



**HAL**  
open science

## Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?

Alix Chagué, Victoria Le Fournier, Manuela Martini, Éric Villemonte de La Clergerie

### ► To cite this version:

Alix Chagué, Victoria Le Fournier, Manuela Martini, Éric Villemonte de La Clergerie. Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?. Colloque DHNord 2019 "Corpus et archives numériques", MESHS Lille Nord de France, Oct 2019, Lille, France. hal-02448921

**HAL Id: hal-02448921**

**<https://inria.hal.science/hal-02448921>**

Submitted on 22 Jan 2020

**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

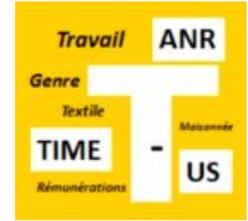
# Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?

---

DHNORD 2019  
18 octobre 2019  
MSH Lille

Alix Chagué (*ALMAnaCH - ICT*)  
Victoria Le Fournier (*ALMAnaCH - ICT*)  
Manuela Martini (*LARHRA*)  
Eric Villemonte de la Clergerie (*ALMAnaCH*)

# « TIME-US »



- ★ « **Rémunérations et usages du temps des femmes et des hommes dans l'industrie textile en France, de la fin du XVIIe siècle au début du XXe siècle** »
- ★ 2017-2020
- ★ Coordonné par **Manuela Martini** (LARHRA, Université Lumière Lyon 2)
- ★ Membres du projet : ALMAnaCH (Inria, Paris), laboratoire ICT (Université de Paris), TELEMMe (Université Aix-Marseille), Centre Maurice Halbwachs, IRHiS (Université de Lille), LARHRA-UMR 5190 (Lyon)
- ★ **Objectif** : collecter, rendre accessibles et analyser des données sérielles sur le travail des femmes dans l'économie du textile et son industrie



# EXPLORATIONS METHODOLOGIQUES

- ★ Approche 1 : analyse empirique approfondie des contextes historiques de production des sources
- ★ Approche 2 : intégration d'une logique "humanités numériques" avec traitement informatique des documents et du texte



# EXPLORATIONS METHODOLOGIQUES

- ★ Approche 1 : analyse empirique approfondie des contextes historiques de production des sources
- ★ Approche 2 : intégration d'une logique "humanités numériques" avec traitement informatique des documents et du texte

**ALMAnACH + ICT + LARHRA**



# EXPLORATIONS METHODOLOGIQUES

- ★ Approche 1 : analyse empirique approfondie des contextes historiques de production des sources
- ★ Approche 2 : intégration d'une logique "humanités numériques" avec traitement informatique des documents et du texte

## ALMAnaCH + ICT + LARHRA

### ★ Objectifs :

- Extraire le texte, des informations annotées et structurées en langage naturel
- Publier et rendre interrogeable le corpus
- Automatiser les tâches



```
graph LR; A[COLLECTE] --> B[SEGMENTATION]; B --> C[TRANSCRIPTION]; C --> D[UNIFORMISATION]; D --> E[ANNOTATION]
```

**COLLECTE**

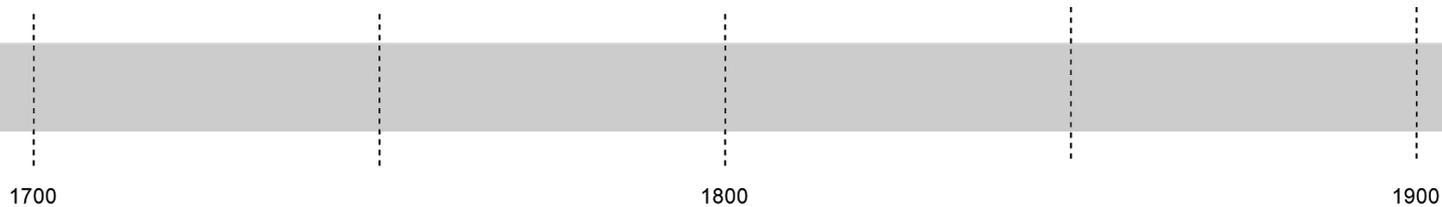
**SEGMENTATION**

**TRANSCRIPTION**

**UNIFORMISATION**

**ANNOTATION**

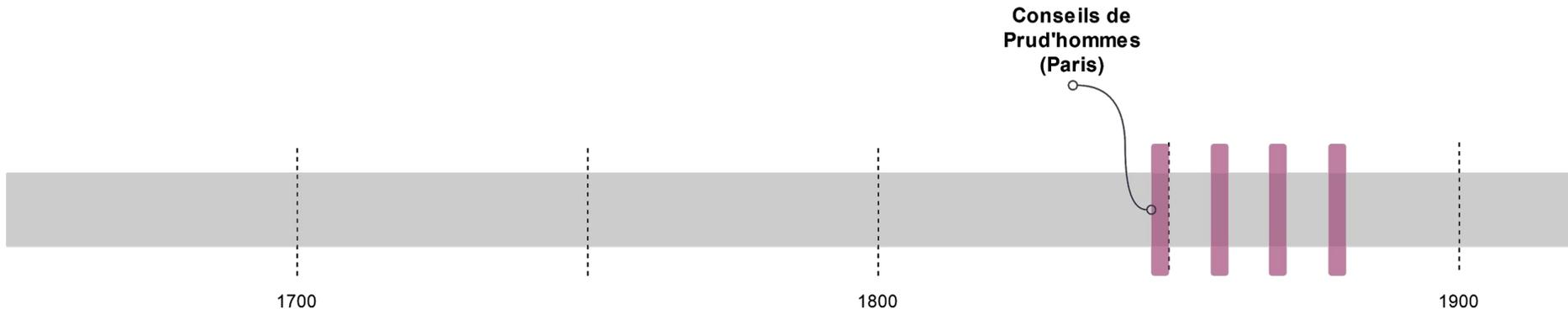
# COLLECTE DES IMAGES



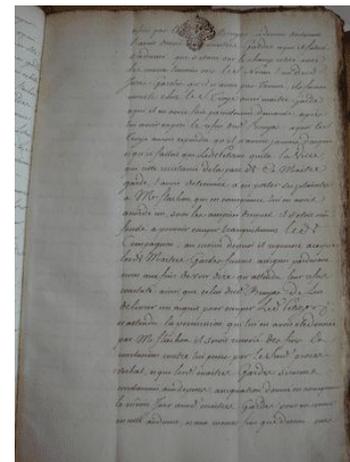
- ★ Sources imprimées et sources manuscrites
- ★ 11 000 images
- ★ De la fin du XVIIIe siècle au début du XXe siècle



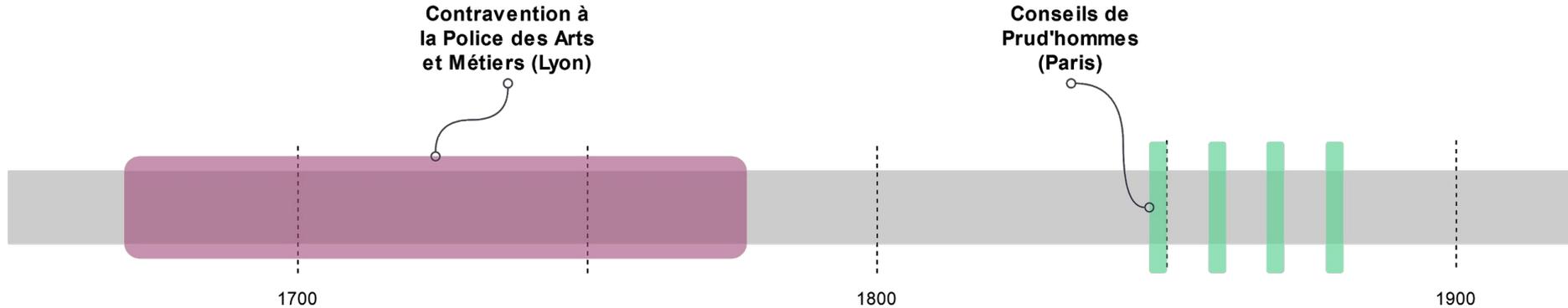
# COLLECTE DES IMAGES



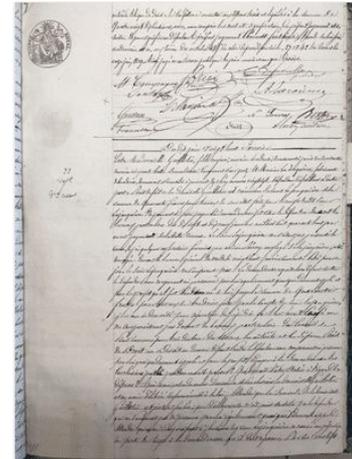
- ★ Minutes du Conseil des Prud'hommes de Paris (sect. Textile)
- ★ 1847-49, 1858, 1868, 1878
- ★ Manuscrit et structuré



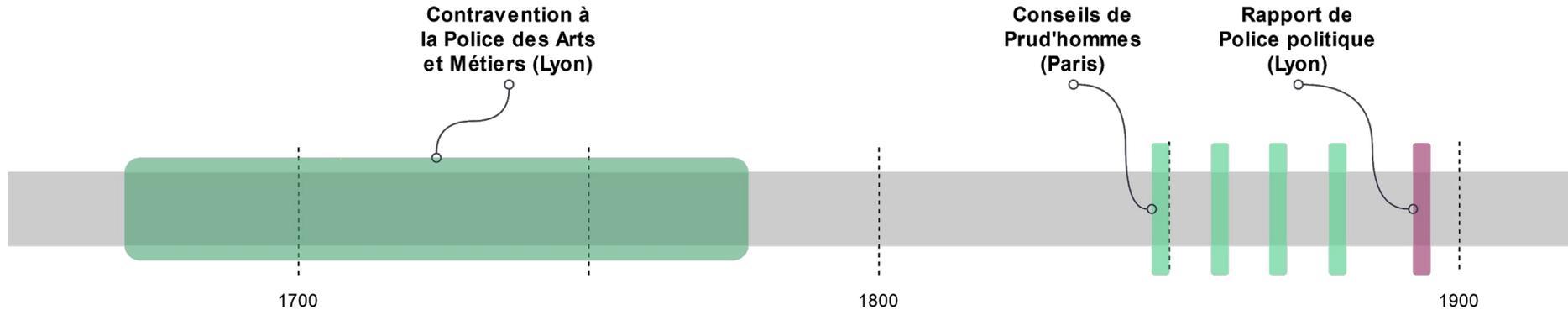
# COLLECTE DES IMAGES



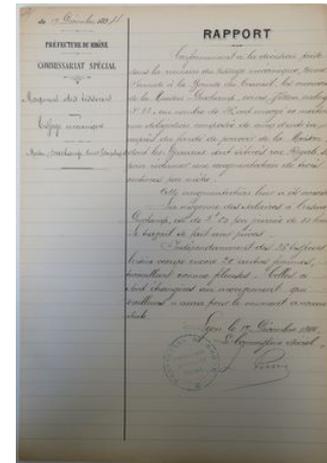
- ★ Contraventions à la Police des Arts et Métiers de Lyon
- ★ De 1667 à 1781
- ★ Focus sur 1760
- ★ Manuscrit, peu structuré



# COLLECTE DES IMAGES

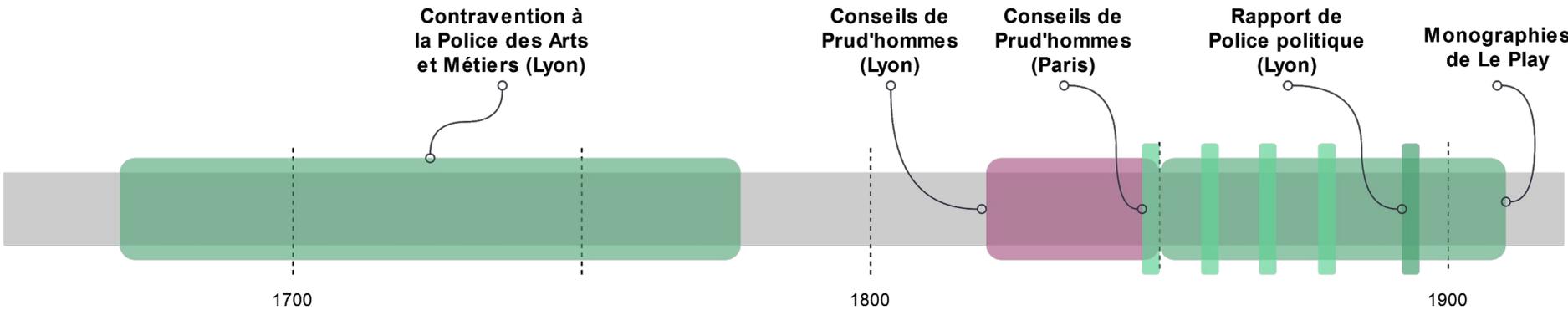


- ★ Rapports de police de la Préfecture de Lyon
- ★ Grèves des ouvriers de la soie de 1894
- ★ Manuscrit et hétérogène





# COLLECTE DES IMAGES



- ★ Comptes rendus des audiences du Conseil de Prud'hommes de Lyon publiés dans la presse ouvrière
- ★ 1831-1851
- ★ Imprimé et hétérogène, impression de qualité variable

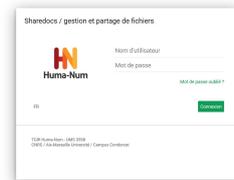
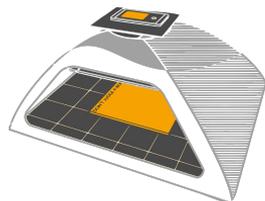


# TPOLOGIE DES DOUBLES NUMERIQUES

- ★ A.D. du Rhône
- ★ A.M. de Lyon
- ★ A.D. de la Seine
- ★ Total : 10 000 photos



- ★ Numelyo
  - 390 PDF
- ★ Internet Archives (Université de Toronto)
  - 6 500 JP2



```
graph LR; A[COLLECTE] --> B[SEGMENTATION]; B --> C[TRANSCRIPTION]; C --> D[UNIFORMISATION]; D --> E[ANNOTATION]
```

**COLLECTE**

**SEGMENTATION**

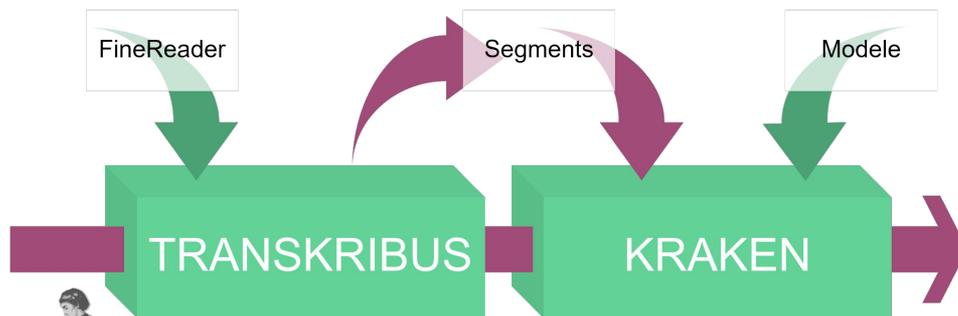
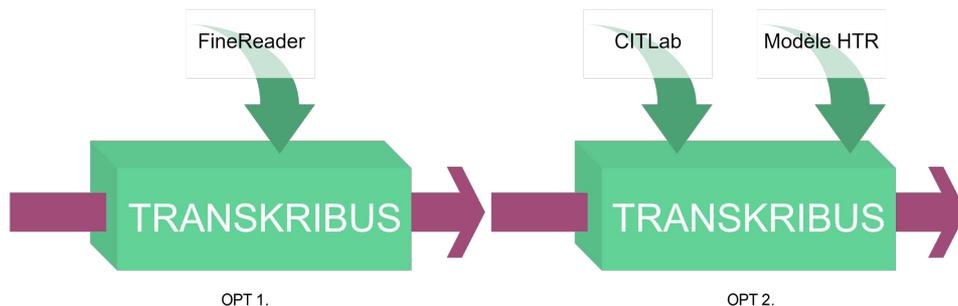
**TRANSCRIPTION**

**UNIFORMISATION**

**ANNOTATION**



# TRANSCRIPTION AUTOMATIQUE



## Transkribus : HTR+

- ★ **Prud'hommes\_1858\_M4+**
  - Taux d'erreur (CER) : 5.2%
  - 4 577 lignes de GT
- ★ **Comb\_French\_Admin\_XIX\_M3+**
  - Taux d'erreur (CER) : 8.8%
  - 20 025 lignes de GT

## Kraken : OCR

- ★ **Model\_od2m**
  - Taux d'erreur (CER) : 2.2%
  - 1 300 lignes de GT



# BILAN SUR LA TRANSCRIPTION

	 			
Prud'hommes Paris	<b>3 439</b>		<b>1 254</b>	<b>35 %</b>
Rapports de police	<b>451</b>		<b>126</b>	<b>27 %</b>
Contraventions	<b>2 525</b>		<b>1 093</b>	<b>43 %</b>
Prud'hommes Lyon (presse)	<b>520</b>		<b>520</b>	<b>100 %</b>
Monographies	<b>6 500</b>		<b>6 500</b>	<b>100 %</b>



# UNIFORMISATION

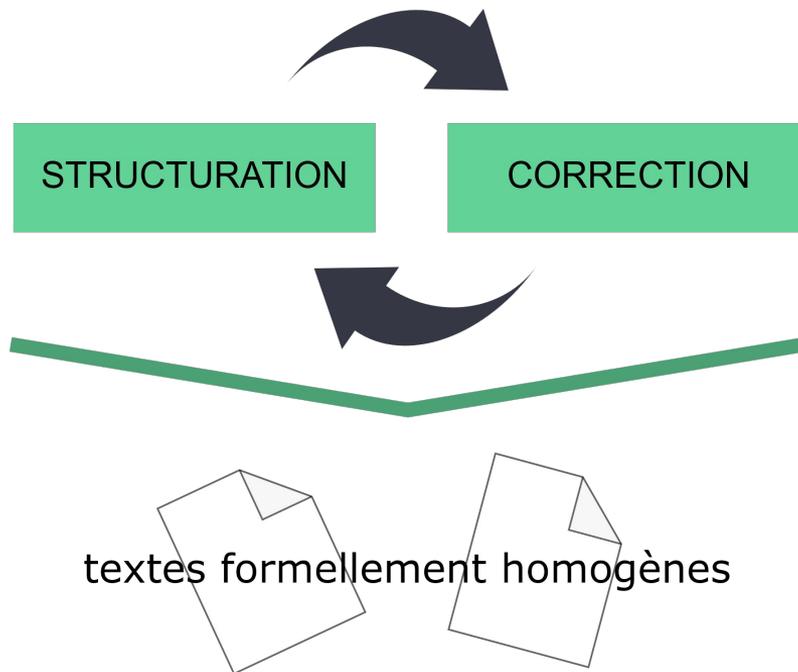
COLLECTE

SEGMENTATION

TRANSCRIPTION

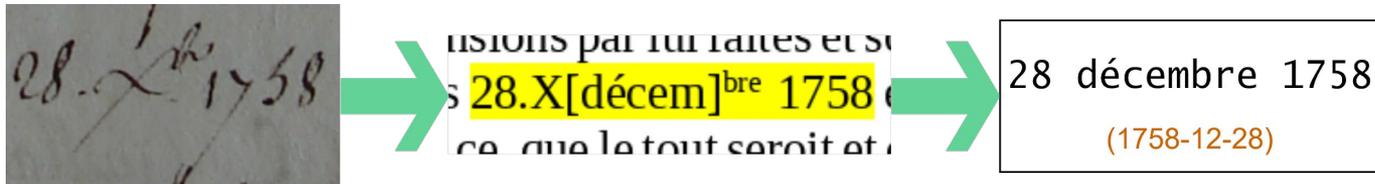
UNIFORMISATION

ANNOTATION

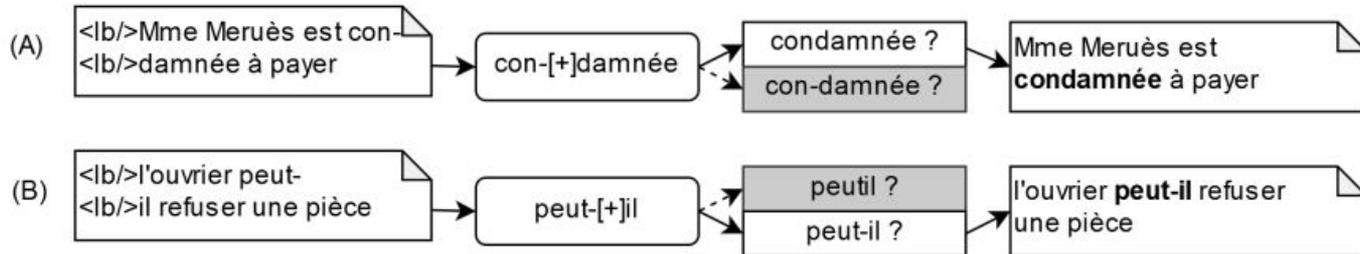


# UNIFORMISATION : CORRECTION

- ★ Résoudre les abréviations (manuscrits principalement)



- ★ Résoudre les césures (imprimés principalement)



- ★ Correction post-transcription



# UNIFORMISATION : STRUCTURATION

## Sources peu structurées

- fond
  - rapport
- registre
  - affaire

## Sources structurées

- journal
  - article
    - audience
      - affaire
- volume
  - chapitre
    - partie
      - sous-partie
        - ...
- registre
  - audience
    - affaire
      - partis
      - point de droit
      - point de fait
      - ...

- ★ Détecter les indices typographiques et textuels ;
- ★ Modéliser sous la forme d'un schéma TEI (ODD) ;
- ★ Faciliter la navigation et le ciblage de portions dans les documents.



```
graph LR; A[COLLECTE] --> B[SEGMENTATION]; B --> C[TRANSCRIPTION]; C --> D[UNIFORMISATION]; D --> E[ANNOTATION];
```

**COLLECTE**

**SEGMENTATION**

**TRANSCRIPTION**

**UNIFORMISATION**

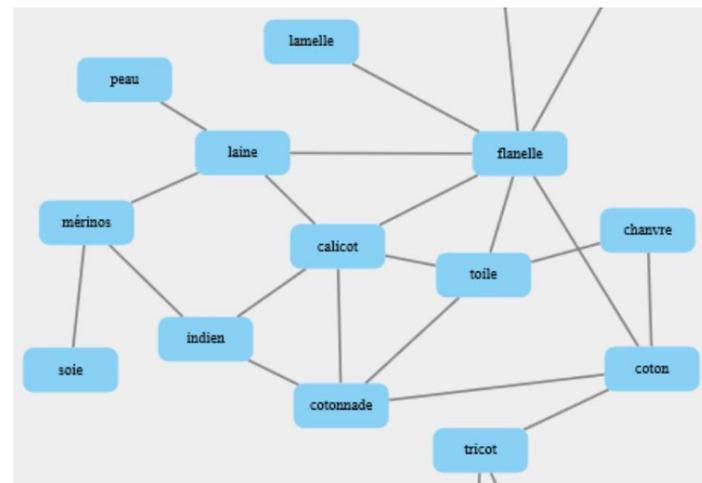
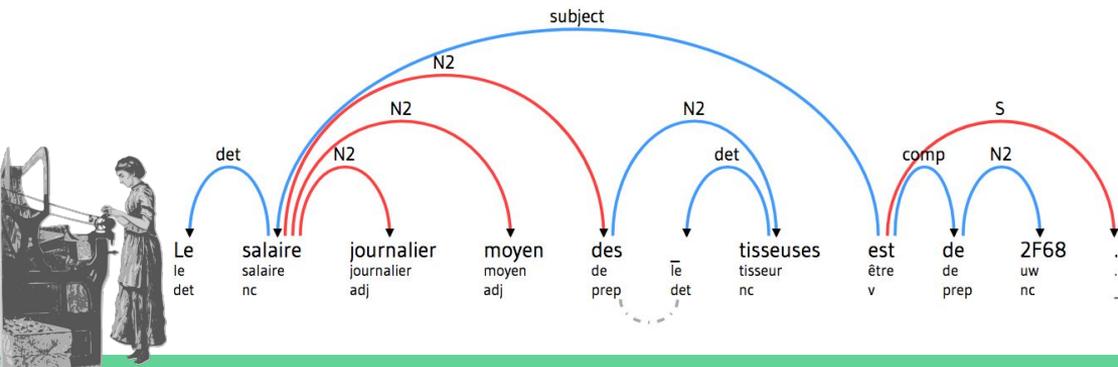
**ANNOTATION**

# ANNOTATION SÉMANTIQUE

- ★ Définir les informations à annoter (personnes, rémunérations, temps...)
- ★ Modéliser sous la forme d'un schéma TEI (ODD chaînée)

```
<p>Le <rs type="salary">salaire <measure type="frequency" quantity="1" unit="day">
journalier</measure> moyen</rs> des <rs type="occupation">tisseuses</rs> est de
<measure type="sum" quantity="2.68">2<unit normal="francNapoleon">F</unit>68</measure>.</p>
```

- ★ Élargissement progressif du modèle
- ★ Automatisation avec un analyseur syntaxique (FrMG)



# Bilan en guise de conclusion

- ★ Un gain de temps ?
  - Oui, mais aussi une nouvelle répartition du temps
  - Oui, mais tout le monde doit se former
- ★ Un gain méthodologique ?
  - Oui, plutôt : méthodologie reproductible, résultats distribuables pour d'autres projets, données interconnectées/ables
- ★ Une chaîne de traitement opérationnelle ?
  - C'est bien parti pour, malgré la variété des sources constituant notre corpus !





Des questions ?

Merci :)

# Quelques liens



★ Accès aux codes sources :

➔ <https://gitlab.inria.fr/almanach/time-us>

★ Interface visualisation et de documentation du projet :

➔ <http://timeusage.paris.inria.fr/>

★ Carnet de recherche du projet :

➔ <https://timeus.hypotheses.org/>