



HAL
open science

Initiation to Handwritten Text Recognition with eScriptorium

Alix Chagué

► **To cite this version:**

Alix Chagué. Initiation to Handwritten Text Recognition with eScriptorium. Horizons of digital philology, Università degli Studi di Napoli Federico II; Université de Montréal, Apr 2024, Naples, Italy. hal-04697560

HAL Id: hal-04697560

<https://inria.hal.science/hal-04697560v1>

Submitted on 13 Sep 2024

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



Initiation to Handwritten Text Recognition

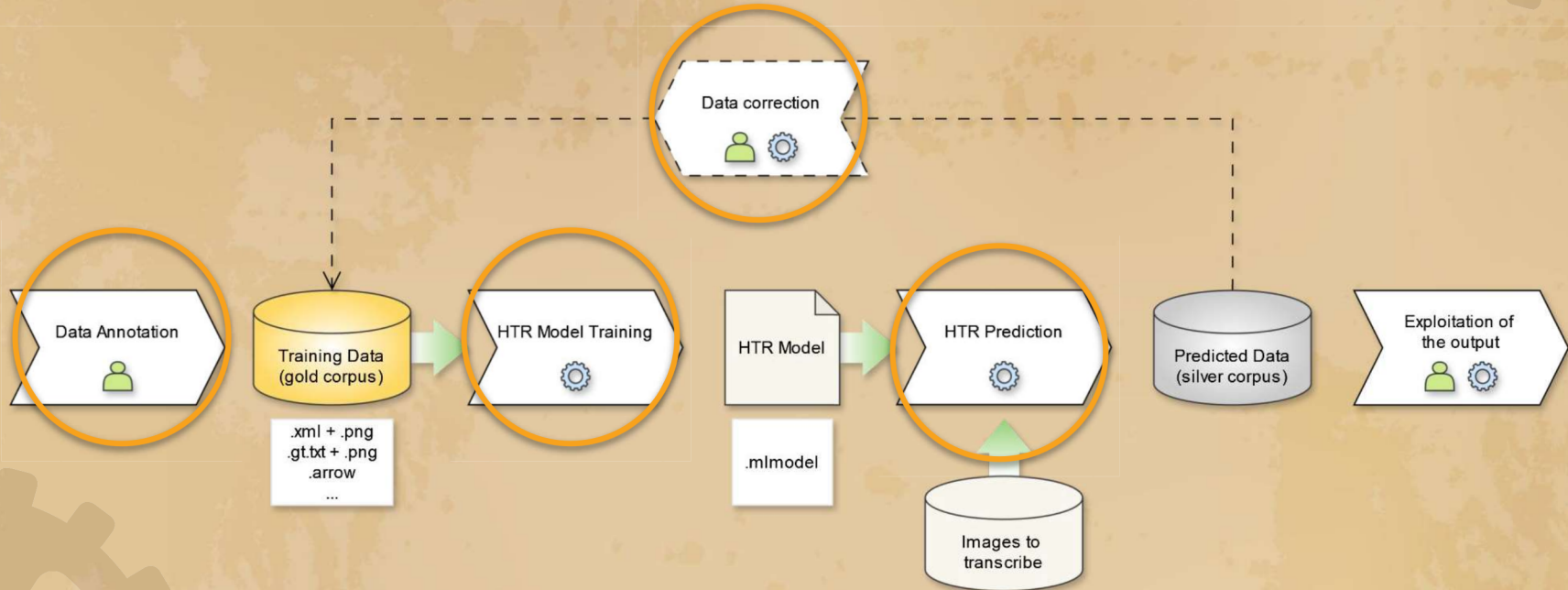
with eScriptorium

Alix Chagué - alix.chague@inria.fr
(Inria; UdeM; EPHE)

Naples - April 16, 2024

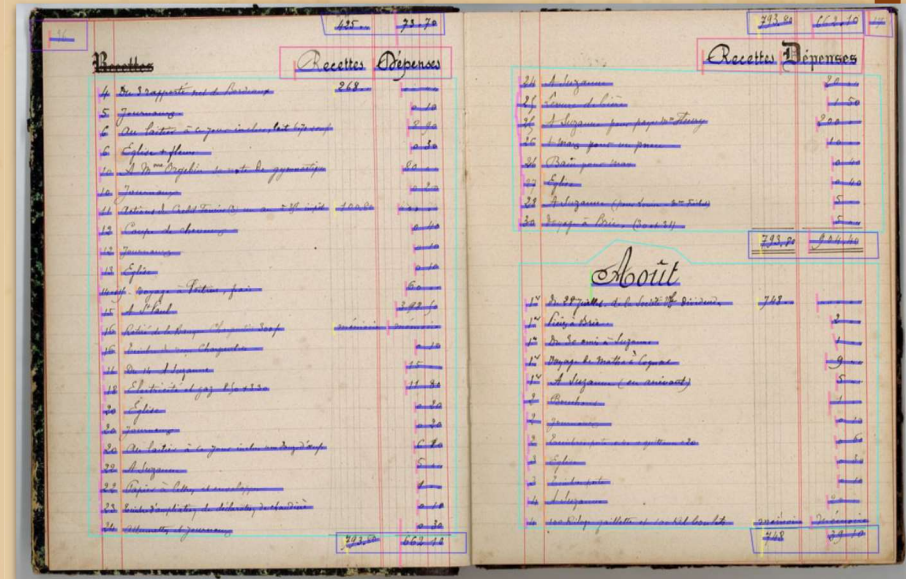
Horizons of digital philology

Creating data and training a model



Steps to apply HTR

1. Retrieve images
2. Process images (optional)
3. Detect the layout of the image
 - a. can be project specific
 - b. can be generic or very detailed
4. Detect text lines ("segment") and line order



Steps to apply HTR

1. Retrieve images
2. Process images (optional)
3. Detect the layout of the image
 - a. can be project specific
 - b. can be generic or very detailed
4. Detect text lines ("segment") and line order
5. Predict the text in the lines

16	425	73.70	793.80	662,10	17
		Recettes		Dépenses	
Du 3 rapporté net de Bordeaux	268	0110	24 A Suzanne	261	
Jeuneaux		0110	25 Levure de bière	1150	
Au laitier à ce jour inclus, lait 6.70 et oeufs		8190	25 A Suzanne pour payer M ^{me} Fleury	2001	
Eglise + fleurs		0130	26 A Max pour un pneu	101	
10 A M ^{me} Orgebin sa note de gymnastique	261		26 Bain pour Max	0140	
10 Jeuneaux		0120	27 Eglise	0140	
11 Actions du Crédit Foncier (3) un an à 35% impôt 100.80		0110	28 A Suzanne (pour Louise - M ^{me} Froiset)	51	
12 Coupe de cheveux		0140	30 Voyage à Brie, (30 et 31),	51	
12 Jeuneaux		0110		793.80	904.40
13 Eglise		0110			
14 et 15 Voyage à Poitiers, frais.		601	Août		
15 A St Paul.		398,50	1 ^{er} du 28 Juillet de la Société V ^{ie} le dividendes	748	
16 Retiré de la Banque Charpentier 300 fr.		mémoire	1 ^{er} Pièce à Brie.	21	
16 Timbre du reçu Charpentier.		0110	1 ^{er} Du 30 omis à Suzanne.	11	
16 Du 14 A Suzanne		1151	1 ^{er} Voyage de Matha à Cognac	91	
18 Electricité et gaz 8.50 + 3.30		11.80	1 ^{er} A Suzanne (en arrivant.)	51	
20 Eglise		0120	Beuchons.	11	
20 Jeuneaux		0120	Jeuneaux	0110	
20 Au laitier à ce jour inclus une douz. d'oeufs		0120	Timbres-poste, 0.40 + quittance 0.20	0100	
22 A Suzanne		51	Eglise	0130	
22 Papier à lettres et enveloppes		11	Timbres poste	0110	
23 Timbre d'amplification de déclaration de chaudière		0110	A Suzanne	201	
24 Allumettes et journaux			100 kilogr. galette et 100 kil. boulets.	mémoire	
		793.80		748	39110

Steps to apply HTR

1. Retrieve images
2. Process images (optional)
3. Detect the layout of the image
 - a. can be project specific
 - b. can be generic or very detailed
4. Detect text lines ("segment") and line order
5. Predict the text in the lines
6. Export the data (XML ALTO, XML PAGE, TXT...)

```
<Layout>
  <Page WIDTH="4284" HEIGHT="2772" PHYSICAL_IMG_NR="9" ID="eSc_dummyspage_">
    <PrintSpace HPOS="0" VPOS="0" WIDTH="4284" HEIGHT="2772">
      <TextBlock HPOS="2414" VPOS="1133" WIDTH="1702" HEIGHT="1364" ID="eSc_textblock_074cf1c7"
        TAGREFS="BT11982">
        <Shape>
          <Polygon POINTS="2414 1244 2414 2497 4112 2497 4116 1244 ... 1244" />
        </Shape>
        <TextLine ID="eSc_line_7588489e" TAGREFS="LT407" BASELINE="2754 1346 3177 1342" HPOS="2749"
          VPOS="1182" WIDTH="423" HEIGHT="183">
          <Shape>
            <Polygon POINTS="2752 1328 2749 1221 2788 1235 2837 1235 ... 1358" />
          </Shape>
          <String CONTENT="Août" HPOS="2749" VPOS="1182" WIDTH="423" HEIGHT="183" />
        </TextLine>
      </TextBlock>
    </PrintSpace>
  </Page>
</Layout>
```

The data you create can be used by others!

Gold Data

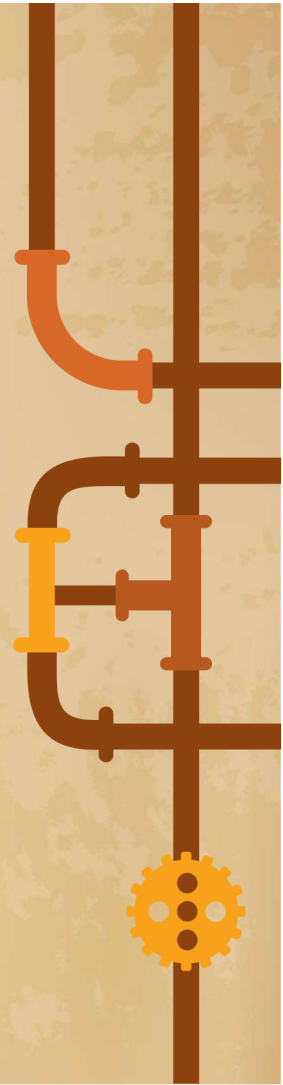
Even if you don't train a model yourself, you can create training data for HTR!

1. define a portion of your corpus (even just 5-10 pages)
2. pay special attention to the correctness of their annotation (text and layout)
3. document the transcription rules you followed
4. publish it with appropriate licencing and credits
5. reference it in HTR-United

It doesn't have to take more than one day of work and it can be very useful to others!

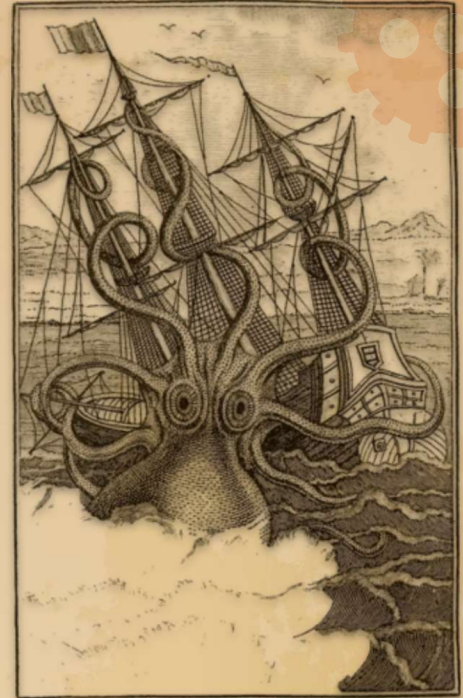


Software environment



Kraken

- Python-based software to train and apply
 - transcription models
 - layout detection models
 - (more since Kraken 5.0.0?)
- Developed by Ben Kiessling
- Available in command line, as a Python module, or through eScriptorium
- **Documentation:** <https://kraken.re>
- **Source code:** <https://github.com/mittagessen/kraken>



eScriptorium

- **Web application** available as a free software in open source
- Manage transcription campaign (train/apply Kraken models, load documents, manually annotate documents, team work, etc)
- Developed by the SCRIPTA PSL project
- Requires a server (locally emulated with Docker or a public server)
- **Documentation:** <https://escriptorium.readthedocs.io/>
- **Source code:** <https://gitlab.com/scripta/escriptorium>



What do you need to use eScriptorium?

- install a local instance or obtain an access (account) to an existing eScriptorium server
- have digitizations of the documents you wish to transcribe
 - local files (.png, .jpeg, .pdf)
 - files stored by an institution on a IIIF server
 - then you need a link to a IIIF manifest:

.tif are sometimes too heavy

<https://gallica.bnf.fr/iiif/ark:/12148/btv1b84259980/manifest.json>

Demo of eScriptorium

- <https://escriptorium.inria.fr>
- login:
- password:

eScriptorium: loading images

The screenshot displays the eScriptorium web interface. At the top, the navigation bar includes the eScriptorium logo, links for Home and Contact, a search bar with the text "Search in Livre Recettes e", and user-specific links for My Projects, My Models, and Hello Alix. Below the navigation bar, a breadcrumb trail shows "Description", "Ontology", "Images", "Edit", "Models", and "Reports", with "Images" selected. The main content area is a large dashed blue box containing the text "Drop images here or click to upload." Below this area, a row of controls includes "Select all", "Unselect all", "Selected 0/84", "Import", "Export", "Train", "Binarize", "Segment", "Transcribe", and "Align".

Below the controls, a series of eight image thumbnails are displayed, numbered 1 through 8. Thumbnail 1 shows the book cover with the title "RECETTES & DÉPENSES". Thumbnails 2 through 8 show pages of handwritten text from the book. Each thumbnail has a small "x" icon in the top right corner and a status bar at the bottom. The status bars for thumbnails 2 through 8 show a "100%" progress indicator. The interface is set against a background with decorative orange gears.

Transkribus

- Software as a service relying on some open source contents (like PyLaia)
- Desktop application (not sustained anymore) and web application
- Other services in the ecosystem (publication, querying, etc)
- Several changes in the business plan over the past few years
 - <https://www.transkribus.org/plans>
- Developed and maintained by READ-COOP
- Allow to train/apply transcription models, apply segmentation models
- Does not allow exporting models

Transkribus[®]