

# Nouvelles approches pour l'extraction d'informations dans les documents (manuscrits) numérisés

Mélodie Boillet · Solène Tarride · [Christopher Kermorvant](#)

TEKLIA, Paris, France

**T E K L I A**

# New approaches?

- Standard sequential workflow
- Integration of the different steps into a single model
- End-to-end goal oriented models
- Continuous improvement





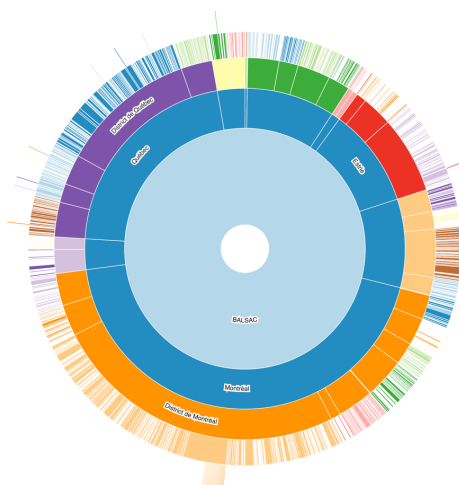
# The e-Balsac project



Genealogical database manually extracted from parish and civil registers for more than 50 years

Goal : Automatic processing of the Québec parish registers 1850-1920

10 Centers  
36 Districts  
1,985 Parishes  
44,742 Registers  
1,995,646 images



M. 71 Le sept septembre mil neuf cent quatorze, en la publication de Pierre Chénier trois bans de mariage, faite au prône de nos Messes paroissiales et entre Pierre Chénier, fils mineur de Jeanne Chénier et d'Elphondine Bernadette St. Louis Roy de cette paroisse d'une part, et Bernadette St. Louis, fille mineure de feu Louis St. Louis et de Joséphine Vaudric, aussi de cette paroisse d'autre part; m. s'étant découverts aucun empêchement de mariage et du consentement des parents de dite mineurs, nous, prêtre soussigné, avons reçu leur mutuel consentement de mariage et leur avons donné la bénédiction nuptiale en présence de Jeanne Chénier, mère de l'époux et de Louis Lévesque, beau-frère de l'épouse, les époux seuls et d'eux avec nous. Lecture faite.  
Pierre Chénier  
Bernadette St. Louis

M. 72 Le sept septembre mil neuf cent quatorze, en la dispense Marie Renaud de deux bans de mariage, par nous accordée en vertu de pouvoirs à nous conférés par sa grandeur Monseigneur Charles Thurot Paquette Hugues Gauthier, archevêque d'Osaka, en aussi la publication d'un ban faite au prône de notre messe paroissiale ainsi qu'à celui de Notre Dame d'Osaka, entre Marie Renaud, fille majeure d'Alarie Renaud et d'Abala Lacroixmouille de Notre Dame d'Osaka d'une part, et Aurora Paquette, fille mineure d'Olivier Paquette et d'Anna Morin de cette paroisse d'autre part; m. s'étant découverts aucun empêchement de mariage et du consentement des parents de la dite mineure, nous, prêtre soussigné, avons reçu leur mutuel consentement de mariage et leur avons donné la bénédiction nuptiale en présence d'Alarie Renaud, père de l'époux et d'Olivier Paquette, père de l'épouse. Tous ont, d'eux avec nous. Lecture faite.  
A. Renaud  
A. Paquette

M. 73 Le sept septembre mil neuf cent quatorze, en la dispense Oscar Anderson de deux bans de mariage, par nous accordée en vertu de pouvoirs à nous conférés par sa grandeur Monseigneur Charles Thurot Paquette Hugues Gauthier, archevêque d'Osaka, en aussi la publication d'un ban faite au prône de notre messe paroissiale ainsi qu'à celui de Notre Dame d'Osaka, entre Oscar Anderson, fils majeur d'Oscar Anderson et d'Anna Anderson de cette paroisse d'une part, et Anna Anderson, fille mineure d'Oscar Anderson et d'Anna Anderson de cette paroisse d'autre part; m. s'étant découverts aucun empêchement de mariage et du consentement des parents de la dite mineure, nous, prêtre soussigné, avons reçu leur mutuel consentement de mariage et leur avons donné la bénédiction nuptiale en présence d'Oscar Anderson, père de l'époux et d'Anna Anderson, mère de l'épouse. Tous ont, d'eux avec nous. Lecture faite.  
Oscar Anderson  
Anna Anderson

# Québec civil and parish registers

Treizieme feuille

B. 25

M.M. ✓  
Annuncia  
Beaumont

Le sept août dix neuf cent dix nous soussigné, prêtre du Séminaire de Québec, avons baptisé Marie-Marguerite-Annoncia, née le mille fille légitime de Pierre Beaumont, marchand et de Agilda Beaumont, de cette paroisse. Parrain: Pierre Beaumont, marchand de cette paroisse, marraine: Marie-Félicité Cantin, épouse du parrain, lesquels ont signé avec nous et le père. Lecture faite.

Marie-Félicité Cantin  
Pierre Beaumont  
Nurse Beaumont

Jos. Saguellet

Birth act

Le 10  
Diana  
Dupresne

Le vingt sept mars mil neuf cent deux, nous prêtre, curé sousigné, avons inhumé dans le cimetière de cette paroisse le corps de Diana Dupresne, décédée l'avant veille dans cette paroisse à l'âge de quatre ans, fille légitime de Victor Dupresne, benêtier et de Albertine Pathin de cette paroisse. Prévôt présent à l'inhumation Victor Dupresne et Alfred Prévost de qui ont déclaré en savoir signer.

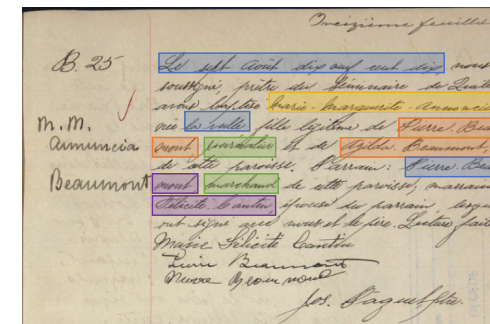
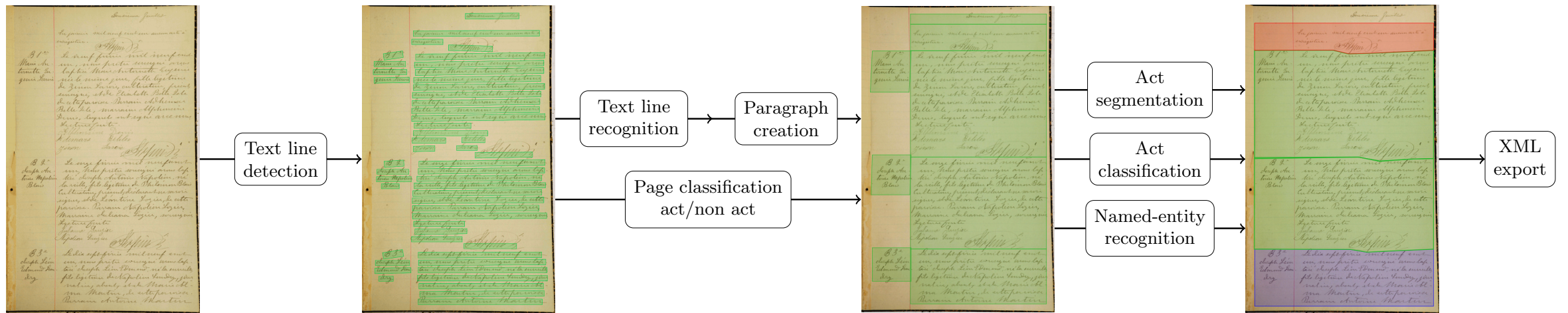
Edm. P. de la Cour

Death of a single person act

■ Date; ■ Subject of the act; ■ Parents; ■ Spouse; ■ Age; ■ Occupation; ■ Godfather/godmother



# Overview of the document processing workflow



NER

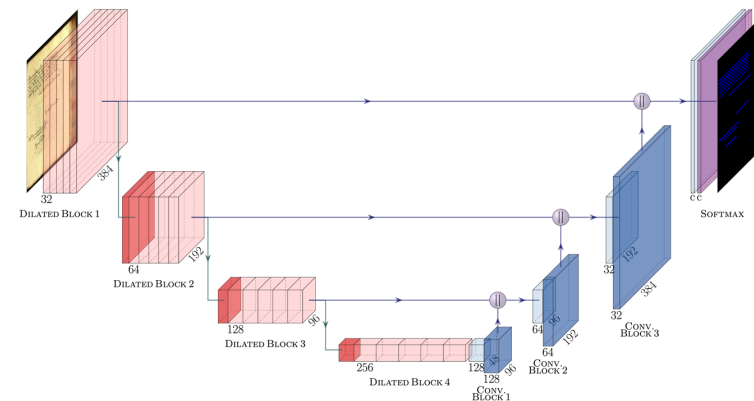
# Data & Technologies

- Automatic line detection + correction
- Manual transcription
- Named-entity annotation + relations
- Annotation : 0,05% of the full corpus

	pages	acts	lines	words	entities
<b>count</b>	896	2,661	45,479	205,165	25,564

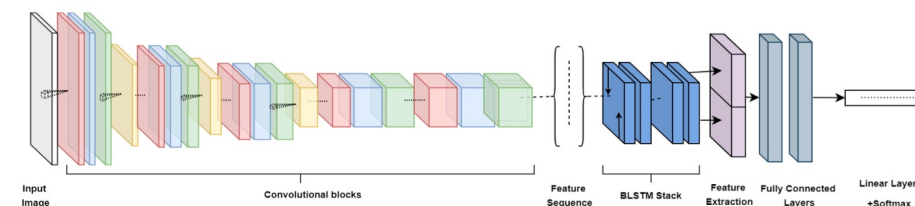
  

	entities			
	PER	LOC	DATE	OCC
<b>count</b>	15,810	2,823	4,551	2,380



DLA: Doc-UFCN model

<https://gitlab.tekليا.com/dla/doc-ufcn>



HTR : PyLaia CNN+biLSTM

<https://gitlab.tekليا.com/atr/pylaia>

NER

flair

Classification





# Results: Automatic data validation with act template

## Quantitative evaluation

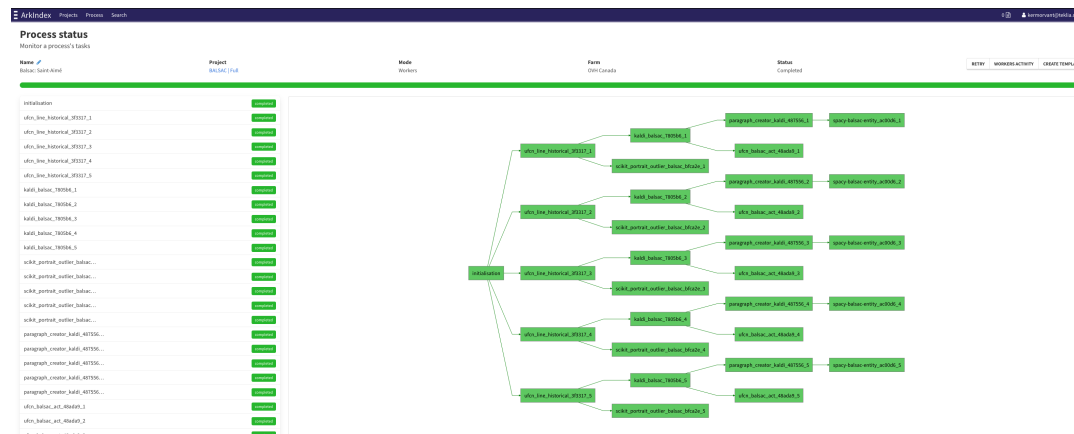
Entities	Count
Occupations	4 624 291
Locations	5 429 807
Dates	8 655 893
Persons	30 986 429

## Qualitative evaluation

type	Birth act	Death act
Valid	75.5%	69.0%
Fusion	19.9%	18.2%
Invalid	7.1%	8.9%
Special	0.4%	3.2%

# Lessons learned

- Complex workflow with one model per task
- Unsupervised metrics of the full process are needed to improve the quality
- Automatic processing but under human monitoring
- The workflow would be simpler now using end-to-end transformer models





# Final goal : key-value information extraction

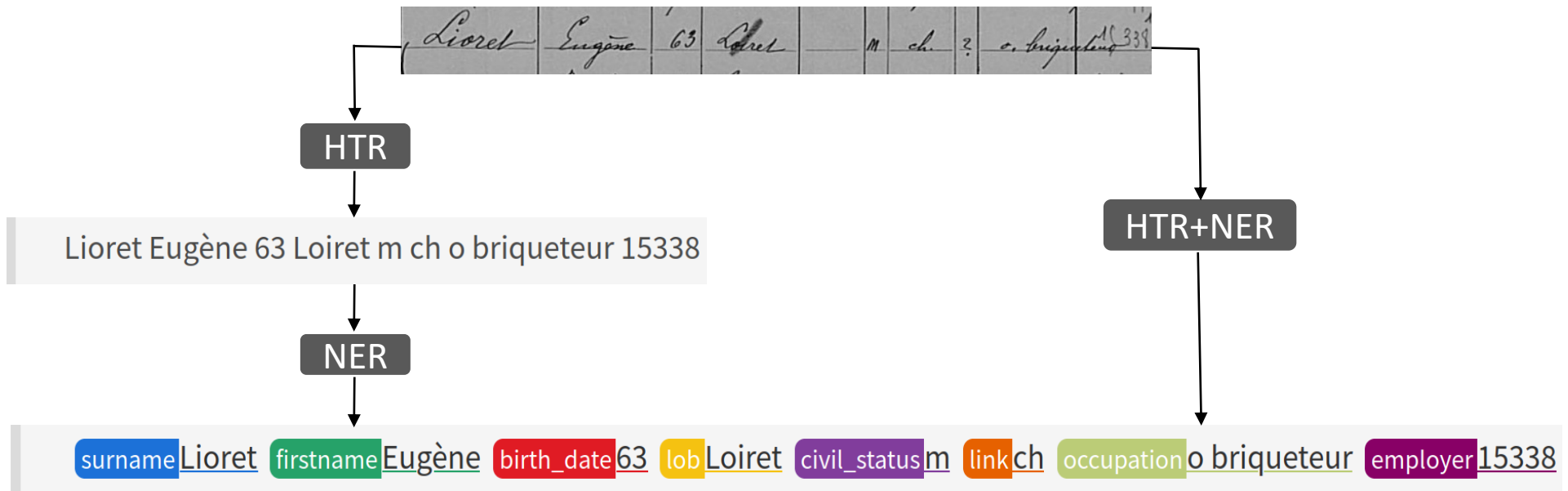
B.4.  
Marie  
Alice  
Cora  
Charlotte  
Dufault.

Le vingt-sept février mil neuf cent un, sous seing privé, prêtre curé, de cette paroisse, avons baptisé Marie Alice Cora Charlotte, née la veille fille légitime de Napoléon Dufault, et de Alice Mercier de cette paroisse. Le parrain a été Pierre Dufault, grand père de l'enfant et la marraine son épouse Marie Larolle qui ainsi que le père ont signé avec nous; le parrain a déclaré ne savoir signer. Lecture faite.

Marie Larolle  
Napoléon Dufault  
J. D. Bernier (pasteur)

Key	Value
Child name	Marie Alice Cora Charlotte
Birth date	26/02/1901
Father name	Napoléon
Father last name	Dufault
Mother name	Alice
Mother last name	Mercier

# Integration of HTR+NER



## Sequential approach

DAN or PyLaia for HTR  
SpaCy for NER

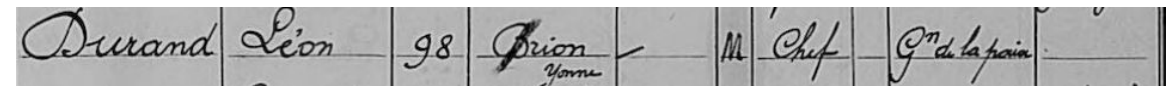
## End-to-end approach

DAN [4] for HTR+NER  
Special tokens are used to tag entities



# Integration of HTR+NER

Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
GT + SpaCy	0.0	96.4	Line
PyLaia + SpaCy	17.19	76.3	Line
DAN + SpaCy	8.18	84.0	Page
DAN end-to-end	7.83	85.9	Page



surname Durand    firstname Léon    birth\_date 98    lob Brion Yonne    civil\_status M  
link Chef    occupation Gardien de la paix

- End-to-end models are competitive
- End-to-end models are easier to train, improve and deploy

# One step beyond : Integration DLA+HTR+NER

REGISTRATION	NUMERO	PRENOM	DATE	LIEU	PROFESSION	REMARQUES
371	Léautaud	François	11/12	Barcelonnette	chef négociant	
372	Marie ép Léautaud	Marie	11/12	Barcelonnette	épouse néant	
373	Léautaud	Berthe	11/12	Barcelonnette	enfant idem	
374	Jaufred	Aglaé	11/12	Barcelonnette	domestique	
375	Brunet	François	11/12	Bayons	chef brig ferrier	
376	Margaillan ep Brunet	Euphanie	11/12	Bayons	épouse néant	
377	Brunet	Fernand	11/12	Bayons	enfant idem	
378	Brunet	Eva Marie	11/12	Bayons	enfant idem	
379	Brunet	Yvonne	11/12	Bayons	enfant idem	
380	Audiffred	Julie	11/12	Barcelonnette	chef néant	
381	Derbez	Césarine	11/12	Revel	culurière	
382	Arnaud	Joseph	11/12	Enchartrayes	cafetier	
383	Risolép Arnaud	Marie	11/12	St Pons	épouse néant	
384	Jaubert	Jean Baptiste	11/12	Nvnet	domestique	
385	Allard vve Bellon	Joséphine	11/12	Barcelonnette	blanchisseuse	
386	Bellon	Félix	11/12	idem	journalier	
387	Bellon	Clémentine	11/12	idem	blanchisseuse	
388	Bellon	Julie	11/12	idem	repasseuse	
389	Abel	Edouard	11/12	Faucon	tailleur	
390	Jean ép Abel	Odile	11/12	Condarnine	épouse néant	
391	Abel	Paul	11/12	Barcelonnette	tailleur	
392	Abel	Marie	11/12	idem	épouse néant	
393	Abel	Alfred	11/12	idem	enfant idem	
394	Abel	Fortuné	11/12	idem	enfant idem	
395	Abel	Berthe	11/12	idem	enfant idem	
396	Eymé	Marius	11/12	Puy de Euzèbe	boulangier	
397	Lyme	Marguerite	11/12	Barcelonnette	enfant néant	
398	Maurel	Adrien	11/12	Fours	cant chef	
399	Signout ep Maurel	Constantine	11/12	S. Pon	épouse néant	
400	Maurel	Jean	11/12	Barcelonnette	enfant idem	



surname Léautaud, firstname François, occupation négociant, link chef, birth\_date 1839, nationality française, lob Barcelonnette  
 surname Caire ep Léautaud, firstname Marie, occupation néant, link épouse, birth\_date 1849, nationality idem, lob Encharteaues  
 surname Léautaud, firstname Berthe, occupation idem, link enfant, birth\_date 1876, nationality idem, lob Barcelonnette  
 surname Jaufred, firstname Aglaé, occupation idem, link domestique, birth\_date 1875, nationality idem, lob Gnetrartrayes  
 surname Brunet, firstname François, occupation brig ferrier, link chef, birth\_date 1852, nationality idem, lob Bayons  
 surname Margaillan ep Brunet, firstname Euphanie, occupation néant, link épouse, birth\_date 1865, nationality idem, lob Selonnet  
 surname Brunet, firstname Fernand, occupation idem, link enfant, birth\_date 1892, nationality idem, lob Bayons  
 surname Brunet, firstname Eva Marie, occupation idem, link idem, birth\_date 1894, nationality idem, lob idem  
 surname Brunet, firstname Yvonne, occupation idem, link idem, birth\_date 1899, nationality idem, lob idem  
 surname Audiffred, firstname Julie, occupation néant, link chef, birth\_date 1825, nationality idem, lob Barcelonnette  
 surname Derbez, firstname Césarine, occupation culurière, link idem, birth\_date 1872, nationality idem, lob Revel  
 surname Arnaud, firstname Joseph, occupation cafetier, link idem, birth\_date 1865, nationality idem, lob Enchartrayes  
 surname Risolép Arnaud, firstname Marie, occupation néant, link épouse, birth\_date 1865, nationality idem, lob St Pons  
 surname Jaubert, firstname Jean Baptiste, link domestique, birth\_date 1876, nationality idem, lob Nvnet  
 surname Allard vve Bellon, firstname Joséphine, occupation blanchisseuse, link chef, birth\_date 1834, nationality idem, lob Barcelonnette  
 surname Bellon, firstname Félix, occupation journalier, link enfant, birth\_date 1864, nationality idem, lob idem  
 surname Bellon, firstname Clémentine, occupation blanchisseuse, link idem, birth\_date 1876, nationality idem, lob idem  
 surname Bellon, firstname Julie, occupation repasseuse, link idem, birth\_date 1875, nationality idem, lob idem  
 surname Abel, firstname Edouard, occupation tailleur, link chef, birth\_date 1849, nationality idem, lob Faucon  
 surname Jean ép Abel, firstname Odile, occupation néant, link épouse, birth\_date 1856, nationality idem, lob Condarnine  
 surname Abel, firstname Paul, occupation tailleur, link enfant, birth\_date 1878, nationality idem, lob Barcelonnette  
 surname Abel, firstname Marie, occupation néant, link idem, birth\_date 1879, nationality idem, lob idem  
 surname Abel, firstname Alfred, occupation idem, link idem, birth\_date 1890, nationality idem, lob idem  
 surname Abel, firstname Fortuné, occupation idem, link idem, birth\_date 1892, nationality idem, lob idem  
 surname Abel, firstname Berthe, occupation idem, link idem, birth\_date 1900, nationality idem, lob idem  
 surname Eymé, firstname Marius, occupation boulangier, link chef, birth\_date 1842, nationality idem, lob Puy de Euzèbe  
 surname Lyme, firstname Marguerite, occupation néant, link enfant, birth\_date 1880, nationality idem, lob Barcelonnette  
 surname Maurel, firstname Adrien, occupation cant, link chef, birth\_date 1870, nationality idem, lob Fours  
 surname Signout ep Maurel, firstname Constantine, occupation néant, link épouse, birth\_date 1868, nationality idem, lob S. Pon  
 surname Maurel, firstname Jean, occupation idem, link enfant, birth\_date 1904, nationality idem, lob Barcelonnette

# One step beyond : DLA+HTR+NER

Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
DAN end-to-end	7.83	85.9	Line*
	-	-	-
	11.74	85.3	Page

- Full page end-to-end models are competitive
- They are even easier to train, improve and deploy
- Ground truth generation is easier (can be trained from page transcription)

<https://socface.site.ined.fr/fr/collaboration/progressions-des-collectes/>





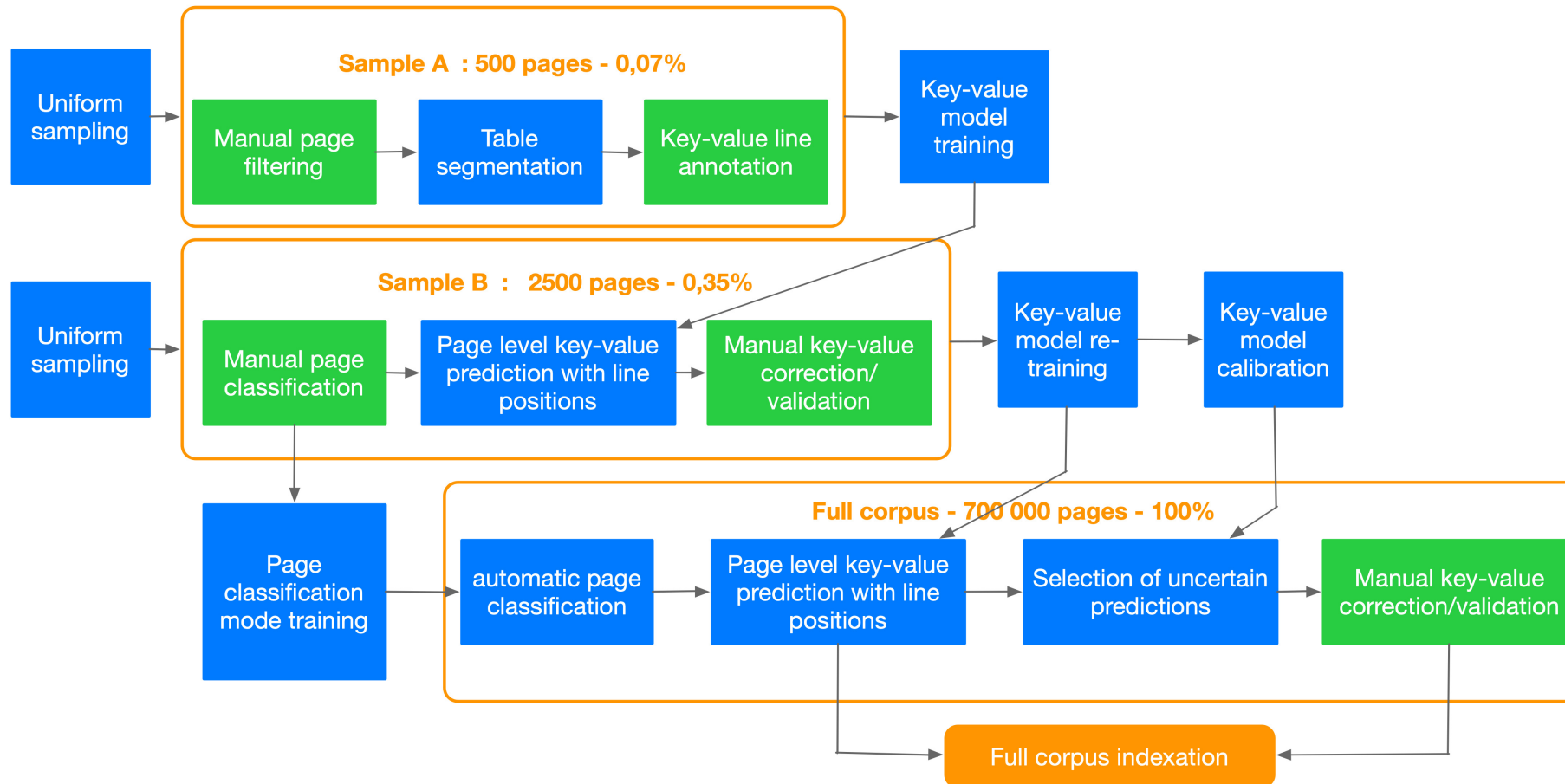
# ICRC

## In practice : the French WWII POW lists

- 36 million names of POWs from WWII in the ICRC Tracing Archives
- 3000 requests can be processed each year
- Automatic indexing would considerably increase the number of beneficiaries

Staats- ange- hörig- keit	Nr. der Er- tennungs- marke	Pr a m e	Borname	Geburts- tag	Geburts- ort	Bor- name des Vaters	Familien- name der Mutter	Name und Anschrift des zu benachrichtigenden Person	Dienstgrad	Truppenteil (m. Rp. ufm.)	Matrikel- Nr.	Ort und Tag der Gefangennahme	Verwundungen, Berlegungen, Tod (Beerdigungsplatz)	Bemerkungen (z. B. Zugänge von anderen Lagern, bei Gläuberverlust über die Eltern erfahren)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FRANZ	38565	VANDESTEEM -DAM	Jules Cornis	19-5 1913	Toulon (Viel)	Rene	LEPIN	Mme Jules VANDESTEEM-DAM à la Croix Verte St Paul d'Eyrieux Hte Vienne	SOLDAT	121 Infa tenu gaw C	1101 Limoge	BOULOGNE 25-5-40	3190	
FRANZ	38566	MAURY	Fernand	7-3 1913	Montignat	Georges	VAIT- TIERES	Mme MAURY Julia Limogne par Puyfauque fam. et nomme	SOLDAT	417 proumors	142	BOULOGNE 24-5-40		
FRANZ	38567	TERRASSE	Maxime	16-10- 1906	Espaly	Hemi	LAZER	Mme Louise TERRASSE Aux Basses Puyfauque fam. et nomme	SOLDAT	F.A.R.C.	2000 Lo Puy	BOULOGNE 24-5-40		
FRANZ	38568	MOHAMED BEN SOUGLATI	Mohamed	1921	Kenitra	Ahmed Ben Souglati	ITTO BENT ALI OU HACEM	ITTO BENT ALI OU HACEM Douar Oulaidi Kenitra Maroc	SOLDAT	264 R.A.D.	1939 T306 Maroc	Belgique 25-5-40		
FRANZ	38569	MOHAMED BEN LAHSEN	Mohamed	1918	douar oued Jlmane	Rhego- mani	BOU- CHAÏB	LAHSEN Ben RHEZONANI Douar oued Jilmane Touat Maroc	SOLDAT	264 RAD	T299 Maroc	Belgique 15-5-40		
FRANZ	38570	ZILLALI HAMED BEN ALI	ZILLALI -LI	17-4 1917	Adjadi	AHMED BEN ALI	FATMA BENT KEM- RADAM	Mme HAMED Fatma Bent KEMRADAM Douar Sem che Mekrida Toud- Touat Maroc	SOLDAT	26 RAD	26 RAD	Belgique 13-5-40		
FRANZ	38571	PERINET	Germain Abuch	22-1 1916	Epuss	Emile	GUILLE -MIN	Mme Germaine Ches M. O. tave L'ARCENTIER Somme Velle pour cont. 15.15 Marne	SOLDAT	106 R. I.	118 Chalon	Seclin 28-5-40		
FRANZ	38572	BDURQUIN	Pierre	10-8 1913	PARIS	Pierre	LUCAS	Mme Pierre BDURQUIN 72 rue de Colombes Asnières	SOLDAT	106 R. I.	7015 Bramonides	Seclin 28-5-40		

# In practice : the French WWII POW lists



# Sample A : Automatic line detection + Annotation

Callico Projects / ICRC | List annotation / Campaign Line transcription / Tasks / Annotation table\_row 3

Element numéro de série	N° de la liste	Nom	Prénom	Date de naissance	Lieu de naissance	Sexe	Statut civil	Profession	Remarques
72892	Harris	Marcel	Amélie	1901	Paris	M	UFG	19113	10/10/1901
72893	Brardel	Léon	André	1901	Paris	M	UFG	19113	10/10/1901
72894	Zwundace	Eloi	Albert	1901	Paris	M	UFG	19113	10/10/1901
72895	Leveux	Paul	André	1901	Paris	M	UFG	19113	10/10/1901
72896	Brondeau	Jacques	Jacques	1901	Paris	M	UFG	19113	10/10/1901
72897	Romain	Louis	André	1901	Paris	M	UFG	19113	10/10/1901
72898	Delcroix	André	André	1901	Paris	M	UFG	19113	10/10/1901

Pending

Numéro de référence / Reference number

Nom / Last name / Name

Prénom / First name(s) / Vorname

Date de naissance / Date of birth / Geburtstag

Skip task Submit



# Sample B : Prediction validation/correction

## Prediction

- Name **Jardoin** firstname **Maurice** Birth date **4.8.09** RF number **4963**
- Name **Faliani** firstname **Joseph** Birth date **19.5.09** RF number **4963**
- Name **Sermoud** firstname **Léon** Birth date **6.10.18** RF number **4963**
- Name **Savouillon** firstname **Félix** Birth date **18.1.08** RF number **4963**
- Name **Méric** firstname **Roger** Birth date **19.1.11** RF number **4963**
- Name **Besson** firstname **Jean** Birth date **7.3.08** RF number **4963**
- Name **Haon** firstname **Marcel** Birth date **20.7.13** RF number **4963**
- Name **Carrasset** firstname **Fernand** Birth date **16.2.18** RF number **4963**

## Validation/correction

Reference number/Numéro de référence 36411

Last name/Nom/Name Chatain

First name(s)/Prénom(s)/Vorname Anbin

Date of birth/Date de naissance/Geburtstag 24-12-1904

Skip task Submit

## Re-training

	Time second	Name F1	Firstname F1	Date F1	RF number F1	All F1
Sample A (500)	32	0.91	0.86	0.86	0.97	0.90
Sample B (2500)	16	0.96	0.93	0.96	0.97	0.95

# New methods for information extraction

- Full page end-to-end methods are effective and easy to deploy
- In standard cases, little experimentation and fine-tuning of models
- The entire pipeline is annotation-dependent: data cleansing is the critical point
- Does it apply to printed documents ?
- The next frontier: continuous training ?





# New methods for information extraction

**#non-answerable**

Q: In which year does the Net Requirement exceed 25,000?

A: None

**#abstractive #counting**

Q: How many attorneys are listed for the plaintiffs?

A: Two

**#layout-navigating #graphic-intensive**

Q: Are the margins of the page uniform on all pages?

A: Yes

**#extractive #list**

Q: What are the Years mentioned in Chart 1?

A: [2020, 2021, 2022]

**Page 1**

**Page 2**

**Page N**

**#multi-hop #layout-navigating**

Q: From the list of Top 10 Key Recovery Components, which is the last component listed on the second page?

A: Hope

**#abstractive #graphic-intensive**

Q: Does this document contain any checkboxes?

A: No

## Document Visual Question Answering

- Foundation models ?
- It is adapted to large batches of documents ?





Thank you !

[kermorvant@tekliia.com](mailto:kermorvant@tekliia.com)

**T E K L I A**