Initiation to Handwritten Text Recognition with eScriptorium

Alix Chagué

Initiation to Handwritten Text Recognition

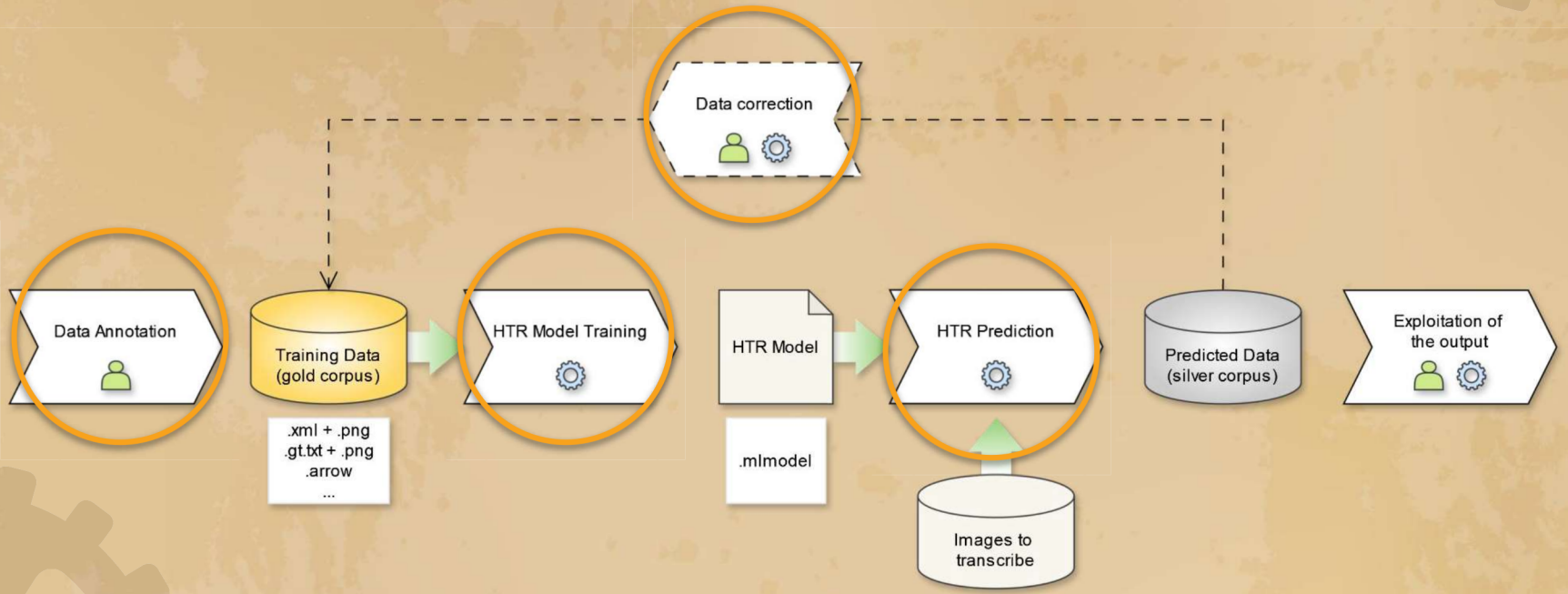
with eScriptorium

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

Naples - April 16, 2024

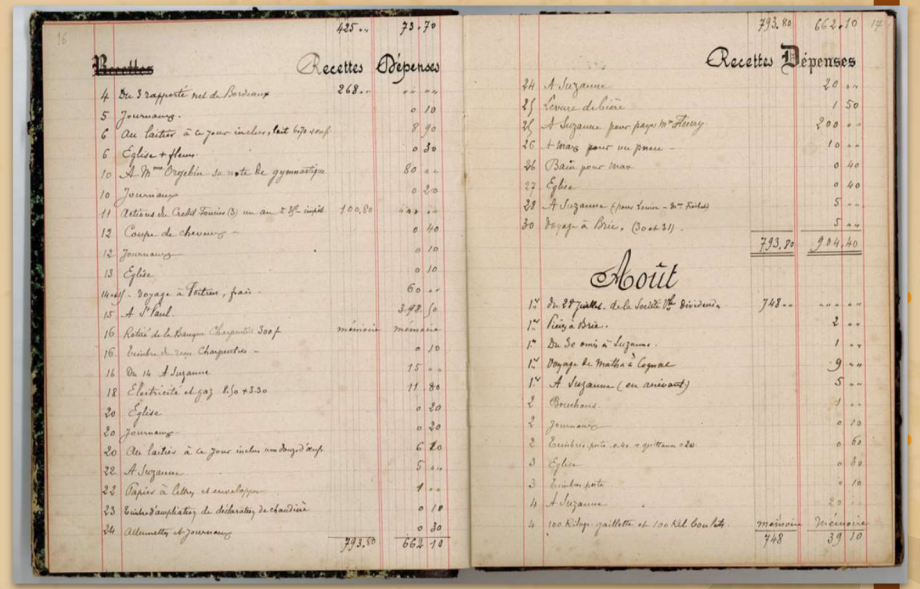
Horizons of digital philology

Creating data and training a model



Steps to apply HTR

1. Retrieve images

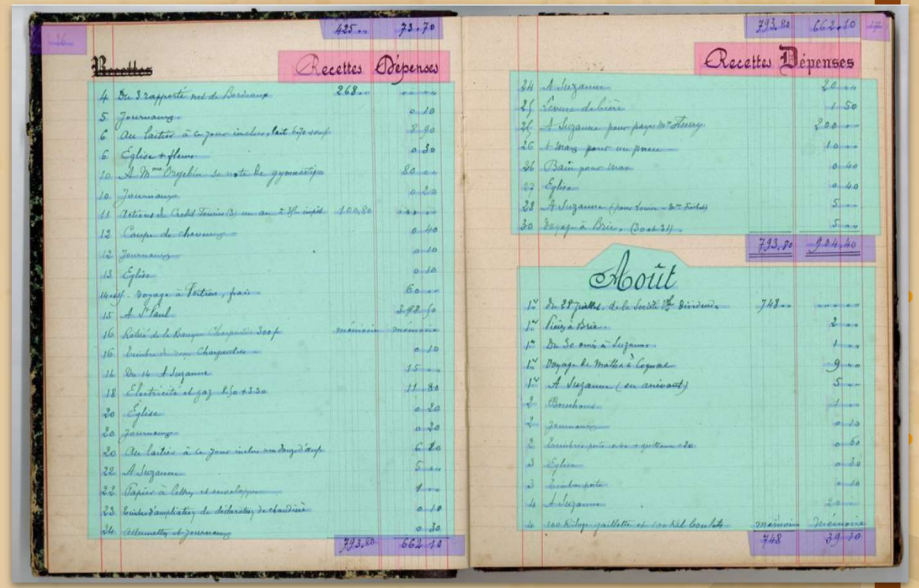


Steps to apply HTR

1. Retrieve images
2. Process images (optional)

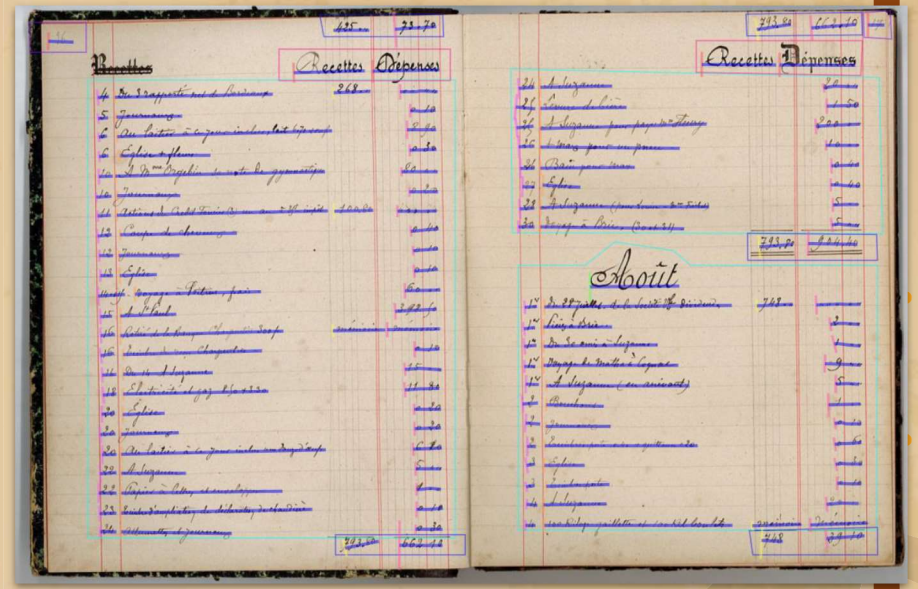
Steps to apply HTR

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



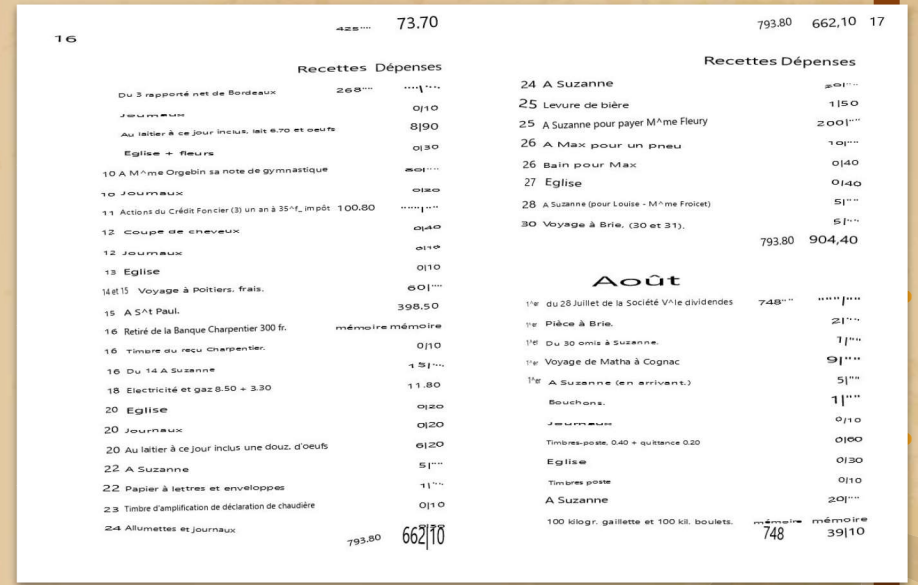
Steps to apply HTR

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

Recettes Dépenses

Du 3 rapporté net de Bordeaux 268···· ····|····
Journaux 0|10
Au laitier à ce jour inclus, lait 6.70 et oeufs 8|90
Eglise + fleurs 0|30
10 A M^me Orgebin sa note de gymnastique 60|····
10 Journaux 0|20
11 Actions du Crédit Foncier (3) un an à 35^f_ impôt 100.80 ·····|···
12 Coupe de cheveux 0|40
12 Journaux 0|10
13 Eglise 0|10
14 et 15 Voyage à Poitiers, frais. 60|····
15 A S^t Paul. 398.50
16 Retiré de la Banque Charpentier 300 fr. mémoire mémoire
16 Timbre du reçu Charpentier. 0|10
16 Du 14 A Suzanne 15|···
18 Electricité et gaz 8.50 + 3.30 11.80
20 Eglise 0|20
20 Journaux 0|20
20 Au laitier à ce jour inclus une douz. d'oeufs 6|20
22 A Suzanne 5|···
22 Papier à lettres et enveloppes 1|···
23 Timbre d'amplification de déclaration de chaudière 0|10
24 Allumettes et journaux 0|10

793.80 662|10

793.80 662,10 17

Recettes Dépenses

24 A Suzanne 20|···
25 Levure de bière 1|50
25 A Suzanne pour payer M^me Fleury 200|···
26 A Max pour un pneu 10|···
26 Bain pour Max 0|40
27 Eglise 0|40
28 A Suzanne (pour Louise - M^me Froicet) 5|···
30 Voyage à Brie, (30 et 31). 5|···

793.80 904,40

Août

1er du 28 Juillet de la Société V^le dividendes 748·· ····|···
1er Pièce à Brie. 2|···
1er Du 30 omis à Suzanne. 1|···
1er Voyage de Matha à Cognac 9|···
1er A Suzanne (en arrivant.) 5|···
Bouchons. 1|···
Journaux 0|10
Timbres-poste, 0.40 + quittance 0.20 0|60
Eglise 0|30
Timbres poste 0|10
A Suzanne 20|···
100 kilogr. gaillette et 100 kil. boulets. mémoire mémoire

748 39|10

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_dummypage_">
    <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>
```

The data you create can be used by others!

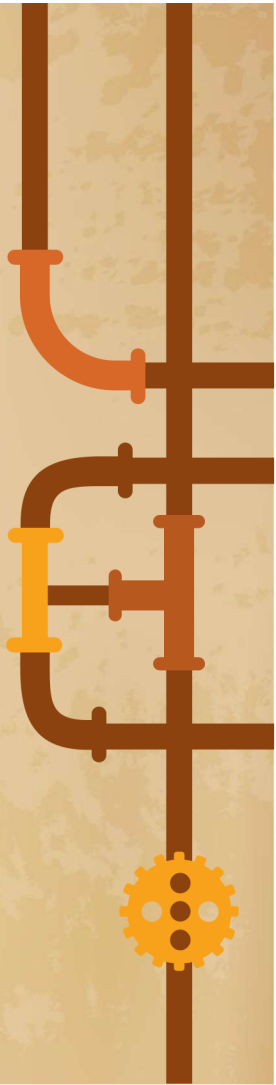
Gold Data

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

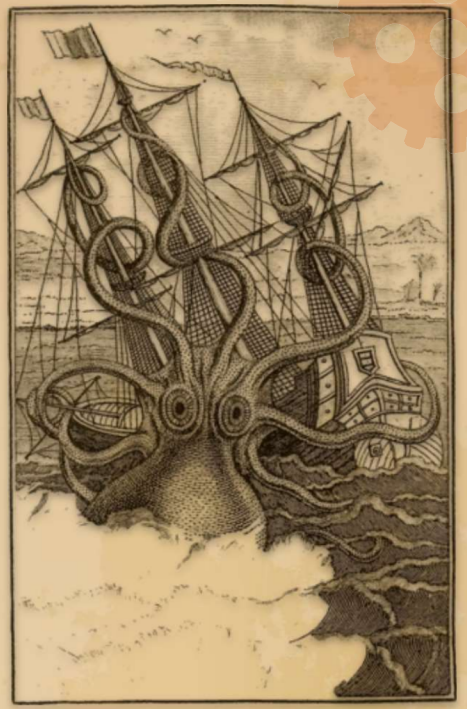
It doesn't have to take more than one day of work and it can be very useful to others!

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?

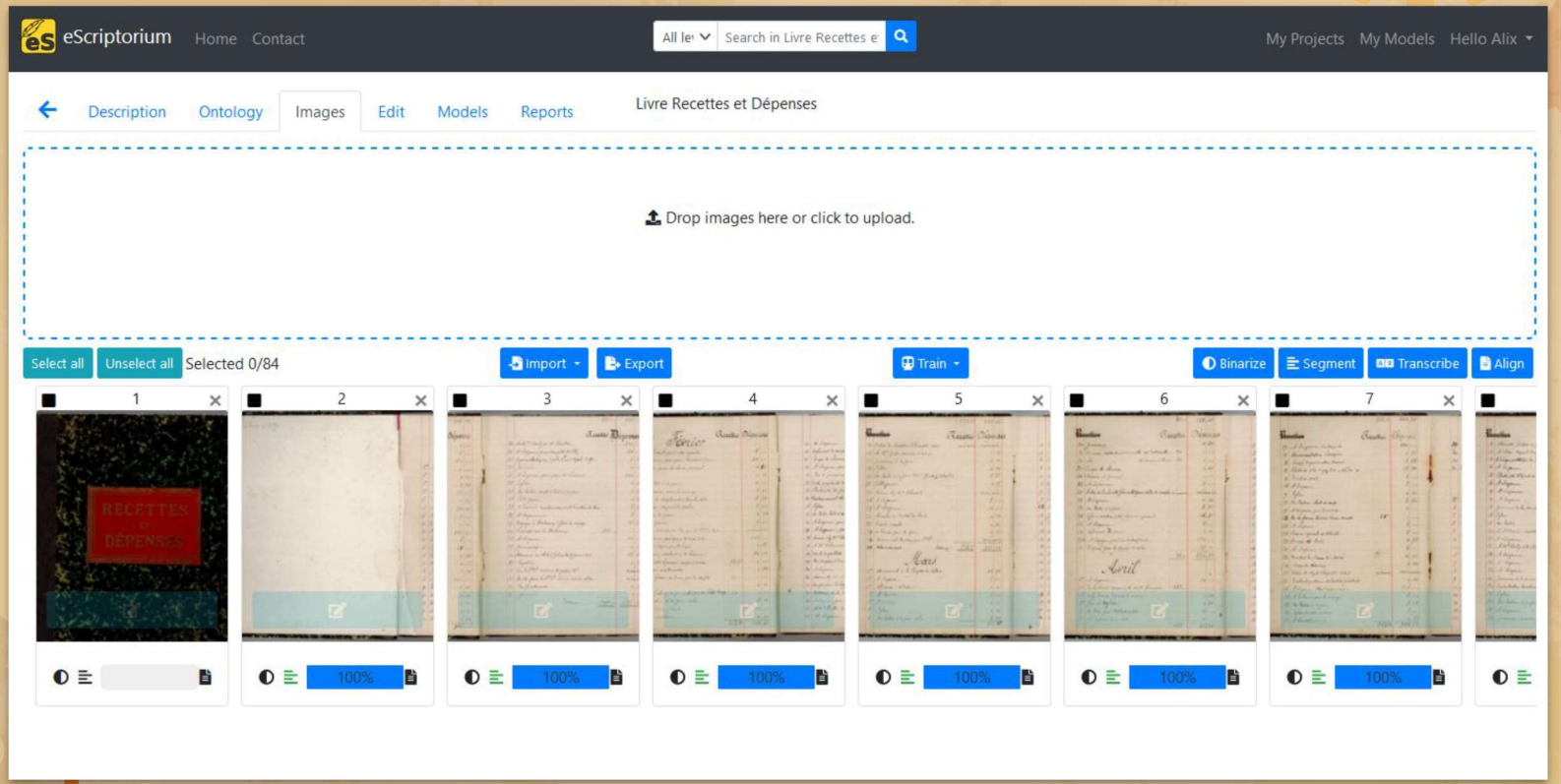
- install a local instance or obtain an access (account) to an existing eScriptorium server
- have digitizations of the documents you wish to transcribe
  - local files (.png, .jpeg, .pdf)
  - files stored by an institution on a IIIF server
    - then you need a link to a IIIF manifest: https://gallica.bnf.fr/iiif/ark:/12148/btv1b84259980/manifest.json

.tif are sometimes too heavy

Demo of eScriptorium

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

eScriptorium: loading images



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