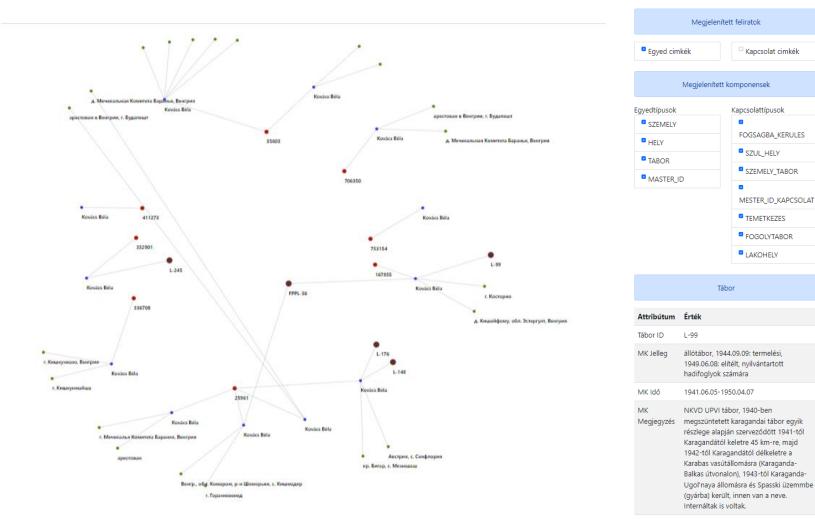
- entity recognition
- text analysis
- text processing
- NLP technologies
- record linking
- semantic linking

#### *Örkény <mark>István</mark> -* zászlós | Budapest, 1912

HU MNL OL X 10874		1912
Azonosító	1361263	
Név	Örkény István	
Vezetéknév (gépi átírás)	Orkonyi [0.77], Alkonyi [0.46], Argényi [0.46], Alkony [0.33] - Оркени	
Utónév (gépi átírás)	István - Иштван	
Apai utónév (gépi átírás)	Hugó [1.00], Csikó [0.00] - <mark>Fyro</mark>	
Nemzetiség	magyar - венгр	
Rendfokozat	zászlós - фенрих	
Születési hely (gépi átírás)	Budapest - Budapest - г. Будапешт	
Születési év	1912	
Fogságba esés helye (gépi átírás)	Bjelgorod közelében - Blagorodovac [0.73], Belgrad [0.71], Belegrád [0.67], Bolgárom [0.67], Bielavoda [0.53] - под Белгородом	
Fogságba esés időpontja	1943.02.01	
Távozás dátuma	1946.12.21	
Fogolytábor	27. sz. tábor - лагерь № 27	
Megjegyzés	További névalakok - Hadifogságba kerülés helye: Belgorod	
Kapcsolódó rekordok (gépi tanulás) *	1. Szovjet táborok magyar foglyai - Örkény István (23152089) [0.99] Gépi tanulással feldolgozott lehetséges személykapcsolatok más adatbázisban	
Képek		
	Ð Ð	
	His Burry with the second state of the second	

## Graph database

Név Kovács Béla	Vezetéknév	Utónév
Apja neve	Anyja neve	Születési év 1908
Életkor	Születési dátum éééé, hh. nn.	Hadifogságba kerülés dátuma éééé. hh. nn.
Távozás dátuma éééé. hh. nn.	Elhalálozás dátuma éééé, hh. nn.	Hazaérkezés dátuma éééé, hh. nn.
Születési hely	Lakhely	Hadifogságba kerülés helye



MK Magyar Összesen 854 fő foglyok

## Some takeaways

- -digital humanities use cases are different from archives'
- -JSON (instead of ALTO or PAGE)
- hypotheses with confidences
- randomized annotation
- -quality over quantity



# Strength and weaknesses

## What we have

- infinite number of data with big diversity
- capacity for building training sets
- domestic collaboration (DHLAB, OSZK, NYTI)
- international cooperation

## What we miss are

- capacity for development
- HPC



## Conclusions

Perhaps Handwritten Text Recognition **may not be perceived as a wonder in the near future anymore**. However, text extraction serves as **an essential cornerstone** in big archives, aiding in our overarching mission to make archives accessible to all. As we continue on this journey, we must establish two key objectives: **fostering collaborative efforts** to find the optimal approaches and **proudly publicizing** our achievements. This groundbreaking advancement in historical research and citizen access **highlights the undeniable relevance of us** in the era of artificial intelligence. Together, we can demonstrate the immense value we hold.



Thank you!