

---

# Collecte des données

Constituer un corpus numérique à partir de sources

Présentation par **Sékolène Albouy**  
Cheffe de projet numérique Observatoire de Paris | CNRS

---

# Plan du cours

## Introduction

Présentation générale

## Pourquoi ?

Intérêt et enjeux de manipuler de la donnée

## Quoi ?

Quelques définitions de termes

## Quelle forme ?

Exemples d'usage

## Comment ?

Manières de collecter

## ***Travaux pratiques !***

---

# Introduction

## Présentation générale

# But du cours

- Apprendre à transformer des sources documentaires en données
- Rassembler les données de manière à former un corpus
- Constituer une base solide pour mener à bien ses recherches

# Ségolène Albouy

Projet de recherche en histoire de l'astronomie

[segolene.albouy@gmail.com](mailto:segolene.albouy@gmail.com)

**GitHub** : Segolene-Albouy

→ [M2-UVSQ-Cours-Humanites-numeriques](#)

---

## Vous

Période	Antiquité	Ve-XVIIIe	XIXe-XXIe	?
Étudiants	3	1	11	2

Projet pro	Recherche	Enseignement	Autre	?
Étudiants	9	7	4	3

---

## Objectif

Obtenir un corpus exploitable pour mener vos recherches

Traiter et organiser les sources primaires

Créer des données pour une étude qualitative et quantitative

---

—

# Pourquoi ?

## De l'intérêt de la donnée



Pourquoi ?

# Quelques questions

**Comment collectez-vous vos sources ?**

**Quelles sont leurs natures ?**

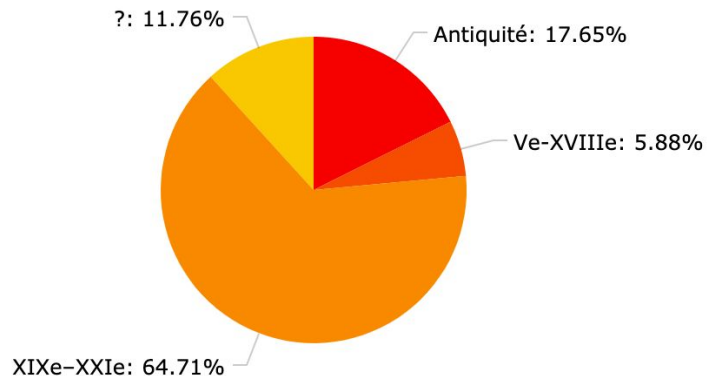
Pourquoi ?

# Quelques questions

Selon vous, quel est l'intérêt  
de travailler avec des  
données plutôt qu'avec un  
document brut ?

## Exploitation et manipulation

Période	Antiquité	Ve-XVIIIe	XIXe-XXIe	?
Étudiants	3	1	11	2

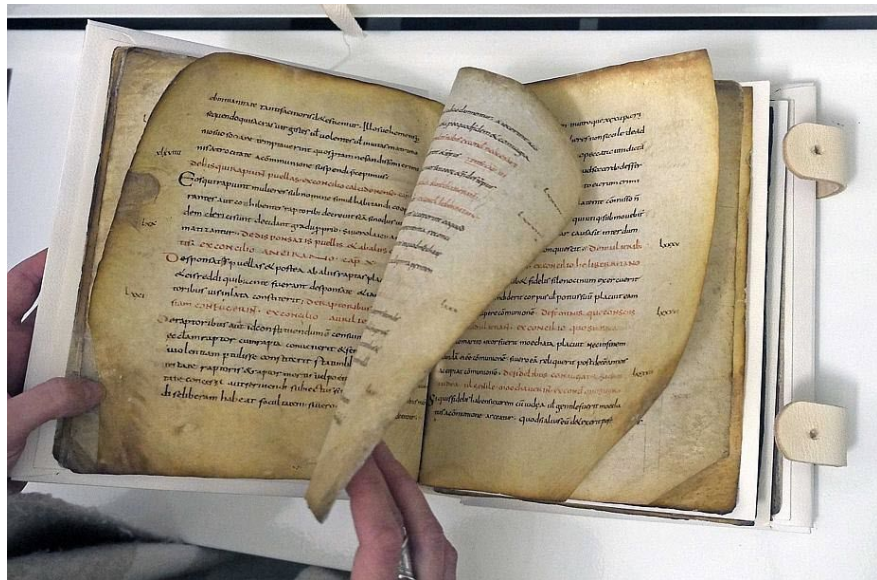


# Indexation et recherche



# Introduction

## Passage à l'échelle



---

Quoi ?

Quelques définitions

---

# Glossaire Digit-Hum

[digithum.huma-num.fr/ressources/glossaire](https://digithum.huma-num.fr/ressources/glossaire)

---

Quoi ?

**Comment définir ?**

**Donnée**



---

# Donnée

Représentation d'**une information** sous une forme conventionnelle destinée à **faciliter son traitement** et sa communication.

*Digit-Hum*

Une donnée est la **représentation d'une information** dans un programme [...] Les données peuvent être **conservées et classées sous différentes formes** : textuelles (chaîne), numériques, images, sons, etc.

*Wikipédia*

---

---

# Caractéristiques

Unité d'information

Lisible et manipulable par une machine

Représentation typée

Traitement systématique et automatisable

Stockable informatiquement

---

Quoi ?

**Comment définir ?**

**Métadonnée**

---

# Métadonnée

Données structurées décrivant une ressource ou une autre donnée [...] qui servent à référencer, identifier et partager correctement un document. Elles permettent la description et le traitement des ressources numériques (ou papier).

*Digit-Hum*

---

---

## Caractéristiques

Donnée descriptive

Utilisée pour classer et retrouver

Pas de donnée utilisable sans métadonnées

Souvent structurées selon des standards (Dublin Core)

Fiches, notices (en langue vernaculaire)

---

Quoi ?

---

## Pour quoi faire ?

### COMPRENDRE L'INFORMATION

Description qui permet  
de décrypter la donnée

Exemple du code de  
sécurité sociale dont  
les chiffres signifient  
des choses

### RETROUVER L'INFORMATION

Dans des corpus vastes,  
une donnée seule est  
introuvable

Comme si on devait  
retrouver Charlie sans  
connaître sa  
description

### GÉRER L'INFORMATION

Organiser des  
migration vers d'autres  
formats

Réutiliser les données  
pour différents usages

---

---

# Classement des métadonnées

En fonction :

- de ce qu'elles décrivent (le contenu)
  - de la façon dont elles sont créées (leur provenance)
  - du moment où on les crée (leur historique)
  - de l'endroit où on les trouve (leur localisation)
  - de l'aspect qu'elles ont (leur forme)
  - de l'usage qu'on en fait (leur objectif)
-

Quoi ?

Quelques

Métaphores





## Donnée et métadonnée

Vin = donnée



Sans bouteille, très difficile de reconnaître un vin

La bouteille nous apporte de nombreuses informations :

- Forme de la bouteille (provenance)
- Étiquette
- Droit d'accise 
- Réglementation nationale 
- etc.

Bouteille = métadonnée

Quoi ?

**Comment définir ?**

**Corpus numérique**

---

## Corpus numérique

Recueil de documents (numériques) relatifs à une discipline ou une thématique, réunis en vue de leur conservation, leur édition ou leur exploitation.

*Digit-Hum*

---

---

## Caractéristiques

Ensemble cohérent de données

Constitué en vue d'un objectif d'analyse

Structuré selon un modèle contraignant

Borné et contraint par des choix éditoriaux

Permet de tisser des liens entre les ressources et grouper des ensembles cohérents de données

---

---

# Quelle forme ?

## Exemples d'usage

Quelle forme ?

# Exemples de Données

Quelle forme ?

---

Quelle source = quelle donnée ?

MANUSCRITS /  
ÉCHANGE  
ÉPISTOLAIRE

COUPURES DE  
PRESSE

TABLEAUX /  
PHOTOS

PERSONNAGES  
HISTORIQUES

---

Quelle forme ?

## Ça dépend de l'usage !

MANUSCRITS /  
ÉCHANGE  
ÉPISTOLAIRE



scans,  
transcription,  
vectorisation de l'  
écriture, etc.

COUPURES DE  
PRESSE



transcriptions,  
photos,  
prosopographie,  
etc.

TABLEAUX /  
PHOTOS



Photos,  
chronologie,  
palettes de  
couleurs

PERSONNAGES  
HISTORIQUES



Fiche  
prosopographique  
, enregistrements  
sonores, photos,  
Textes de discours



Quelle forme ?

# Exemples de Métadonnées

Quelle forme ?

---

Quelle donnée = quelle métadonnée ?

PHOTO

TEXTE

SON

FICHIER  
INFORMATIQUE

---

Quelle forme ?

---

## Beaucoup de métadonnées !

### PHOTO

Dimension,  
couleur,  
technique  
utilisée, auteur,  
sujet, date de la  
prise, etc.

### TEXTE

Longueur du  
texte, nature,  
auteur, lieu et  
date de  
rédaction,  
langue, etc.

### SON

Durée de  
l'enregistrement,  
nature du son,  
technologie de  
prise de son, date,  
lieu, contexte,  
etc.

### FICHIER INFORMATIQUE

Format, taille,  
type de donnée,  
date de création,  
système  
d'exploitation,  
etc.

---

Quelle forme ?

**Exemples de**

**Corpus numérique**

Quelle forme ?

---

## Quelles techniques ?

TABLEUR

ARCHITECTURE  
DE DOSSIERS

LOGICIEL  
SPÉCIALISÉ

BASE DE  
DONNÉES

---

Quelle forme ?

---

## Exemples d'outils

EXCEL

DRIVE

TROPY

HEURIST

---

---

# Comment ?

## Manières de collecte

---

## Méthodes de collecte

Transcription

*Web scrapping*

Récolte dans des entrepôts de données

Utilisation d'API

Enregistrements numériques

Saisie de formulaires

---



---

# Entrepôts de données



## Bibliothèques numériques

[Gallica](#), [Europeana](#), [World Digital Library](#), [Internet Archives](#), etc.

## Archives en ligne

[Salle des inventaires virtuelle](#), [archives départementales](#), [francearchives](#), etc.

## Datasets ouverts

[Wikidata](#), [data.gouv](#), [plateforme pop](#), [Europeana](#), [data.bnf](#), etc.

## Éditions numériques & sites de projets de recherche

[E-Man](#), [projets Zooniverse](#), [sites Huma-Num](#), [Biblissima collections](#), etc.

---

## Attention



Toutes les sources secondaires ne se valent pas

---

---

## Outils de collecte

[Tropy](#) : gestion des photos de documents de recherche

[Heurist](#) : pour faire des bases de données (cours à venir)

[Mirador](#) : pour visualiser des ressources de différentes bibliothèques

[Catalogue OPIDoR](#) : wiki des données de la recherche

[Zotero](#) : stockage de données bibliographiques

---

---

# Travaux pratiques !

## Découverte de Transkribus

# Récupérer une source Sur Gallica

---

# Numérisations Gallica

## Téléchargement

Autorisé jusqu'à 1470 x 1024 (en 96 pp)

## Deux protocoles

**ARK** : id. basé sur la norme URI pour un accès pérenne

**IIIF** : ensemble de spécifications techniques pour définir un **cadre d'interopérabilité** pour la **diffusion d'images HD** en ligne.

---

# Choix d'une ressource

<https://gallica.bnf.fr/ark:/12148/btv1b8448967x>

*Christine de Pizan adresse à la reine Isabeau de Bavière une épître pour convaincre la reine de s'opposer à la bataille qui pourrait éclater entre le duc d'Orléans et le duc de Bourgogne.*

---

# Affichage des métadonnées

<https://gallica.bnf.fr/iiif/ark:/12148/btv1b8448967x/manifest.json>

*Manifeste agréant les métadonnées du document*

---



# Bidouillage d'URL

gallica.bnf.fr/[/iiif](#)/ark:/12148/btv1b8448967x/f111/[/full/full/0/native.jpg](#)

### Coordonnées de la région d'image

*Décalage px vers la droite de l'angle sup. droit, décalage vers le bas, largeur, hauteur =>  
0,50,1000,1000*

### Dimension de l'image

*Largeur, hauteur en pixel => 1500,750*

### Rotation de l'image

*Angle en degrés=> 90*

### Qualité de l'image

*Native, grey (niveaux de gris), bitonal (N&B)*

### Format

*jpg, png, pdf, etc.*

---

# Récupération de l'image



Enregistrer l'image dans un dossier à part

---

# Présentation du logiciel

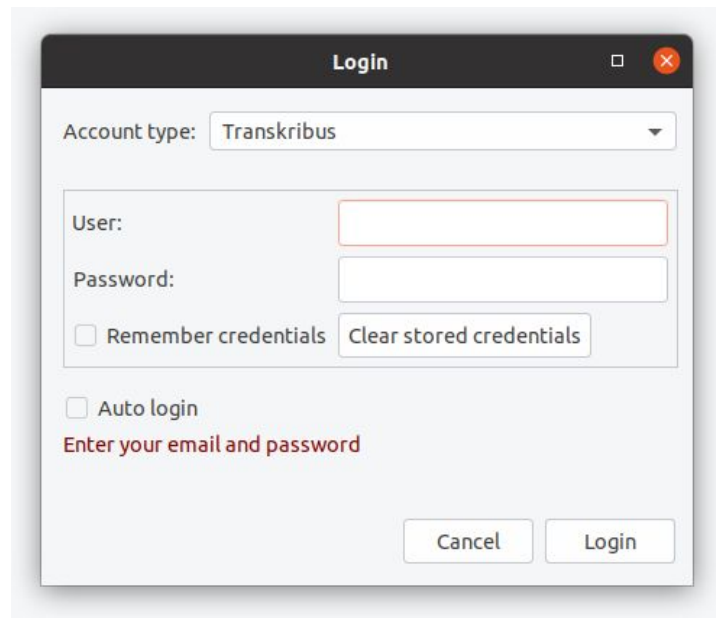
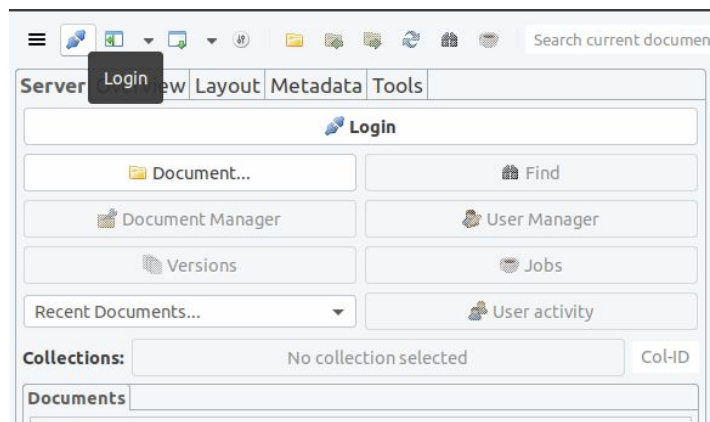
# Transkribus

---

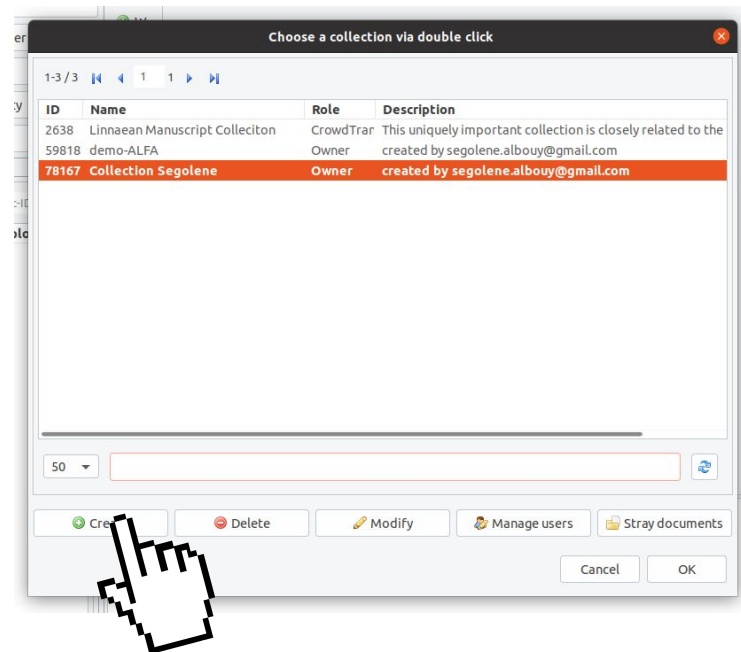
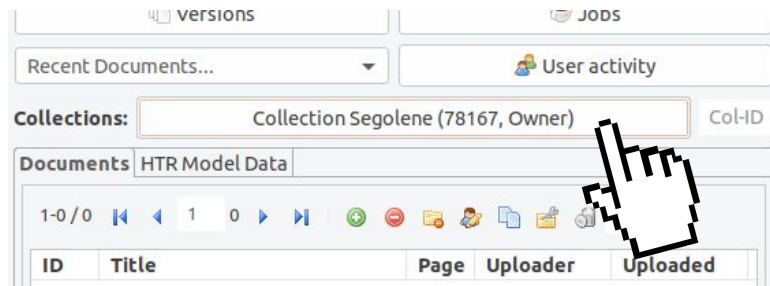
# Utilisations de Transkribus

- Transcrire des documents pour une édition savante.
  - Création de données d'apprentissage pour alimenter le moteur de reconnaissance de texte manuscrit (HTR) afin qu'il puisse apprendre à déchiffrer l'écriture manuscrite.
  - Exécuter HTR sur vos documents et recevoir automatiquement générés transcriptions.
-

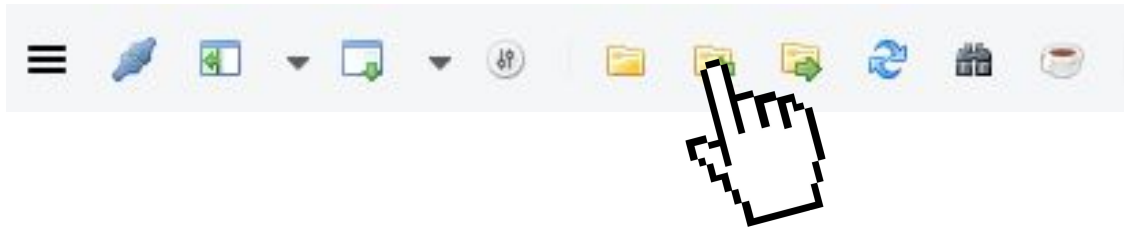
## Se connecter à Transkribus



## Créer une collection



## Importer des documents



## Importer des documents


Document ingest / upload

☐ Upload via private FTP (also PDF files) ☒ Upload single document

☐ Upload via URL of DFG Viewer METS ☐ Upload via URL of IIIF manifest

☐ Extract and upload images from pdf

Single document upload

Local folder:  

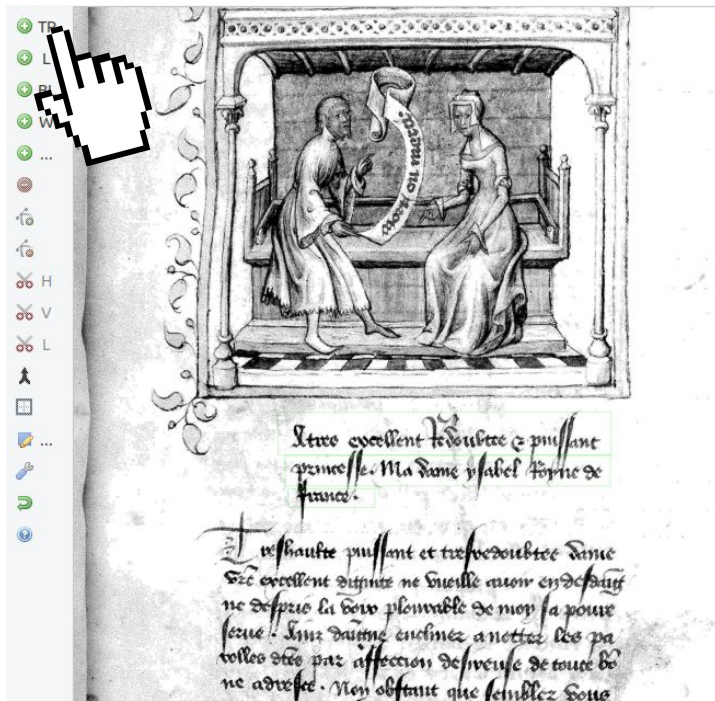
Title on server:

Add to collection:

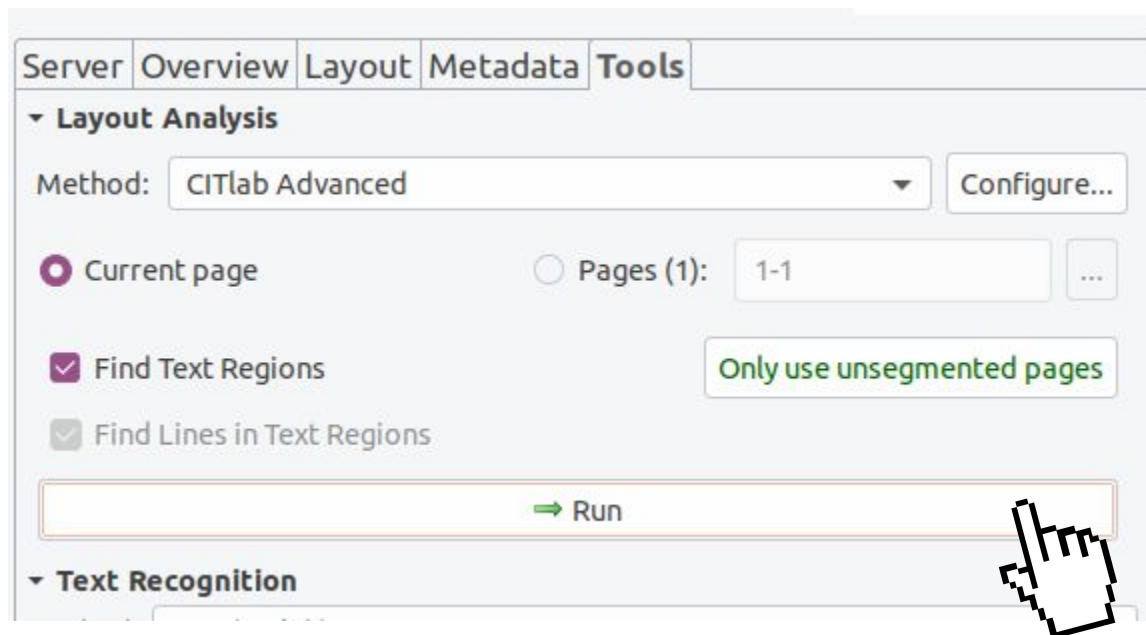
Choisir le dossier où sont conservées les numérisations > Upload



## Segmenter en zone de texte



## (Mieux) segmenter



The screenshot shows the 'Tools' tab of a software interface. The 'Layout Analysis' section is expanded, showing a 'Method' dropdown set to 'CITlab Advanced' with a 'Configure...' button. Below this are radio buttons for 'Current page' (selected) and 'Pages (1):' with a text input '1-1' and an ellipsis button. There are two checked checkboxes: 'Find Text Regions' and 'Find Lines in Text Regions'. A green button labeled 'Only use unsegmented pages' is visible. At the bottom of the section is a large 'Run' button with a green arrow icon. A hand cursor icon is pointing at the 'Run' button. The 'Text Recognition' section is partially visible below.

Server Overview Layout Metadata **Tools**

▼ **Layout Analysis**

Method: CITlab Advanced ▼ Configure...

☒ Current page ☐ Pages (1): 1-1 ...

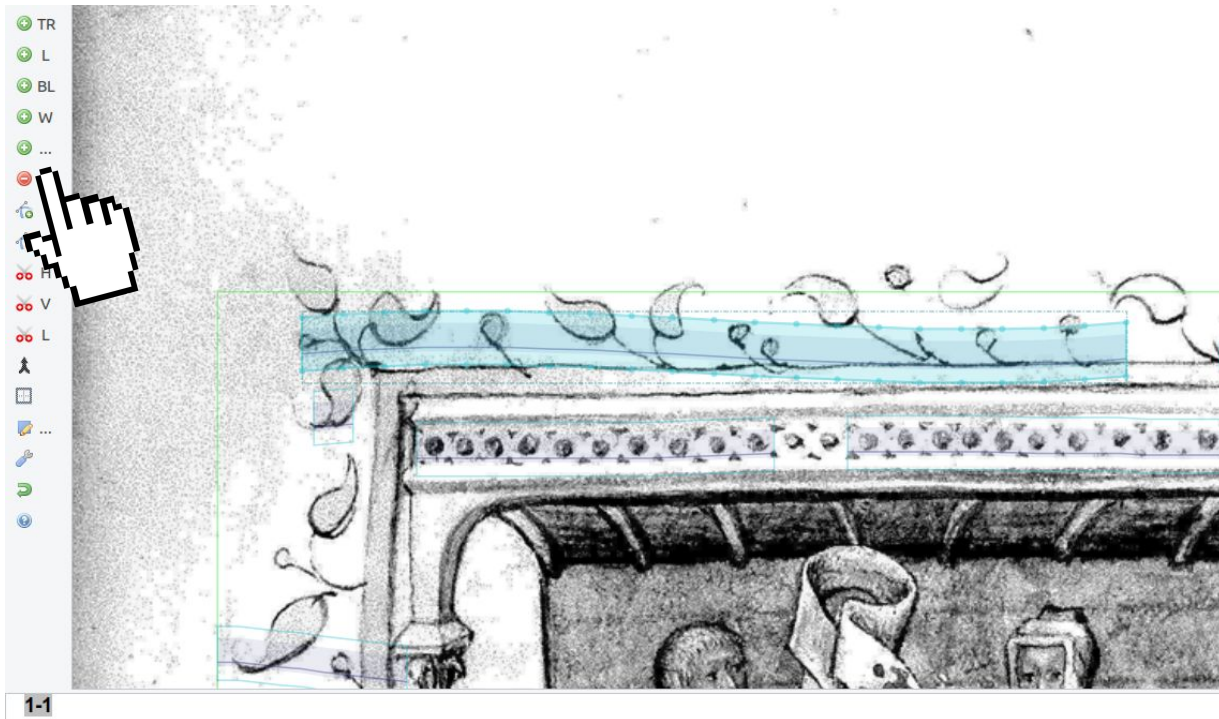
☒ Find Text Regions ☒ Find Lines in Text Regions

Only use unsegmented pages

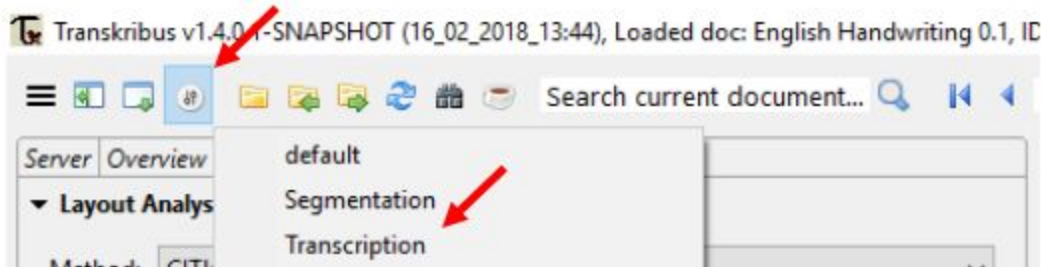
→ Run

▼ **Text Recognition**

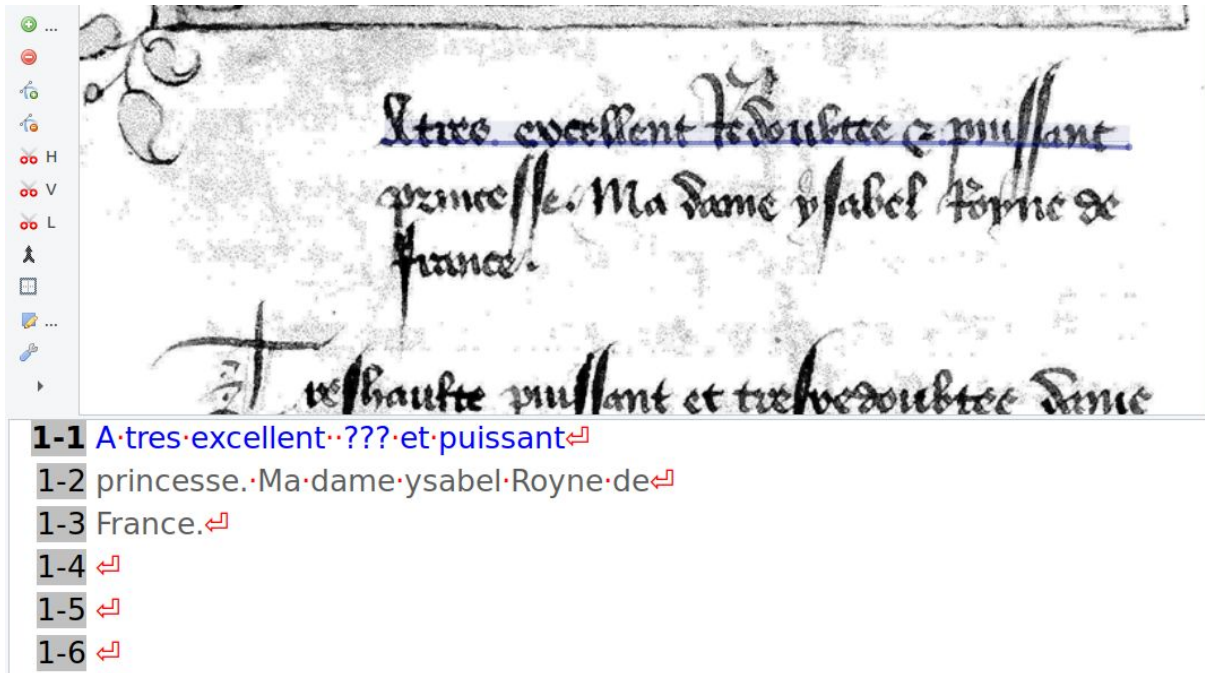
# Nettoyer la segmentation



## Passer en profil transcription



## Transcrire



The image shows a screenshot of a digital transcription interface for a medieval manuscript. The manuscript text is in Gothic script. The first line is "A tres excellent Redoubtee & puissant" and the second line is "princesse. Ma dame ysabel Royne de France." The third line is "Et reshaute puissant et tresredoubtee Dame". The interface includes a toolbar on the left with icons for zooming, panning, and other editing functions. Below the manuscript image, there is a list of transcription steps, each with a label and a red arrow icon.

1-1 A·tres·excellent·???·et·puissant↵

1-2 princesse·Ma·dame·ysabel·Royne·de↵

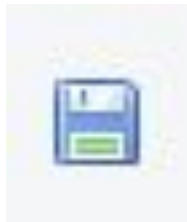
1-3 France.↵

1-4 ↵

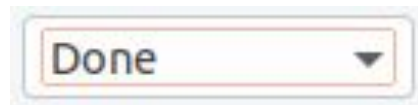
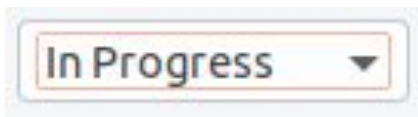
1-5 ↵

1-6 ↵

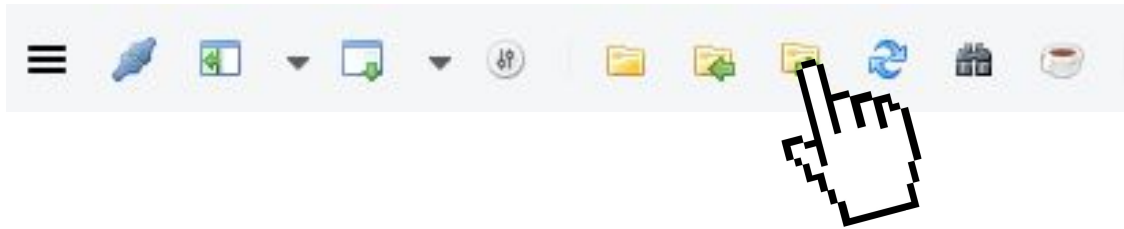
## Sauvegarder la transcription



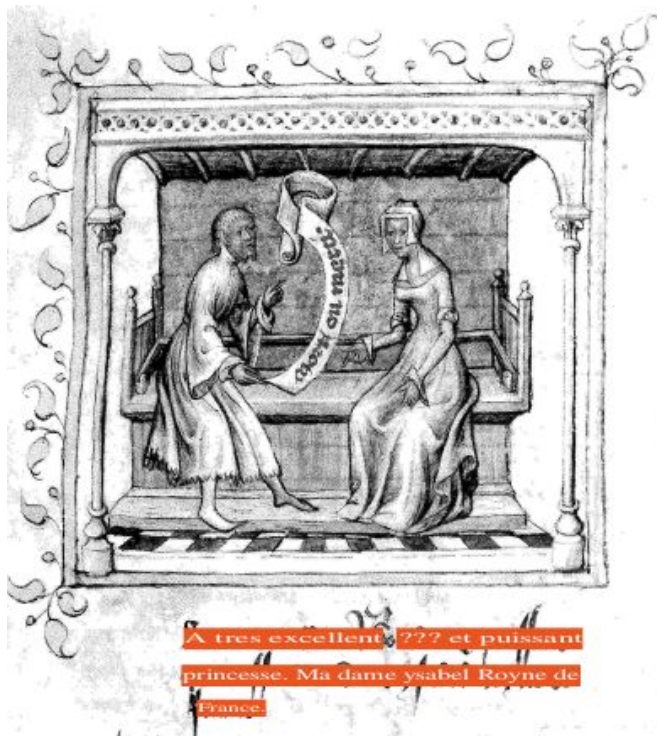
## Achever la transcription



## Exporter le document



## Exporter en pdf

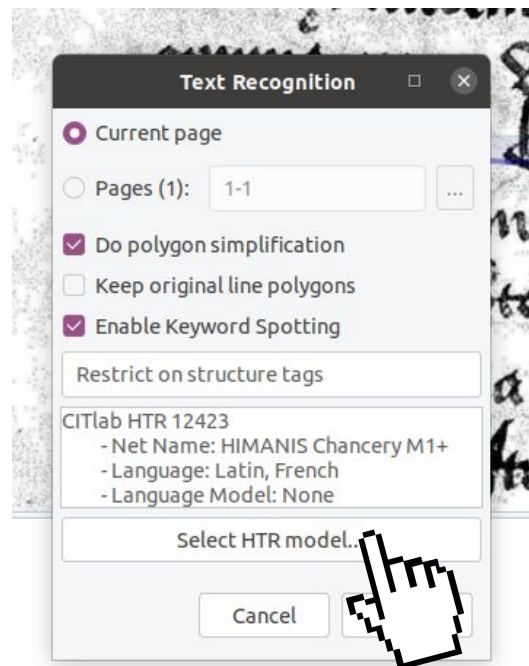
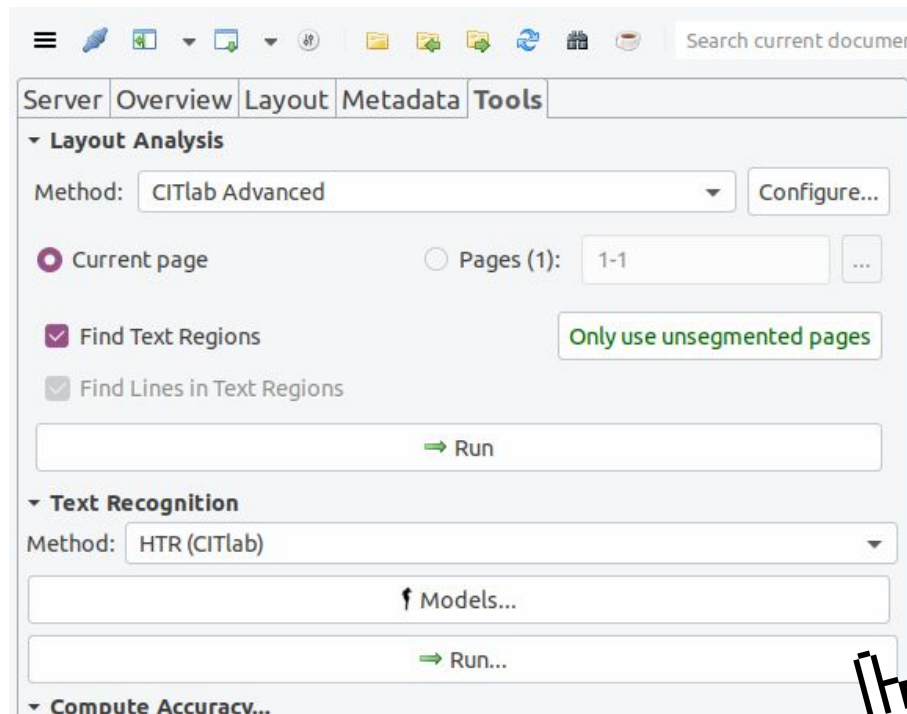




# Passage à l'échelle

- Possibilité d'entraîner un algorithme d'HTR (20 000 mots  $\approx$  100p.)  
→ envoi des documents à l'équipe de Transkribus pour la création du moteur HTR (*Handwritten Text Recognition*)
  - Possibilité d'utiliser un algorithme disponible entraîné sur des textes avec une écriture comparable à vos documents
-

## Transcription automatique



## Choix d'un algorithme

Filtrer par période, langue et CER (Character Error Rate: taux de caractères fautifs)

The screenshot shows the 'Text Recognition Configuration' window. It features a search bar at the top with 'french' entered, and filters for 'All' and 'Show all'. Below is a table of models:

Name	Language	Curator	Technology	Created	ID
French_18thC_Pylaia	French	info@caromein.i	PyLaia	16.09.20	2612
Transkribus print 0.1	Danish, Dutch	guenter	CITlab HTR+	19.05.20	2396
Charter Scripts XIII-XV_M1	German, Latin	tobias.hodel@uz	CITlab HTR+	23.12.19	1987
Français ANOM	French	maxime_gohier@	CITlab HTR+	07.12.19	1924
French_18thC_Print	French	info@caromein.i	CITlab HTR+	05.12.19	1916
LaMOP-Livre_Rouge_1	french	pierre.brochard@	CITlab HTR+	27.11.19	1890
HIMANIS Chancery M1+	Latin, French	guenter.hackl@t	CITlab HTR+	14.04.19	1242
BnF_Newseye_M2+	french	guenter.hackl@t	CITlab HTR+	28.02.19	1120
Parallèle des Anciens et des N	french	guenter.hackl@t	CITlab HTR	12.04.18	2756

On the right, the 'Details' panel shows fields for Name, Language, Description, Parameters, Nr. of Words, and Nr. of Lines. Below these is a 'Learning Curve' graph titled 'Accuracy in CER' vs 'Epochs'. The graph area contains the text 'No data available'. At the bottom of the details panel, it shows 'CER on Train Set: N/A' and 'CER on Validation Set: N/A'. A 'Language Model' list is visible on the far right, including models like 289recto\_M1.dict, 289recto\_M2.dict, etc. Buttons for 'Save', 'Show Train S', 'Show Validat', and 'Show Charac' are located below the details panel.

## Aller prendre un café

