



**HAL**  
open science

# LECTAUREP: Paris Notary Record Books Automated Reading

Alix Chagué, Aurélia Rostaing

► **To cite this version:**

Alix Chagué, Aurélia Rostaing. LECTAUREP: Paris Notary Record Books Automated Reading. Fantastic Futures 2021 / Futures Fantastiques 2021, AI4LAM; BnF; Université Paris Saclay, Dec 2021, Paris, France. hal-03479258

**HAL Id: hal-03479258**

**<https://hal.inria.fr/hal-03479258>**

Submitted on 14 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License



# LECTAUREP

## Paris Notary Record Books Automated Reading

FF21 - BnF - 10/12/2021

Alix Chagué (Inria, Université de Montréal)  
Aurélia Rostaing (Archives nationales)

*Inria*

ARCHIVES  
NATIONALES

  
MINISTÈRE  
DE LA CULTURE  
Liberté  
Égalité  
Fraternité

# LECTAUREP's perimeter

B/W & colored digitizations of notary record books, 1803-1940s

Visionneuse

Cotes : 44 v°-51 r°

Liste chronologique des actes pour la période du 2 janvier au 14 mai 1902

[Permalien](#)

[Télécharger](#)

N <sup>os</sup> DU RÉPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIES  INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE l'Enregistrement.	
		EN BREVETS	EN MINUTES		DATES	DROITS
1196	28	Procurat		An 1901, mois de Décembre Portal (par Marie) à Paris, rue d'Amsterdam 24, et Marie, épouse de Abel Théophile Berger, Bd Hausmann, 104, 1 <sup>er</sup> renoncer à s'oppos	30	3.75
1197	28	Dépot de procurat		Jean teaucu nommée par Charles au s'oppos à Sedan, à Paris le 25 août 1901	30	3.75
1198	28	Inventaire		Treuel (après décès de Henry Dérive), à Paris, rue Legendre 121, décide à Bétiers, le 25 Août 1901	6	12.45
1199	28	Notoriété		Miziel (après le décès de Hortense Henriette Josephine, femme de Alfred Jean Marie)	6	3.75

Z

Zoom       
Luminosité       
Contraste

Verrouiller les paramètres

N° 1

/ 7

Aller à

# Stakes

**Help** end-users read, search, mine notarial registers

- > Turn massively digitized archives into data through AI (recurrent neural networks - LSTM)

**Serve** GLAM communities, academics, genealogists, HTR/AIS/LIS SPs

- > Mutualize & document data, models, platform, methods for interoperability purposes

Consider **digital heritage** as a whole

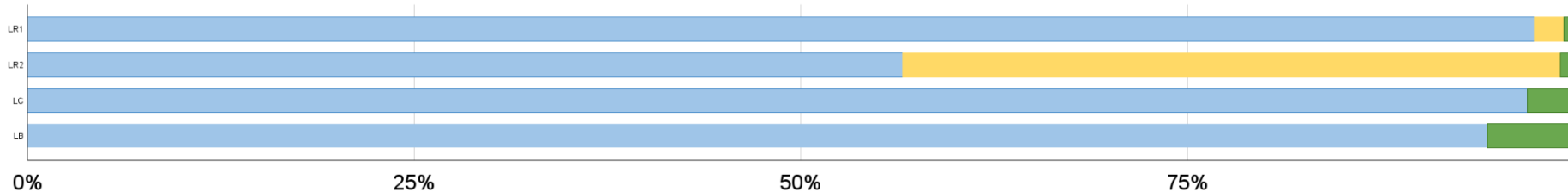
- > Take into account B&W digitized microfilms & digitization from the original



# Sample sets

	Images	Dates	Notaries	Notary Studies	Hands	Hands	Size (Go)	Target (p.)	Lots (p.)	Transcribed (p.)	Revised (p.)
Lot Rép. 1	b/w	1803-1907	12	4	~50	~16	11.6	~174 840 ?	20 800	533	138
Lot Rép. 2	col.	1889-1943	~100	~57	~200	~100	3.75	~1 M ?	1844	800	16
Lot CM-SD	col.	1829-1934	N/A	N/A	~30	~10	42.5	20 000	600	600	218
Lot Bronod	col.	1719-1760	1	1	1	1	1.2	3595	200	200	200

■ remaining in the lot ■ transcribed, not revised ■ transcribed and revised



# Technological environment

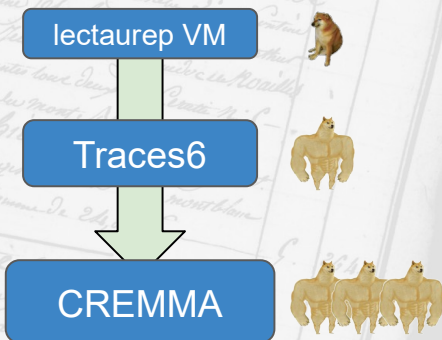
NOI	DATES	NATURE ET ESPECE	NOMS, PRENOMS ET DOMICILES DES PARTIES	RELATION	
du	des	DESACTES		(Engagement)	
REFOR	ACTES	EN MINUTES	INDICATIONS SUR L'OBJET ET PRO DES BENS	ACTU	
576	24	Inventaire	An 1903, mois de Juillet	29	11,25
577		Décharge			
578	24	Procuration	Donation		8,75
580	24				
581	25	Procuration			
582	27	Mainlevée			7,50
584	28	Contratement à			
585	29	Mainlevée			Grats
586	29	Cession de bail			12,75
587	29	Recomptatance			14,85
588	31	Depôt de procès			47,50
589	31				3,75
590	1	Procuration			238,50
591	1 <sup>er</sup>	Bail			3,75
592	3 <sup>es</sup>	Procuration			14,25

## Software

- **Kraken** - open source HTR engine developed by Ben Kiessling since 2015
- **eScriptorium** - open source web application developed by SCRIPTA PSL since 2018

## Hardware

- 2019 - lectaurep.paris.inria.fr (no GPU, virtual machine)
- 2020 - traces6.paris.inria.fr (2 GPU, more storage)
- 2021-2022 - escriptorium.cremma.fr (better architecture, + 2 GPUs, more storage, more RAM)



# Contractual deliverables

## Models

- 1 perfectible model for line detection
- 1 perfectible model for region detection
- at least 4 usable models for text recognition
  - 2 generics (several hands): “Generic” (9 %) & “Random Set” (10 %)
  - 2 specialized (1 hand): “Bronod” (5 %) & “Contrats de Mariage” (3 %)

## Documentation

- Transcription conventions & good practices
- Open Science and Gitlab/Github environments (issues and codes)



# Contractual deliverables

## Data

- 239 images (31 401 lines) gold transcription published and documented as ground truth on HTR-United
  - from Répertoires, CM/SD and Bronod
  - additional publication on [data.archives-nationales.culture.gouv.fr](https://data.archives-nationales.culture.gouv.fr/)?
  - 333 more pages ready to join HTR-United
- 1561 images of silver transcription awaiting revision

[data.culture.gouv.fr](https://data.culture.gouv.fr)

HTR  
United

Quantity of pages in gold, silver and remaining sets

■ Gold ■ Silver ■ Remaining in target





# Bonus deliverables

## Scripta-PSL

- Providing use cases and feedback
- Developing features for eScriptorium (document tags, dashboard...)
- Creating general documentation for eScriptorium (tutorial)

## GLAM & academics

- Sharing expertise with users, project carriers and working groups (smaller projects, CREMMALab, AI4LAM)
- Sharing expertise via scientific publication (Hypotheses blog)

# KaMI: metrics from the lab to the field

A tool to better evaluate the performances of HTR models with:

- filters (digits, punctuation, diacritics, lower/upper case, etc.)
- more metrics (CER, WER, hit and substitutions or deletions, etc.).
- KaMI is agnostic (not limited to Kraken)

The screenshot shows the KaMI web interface. It features two text boxes: 'Reference' and 'Prediction'. Both contain the same text: "Maison Chevalier à la requette de Delacroix acceté à Paris 11:25 2 Certificat devise Le Roy (gouverneur Louise Amable Anais à Paris B<sup>n</sup>d Diderot 902 45, p<sup>r</sup>f resopouster inscription de rente 3% de 1100<sup>r</sup>f n<sup>e</sup> 2". Below the text boxes are several checkboxes: "ignore all digits" (checked), "ignore text case (all in lower case)" (checked), "ignore the punctuation" (checked), and "ignore diacritical signs" (checked). A red "Compare" button is located at the bottom right of the interface.

	Default	Ignoring digits	Ignoring case	Ignoring punctuation	Ignoring diacritics	Combining all options
Levenshtein Distance (Char)	440	433	433	421	210	178
Levenshtein Distance (Words)	228	216	224	219	130	99
Hamming Distance	0	0	0	0	0	0
Word Error Rate (WER)	30.645	33.333	30.107	29.514	17.473	16.202
Char. Error Rate (CER)	9.596	10.253	9.443	9.57	4.554	4.38
Word Accuracy (Wacc)	69.354	66.666	69.892	70.485	82.526	83.797
Match Error Rate (MER)	9.333	9.949	9.183	9.297	4.538	4.363
Char. Information Lost (CIL)	13.52	14.385	13.196	13.425	6.203	5.75
Char. Information Preserved (CIP)	86.479	85.614	86.803	86.574	93.796	94.249
Hits	4274	3919	4282	4107	4417	3901
Substitutions	204	200	195	193	78	57
Deletions	107	104	108	99	116	105
Insertions	129	129	130	129	16	16

# A corpus raising challenges for NLP-ists

A corpus of text with strong potential to raise new challenges for NLP due to its specificities

- many abbreviations
- many named entities
- non verbal sentences

Notoriété | Leroy (après décès de Augustin Jules) demeurant à Paris, avenue de Wagram, 157, du 31<sup>e</sup> décembre 1903

Notoriété | Leroy (après décès de Augustin Jules) demeurant à Paris, avenue de Wagram, 157, du 31 décembre 1903."

Affidavit | Leroy (after death of Augustin Jules) dwelling in Paris, avenue de Wagram, 157, on December 31st 1903.





# Conclusion

Generic HTR models enable fuzzy search in loose lines (B/W 19th c. & part of col. 20th c.).

Segmentation models need to be upgraded in case of tight lines (large part of col. 20th c.).

A segmentation rating tool could be handy.

Target corpus: 3100 notary record books (~14% color)

~1,2 M pages, ~1000s handwritings, 1803-1940s (sample transcribed: 1,11 ‰)

## Scaling for production deployment means

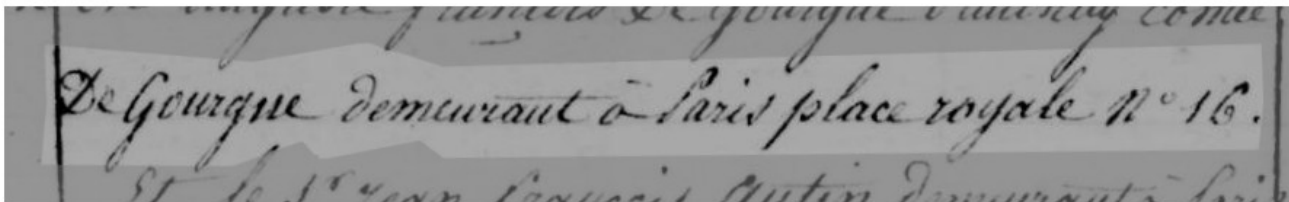
> collaborative logistics, infrastructures, project engineering.

> better sources diplomatic (how many handwritings in a book?)

> better mastering of digital images metadata (B/W-col. distribution).



in order to avoid wasting AI resources.



de Gourgue demeurant à Paris place royale n°16.



# Thank you!

## Links to doc!

- Blog: <https://lectaurep.hypotheses.org/>
- Gitlab: <https://gitlab.inria.fr/almanach/lectaurep>
- Github: <https://github.com/lectaurep>

## Data on HTR-United

- <https://github.com/HTR-United/lectaurep-mariages-et-divorces>
- <https://github.com/HTR-United/lectaurep-bronod>
- <https://github.com/HTR-United/lectaurep-repertoires>

## Contact

- ✉ [alix.chague@inria.fr](mailto:alix.chague@inria.fr)
- ✉ [aurelia.rostaing@culture.gouv.fr](mailto:aurelia.rostaing@culture.gouv.fr)