

Large Scale Genealogical Information Extraction From Handwritten Québec Parish Records

Solène Tarride · Martin Maarand · Mélodie Boillet · James McGrath · Eugénie Capel · Hélène Vézina · Christopher Kermorvant

TEKLIA, Paris, France

UQAC, Chicoutimi, Québec

TEKLIA

ICDAR 2023

August 21, 2023

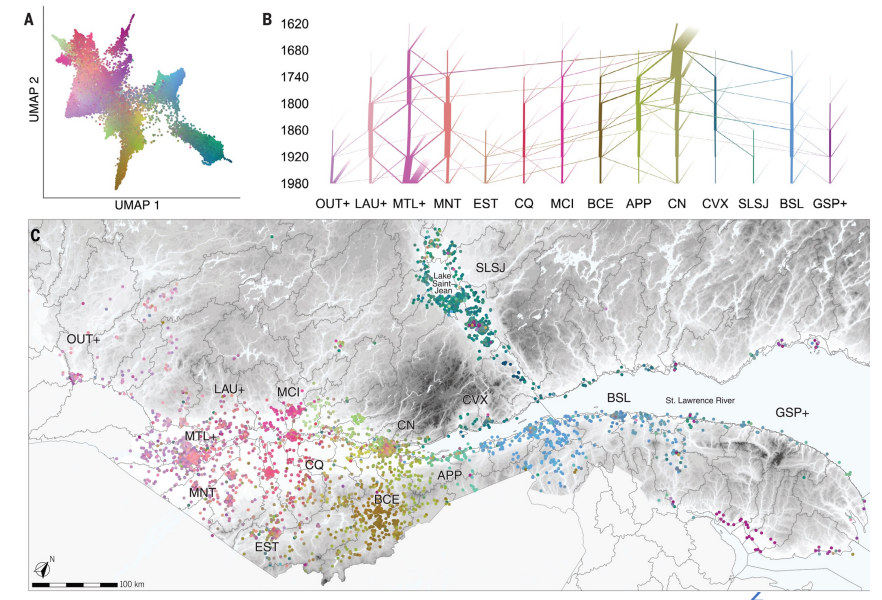
UQAC

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
Science, 25 May 2023



The e-Balsac project

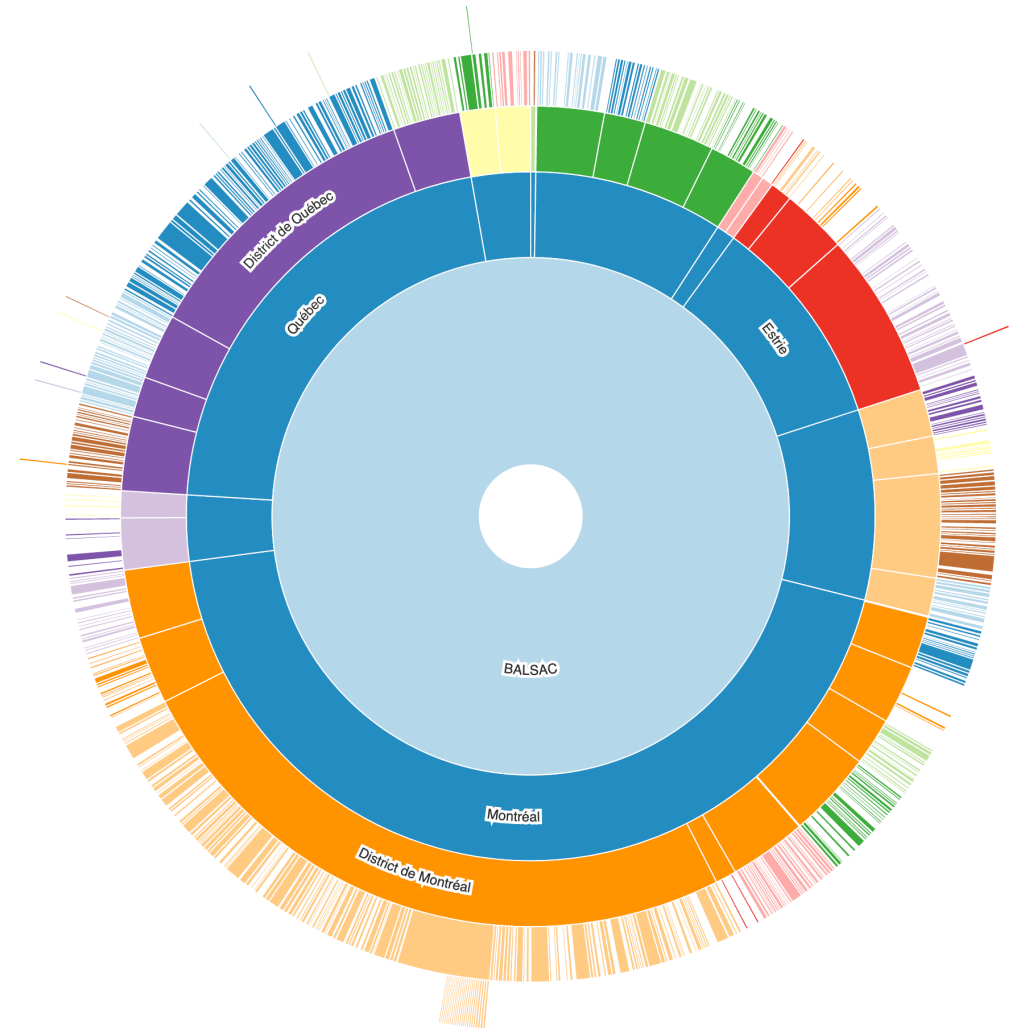
- Database built for more than 50 years
- Extracted from transcribed parish and civil registers
- First only marriages, now birth and death records
- Manual transcription is no longer feasible: the Quebec population was multiplied by 6 between 1851 and 1961

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

210

M. 71 Le sept septembre mil huit cent quarante, ou la publication de
 Pierre Chénier nous avons de mariage, fait au jour de nos paroissiaux
 entre Pierre Chénier, fils mineur de Louis Chénier et d'Elisabeth
 Bonadette, tous deux de cette paroisse d'une part et Bernardette de Louis, fils mineur
 de feu Louis de Louis et de Joséphine Landry, veuve de cette
 paroisse d'autre part; nous n'étant intervenus aucun empêchement
 ni de mariage et de consentement des parents de la dite
 mineurs, nous, prêtre consacré, avons reçu leur mutuel
 consentement de mariage et leur avons donné la béné-
 diction nuptiale en présence de Louis Chénier, père de
 l'époux et de Louis Corage, beau-père de l'épouse, lesdits
 deux ont signé avec nous. Lecture faite.

Lois Chénier
 Bernardette de Louis

M. 72 Le sept septembre mil huit cent quarante, ou la dispense
 Marie Renaud de son beau de mariage, par nous accordée en vertu de provision
 et nous conférés par sa grandeur Honorables Messieurs
 Jacques Sauthier, archevêque d'Acadie, ou aussi le prêtre
 d'un bon fait au jour de notre messe paroissiale
 ainsi qu'à celui de Notre Dame d'Acadie, entre Marie Renaud
 fils majeur de Marie Renaud et d'Albale Robitaille de
 la dite paroisse d'une part et Bernard Paguette, fille
 mineure de Olivier Paguette et d'Anna Miron de cette paroisse
 d'autre part; nous n'étant intervenus aucun empêchement
 de mariage et de consentement des parents de la dite
 mineurs, nous, prêtre consacré, avons reçu leur mutuel
 consentement de mariage et leur avons donné la béné-
 diction nuptiale en présence de Marie Renaud, père de l'époux
 et d'Olivier Paguette, père de l'épouse, tous deux ont signé avec
 nous. Lecture faite.

Marie Renaud
 d'Albale Robitaille

M. 73 Le sept septembre mil huit cent quarante
 nous avons de mariage, fait au jour de nos paroissiaux
 entre André, fils mineur de Louis André et d'Elisabeth
 Florida, tous deux de cette paroisse d'une part et Bernardette de Louis, fils mineur
 de Louis de Louis et de Joséphine Landry, veuve de cette
 paroisse d'autre part; nous n'étant intervenus aucun empêchement
 ni de mariage et de consentement des parents de la dite
 mineurs, nous, prêtre consacré, avons reçu leur mutuel
 consentement de mariage et leur avons donné la béné-
 diction nuptiale en présence de Louis André, père de l'époux
 et de Louis Corage, beau-père de l'épouse, lesdits
 deux ont signé avec nous. Lecture faite.

André
 Bernardette de Louis

Le sept septembre mil huit cent quarante, ou la publication de
 mariage et leur avons donné la ben-
 diction nuptiale en présence de Thomas Bergeron
 père de l'époux et de André-Christophe ami
 de l'épouse qui tous ensemble ont signé avec nous.
 Lecture faite.

André-Christophe
 Thomas Bergeron

M. 31 Marie Emelie
 Dornier

M. 32 Marie
 Wilhelmine
 Legerie

M. 33 Joseph Jean
 Legerie

M. 34. 16
 Anonyme
 Sirenon

Le sept septembre mil huit cent quarante, ou la publication de
 mariage et leur avons donné la ben-
 diction nuptiale en présence de Thomas Bergeron
 père de l'époux et de André-Christophe ami
 de l'épouse qui tous ensemble ont signé avec nous.
 Lecture faite.

André-Christophe
 Thomas Bergeron

Le sept septembre mil huit cent quarante, ou la publication de
 mariage et leur avons donné la ben-
 diction nuptiale en présence de Thomas Bergeron
 père de l'époux et de André-Christophe ami
 de l'épouse qui tous ensemble ont signé avec nous.
 Lecture faite.

André-Christophe
 Thomas Bergeron

Le sept septembre mil huit cent quarante, ou la publication de
 mariage et leur avons donné la ben-
 diction nuptiale en présence de Thomas Bergeron
 père de l'époux et de André-Christophe ami
 de l'épouse qui tous ensemble ont signé avec nous.
 Lecture faite.

André-Christophe
 Thomas Bergeron

Le sept septembre mil huit cent quarante, ou la publication de
 mariage et leur avons donné la ben-
 diction nuptiale en présence de Thomas Bergeron
 père de l'époux et de André-Christophe ami
 de l'épouse qui tous ensemble ont signé avec nous.
 Lecture faite.

André-Christophe
 Thomas Bergeron

Three folios E.P. C.C. B.R.

This book or Register contains three
 two folios or double pages and presented
 by the Reverend R.S. Booy, Incumbent of
 the Church of the Good Shepherd at Bond-
 ville the Bishop Carmichael Memorial
 Church at Boston, to be used for the celebra-
 tion of the Act of Baptism, Marriage &
 Burial to be performed by him or his
 Successor in office for the year one thou-
 sand nine hundred & forty two in said
 County & Vicinity.

Given & sealed at Montreal under my hand
 & the Seal of the Circuit Court of the Pro-
 vince of Quebec in the County of Brome, in the
 District of Bedford, in due & lawful order of the
 13th of the Code of Civil Procedure article
 42-42 B, 42-B-42-C of the said Code
 of the said Province of Quebec.

At Montreal this twentieth day of January
 one thousand hundred & forty two

J. J. J. J.
 Deputy. C.C. B.R.

Notre Dame de Grâce de Hull,
 1914

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

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

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 Felicité Cantin, épouse du parrain, lesquels ont signé avec nous et le père. Lecture faite.

Maria Felicité Cantin
Pierre Beaumont
Nurse Beaumont

Jos. Saguellet

Birth act

Le 10
Diana
Duprene

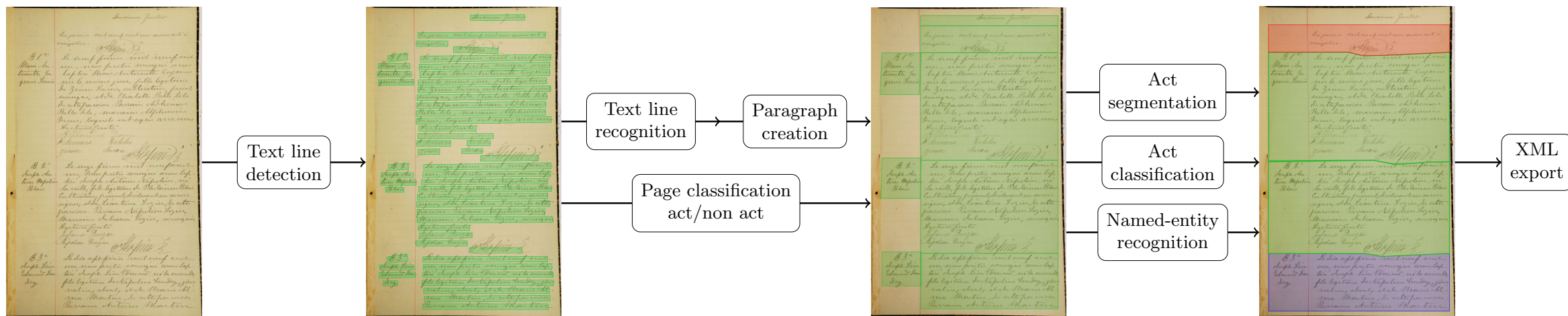
Le vingt sept mars mil neuf cent deux, nous prêtre, curé sousigné, avons inhumé dans le cimetière de cette paroisse le corps de Diana Duprene, décédée l'avant veille dans cette paroisse à l'âge de quatre ans, fille légitime de Victor Duprene, benêtier et de Albertine Pathin de cette paroisse. Prévôt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

Edm. P. de la Cour

Death of a single person act

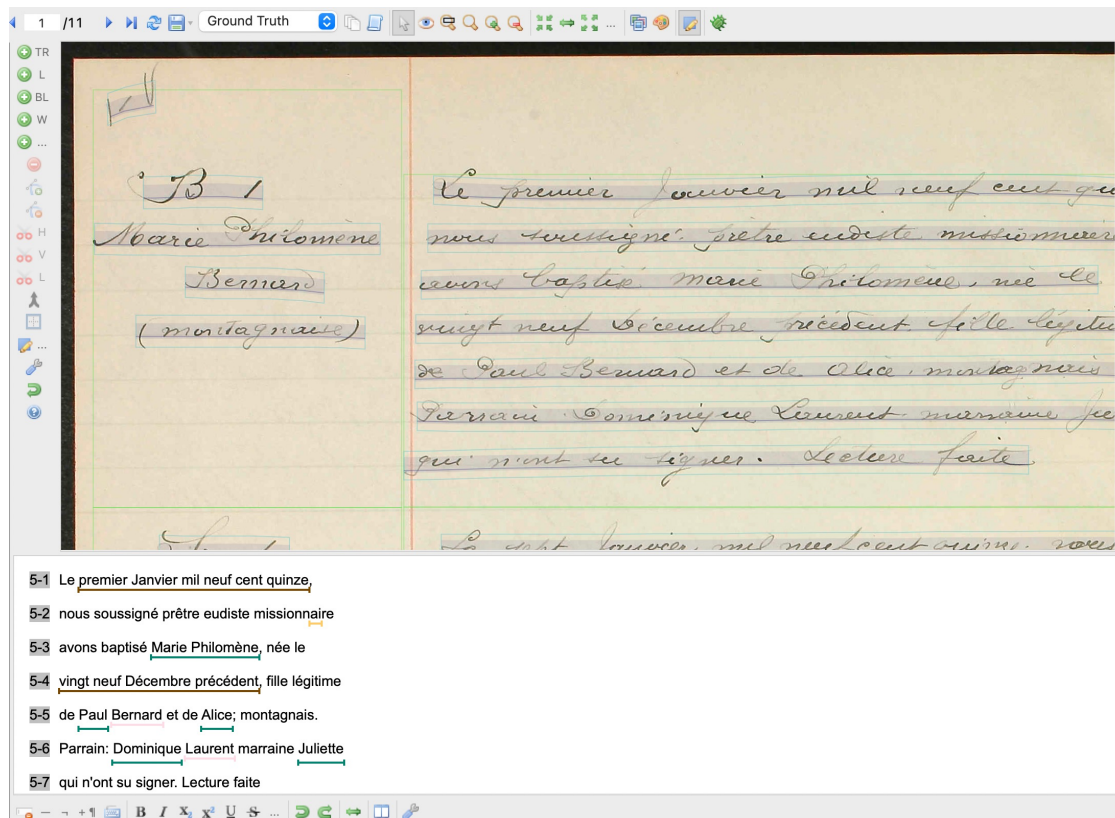
■ Date; ■ Subject of the act; ■ Parents; ■ Spouse; ■ Age; ■ Occupation; ■ Godfather/godmother

Overview of the document processing workflow



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

Ground truth generation



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

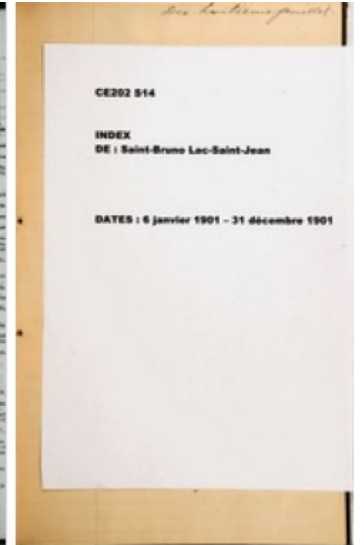
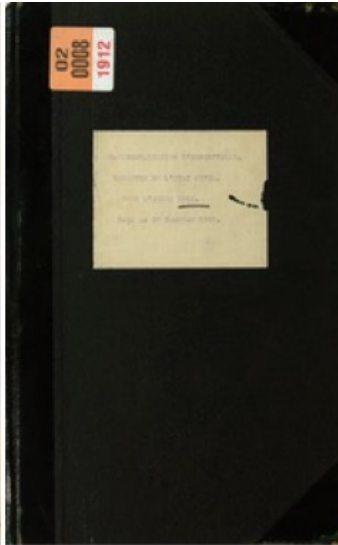
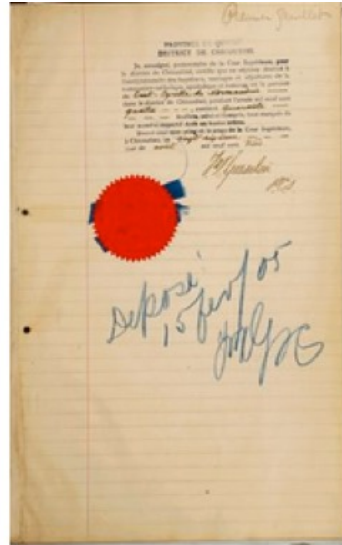
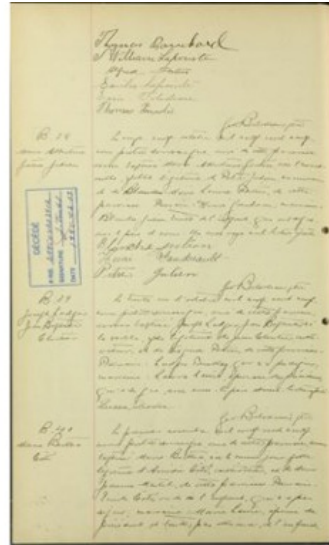
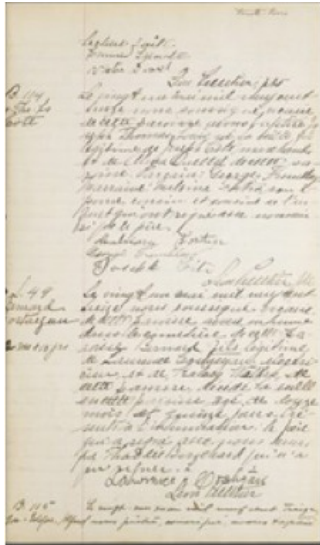
	pages	acts	lines	words	entities
count	896	2,661	45,479	205,165	25,564

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

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

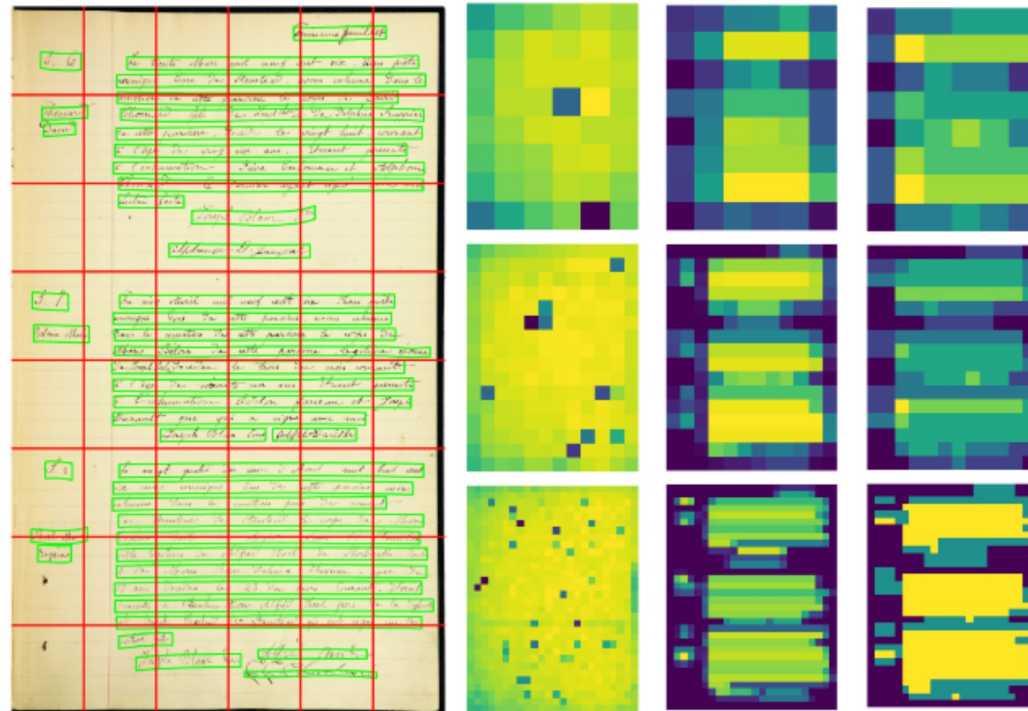
Page types: Act vs Non Act

Pages containing acts are very homogenous

Non act pages are very diverse

Training of an outlier detector with isolation forest

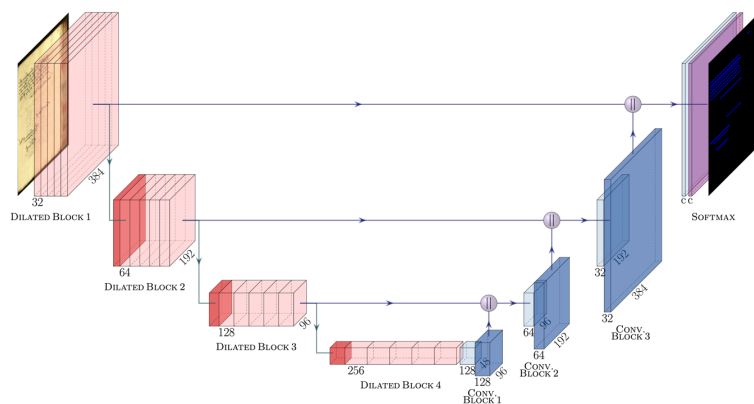
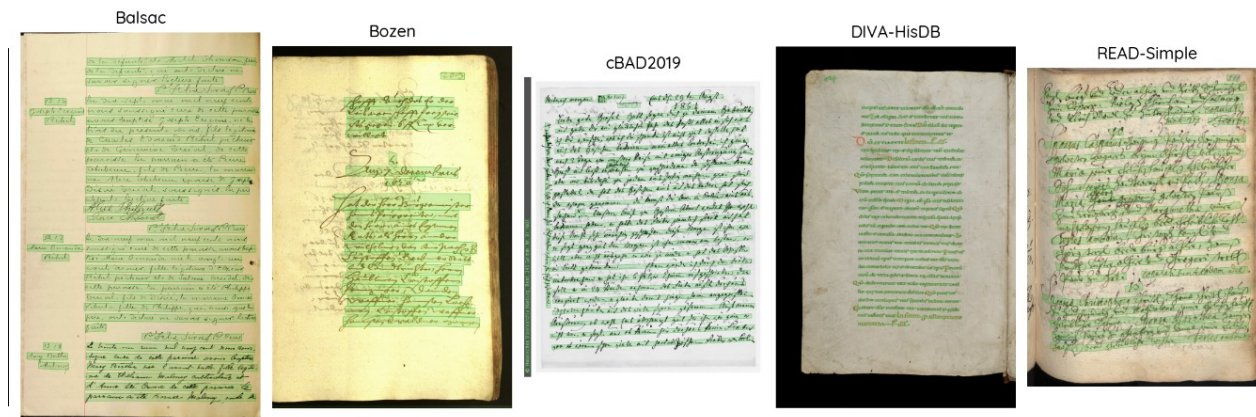
100% recall on acts, 97% f-score



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



Doc-UFCN model

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

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

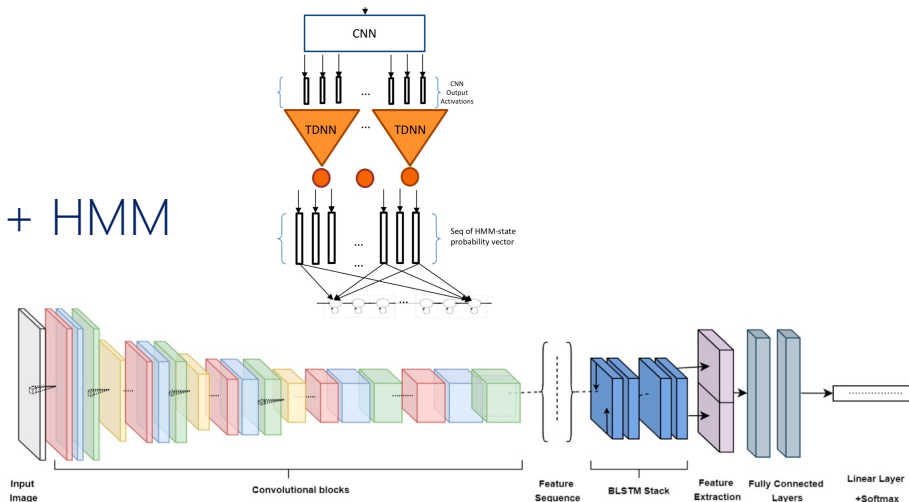
	IoU	AP@.50	AP@.75	mAP
Doc-UFCN	0.87	0.98	0.91	0.76
dhSegment	0.74	0.94	0.54	0.51
ARU-Net	0.98	0.76	0.20	0.34

Handwritten Text Recognition

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

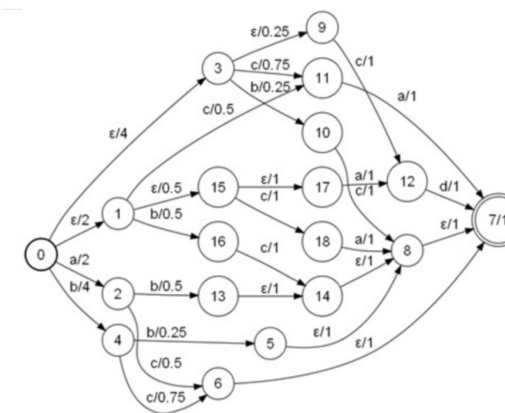
	train		val		test	
	CER	WER	CER	WER	CER	WER
Kaldi	4.13	12.36	6.22	17.10	6.41	17.41
PyLaia (no LM)	2.52	9.67	6.53	19.75	6.92	20.41
PyLaia (with LM)	2.24	8.27	5.90	17.16	6.29	17.92

Kaldi
CNN/TDNN + HMM



PyLaia
CNN+biLSTM

+



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

	Text	Precision	Recall	F1-score
Stanza	manual	77.9	85.5	81.5
Stanza	auto.	69.7	78.3	73.7
Flair	manual	93.7	93.1	93.4
Flair	auto.	86.0	85.6	85.8
Spacy	manual	83.1	83.6	83.4
Spacy	auto.	74.5	76.7	75.6

Manual transcription versus HTR (auto)

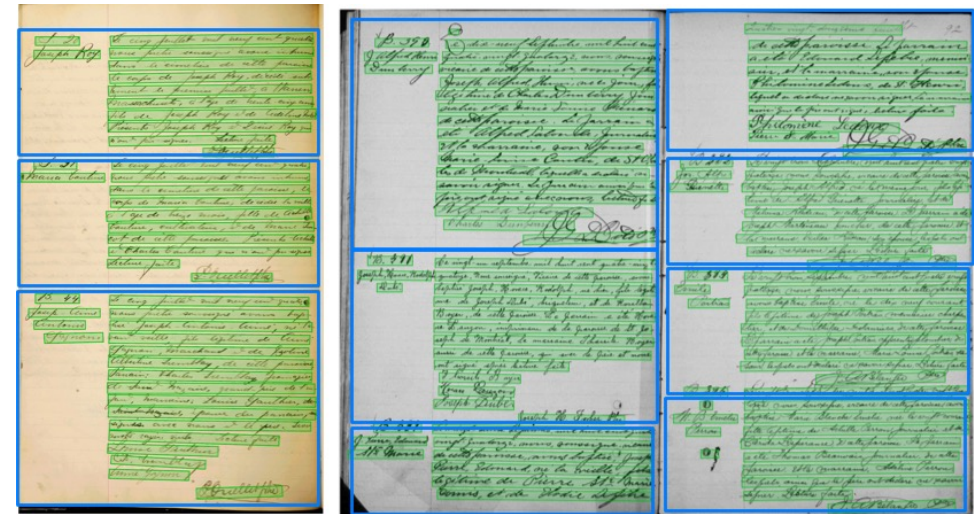
A Comprehensive Study of Open-source Libraries for Named Entity Recognition on Handwritten Historical Documents; Bizon, Miret, Bonhomme and Kermorvant. DAS 2022

Pierre H . Marchand , ptre Le **_DATE** 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 **_DATE** 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** Cultivateur , de **_LIEU_RESIDENCE** Baliscoeu , marraine **_PERSONNE** 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 cinq a Emile Messard ptre

<https://demo.arkindex.org/element/7fe3d786-068f-48c9-ae63-86db2f986c4c>

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

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Register id="9d6e041c-d510-4db8-b646-6106e4c3a5b0" name="test_balsac_export_v2">
  <Acts>
    <Act id="003f73e1-33e7-47c3-8e38-e87d75c63b5a" elem_type="act" name="4"
      act_type="act_start" act_type_score="0.78" act_class="N/A" act_class_score="0"
      page_id="54adb4cc-590e-488c-94b1-c9466d43c5bd"
      balsac_ref="02C_CE202S18_1913_018.jp2"
      polygon="[[34, 5424], [3273, 5461], [3278, 5004], [39, 4966], [34, 5424]]">
      <Paragraphs>
        <Paragraph id="13fb5714-7150-417e-ad83-06850858db85" score="0.465"
          rect="(974, 5069, 2276, 230)">
          <Lines>
            <Line id="6467f9ed-37cc-46ea-992a-9c8254d24d44" score="0.5816"
              rect="(978, 5069, 2276, 89)">Le dix neuf aout mil neuf cent</Line>
            <Line id="26327126-d1c5-430d-ab9e-7f3da533e302" score="0.5942"
              rect="(974, 5211, 2280, 88)">treize , nous curé soussigné avons</Line>
          </Lines>
          <text_with_entities>Le <date id="23b814af-f856-42eb-a0b5-61f98591712f"
            name="dix neuf aout mil neuf cent
            treize">dix neuf aout mil neuf cent
            treize</date> , nous curé soussigné avons</text_with_entities>
          </Paragraph>
        </Paragraphs>
      </Act>
    </Acts>
  </Register>
```

Results: quantitative information

Type	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	(✓)	(✓)	(✓)
Subject	Name of the subject	✓	✓	✓
	Date of the event	(✓)	(✓)	(✓)
	Age of the subject	(✓)	(✓)	(✓)
Father	Name of the father	✓		✓
	Occupation of the father	(✓)		(✓)
	Residency of the father	(✓)		(✓)
Mother	Name of the mother	✓		✓
	Occupation of the mother	(✓)		(✓)
	Residence of the mother	(✓)		(✓)
Spouse	Name of the spouse		✓	
	Occupation of the spouse		(✓)	
	Residency of the spouse		(✓)	
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		(✓)	(✓)

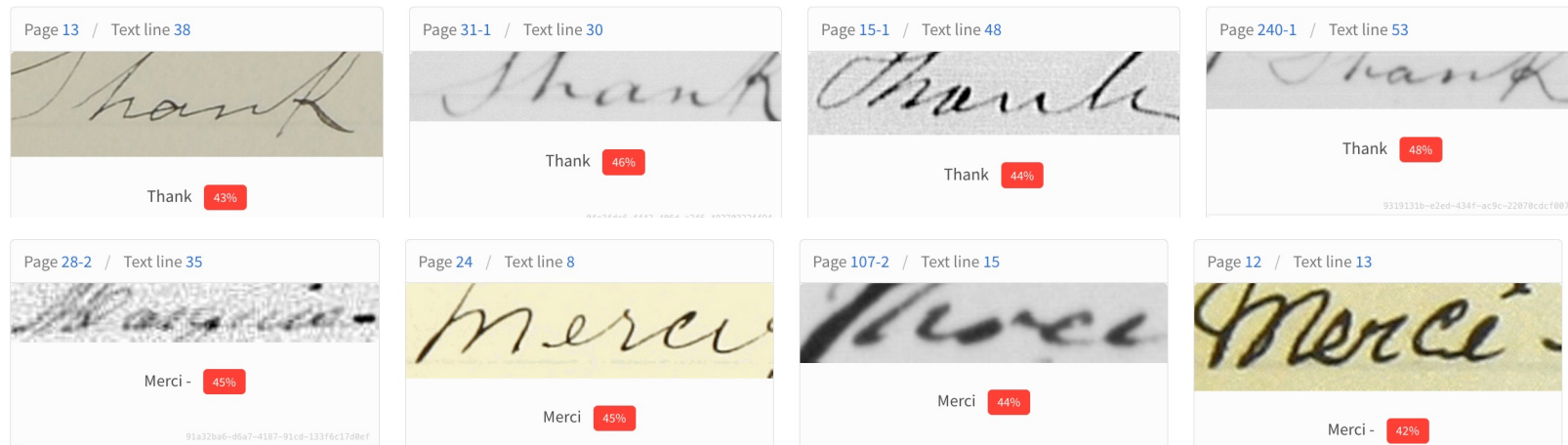
type	Birth act	Death act
Valid	75.5%	69.0%
Fusion	19.9%	18.2%
Invalid	7.1%	8.9%
Special	0.4%	3.2%

Data presence validation: ✓ mandatory, (✓) optional

Conclusions

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

Thank you !



kermorvant@tekliia.com

TEKLIIA