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

Mélodie Boillet · Solène Tarride · Christopher Kermorvant

TEKLIA, Paris, France

New approaches?

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

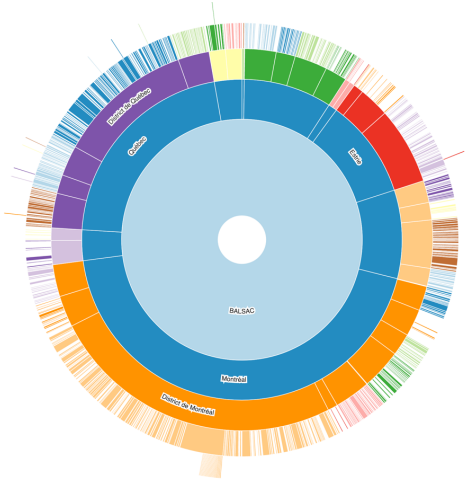
The e-Balsac project

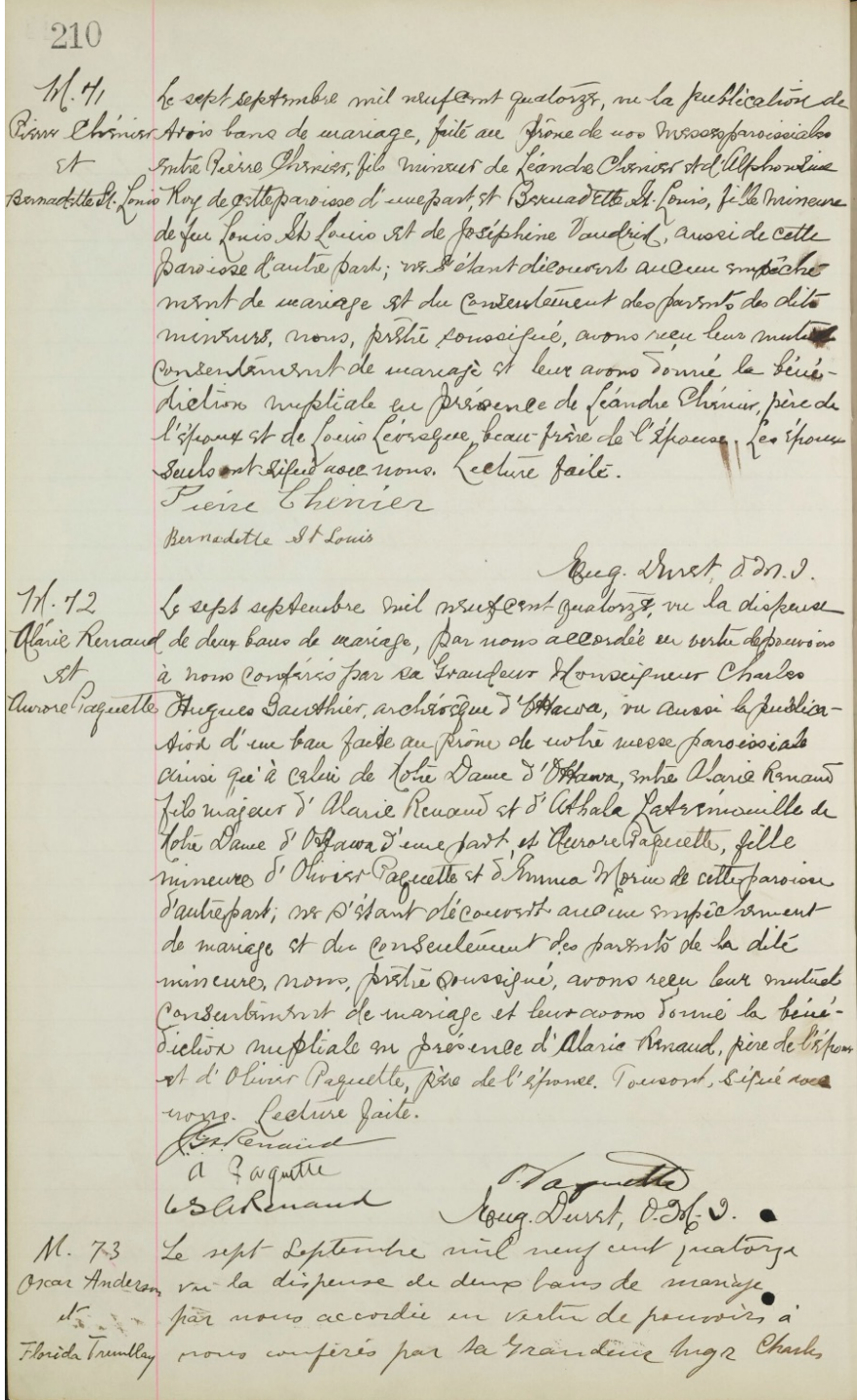
UQAC
Université du Québec
à Chicoutimi

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




210

M. 71 Pierre Chénier et Bernadette St. Louis

Le sept septembre mil neuf cent quatorze, vu la publication de trois bans de mariage, faite au prône de nos messes paroissiales entre Pierre Chénier, fils mineur de Léandre Chénier et d'Alphonsine Roy de cette paroisse d'une part, et Bernadette St. Louis, fille mineure de feu Louis St Louis et de Joséphine Vaudrieul, aussi de cette paroisse d'autre part; ne s'étant découvert aucun empêchement de mariage et du consentement des parents des dits mineurs, nous, prêtre soussigné, avons reçu leur mutuel consentement de mariage et leur avons donné la bénédiction nuptiale en présence de Léandre Chénier, père de l'épouse et de Louis Lévesque, beau-frère de l'épouse, les époux seuls ont signé avec nous. Lecture faite.

Pierre Chénier
Bernadette St Louis

Aug. Durest, ptre V.

M. 72 Alarie Renaud et Aurore Paquette

Le sept septembre mil neuf cent quatorze, vu la dispense de deux bans de mariage, par nous accordée en vertu de pouvoirs à nous conférés par sa Grandeur Monseigneur Charles Hugues Gauthier, archevêque d'Ottawa, vu aussi la publication d'un ban faite au prône de notre messe paroissiale ainsi que à celle de Notre Dame d'Ottawa, entre Alarie Renaud fils majeur d'Alarie Renaud et d'Athala Laderoute de Notre Dame d'Ottawa d'une part, et Aurore Paquette, fille mineure d'Olivier Paquette et d'Emma Morin de cette paroisse d'autre part; ne s'étant découvert 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. Tousont, signé avec nous. Lecture faite.

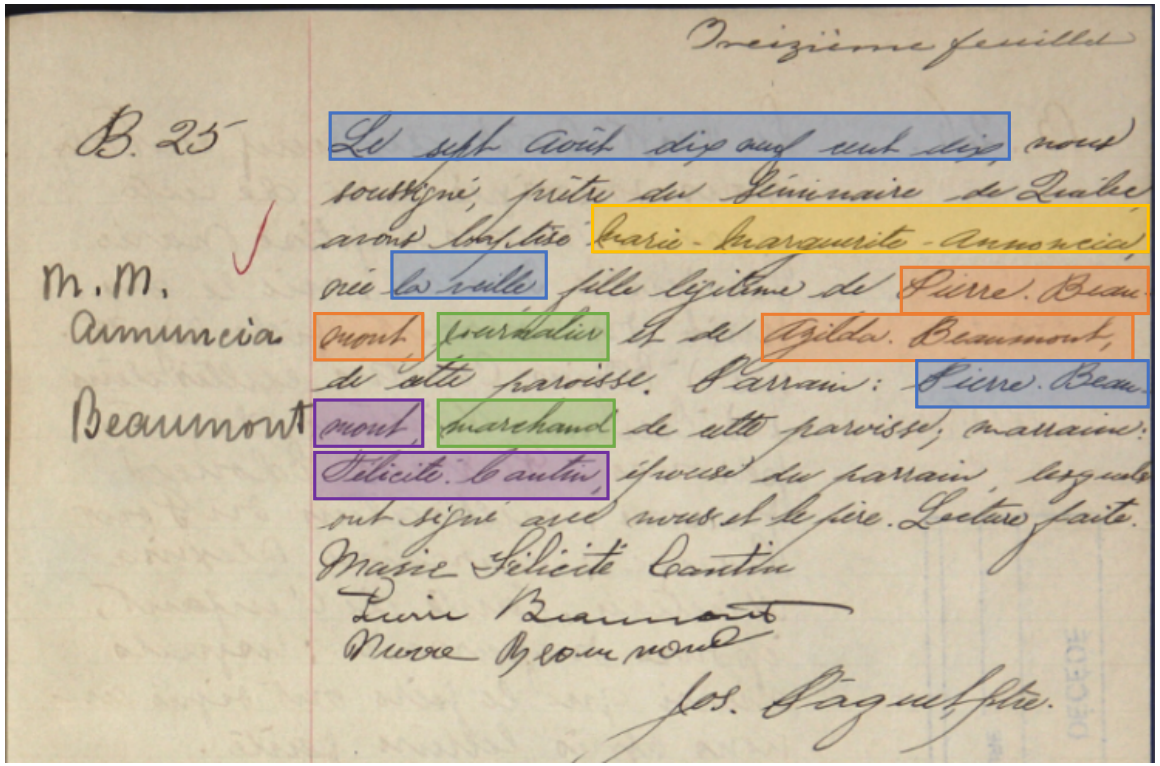
A. Renaud
A Paquette
A.A. Renaud

Aug. Durest, ptre V.

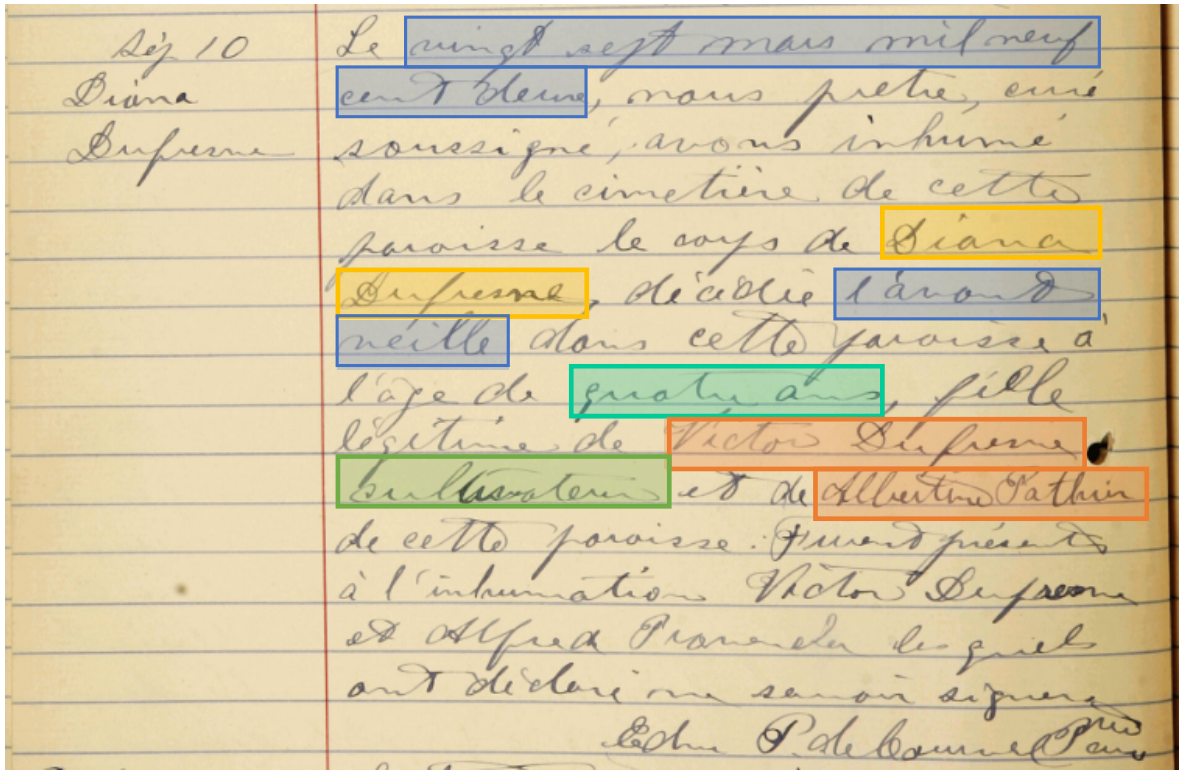
M. 73 Oscar Anderson et Florida Trumbley

Le sept Septembre mil neuf cent quatorze vu la dispense de deux bans de mariage par nous accordée en vertu de pouvoirs à nous conférés par sa Grandeur Mgr Charles

Québec civil and parish registers



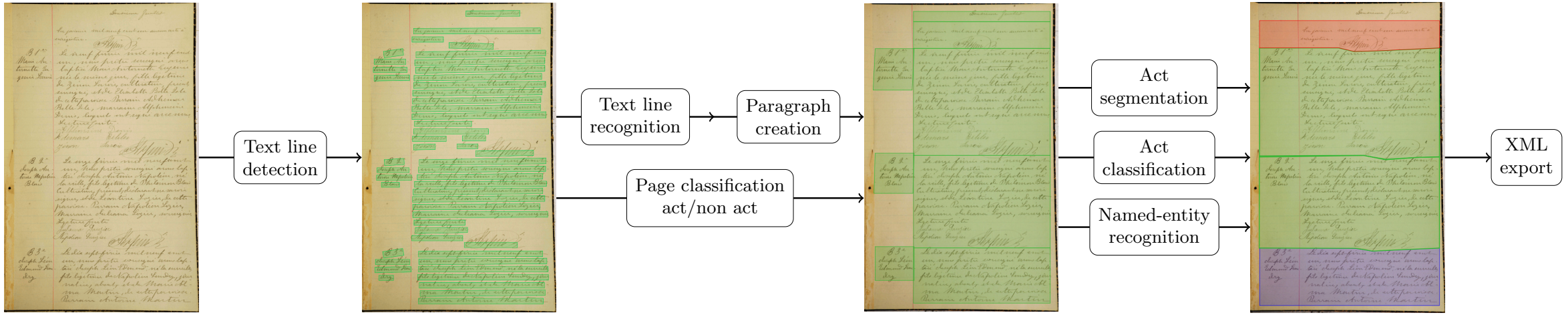
Birth act

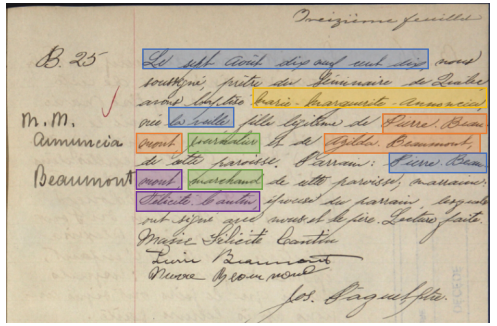


Death of a single person act

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

Overview of the document processing workflow





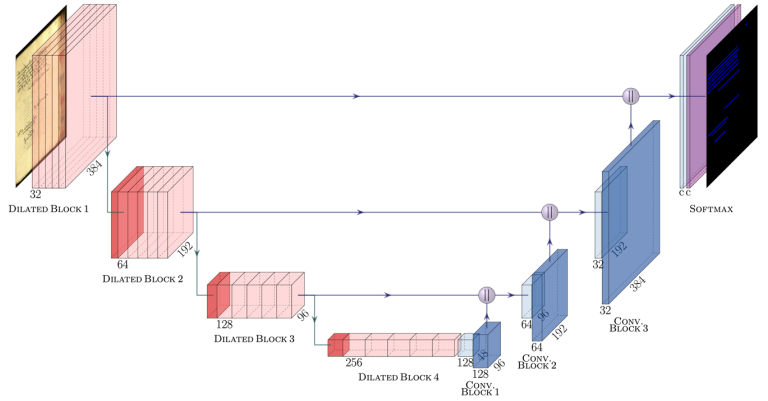
NER

Data & Technologies

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

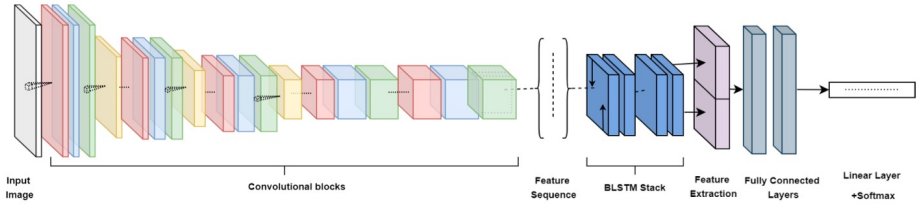
	pages	acts	lines	words	entities
count	896	2,661	45,479	205,165	25,564

	entities			
	PER	*LOC*	*DATE*	*OCC*
count	15,810	2,823	4,551	2,380



DLA: Doc-UFCN model

https://gitlab.teklia.com/dla/doc-ufcn



HTR : PyLaia CNN+biLSTM

https://gitlab.teklia.com/atr/pylaia

NER flair Classification scikit learn

Results: Automatic data validation with act template

Quantitative evaluation

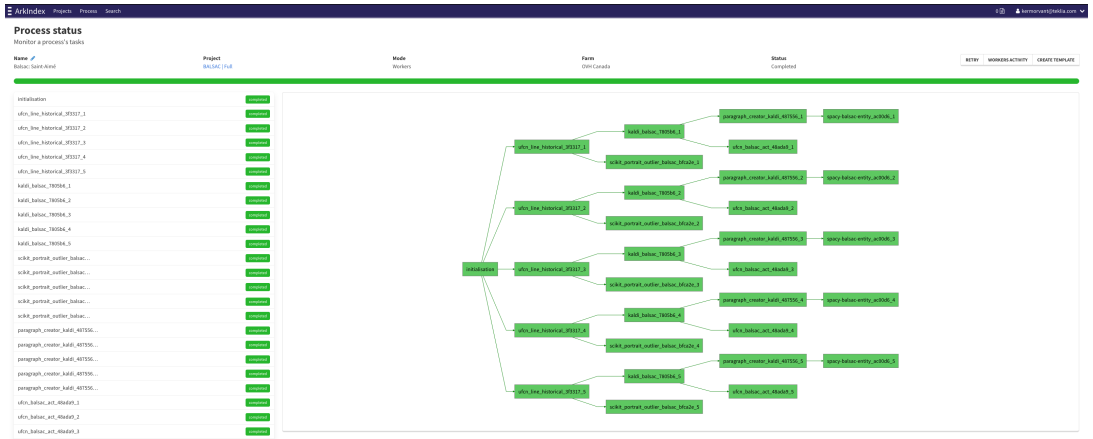
Entities	Count
Occupations	4 624 291
Locations	5 429 807
Dates	8 655 893
Persons	30 986 429

Qualitative evaluation

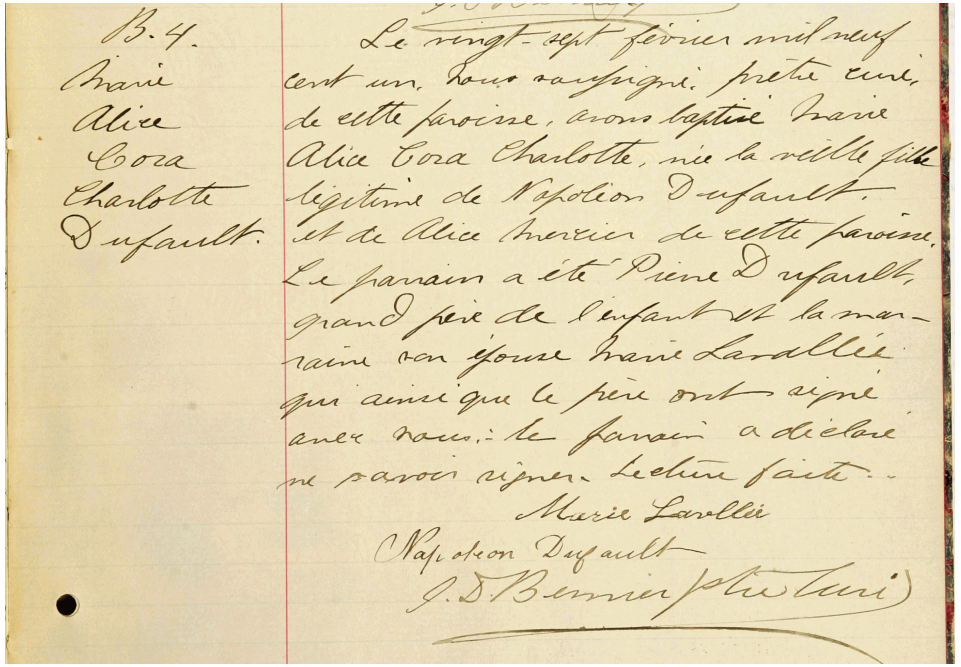
type	Birth act	Death act
Valid	75.5%	69.0%
Fusion	19.9%	18.2%
Invalid	7.1%	8.9%
Special	0.4%	3.2%

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

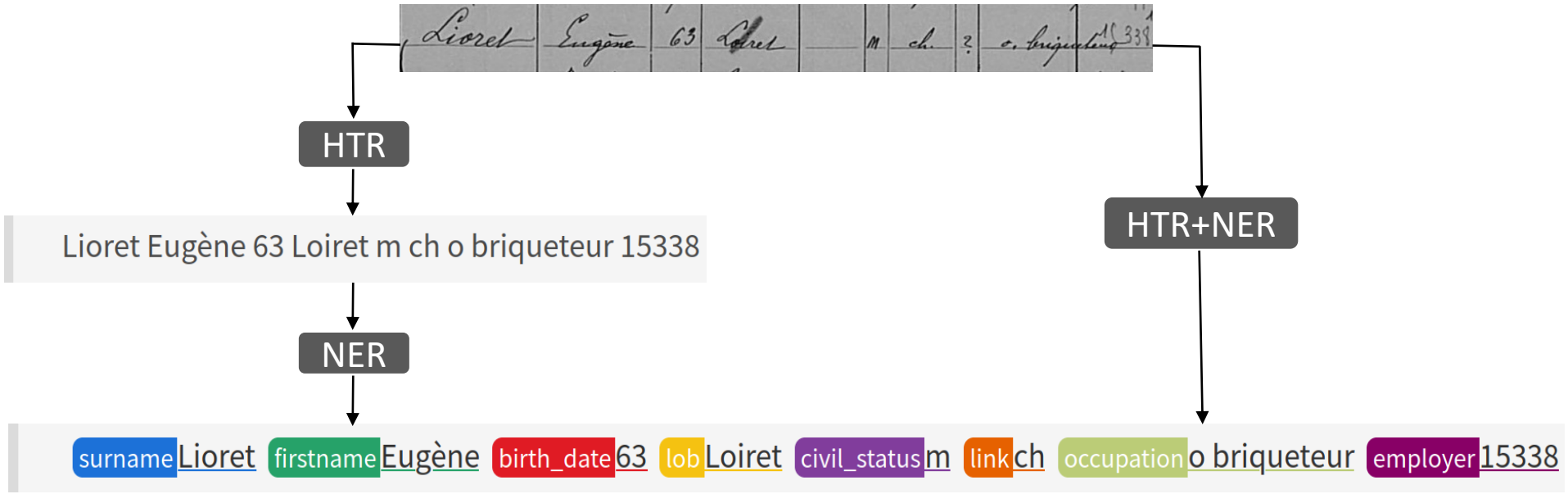


Final goal : key-value information extraction



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

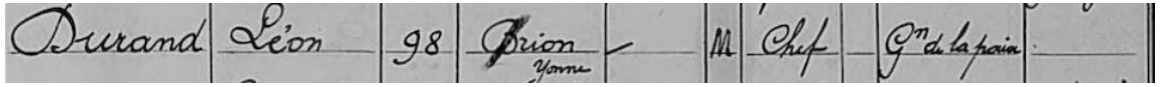
- DAN or PyLaia for HTR
- SpaCy for NER

End-to-end approach

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

Integration of HTR+NER

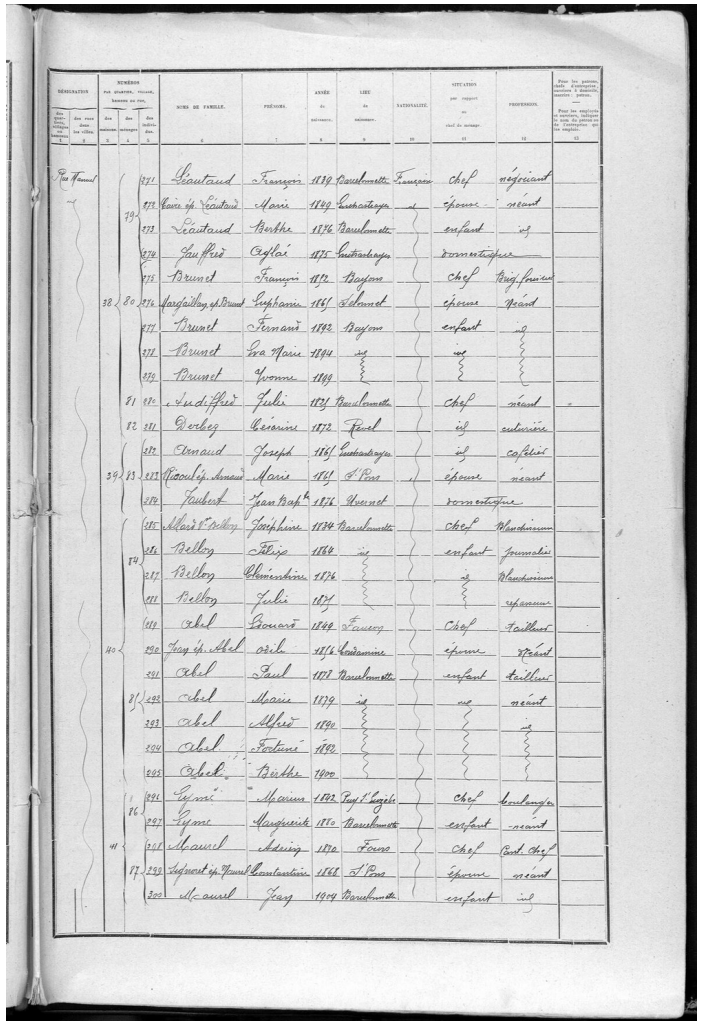
Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
GT + SpaCy	0.0	96.4	Line
PyLaia + SpaCy	17.19	76.3	Line
DAN + SpaCy	8.18	84.0	Page
DAN end-to-end	**7.83**	**85.9**	Page



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



surname Léautaud firstname François occupation négociant link chef birth_date 1839 nationality française lob Barcelonnette
surname Caire ep Léantaud 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 Barcelomette
surname Jauffred 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 Risoulép 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 Nvernet
surname Allard vve Bellon firstname Joséphine occupation blanchisseuse link chef birth_date 1834 nationality idem lob Barcelomette
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 boulanger 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 Barcelomnette
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 Barcelomarle

One step beyond : DLA+HTR+NER

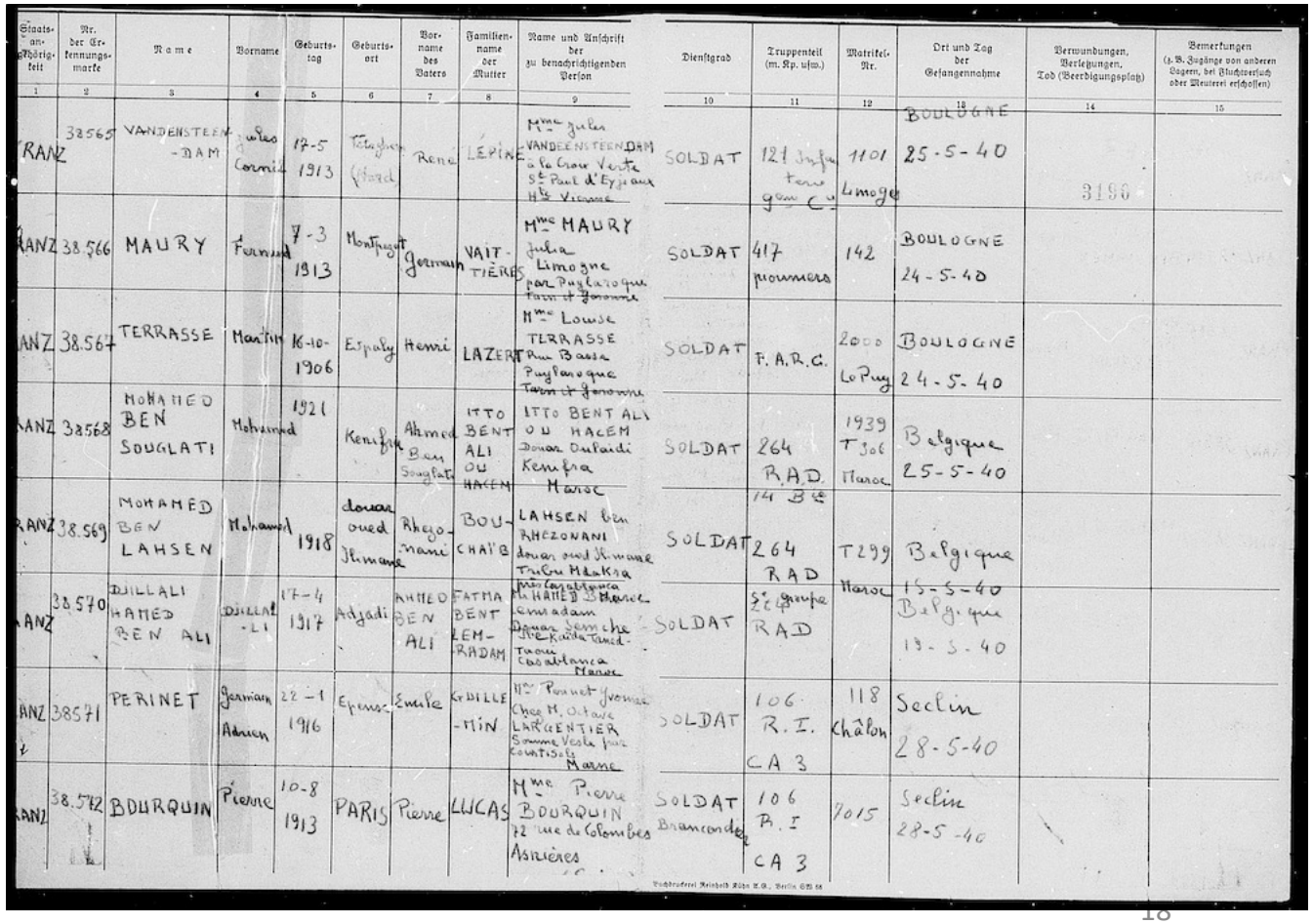
Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
DAN end-to-end	7.83	85.9	Line*
	-	-	-
	11.74	85.3	Page

- 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)

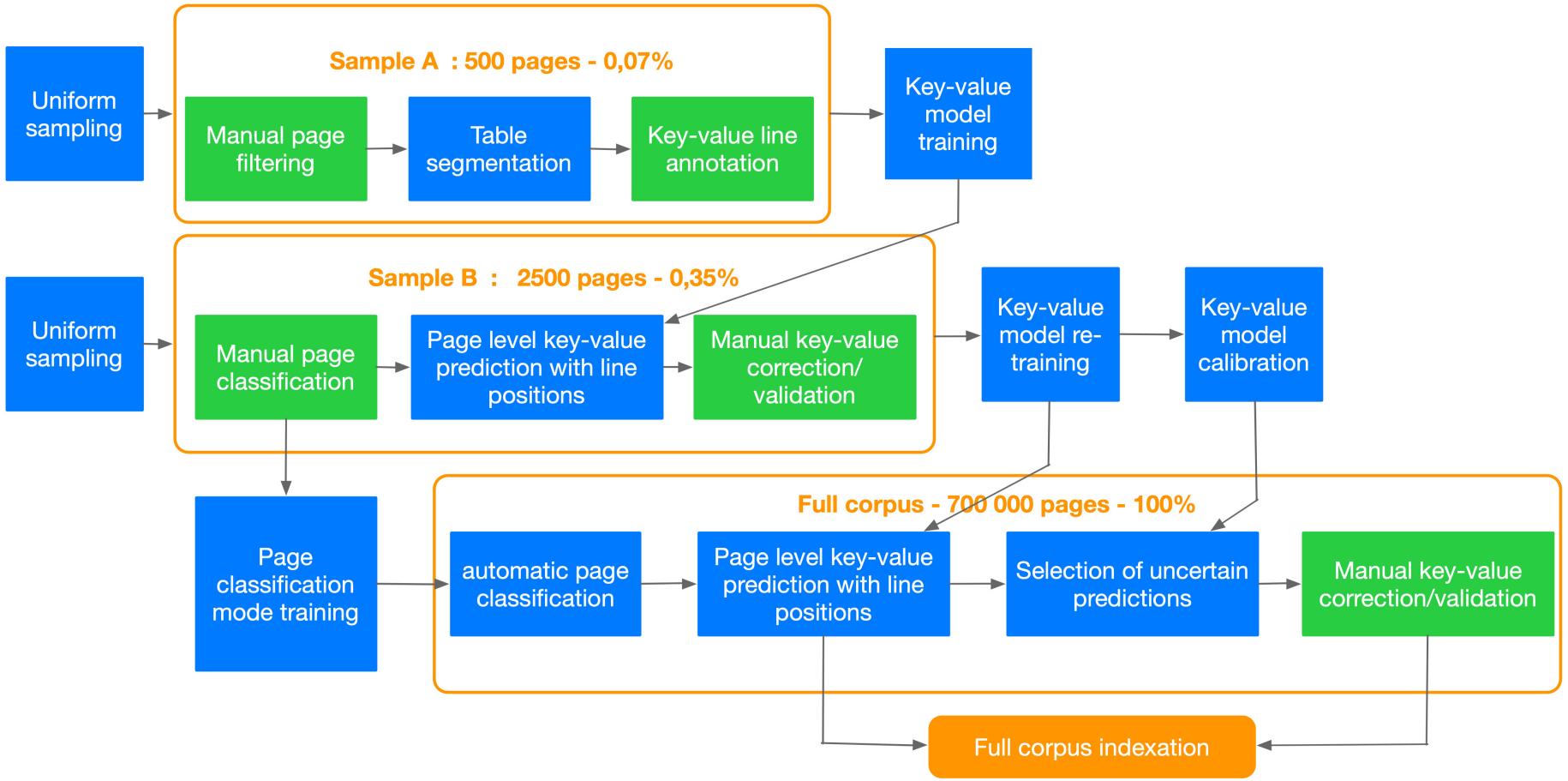
https://socface.site.ined.fr/fr/collaboration/progressions-des-collectes/

In practice : the French WWII POW lists

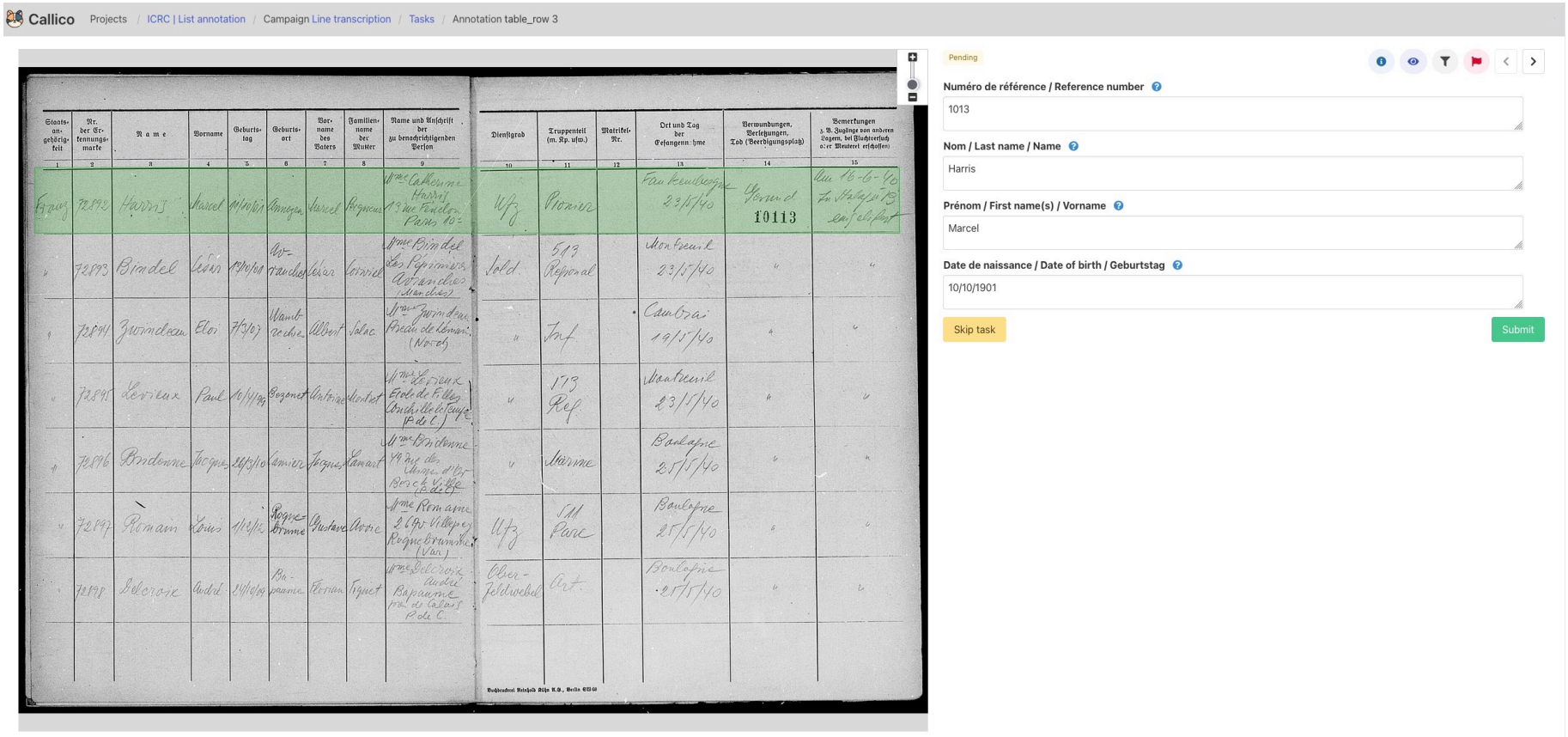
- 36 million names of POWs from WWII in the ICRC Tracing Archives
- 3000 requests can be processed each year
- Automatic indexing would considerabely increase the number of beneficiaries



In practice : the French WWII POW lists



Sample A : Automatic line detection + Annotation



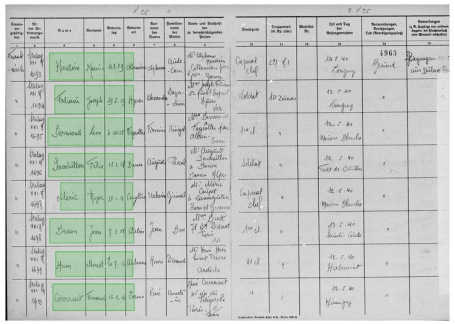
Callico Maintainer

Annnotation platform, companion of Arkindex

https://gitlab.teklia.com/callico/

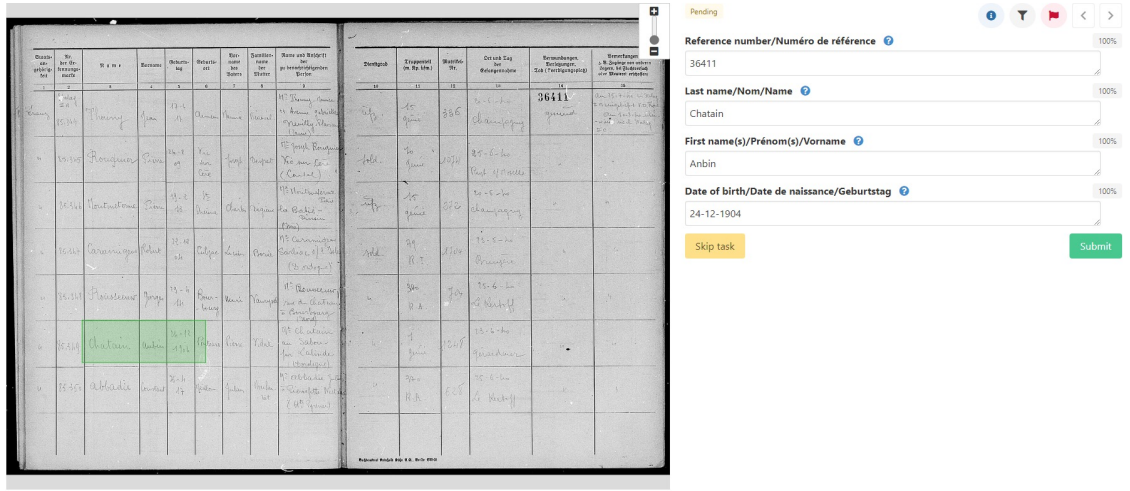
Sample B : Prediction validation/correction

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



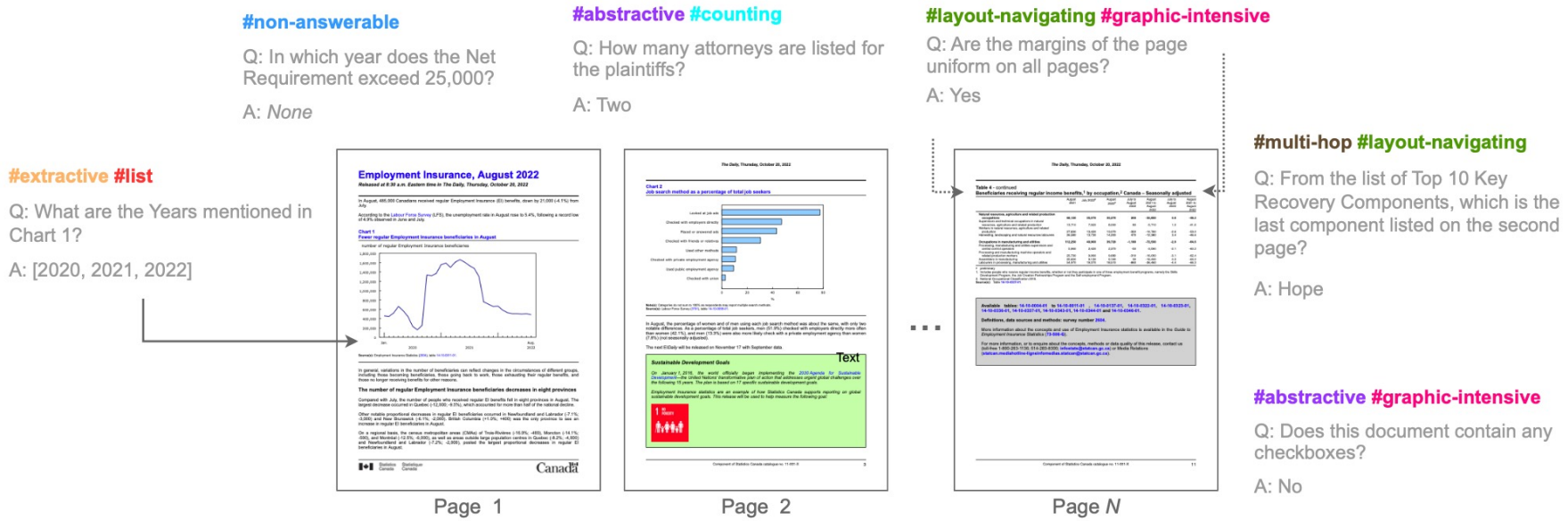
Re-training

	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



Document Visual Question Answering

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

ANR FINLAM
LITIS -BNF - TEKLIA

Thank you !

kermorvant@teklia.com