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

Alix Chagué, Victoria Le Fourner, 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 ?

DHNORD 2019
18 octobre 2019
MSH Lille

Alix Chagué *(ALMAnaCH - ICT)*
Victoria Le Fourner *(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

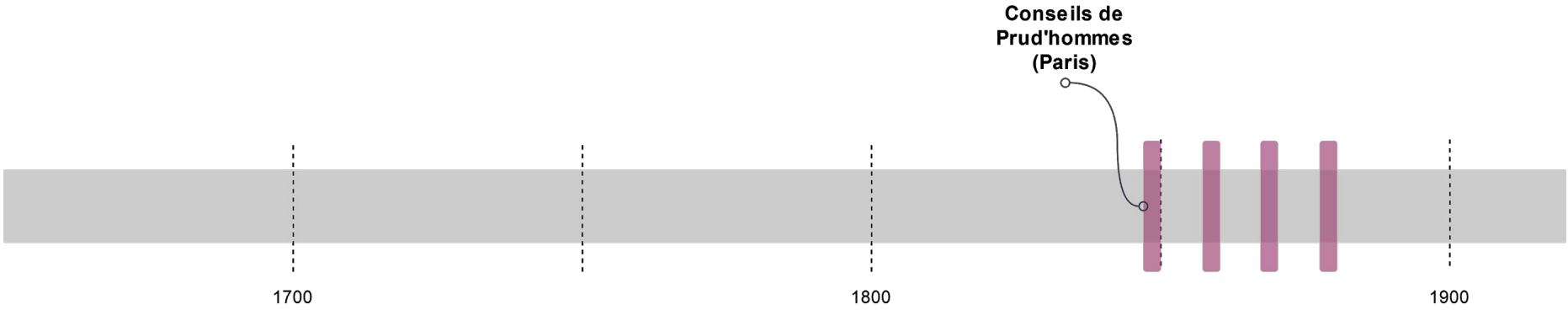
- ★ **Objectifs :**
  - Extraire le texte, des informations annotées et structurées en langage naturel
  - Publier et rendre interrogeable le corpus
  - Automatiser les tâches


COLLECTE
SEGMENTATION
TRANSCRIPTION
UNIFORMISATION
ANNOTATION

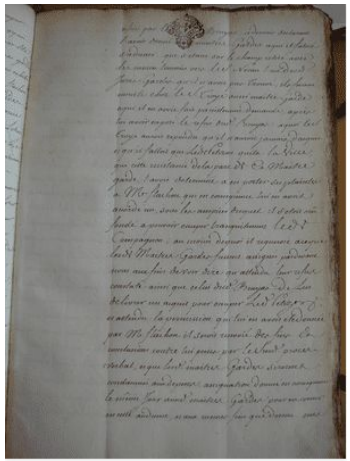
COLLECTE DES IMAGES

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

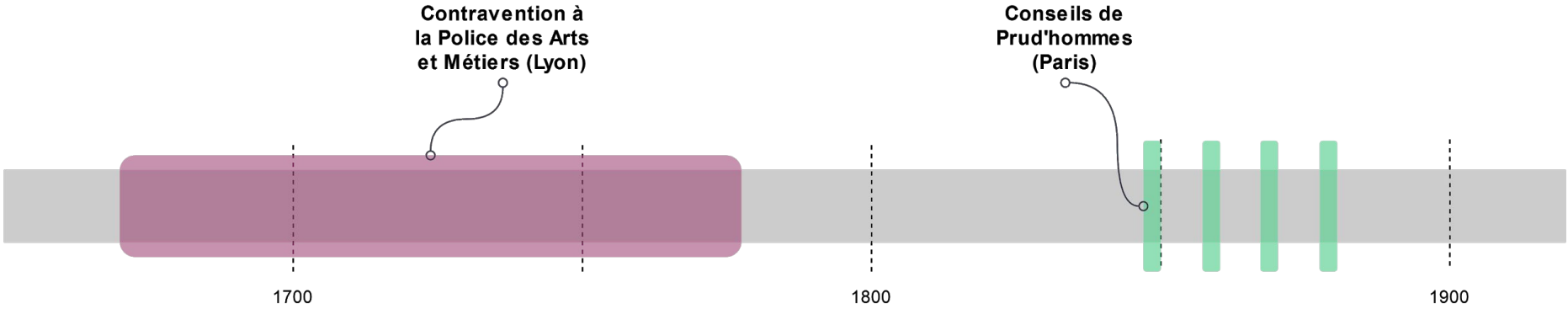
COLLECTE DES IMAGES

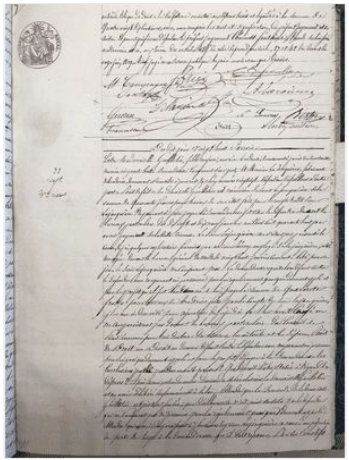


- ★ Minutes du Conseil des Prud’hommes de Paris (sect. Textile)
- ★ 1847-49, 1858, 1868, 1878
- ★ Manuscrit et structuré



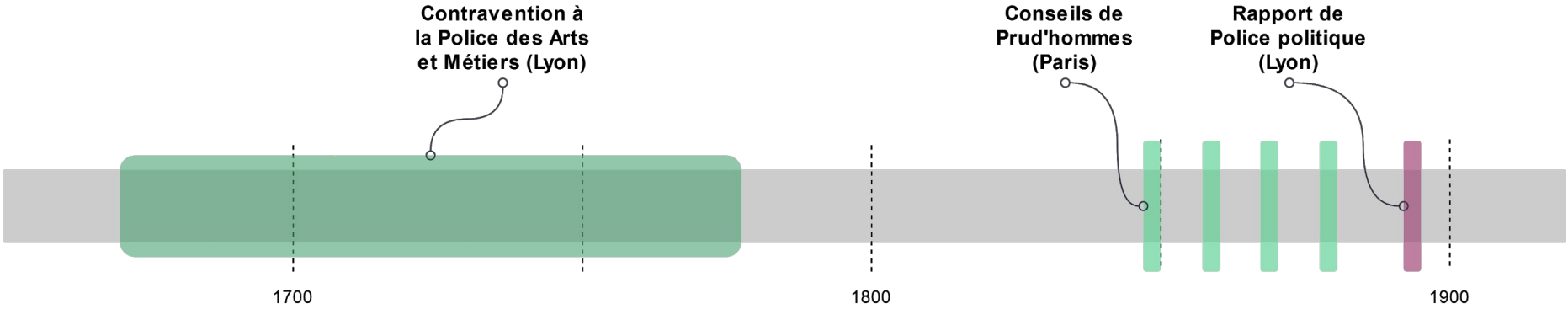
COLLECTE DES IMAGES



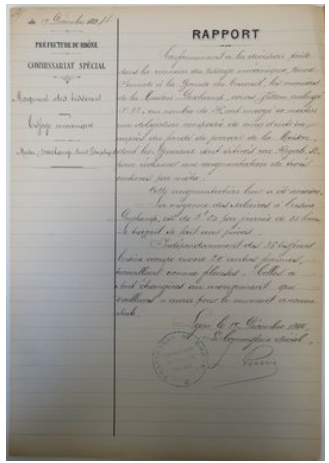


- ★ Contraventions à la Police des Arts et Métiers de Lyon
- ★ De 1667 à 1781
- ★ Focus sur 1760
- ★ Manuscrit, peu structuré

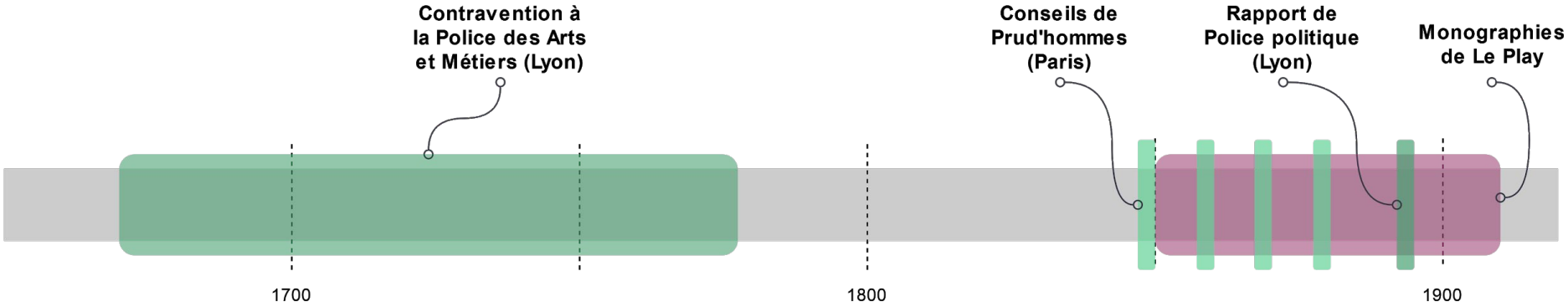
COLLECTE DES IMAGES



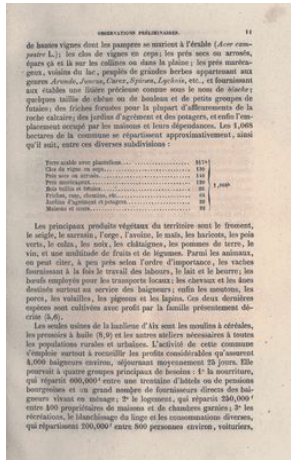
- ★ Rapports de police de la Préfecture de Lyon
- ★ Grèves des ouvriers de la soie de 1894
- ★ Manuscrit et hétérogène



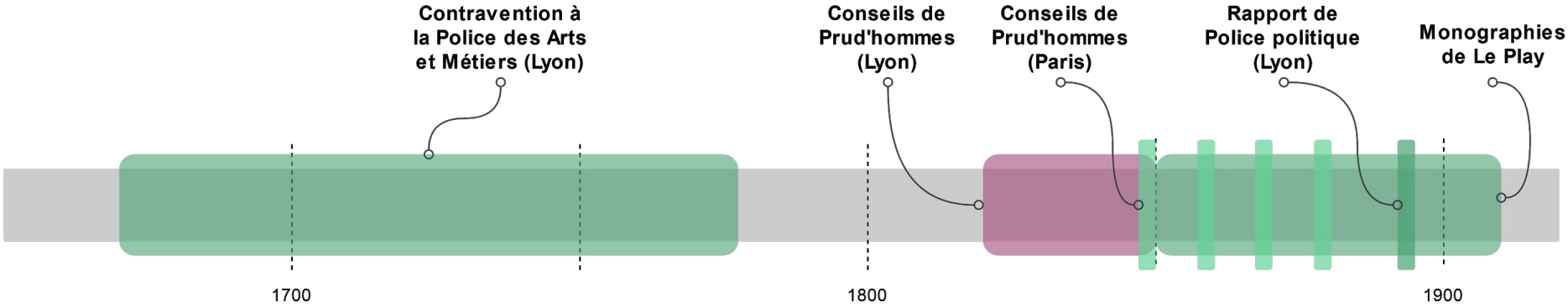
COLLECTE DES IMAGES



- ★ Monographies familiale de Le Play :
  - ○ Les Ouvriers des Deux Mondes
  - ○ Les Ouvriers Européens
- ★ 1851-1908
- ★ Imprimé et structuré



COLLECTE DES IMAGES

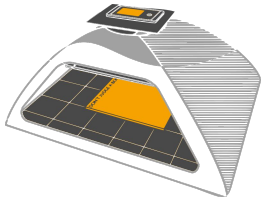


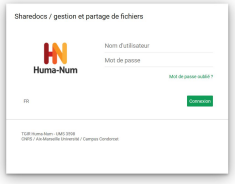
- ★ Comptes rendus des audiences du Conseil de Prud'hommes de Lyon publiés dans la presse ouvrière
- ★ 1831-1851
- ★ Imprimé et hétérogène, impression de qualité variable

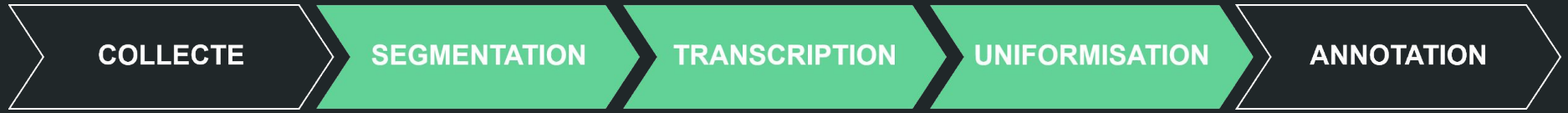
TYPOLOGIE DES DOUBLES NUMERIQUES

- ★ 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



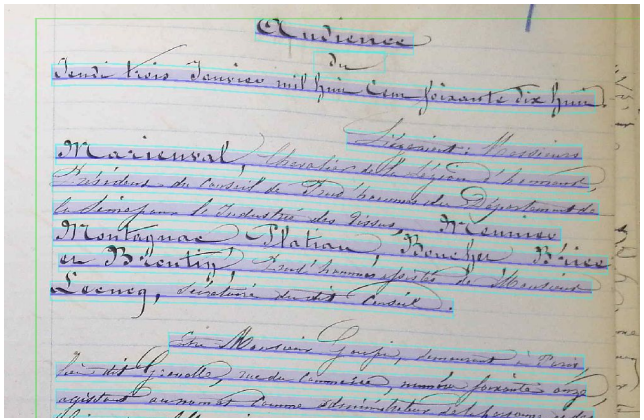



COLLECTE
SEGMENTATION
TRANSCRIPTION
UNIFORMISATION
ANNOTATION

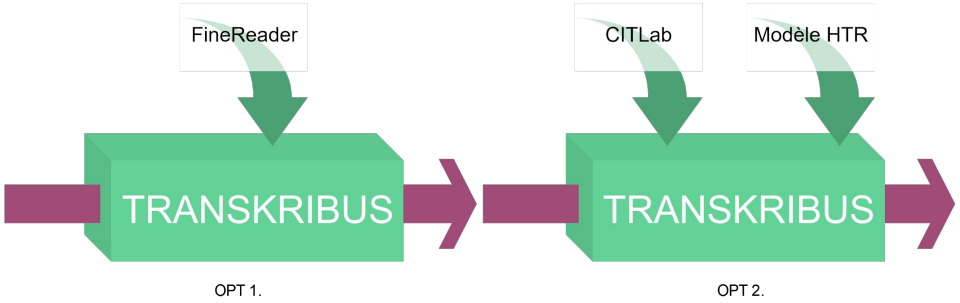
SEGMENTATION

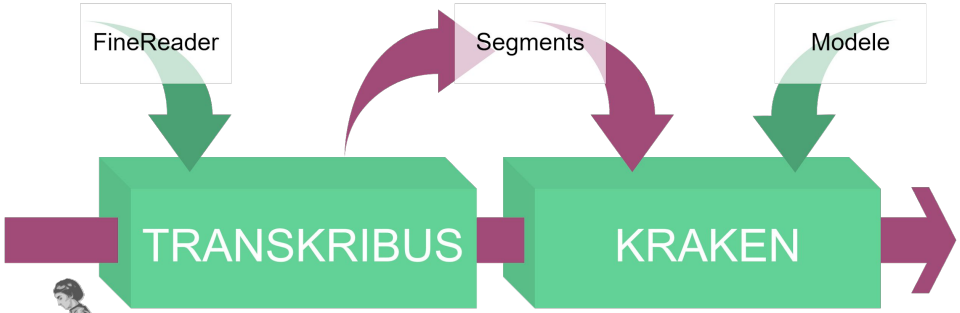
Transkribus

- ★ FineReader
  - ✅ très bon pour la détection des zones, le typage et l'ordonnancement
  - ❌ segmentation et transcription indissociables ; pas d'HTR
  - 📝Conseils des Prud'hommes de Paris ;
  - 📝Rapports de police lyonnaise
- ★ CITLab
  - ✅ adaptation au type de document ; entraînable
  - ❌ moins bon que FineReader
  - 📝Conseils des Prud'hommes de Lyon ;
  - 📝Monographies de Le Play



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

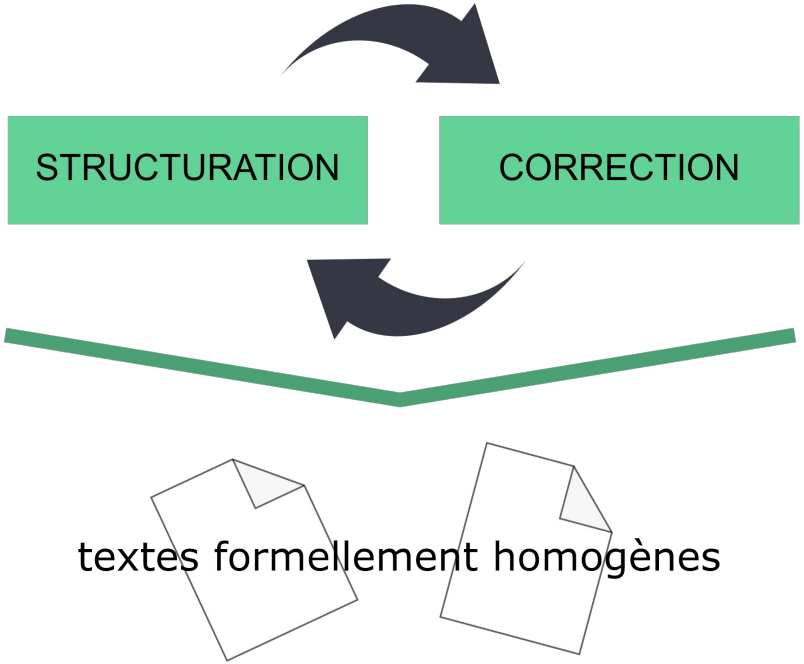
Kraken : OCR

★ **Model_od2m**
- Taux d'erreur (CER) : 2.2%
- 1 300 lignes de GT

BILAN SUR LA TRANSCRIPTION

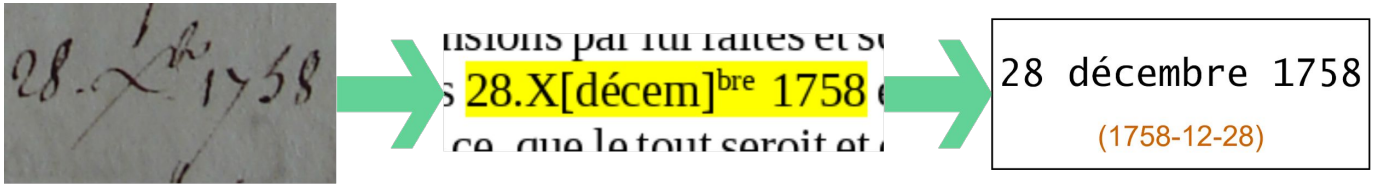
Prud'hommes Paris	3 439		1 254	35 %
Rapports de police	451		126	27 %
Contraventions	2 525		1 093	43 %
Prud'hommes Lyon (presse)	520		520	100 %
Monographies	6 500		6 500	100 %

UNIFORMISATION

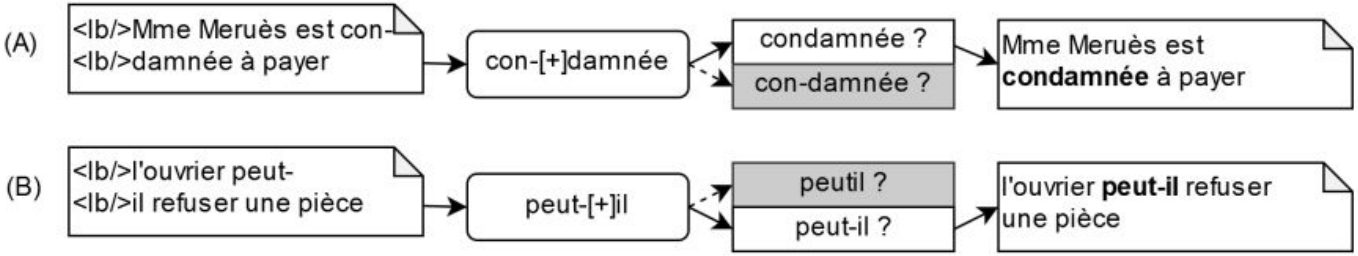


UNIFORMISATION : CORRECTION

★ 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.

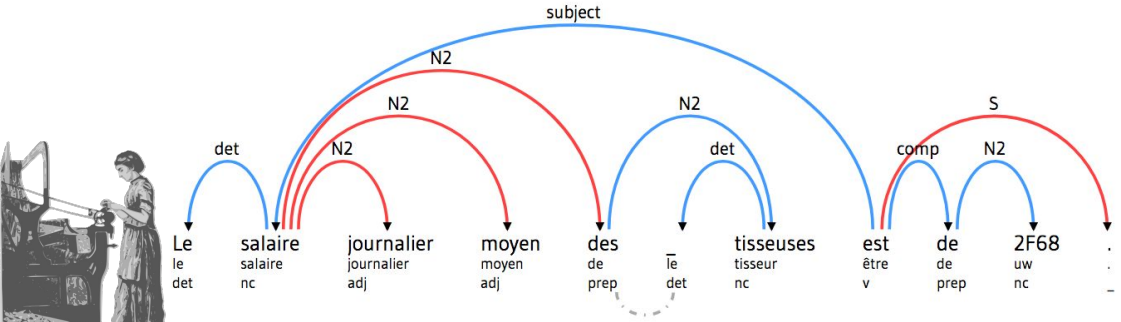

COLLECTE
SEGMENTATION
TRANSCRIPTION
UNIFORMISATION
ANNOTATION

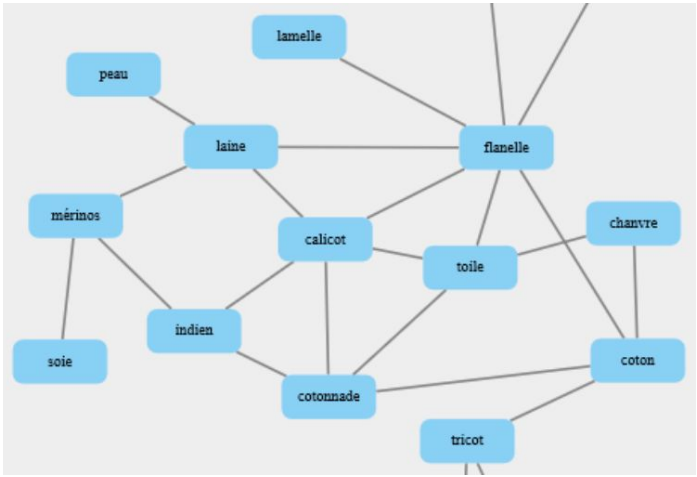
ANNOTATION SÉMANTIQUE

- ★ 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/