Recognition and information extraction in historical handwritten tables: toward understanding early 20th century Paris census

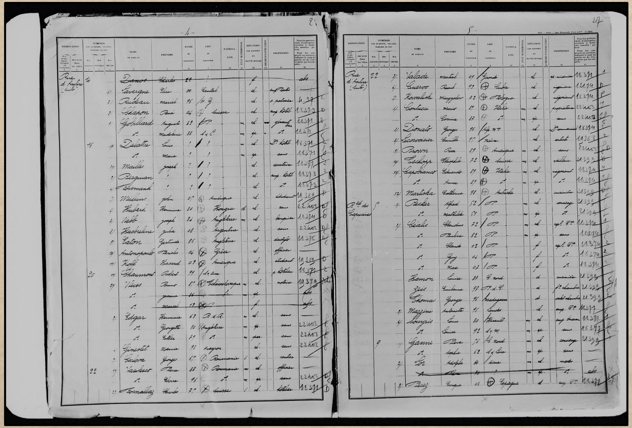
Thomas CONSTUM[1] , Nicolas KEMPF[1] , Thierry PAQUET[1] ,
Pierrick TRANOUEZ[1] , Clément CHATELAIN[2] , Sandra BREE[3] ,
François MERVEILLE[4]

1: LITIS EA4108, University of Rouen Normandy, France
2: LITIS EA4108, INSA Rouen Normandy, France
3: LARHRA, UMR 5190, CNRS, France
4: Campus Condorcet - GED, France

DAS, 23rd May 2022

Introduction

- The 20th century census of Paris (1926, 1931 and 1936) contain information on approximately 9 million individuals in total.
- Demograph historians could use the content of these census to answer questions such as:
  - What is the proportion of divorced, married, or cohabiting individuals in Paris in 1926 by district?
  - How did the structure of households in Paris evolve between 1926 and 1936 in terms of number of individuals and number of children?

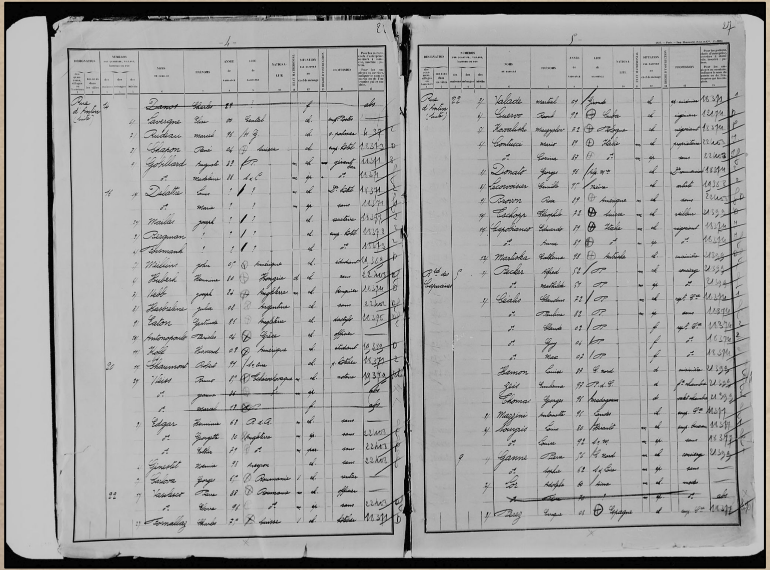


Example of a double page from the Paris census.
(1926 census - Gaillon district)

Introduction

The POPP project (Project of Ocerization of the Parisian Population)

Aim: Get tabular data from double page scans in order to create a database containing information of about 9 million individuals.

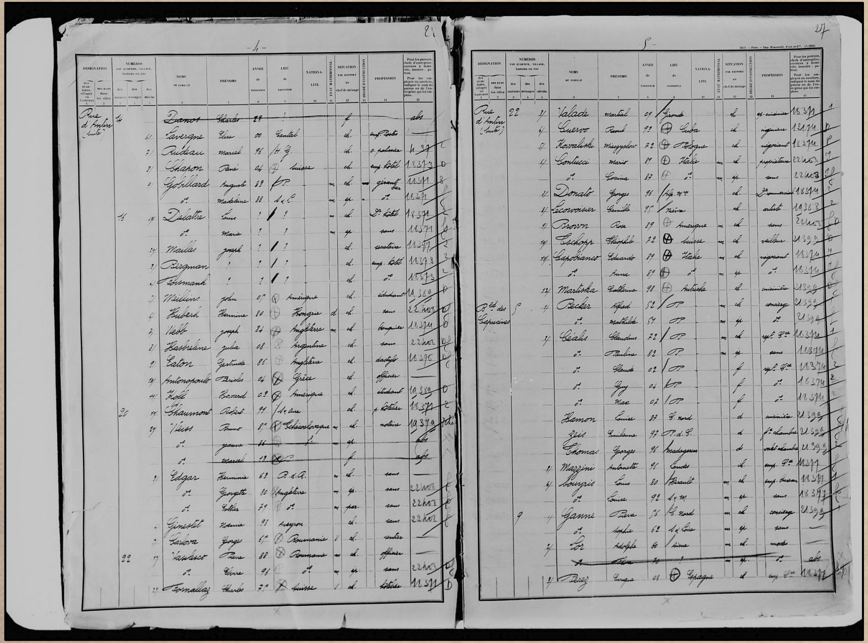


Example of a double page from the POPP corpus.
(1926 census - Gaillon district)

Noms	Prenom	Sexe	Annee de naissance	Ville de naissance	Departement de naissance	Pays de naissance	Nationalite	Etat matrimonial	Situation par rapport au chef de menage	Statut profession	Profession	Code metier
VALADE	MARTIAL	H	1909		GIRONDE	FRANCE	FRANCAISE	C	CH	APP	APPRENTI CUISINIER	18371
CUERVO	RAOUL	H	1893			CUBA	CUBA	C	CH		INGENIEUR	12174
KOVVALISKI	$MIEEZYSLOV$	A	1872			POLOGNE	POLOGNE	C	CH		NEGOCIANT	18374
CONTUCCI	MARIO	H	1889			ITALIE	ITALIE	M	CH		PROPRIETAIRE	22408
CONTUCCI	$CORRINA$	F	1887			ITALIE	ITALIE	M	EP		SANS	22408
DONATO	GEORGES	H	1896		ALPES-MARITIMES	FRANCE	FRANCAISE	C	CH	D	DOMESTIQUE COMMERCIAL	18374
LECORVOISIER	$CAMELLE$	A	1895		NIEVRE	FRANCE	FRANCAISE	C	CH		ARTISTE	19388
BROVN	ROSE	F	1889			AMERIQUE	AMERIQUE	M	CH		SANS	22408
ESCHOPP	THEOPHILE	H	1872			SUISSE	SUISSE	M	CH		VEILLEUR	21399
CAPOBIANCO	EDUARDO	H	1889			ITALIE	ITALIE	M	CH		NEGOCIANT	18374
CAPOBIANCO	ANNA	F	1889			ITALIE	ITALIE	M	EP		NEGOCIANT	18374
MARTISKA	CATHERINE	F	1898					C	CH		CUISINIERE	21399
BECKER	ALFRED	H	1852	PARIS	SEINE	FRANCE	FRANCAISE	M	CH		CONCIERGE	21399
BECKER	MATHILDE	F	1851					M	EP		CONCIERGE	21399
CEALIS	CLAUDIUS	H	1872	PARIS	SEINE	FRANCE	FRANCAISE	M	CH		REPT CCE	18374
CEALIS	PAULINE	F	1882	PARIS	SEINE	FRANCE	FRANCAISE	M	EP		SANS	18374
CEALIS	CLAUDE	H	1902	PARIS	SEINE	FRANCE	FRANCAISE	C	F		REPT CCE	18374
CEALIS	GUY	H	1904	PARIS	SEINE	FRANCE	FRANCAISE	C	F		REPT CCE	18374
CEALIS	MAX	H	1907	PARIS	SEINE	FRANCE	FRANCAISE	C	F		REPT CCE	18374
HAMON	LOUISE	F	1887		COTES-DU-NORD	FRANCE	FRANCAISE	C	D		CUISINIERE	21399
ZEIS	EMILIENNE	F	1897		PAS-DE-CALAIS	FRANCE	FRANCAISE	C	D		FE CHAMBRE	21399
THOMAS	GEORGES	H	1896			MADAGASCAR	MADAGASCAR	C	D		VALET CHAMBRE	21399
MAZZINI	ANTOINETTE	F	1896		LANDES	FRANCE	FRANCAISE	C	CH	E	EMPLOYE CCE	18377
SOUYRIS	LOUIS	H	1880		HERAULT	FRANCE	FRANCAISE	M	CH	E	EMPLOYE BUREAU	18377
SOUYRIS	LOUISE	F	1892		SEINE-ET-MARNE	FRANCE	FRANCAISE	M	EP		SANS	18377
GANNE	PIERRE	H	1856		COTES-DU-NORD	FRANCE	FRANCAISE	M	CH		CONCIERGE	21399
GANNE	SOPHIE	F	1862		SAONE-ET-LOIRE	FRANCE	FRANCAISE	M	EP		SANS	
LOR	ADOLPHE	H	1860		AISNE	FRANCE	FRANCAISE	M	CH		$MEDES$	
LOR	FLORE	F	1870	ABS	ABS	FRANCE	FRANCAISE	M	EP	ABS	$MEDES$	ABS
PEREZ	$EURIQUE$	A	1908			ESPAGNE	ESPAGNE	C	CH	E	EMPLOYE CCE	18377
PRETOT	ERNEST	H	1872	ALGER		ALGERIE	ALGERIE	M	CH		NEGOCIANT	18374
PRETOT	LOUISE	F	1878					M	EP		SANS	18374
PRETOT	MARTHE	F	1905	ALGER		ALGERIE	ALGERIE	C	F		SANS	18374
FERRASSON	ROLANDE	F	1909	ALGER		ALGERIE	ALGERIE	C	D		FE CHAMBRE	21399
CLEMENT	ALBERTINE	F	1894	ALGER		ALGERIE	ALGERIE	C	D		CUISINIERE	21399
GOUJON	LOUIS	H	1873		AISNE	FRANCE	FRANCAISE	M	CH		COMPTABLE	18375
JOUJON	LOUIS	H	1873		AISNE	FRANCE	FRANCAISE	M	CH		COMPTABLE	18375
JOUJON	MARIE	F	1874		AISNE	FRANCE	FRANCAISE	M	EP		SANS	18375
JOUJON	$RENE$	A	1908		AISNE	FRANCE	FRANCAISE	C	F	APP	APPRENTI EBENISTE	11158
ARIVE	FERNANDO	H	1900			ESPAGNE	ESPAGNE	C	CH	E	EMPLOYE CCE	18377
MARAIS	CONSTANT	H	1859		LOIR-ET-CHER	FRANCE	FRANCAISE	C	CH		MD HOTEL	18373
NOYAL	ROBERT	H	1894					M	CH	P	PATRON FLEURISTE	9139
NOYAL	LOUISE	F	1895		NIEVRE	FRANCE	FRANCAISE	M	EP		PATRON FLEURISTE	9139
NOYAL	DENISE	F	1925					C	F			
MARIA	$MIRRA$	A	1883			ITALIE	ITALIE	M	CH	O	OUVRIER COUTURIERE	91132
ARTHEZ	RAOUL	H	1887			BRESIL	BRESIL	C	CH		SANS	22408
MELOUI	$ERNESTA$	A	1851			ITALIE	ITALIE	V	CH		SANS	

Corresponding result in the POPP database.

The POPP corpus



Example of a double page from the POPP corpus.
(1926 census - Gaillon district)

- 3 different census: 1926, 1931 and 1936
- For each census:
  - 100 000 pages
  - 3 million individuals
  - 80 districts
  - between 80 and 500 writers
- Handwritten tabular data with for each individual:
  - 10 columns
  - one row

The POPP processing pipeline

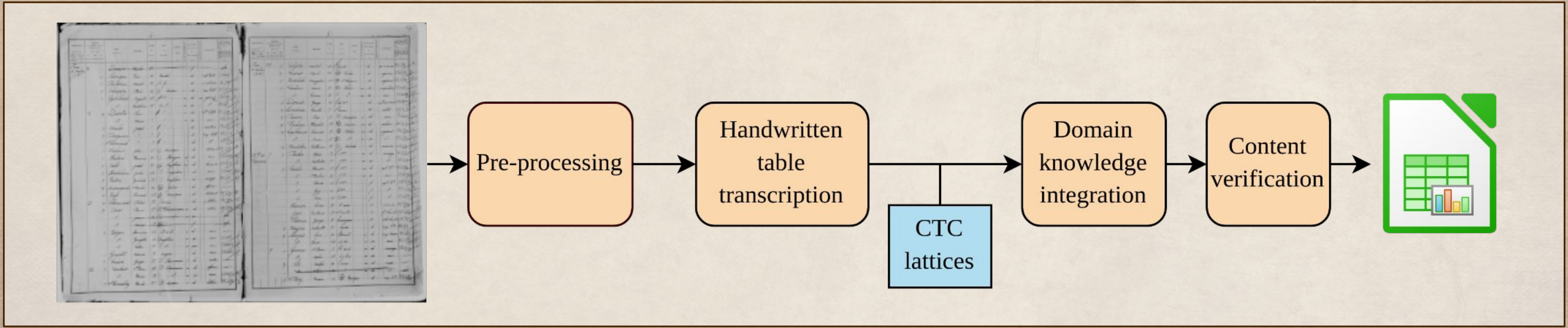


Diagram of the POPP processing pipeline.

Pre-processing steps: information localization

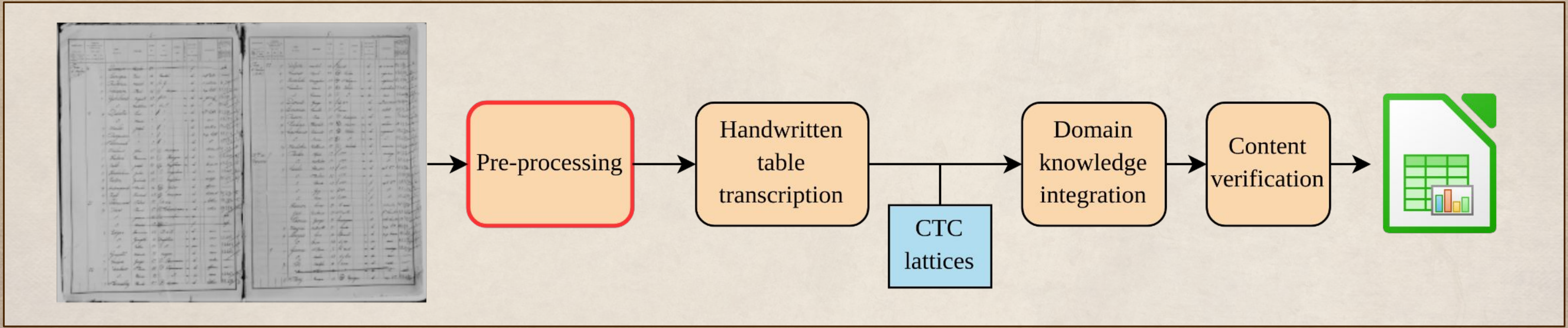
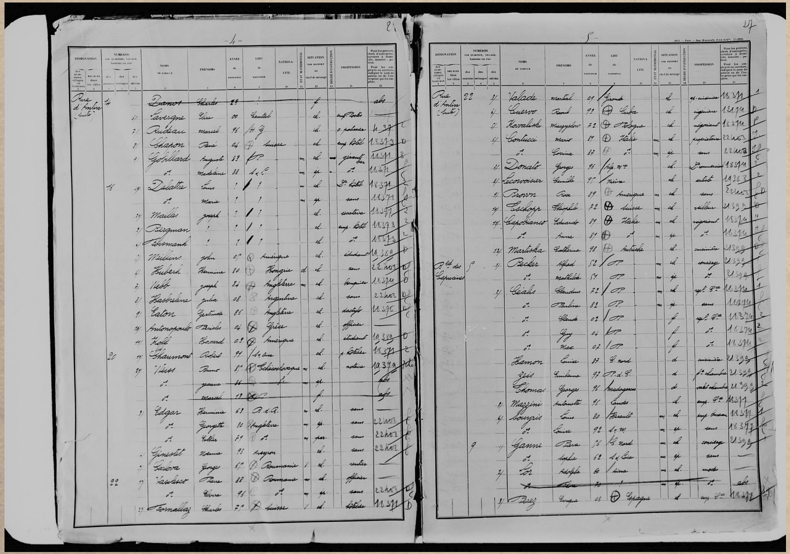


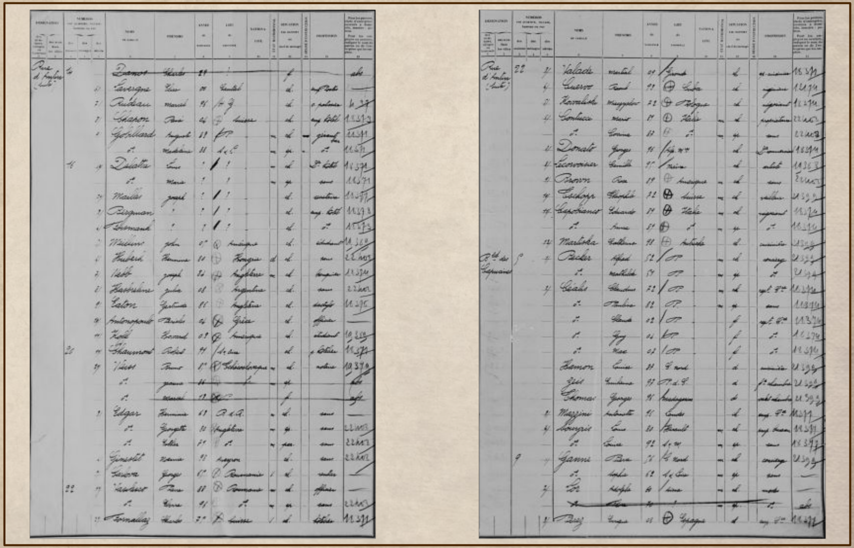
Diagram of the POPP processing pipeline.

Information localization: from double pages to tables

- Segmentation of the tables located in the image using dhSegment[1], a pixelwise predictor, which assigns the class "table" or "background" to each pixel of the image.
- Dewarping of the segmented quadrilaterals to obtain straight tables.



Double page scan.

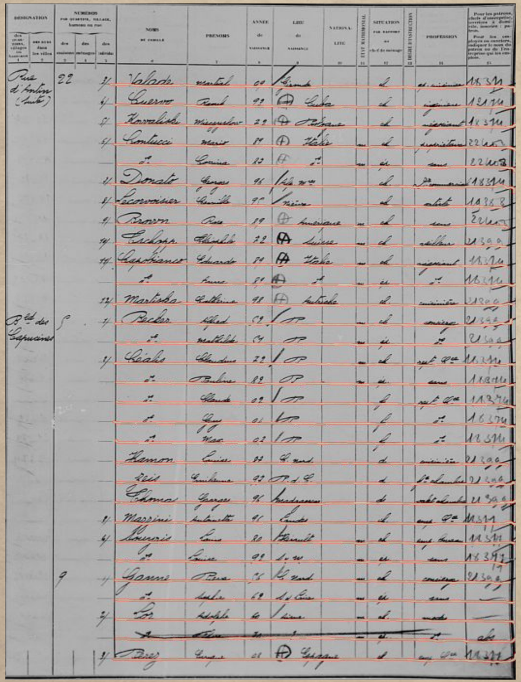


Tables obtained after segmentation and dewarping.

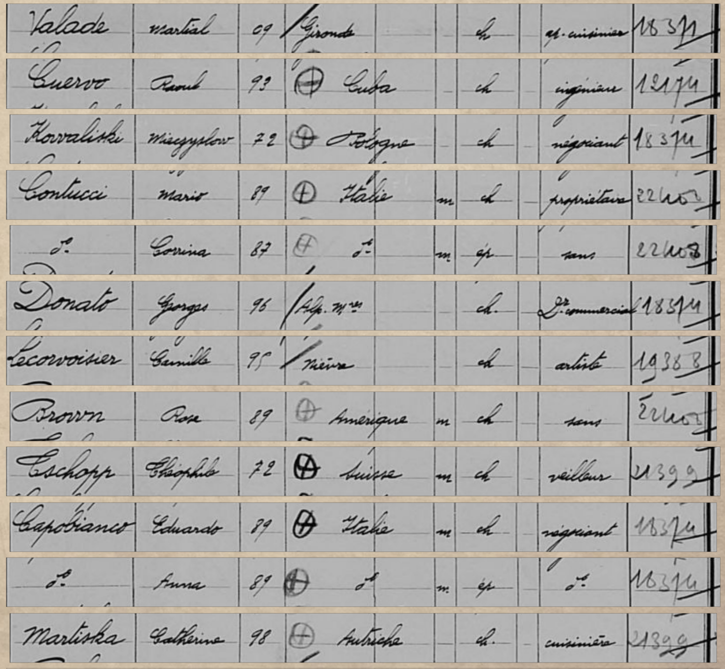
1: S. Ares Oliveira, B.Seguin, and F. Kaplan, “dhSegment: A generic deep-learning approach for document segmentation,” in Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on, pp. 7-12, IEEE, 2018.

Information localization: from tables to table rows

- Detection of the baselines in the table.
- Extraction of a rectangle for each detected baseline, knowing the height of the rows.



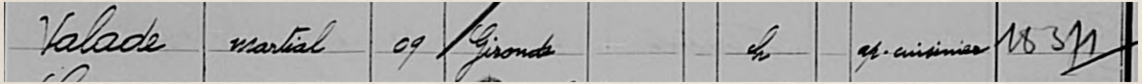
Results of baseline detection on a table.



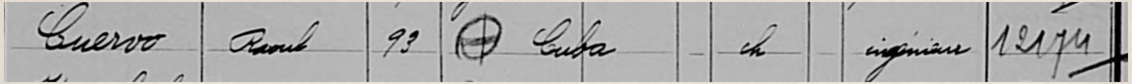
Line images extracted using detected baselines.

Data annotation and logical column segmentation

- Encoding of the logical separation into columns using a semantic token “/”.
- Avoid the need for column segmentation by teaching the HTR model to predict the logical separation into columns beside predicting the handwritten text.

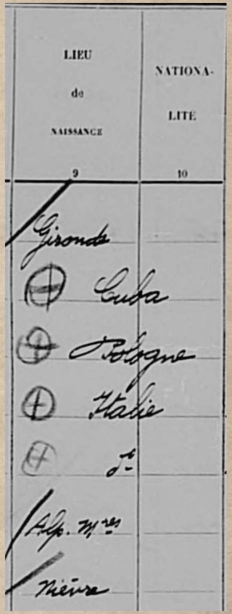


Valade / martial / 09 / Gironde / ¤ / ¤ / ch / ¤ / ap. cuisinier / 18371



Cuervo / Raoul / 93 / Cuba / ¤ / ¤ / ch / ¤ / ingénieur / 12174

Examples of labels.



Example of table where words are overlapping columns.

Dataset	Train # of lines	Validation # of lines	Test # of lines	# of writers
POPP (Generic)	3840 (128 pages)	480 (16 pages)	480 (16 pages)	80
Belleville	1140 (38 pages)	150 (5 pages)	180 (6 pages)	1
Chaussée d’Antin	625	78	77	10

Details about the POPP datasets.

Handwriting recognition

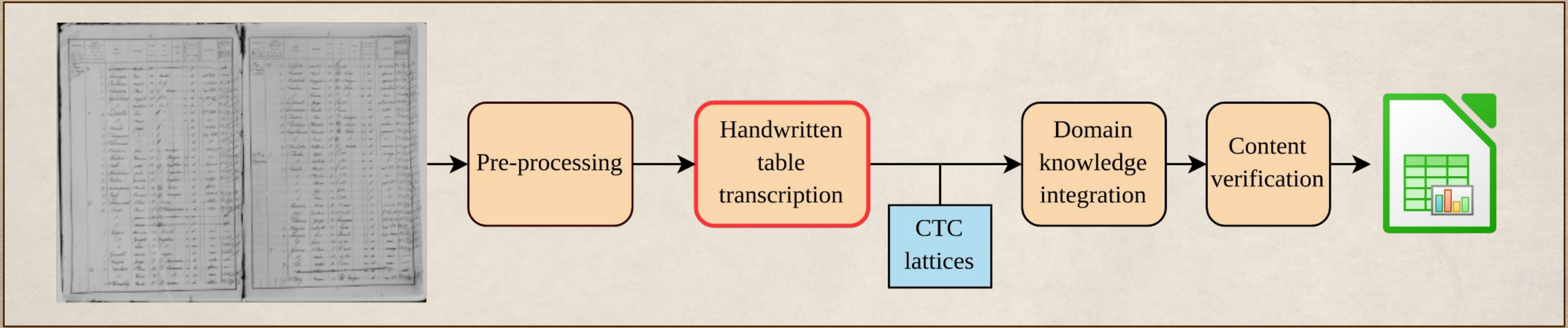


Diagram of the POPP processing pipeline.

Handwriting recognition: supervised learning

- **Architecture:**
  - Line HTR architecture from our team at LITIS described in Coquenet-2022[2].
- **Recognition results:**
  - Generic dataset: Not as good as the results obtained with the same architecture on IAM or RIMES, this dataset seems more challenging.
  - Belleville: shows the improvement that can be expected from writer specialization.

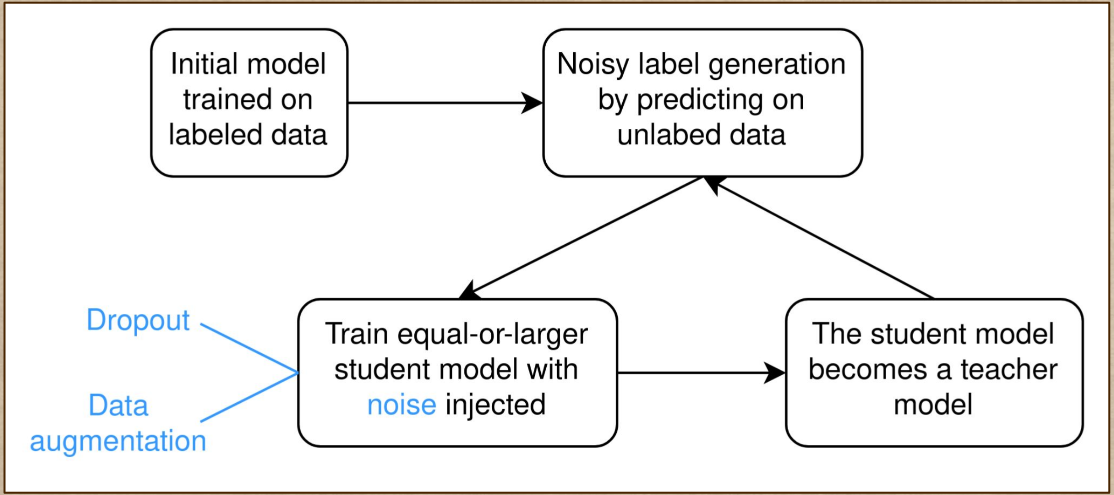
Dataset	Val CER	Val WER	Test CER	Test WER
Generic (multi-writer)	6.86 %	18.66 %	7.08 %	19,05 %
Belleville (writer-specific)	3.36 %	7.47 %	3.65 %	8.65 %

Recognition results obtained on the POPP datasets using supervised learning.

2: D. Coquenet, C. Chatelain and T. Paquet. "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network." *IEEE transactions on pattern analysis and machine intelligence* PP (2022)

Handwriting recognition: self-training

- Method inspired by Q. Xie's 2019[3], which combines self-study and noisy student techniques.
- Useful when a large amount of unlabeled data is available.
- In our case, we use 2.4 million line images selected randomly from the 1926 census.

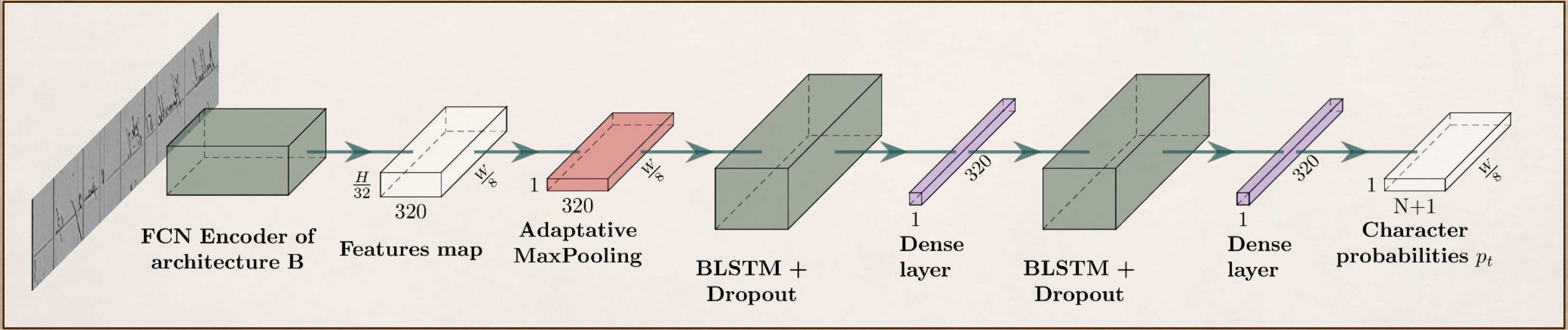


3: Q. Xie, M. -T. Luong, E. Hovy and Q. V. Le, "Self-Training With Noisy Student Improves ImageNet Classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10684-10695

Handwriting recognition: model architectures

During our experiments, 3 architectures were used:

- Architecture A: the line HTR architecture described in Coquenet-2022.
- Architecture B: Architecture A with scaling factor of 1.5 for depth and 1.25 for width.
- Architecture C: Encoder of architecture B with a BLSTM-based decoder.



Schema of architecture C.

Handwriting recognition: results

- A more powerful architecture can further improve the results when a self-training iteration bring no improvement.

- The addition of an LSTM part allows, compared to a fully convolutional model, to implicitly learn a language model from the large volume of data encountered.

Model	Architecture	Dataset	CER (%) test	WER (%) test
Initial (Model 0)	A	Generic	7.08	19.05
Student 1	A	Generic	6.12	17.12
Student 2	A	Generic	5.97	16.83
Student 2*bis*	A	Generic	6.02	16.89
Student 3	B	Generic	5.43	15.50
Student 4	C	Generic	**4.52**	**13.57**
Student 4 specialized	C	Mono-writer	**2.66**	**6.37**

Recognition results obtained using self-training.

Grammar and post-processing steps

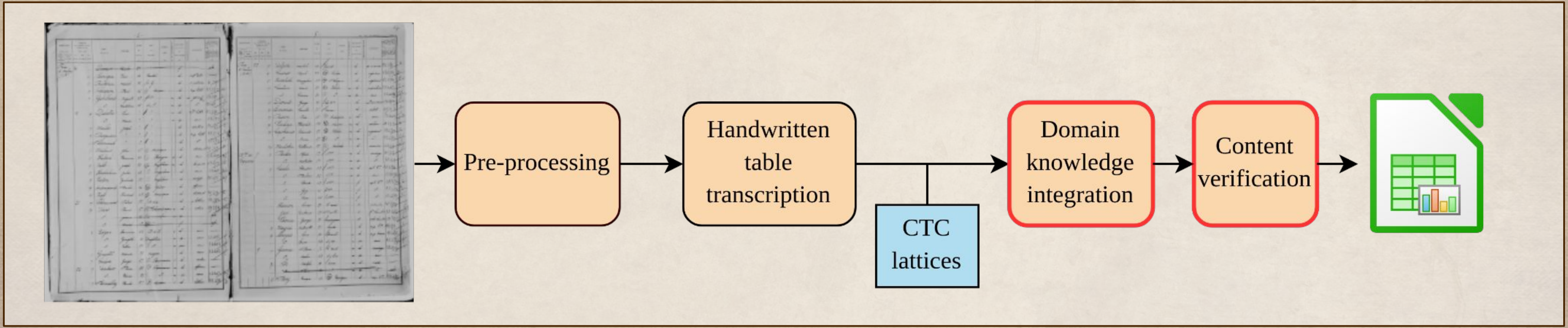


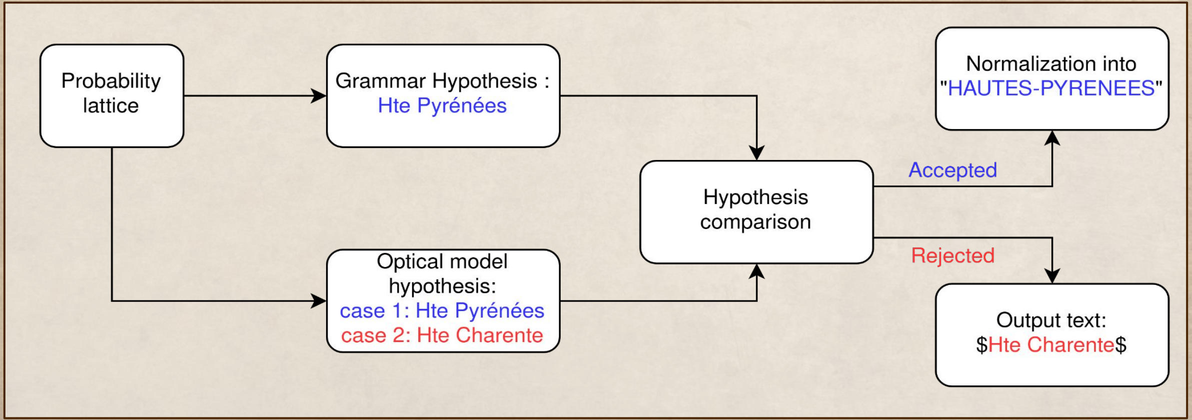
Diagram of the POPP processing pipeline.

Grammar and post-processing steps

- The output of the recognition model is processed by grammars written in Thrax, compiled into Weighted Finite State Transducers[4] and decoded by Kaldi.

- We have created SIGRA, a Python Framework that facilitates the use of grammars for handwriting recognition by linking Thrax and Kaldi.

- The grammars apply a rejection process for each cell and normalize the accepted fields.



4: M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. Computer Speech & Language 16, 1 (2002), 69–88

Grammar and post-processing steps

Valade / martial / 09 / Gironde / ¤ / ¤ / ch / ¤ / ap. cuisinier / 18371
Cuervo / Raoul / 93 / Cuba / ¤ / ¤ / ch / ¤ / ingénieur / 12174

Prediction of the model.

↓

Noms	Prenom	Sexe	Annee de naissance	Ville de naissance	Departement de naissance	Pays de naissance	Nationalite	Etat matrimonial	Situation par rapport au chef de menage	Statut profession	Profession	Code metier
VALADE	MARTIAL	H	1909		GIRONDE	FRANCE	FRANCAISE	C	CH	APP	APPRENTI CUISINIER	18371
CUERVO	RAOUL	H	1893			CUBA	CUBA	C	CH		INGENIEUR	12174

Output of the grammar and post-processing steps.

Conclusion

- **The POPP datasets**[5]
    - We publish our annotated datasets containing ground truth for handwriting recognition and line coordinates.

Dataset	Train # of lines	Validation # of lines	Test # of lines	# of writers
POPP (Generic)	3840 (128 pages)	480 (16 pages)	480 (16 pages)	80
Belleville	1140 (38 pages)	150 (5 pages)	180 (6 pages)	1
Chaussée d'Antin	625	78	77	10

Details about the POPP datasets.

- **SImple GRAmmar toolkit (SIGRA)**[6]
    - We open source the code of SIGRA, a Python framework that facilitates the use of WFSTs with **Kaldi** for handwriting recognition.

QR code for POPP datasets

QR code for SIGRA

5: https://github.com/Shulk97/POPP-datasets/
6: https://gitlab.com/projet-popp/sigra

Conclusion

- **POPP Project**
  - Our pipeline processed the complete census of 1926, 1931, 1936 with a total of 300k pages, 9 million lines and several hundred writers.
  - The POPP database is currently being exploited by a team composed of historians, sociologists, and economists.
- **Perspectives**
  - The processing pipeline could be adapted easily for the other French population census of the time period because the same census procedures and the same table templates were used.
  - The self-training method could be further explored regarding handwriting recognition.

Bibliography

- S. Ares Oliveira, B.Seguin, and F. Kaplan, “dhSegment: A generic deep-learning approach for document segmentation,” in Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on, pp. 7-12, IEEE, 2018.

- D. Coquenet, C. Chatelain and T. Paquet. “End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network.” IEEE transactions on pattern analysis and machine intelligence PP (2022)

- Q. Xie, M. -T. Luong, E. Hovy and Q. V. Le, "Self-Training With Noisy Student Improves ImageNet Classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10684-10695

- M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. Computer Speech & Language 16, 1 (2002), 69–88

Writer specialization combined with self-training

1. A model is initialized with the weights of the best student model
2. The model is trained on the Belleville dataset (mono-writer)
3. The model perform inference on every unlabeled data of the Belleville district
4. A second model is trained on the generated pseudo-labels