#### Large Scale Genealogical Information Extraction From Handwritten Québec Parish Records

Solène Tarride · Martin Maarand · Mélodie Boillet · James McGrath · Eugénie Capel · Hélène Vézina · <u>Christopher</u> <u>Kermorvant</u>

TEKLIA, Paris, France U

UQAC, Chicoutimi, Québec



ICDAR 2023 August 21, 2023

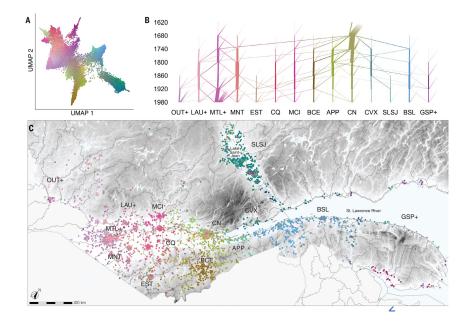
> UQAC Université du Québec à Chicoutimi

#### The Balsac project

Goal: a large genealogical database of the Quebec population

On the genes, genealogies, and geographies of Quebec <u>Science</u>, 25 May 2023





## The e-Balsac project

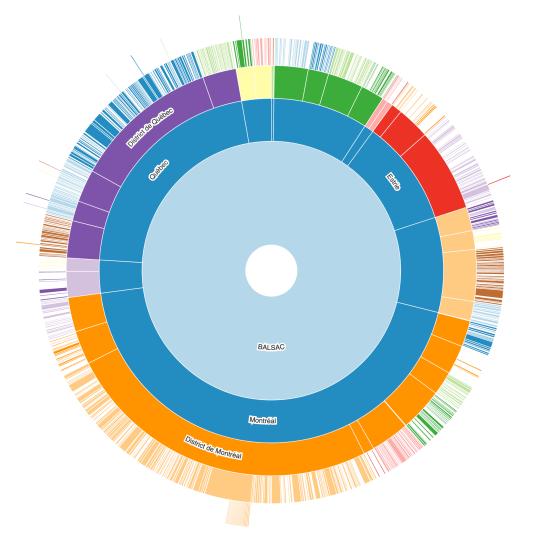
- Database built for more than 50 years
- Extracted from transcribed parish and civil registers
- First only marriages, now birth and death records
- Manual transcription is no longer feasible: the Quebec population was multiplied by 6 between 1851 and 1961

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

### Québec civil and parish registers

Birth, marriage and death records for the Quebec population between 1850 and 1916

> 10 centers 36 districts 1,985 parishes 44,742 registers 1,995,646 images



### Québec civil and parish registers

prêtre curé somsigné avons regu leur mutuel con

M.HI he sept septimbre mit neufant qualongs, me la publication de in Chrinist Atois bans de mariage, fuit au frone de nos mesorspraciona Intre Riero Chriser, file minuer de feandre Chanier stal alphonetice Bernadsteed Inio Koy de osteparo isse d'anepart st Bernad Ette St. Louis, file mineure de fue tonio et louis et de Josephine Vaudrid, aussi de cette paroisse d'autre part ; W. & chant dieouour aucun smpeche ment de mariage st du consentement des parento des dito mentures, nons, pretie conscipie, avons rece leus mater consentancent de mariaje et leur avono Tonnie la binediction meptiale que présence de feandre Chinis, pire de l'épour st de louis l'évague, seau frire de l'épourse, les spour Such and Sifer wave none. Lecture faile. Pierre Chinier Bernadette St Louis

a gagnette

Bug. Smet. Jon. J. Le sept septembre mil mentem qualons vu la dispanse Clarie Renaud de deux baus de mariege, par nons accordée en perte depoursons à non conferco par la transfeur Honseigneur Charles Aurore aquette Huques Gausthier, archiveque I Stawa, nu auroi la pudiasich d'un ban faite au prone de untre messe paroiosiale divisi qu'à calui de tore Dame & Ottama, antre Carie Ranand filo majour & Marie Renand of & athale Latronomille de Tota dance & Agawa Deme fait of Aurore Paquette, fille mineuro d'Olivier Paquette et d'Ammea Morio de cetterparoione d'autrepart; no p'shant de courset aucun sompic sament de mariege et du conseclement des parento de la dele mineure, nono, pretie prussique, avono recu leur mutuel Consentiment de mariage et leur avono Tonne la binediction meptiale en présence d'Alarie Renaud, piere de l'épon A d'Oliver Prquette, pitre de l'éponse. Tousont, l'épie roce oggo. Cecture faite.

isakenand Loug. Durt, 0.34. V. M. 73 Le sept Leptembre mind neug cut jucatorya man Anderen nu la dispense de deux lans de manige Florida Trumbley nous confires par ba Tranderic high Charles

renter cui somaigne or leur anno domie dei entement de maiage et leur anno domie la leur diction nuptuele en présence de Thomas Bergers prère de l'époure de chadreur christernen ant de l'époure qui teux ensemble ent signé once neur. Lecture faite d'élime Magaron Claire Peletre Andras Christensen Thomas Bergeron Darahe Ster Wous recormainons comme notre enfant legitime celui qui a été l'esptié le vingt cinquième jour de Mai de la même année, sans le nom de Marie flete ne Incomme (B. 28. Secultet June Recto) en 13.32 fai de quai nous avons signé en présence de Renerend Louis Willbrad Barake, prêtre curé . Manie Pelpelmine de cette paraisse, de Thomas Bergeron et de Andrew Christensen soussignes Lecture faite. Jaito Jant Alexis le diaire jour de Lum de l'ame mil lergerie huit cent grate vingt dence Helone Hergeron Thogmany Stegeton Clara Tellitica Andreas Christensen ABaraher to 1.15 Le diadein mil huit cent quatre singt deux, mars prêtre cure soussigné avons inhume dans le ci Francois, 03.33 Bouchard metiere de cette parsine le corps de François oseph Lean (69 ans) Bouchard journalier spour de Marie Duches Lergerie ne de cette paroisse décédé l'avant neille in cette paroise agé de soisante neuf ans. Mes cento Louis Lesperance et Ariene Gremon qui betwee faite n'ant au signer Al Barahertre 13.30 Le ange dim mil huit cent grate singt deux news uns consegne prêtre cure servicigne como baptite levis consegne I aurmer "ne l'avant-reilles fils iseue du britme meriage de Touis donnier mennier et de Elde d'orrade de att pareises Parrain Iavier d'orton tameur, marsaine

Acuriano fuellas Imerentienne Godreautt, époure de Desac laté jeur nalier, tano deva de cette pareire et qui, aimi que le pire n'ont su signer Lecture faite B. 31 Le onze deine mil huit cent quatre vingt deux saus Marie Emélie prêtre cure senseigne erono begtise Marie Emelie I ournier nece l'avant-wille fille usue da légitime marige de Lauis Dournier mennier et de Stide Dorcade de cette pareises, Parrain Lean Dertrino culturateur mariame Hermine Lagnon, épasse de florace das cade cultivateur tons devo de cette paraise et qui aimi que le père n'ont en aiguer setue fait Le seize Luin mil huit cent quatre vingt deux nono prêtre curé sousigné avous legatisé Marie Wilhelmine nee, ce jour fille issue du legitime mariage de Méodore Lergerie journahier et de Alexandrine Lavrie Le cette paroisse Par sain Irançois Tremblay fils de Barchal Trem May cultivateur marraine Louise Cergerie fille de George Sergerie culturateur tous deux de cette paranes et qui ainei que le père-noit en signer secture faite. Monship tre Le dia huit Lim mil huit cent grate myt den nous preto cure sensigne arous bog tree Lough Dean ne ce jour fils isen du legitime mariger de de lie Dergene culturateur et de feline Santhier de cette porsisse Parrain Irançais Pentin cultivateur marraine Seorginna Bolduc son epenae tons deux de cette fa raise et qui lectore faite n'ent en aguer Le pire absent. Blanalite B. 34 S. 16 Le myt deux duin mil huit cent grate mig Anonyme deva nang prêtre cure samsigne anno interme Gremon dans le cimetier de cette parisse le corpo

Fint folio EPD COOR Misbook or Register contaminighte two folios and and le feages are presented by the Revour R.S. Boog Lacemberty The Church of the Good Shephurd attBons ville ste Bishop Canuschael Aneniorial Church, at Foster, The used foothe energies Tiation of the acts of Baletisues, marriageor Bureals to be performed by him this Successors inoffice for the year one thou Sand une hundred fourten instand County Wienity. Generaled a Munther more servery hund The Seal of the Circuit Country the From. a AQuelier, withe County of Brome, withe Destud of Bedford, underthy entire of arti che 1311 of the Code of Civil Procedure varlache 42-42 A, H 2-B-H2-C. Jorty six y the Willow of the said Promee of Duckee. atthoutton this tweestell day of January princteau hundred of fourleers Stelledy Dept. CCCBrow

Notre Dame de Grâce de Hull, 1914

District of Chicoutimi, Saint-Alexis-de-Bagot, 1882

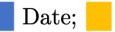
Foster's Bishop Carmichael Memorial Church et Bondville's Christ Church of the Good Sheperd, 1914

### Québec civil and parish registers

sept mais m Dreizieme fesillet Deg 10 Diana nous f B. 25 ussigne, avous in arif dis and with Dupreme le cimetrere, de pritre du génimaire le conto de Wallie Marie - marguerite m.m. alle ligitime de dicalle 1 ano ammeia cer 10 Jarasse arrain: graty Deaumont epouse du parrain et de assaleren albertino and wouse of the fire. Lecture faile 6 paraisse : Furent prese marie Filiente intrimation Victor Dufrem et Alfred Transle Mune geom mont los. Saguelphe. and declarge me serion de ledu de la la la la la and

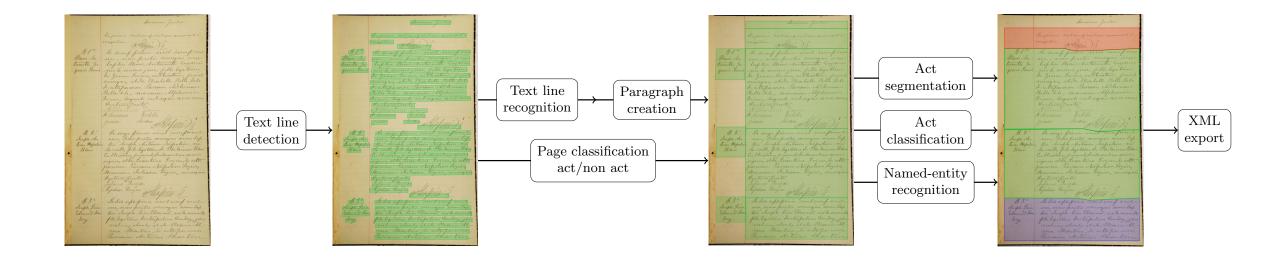
#### Birth act

Death of a single person act



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

## Overview of the document processing workflow



- A classical workflow
- Late decision strategy: all information is consolidated at export stage

## Ground truth generation

<ul> <li>↓ 1</li> <li>○ TR</li> <li>○ L</li> <li>○ BL</li> <li>○ W</li> <li>○ W<th></th><th>Le premier louvier nuit reuf ceut que nous toussaigne. prêtre cudiste missionneis cours bastie marie Prétomene, nee le suigt neuf Soccubre recédent feile legite de Paul Benard et de Olice montagnais Derraie Comercique Lourent marieur je gui mont su signer. Lecter faite</th></li></ul>		Le premier louvier nuit reuf ceut que nous toussaigne. prêtre cudiste missionneis cours bastie marie Prétomene, nee le suigt neuf Soccubre recédent feile legite de Paul Benard et de Olice montagnais Derraie Comercique Lourent marieur je gui mont su signer. Lecter faite
	f.	Le uppt louvier, mel neer locut ouines. roces
5-1	Le premier Janvier mil neuf cent quinze,	
5-2	nous soussigné prêtre eudiste missionnaire	
5-3	avons baptisé Marie Philomène, née le	
5-4	vingt neuf Décembre précédent, fille légitime	
5-5	de Paul Bernard et de Alice; montagnais.	
5-6	Parrain: Dominique Laurent marraine Juliette	
5-7	qui n'ont su signer. Lecture faite	
	$\neg$ +1 $\equiv$ B I X <sub>2</sub> X <sup>2</sup> U S $\supset$ C $\Leftrightarrow$ $\square$	ge and the second se

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

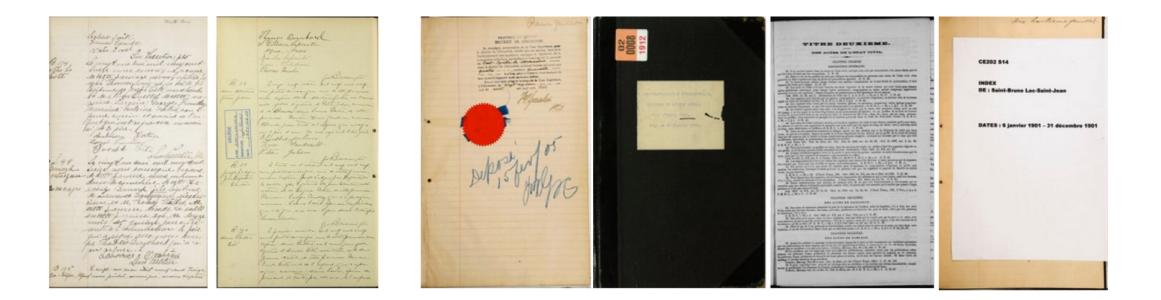
	pages	$\mathbf{acts}$	lines	words	entities
$\operatorname{count}$	896	$2,\!661$	$45,\!479$	205,165	$25,\!564$

	entities					
	PER LOC DATE OCC					
count	$15,\!810$	2,823	4,551	2,380		

#### CoNLL (Eng)

1390 pages - 301,418 words - 35,089 entities

### Page types: Act vs Non Act



Acts

Non acts

Large corpus contains both relevant and irrelevant pages, but both must be annotated

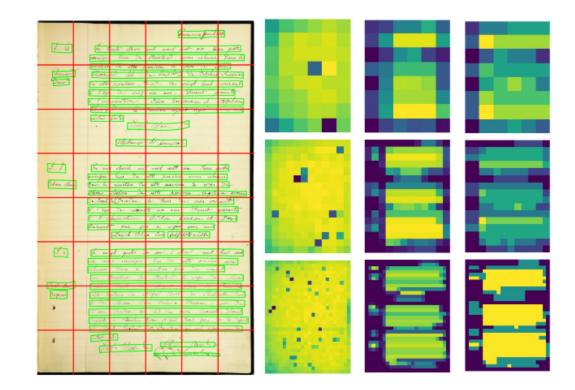
## Page types: Act vs Non Act

Pages containing acts are very homogenous

Non act pages are very diverse

Training of an outlier detector with isolation forest

100% recall on acts, 97% fscore



Pixels vs Lines density vs Line counts for different grid sizes (8x6), (16x12), (32x24

## Text line detection

- Different models tested
- Trained on a large corpus of multiple documents



Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks, Boillet, Kermorvant and Paquet, ICPR, 2020

	IoU	<b>AP@.50</b>	AP@.75	mAP
Doc-UFCN	0.87	0.98	0.91	0.76
${f dhSegment} \\ {f ARU-Net}$	0.74 <b>0.98</b>	$\begin{array}{c} 0.94 \\ 0.76 \end{array}$	$\begin{array}{c} 0.54 \\ 0.20 \end{array}$	$\begin{array}{c} 0.51 \\ 0.34 \end{array}$



DILATED BLOCK 4

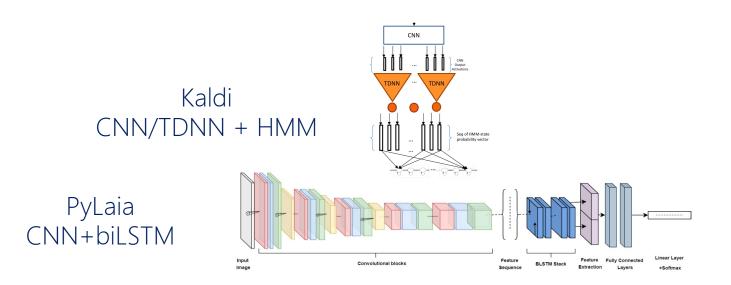
DILATED BLOCK

## Handwritten Text Recognition

- Kaldi and PyLaia tested
- Trained/tested on Balsac annotated corpus only

	tr	train		val		est
	CER	CER WER		CER WER		WER
Kaldi	4.13	12.36	6.22	17.10	6.41	17.41
PyLaia (no LM)	2.52	9.67	6.53	19.75	6.92	20.41
PyLaia (with LM)	2.24	8.27	5.90	17.16	6.29	17.92

+



 $\begin{array}{c} \varepsilon/0.25 & 9 \\ \hline & & & \\$ 

Language model Ngrams

Integration of LM in PyLaia: https://github.com/jpuigcerver/PyLaia

## Named-Entity Recognition

- Spacy, Stanza and Flair libraries tested
- Trained/tested on Balsac annotated corpus only
- Prediction at paragraph level

	$\mathbf{Text}$	Precision	Recall	F1-score
Stanza	manual	77.9	85.5	81.5
$\mathbf{Stanza}$	auto.	69.7	78.3	73.7
Flair	$\operatorname{manual}$	93.7	<b>93.1</b>	93.4
Flair	auto.	86.0	85.6	85.8
Spacy	$\operatorname{manual}$	83.1	83.6	83.4
Spacy	auto.	74.5	76.7	75.6

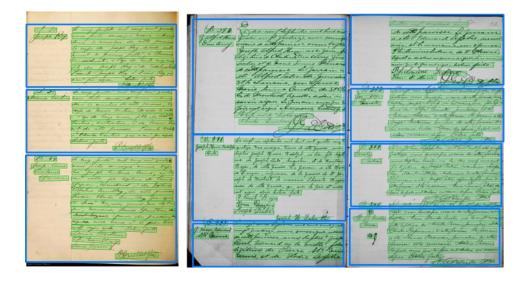
#### Manual transcription versus HTR (auto)

A Comprehensive Study of Open-source Libraries for Named Entity Recognition on Handwritten Historical Documents; Bizon, Miret, Bonhomme and Kermorvant. DAS 2022

Pierre H . Marchand , ptre Le \_\_\_\_\_ seize janvier mil neuf cent trois nous prêtre soussigné avons baptisé **PRENOM** Joseph Louis Amand né DATE la veille, fils légitime de PERSONNE Joseph Toupin \_\_\_\_\_PROFESSION cultivateur et de \_\_\_\_\_PERSONNE Huard Toupin de LIEU\_RESIDENCE cette paroisse . Parrain \_PERSONNE Louis Toupin PROFESSION journalier, marraine PERSONNE Jeanne Toupin épouse du parrain tous deux de LIEU\_RESIDENCE cette paroisse les sec ainsi que le père ont signé avec nous. Lecture faite . Jeanne Toupin Louis Toupin Joseph Coupin H . Shille Lessard Ptre Le **\_\_\_\_\_** dix neuf janvier , mil neuf cent trois , nous prêtre soussigné avons baptisé **PRENOM** Jean Charles Audre, né DATE ce jour , fils légitime de PERSONNE Aimé Turcotte \_PROFESSION cultivateur et de \_\_PERSONNE Adrienne machand de LIEU\_RESIDENCE cette paroisse , du parrain , \_prenom Sémin \_PROFESSION marchand, fils de Ferdinand Blanche \_PROFESSION <u>Cultivateur</u>, de <u>LIEU\_RESIDENCE</u> <u>Balisceau</u>, marraine <u>PERSONNE</u> Blanche Mongrain fille de Jec Louis Mongrain Auavigateur de LIEU\_RESIDENCE cette paroisse, lesquels ainsi que le père ont signé avec nous. Lecture faite - Blanche Mongrain Dem Marchand Aimé Turcotte cing a Emile Messard ptre

## Act Segmentation

- Text lines and entities must be grouped by act
- On a page, one can find one/several full acts, a beginning or and end of act
- Act detection model combining text line layout and keywords (date, and "lecture faite")



Including Keyword Position in Image-based Models for Act Segmentation of Historical Registers; Boillet, Maarand, Paquet & Kermorvant; HIP2021

Structuring the result in meaningful units is crucial and can be tricky

## Deliver the results

- Consolidation of the text and entities at act level for 44 742 registers
- Export to gitlab to benefit from version management
- Issue on export time: 800 hours ! Raw sql optimization for queries + SSD: 8 hours



## Results: quantitative information

Туре	Count
Parishes	1 985
Registers	44 742
Pages	2 635 038
Acts	5 591 535
Paragraphs	15 668 671
Text lines	111 270 929

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

## Results: automatic data validation with act template

	Field	Birth	Death (married person)	Death (single person)	
Record	Date of the record Place of the record	✓ (✓)	✓ (✓)	✓ (✓)	type
Subject	Name of the subject Date of the event Age of the subject	✓ (✓) (✓)	✓ (✓) (✓)	✓ (✓) (✓)	Valid
Father	Name of the father Occupation of the father Residency of the father	✓ (✓) (✓)		✓ (✓) (✓)	Fusion
Mother	Name of the mother Occupation of the mother Residence of the mother	✓ (✓) (✓)		() () () ()	Invalid
Spouse	Name of the spouse Occupation of the spouse Residency of the spouse		✓ (✓) (✓)		Special
Godfather	Name of the godfather Occupation of the godfather Residence of the godfather	(✔) (✔) (✔)			
Godmother	Name of the godmother Occupation of the godmother Residence of the godmother	(✔) (✔) (✔)			
Witnesses	Names of the witnesses		(✔)	(✔)	

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%

#### Data presence validation: $\checkmark$ mandatory, ( $\checkmark$ ) optional

## Conclusions

- Annotation of a <u>representative sample</u> of the full corpus is crucial (it means annotating irrelevant pages)
- <u>Structuring</u> the result in meaningful units is crucial and can be tricky
- <u>Unsupervised metrics</u> 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 : See "*Key-value information extraction from full handwritten pages*" at 5:30pm

# Thank you !



#### kermorvant@teklia.com

