ATR: What can eScriptorium do for you?

Alix Chagué, Floriane Chiffoleau

ATR: What can eScriptorium do for you?

Alix Chagué, Floriane Chiffoleau
From Source to Full Text: Workshop on Using Automatic Text Recognition (ATR)
September 8th 2023

ALMAnaCH project-team

Inria

Le Mans Université

Summary of the presentation

1. Introduction to **ATR software**
2. Guided tour of **eScriptorium**
3. The essence of a good model: **training data**
4. Fantastic **models** (and where to find them)
5. **Predict and assess**: the cornerstone of ATR
6. **Exercise**: Getting acquainted with **eScriptorium**
7. The **eScriptorium documentation**: a helping hand for users
8. **Resources**

Introduction to ATR software

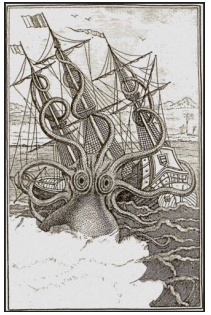
Introduction of the ATR software

- ATR or Automatic Text Recognition is the process of transforming a digitized document into a machine-readable text, through the use of **segmentation** and **text recognition** tools
- Various software for ATR with some specificities:
  - Proprietary or open source
  - Printed documents, handwritten documents or both
  - Charged, freemium, free
- Example of open-source, semi-free OCR engine:
  - Tesseract (no interface)
  - Transkribus (interface) + PyLaia (transcription engine)
  - eScriptorium (interface) + Kraken (transcription engine)

Guided tour of eScriptorium

Guided tour of eScriptorium

- eScriptorium is an open-source software (web application), for transcribing textual documents, developed by the research team Scripta (PSL).
- It provides a **graphical user interface** (GUI) for:
  - **document management**,
  - **layout annotation (aka segmentation)**,
  - **transcribing** (manually or automatically),
  - and **training models**.
  - and more!
- It uses Kraken (Benjamin Kiessling - PSL) as a transcription engine. Kraken is language-agnostic and fully open source.



eScriptorium Home Contact Login

eScriptorium

eScriptorium: A Digital Text Production Pipeline for Print and Handwritten Texts using machine learning techniques.

Automatic Transcription

Apply OCR/HTR to images of printed and handwritten documents using shared open models.

Manual transcription

Make use of an ergonomic user interface leveraging modern browser technology to edit segmentations and transcriptions.

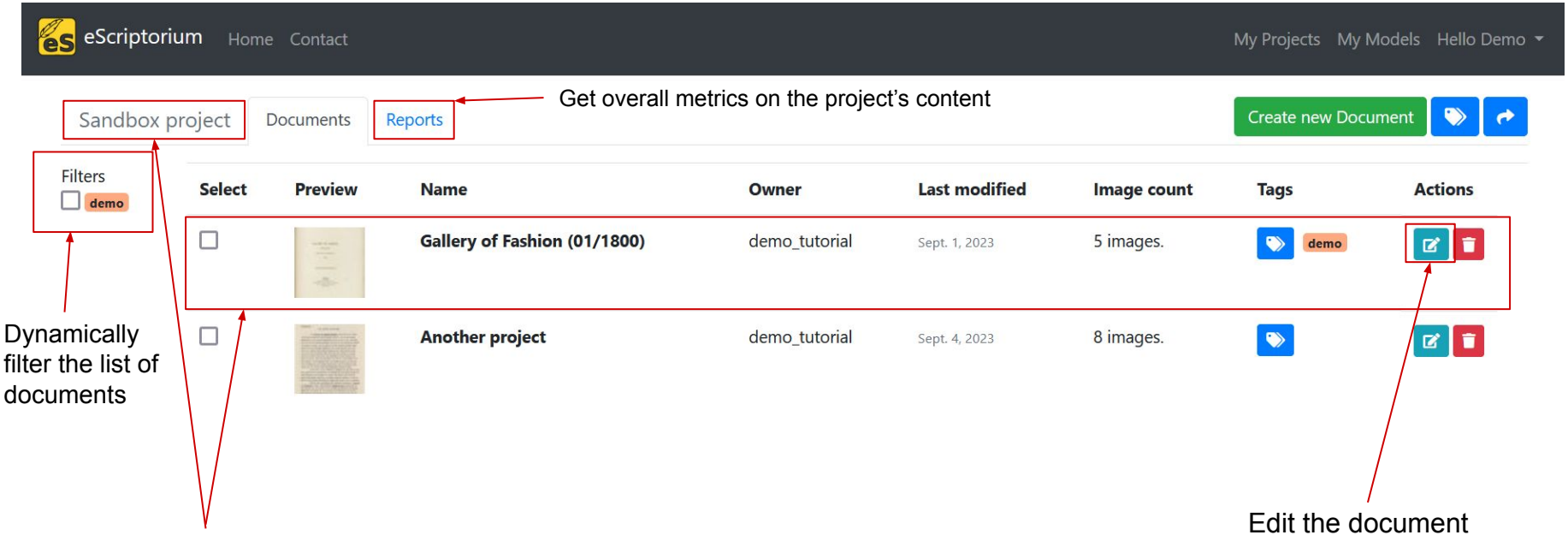
Train Models

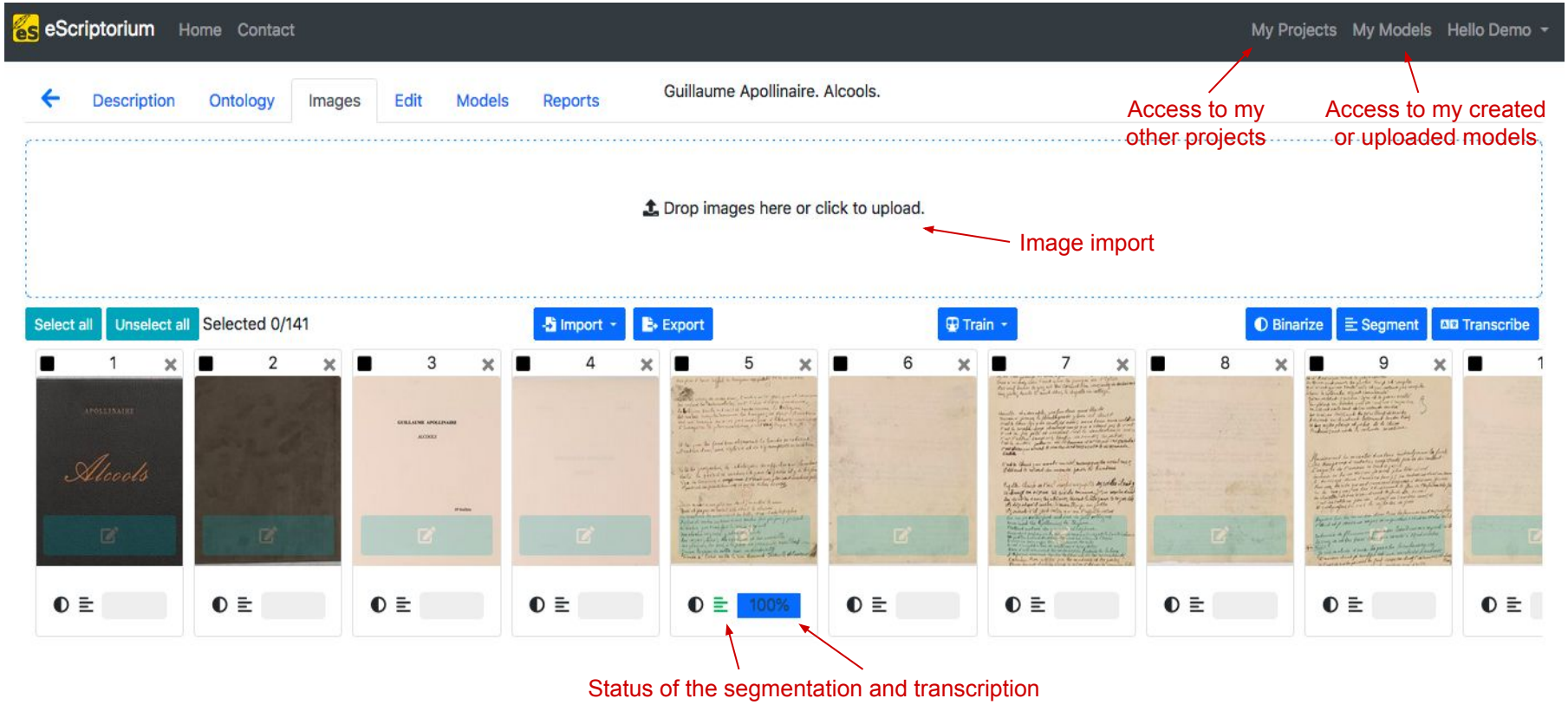
Create new models or finetune existing ones to improve automatic recognition.

Import/ Export

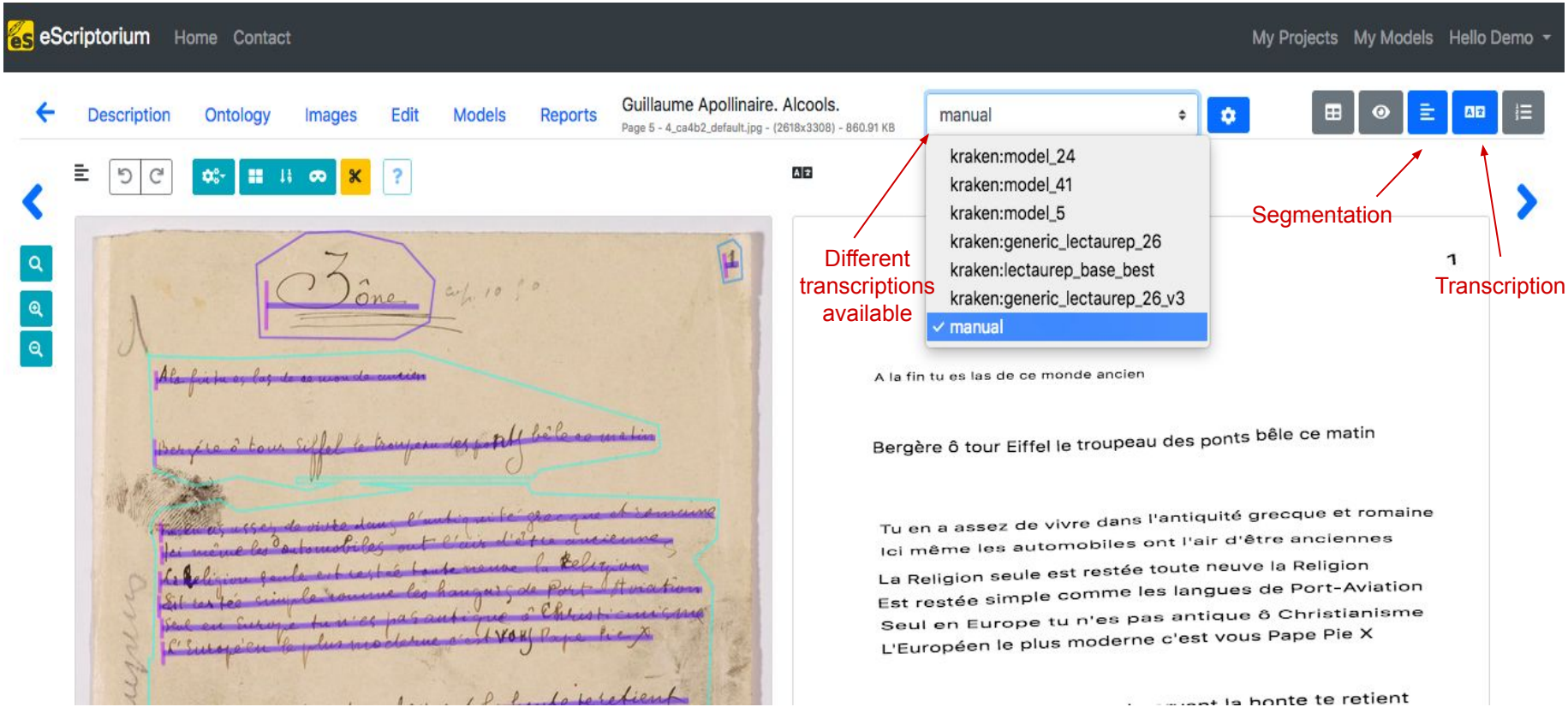
Import and export models and texts transcriptions in a variety of formats. Access data through a full REST API.

Homepage of eScriptorium

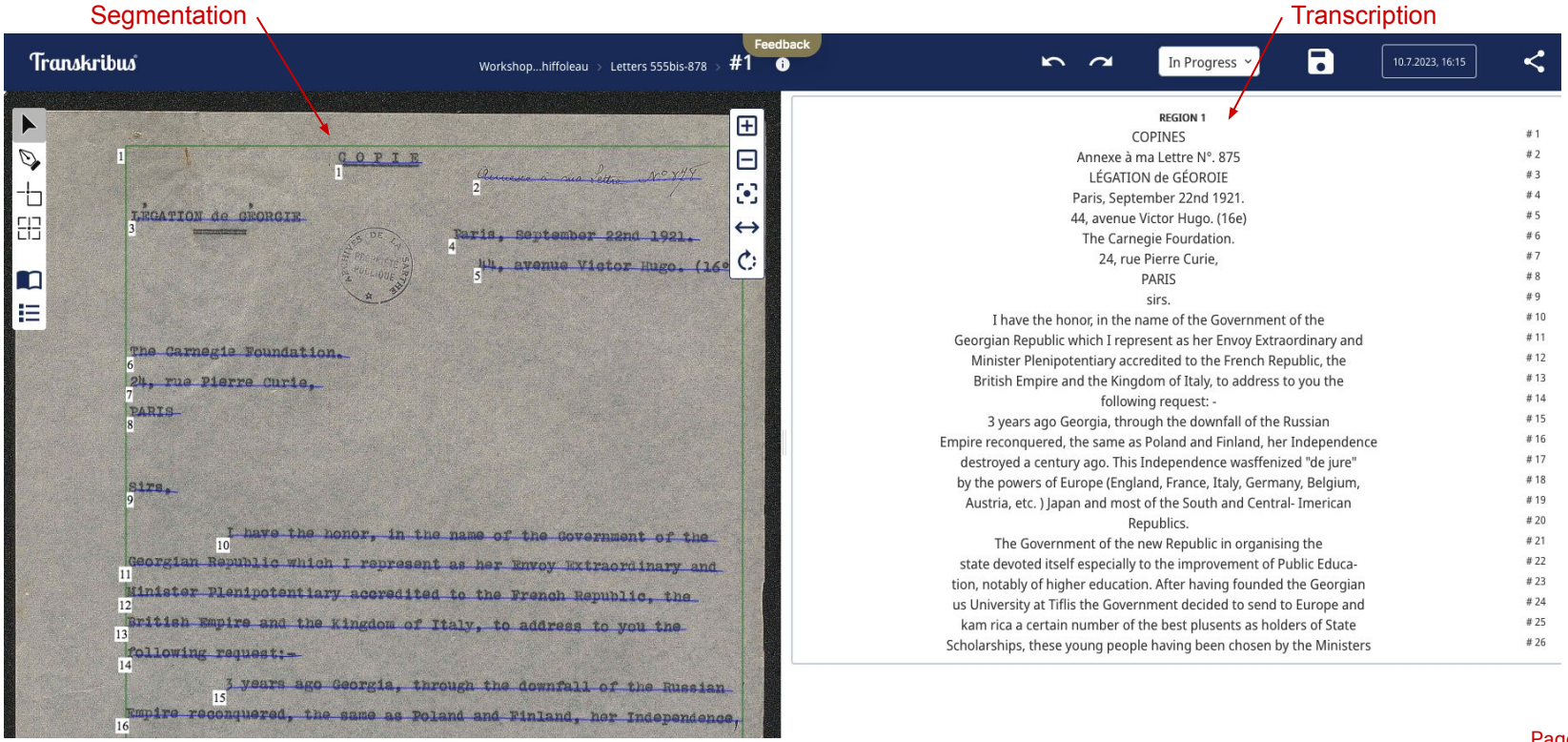

eScriptorium
Home
Contact
My Projects
My Models
Hello Demo
Sandbox project
Documents
Reports
Get overall metrics on the project's content
Create new Document
Filters
demo
Select
Preview
Name
Owner
Last modified
Image count
Tags
Actions
Gallery of Fashion (01/1800)
demo_tutorial
Sept. 1, 2023
5 images.
demo
Another project
demo_tutorial
Sept. 4, 2023
8 images.
Dynamically filter the list of documents
Projects contain documents, which contain images (or document parts)
Edit the document (transcribe, add images, add metadata, etc…)



Document's images dashboard


eScriptorium
Home
Contact
My Projects
My Models
Hello Demo
Description
Ontology
Images
Edit
Models
Reports
Guillaume Apollinaire. Alcools.
Page 5 - 4_ca4b2_default.jpg - (2618x3308) - 860.91 KB
manual
kraken:model_24
kraken:model_41
kraken:model_5
kraken:generic_lectaurep_26
kraken:lectaurep_base_best
kraken:generic_lectaurep_26_v3
manual
Different transcriptions available
Segmentation
Transcription
A la fin tu es las de ce monde ancien
Bergère ô tour Eiffel le troupeau des ponts bêle ce matin
Tu en a assez de vivre dans l'antiquité grecque et romaine
Ici même les automobiles ont l'air d'être anciennes
La Religion seule est restée toute neuve la Religion
Est restée simple comme les langues de Port-Aviation
Seul en Europe tu n'es pas antique ô Christianisme
L'Européen le plus moderne c'est vous Pape Pie X

ATR software are similar because the workflow is close



Transkribus App (09/2023)

The essence of a good model: training data

The essence of a good model: training data

- What is training data ? → Sets of images and their exact transcription (for text recognition) or layout structure (for layout recognition), used to train models
- There are two categories of training data:

GOLD CORPUS

manual annotation (layout or text) or corrected prediction

SILVER CORPUS

automatic prediction (with a model) without correction

The essence of a good model: training data

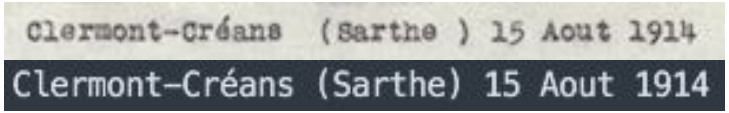
- Three formats for the training data:
  - Sets of ALTO XML files and their images
  - Sets of PAGE XML files and their images
  - Sets of text line images and their transcription
- In each case, same filename, different extension (xml/txt, jpg/png/tiff)

Example of a line from an ALTO XML

```
    <TextLine ID="eSc_line_3b2a0fba"
              BASELINE="703 102 1507 108" HPOS="701.0" VPOS="73.0" WIDTH="806.0" HEIGHT="69.0">
      <Shape><Polygon POINTS="703 102 701 135 1052 129 1080 139 1082 139 1084 139 1085 139 1106 129 1124 140 1126 140 1128 140
      1129 140 1144 140 1267 133 1413 142 1432 142 1433 142 1439 142 1481 133 1501 142 1507 108 1505 76 703 73 703 102"/></
      Shape>
<String CONTENT="Clermont-Créans (Sarthe) 15 Aout 1914" HPOS="701.0" VPOS="73.0" WIDTH="806.0" HEIGHT="69.0"></String>
    </TextLine>
```

Example of a line from a PAGE XML

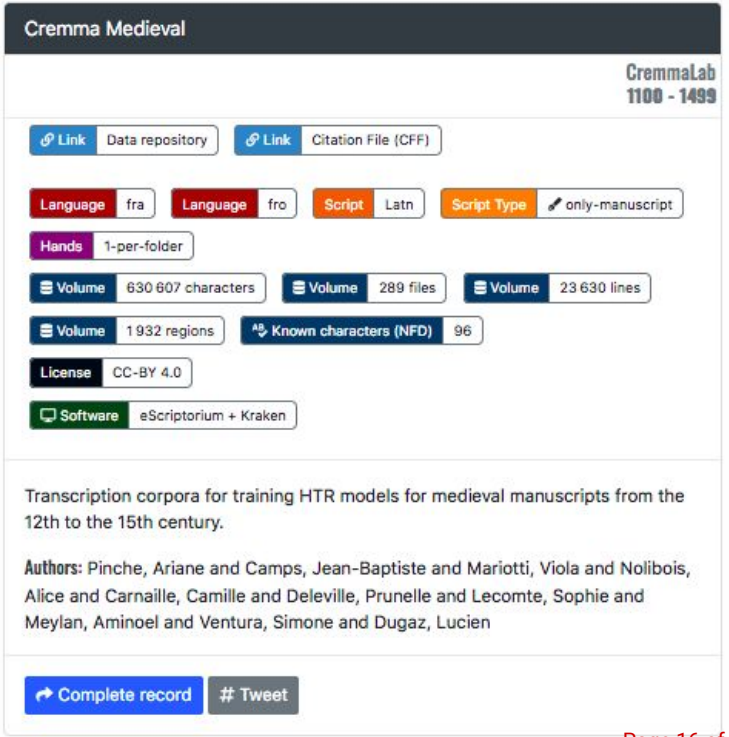
```
<TextLine id="eSc_line_3b2a0fba" >
  <Coords points="703,102 701,135 1052,129 1080,139 1082,139 1084,139 1085,139 1106,129 1124,140 1126,140 1128,140 1129,140
  1144,140 1267,133 1413,142 1432,142 1433,142 1439,142 1481,133 1501,142 1507,108 1505,76 703,73 703,102"/>
  <Baseline points="703,102 1507,108"/>
  <TextEquiv>
    <Unicode>Clermont-Créans (Sarthe) 15 Aout 1914</Unicode>
  </TextEquiv>
</TextLine>
```



Set of a text line image and its transcription

The essence of a good model: training data

- Additionally to your own training data, you can find other training data online, such as thanks to the HTR-United catalog
- It contains metadata on training datasets available for the creation of transcription or segmentation models



Fantastic models (and where to find them)

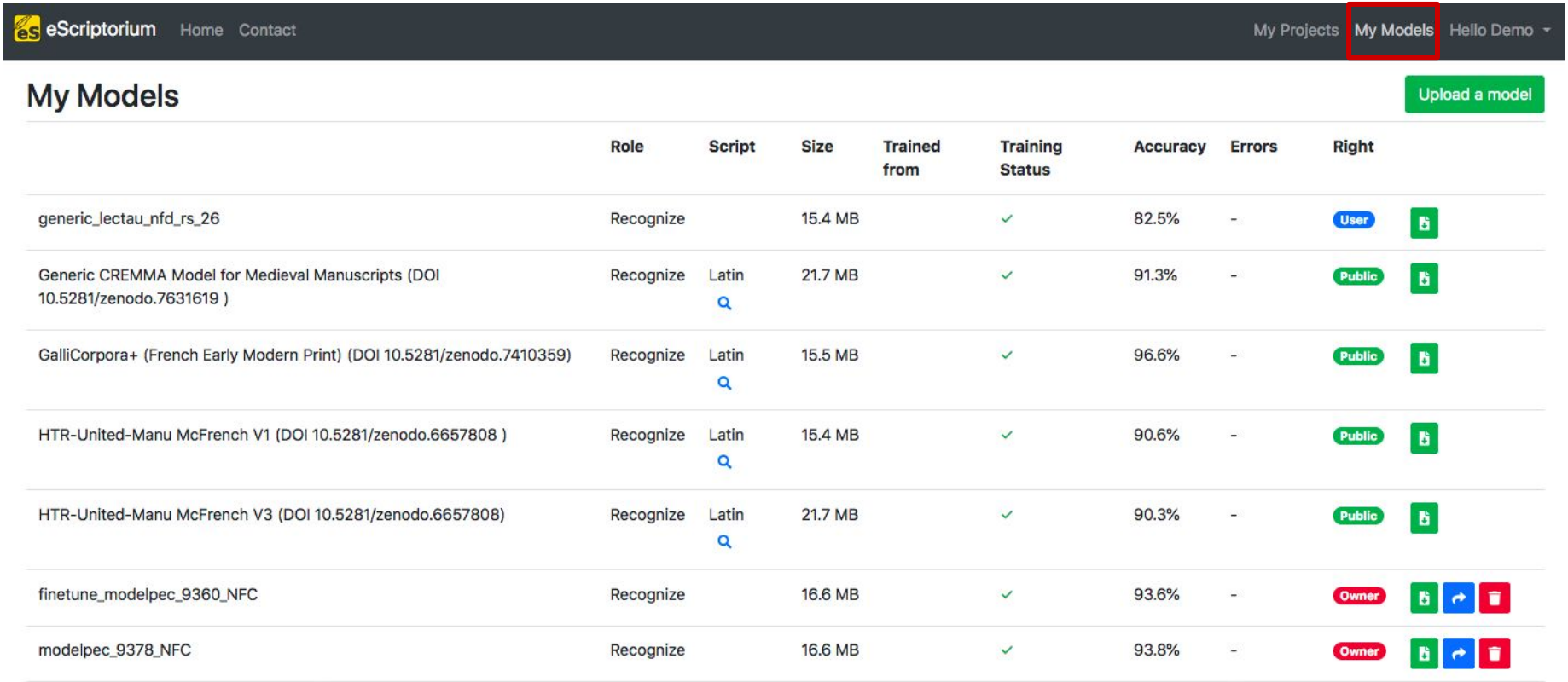
Fantastic models (and where to find them)

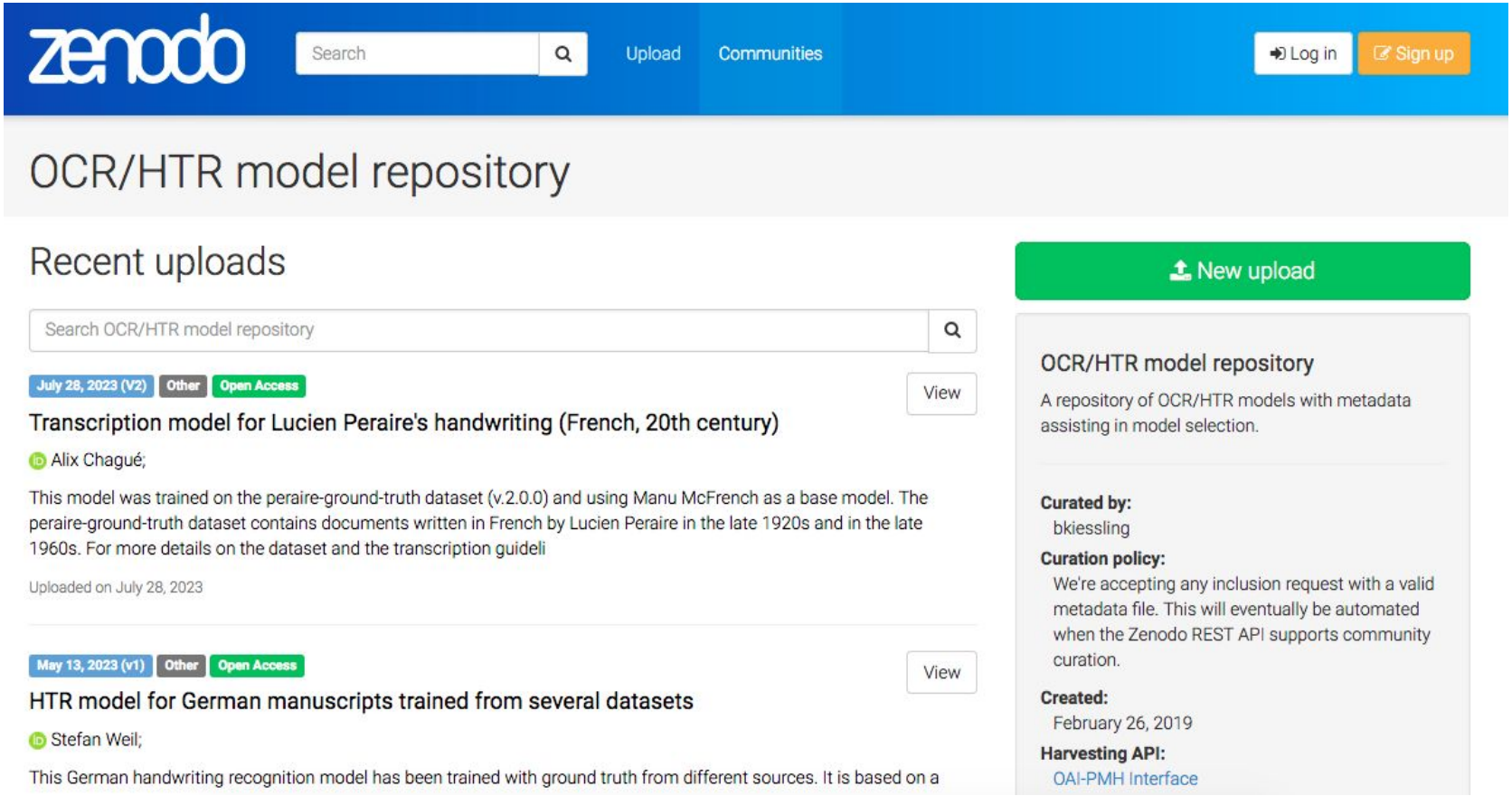
- What is a model ? → It is a file that has been trained to recognize certain types of patterns.
- A model can be developed for 2 types of process
  - **Layout recognition** → Process that determines the constituents of an image and locates the regions and lines of the document where data have been written
  - **Text recognition** → Process that recognizes the characters in document lines and produces a machine-readable text
- Specific to the software that produced them (no standardization)

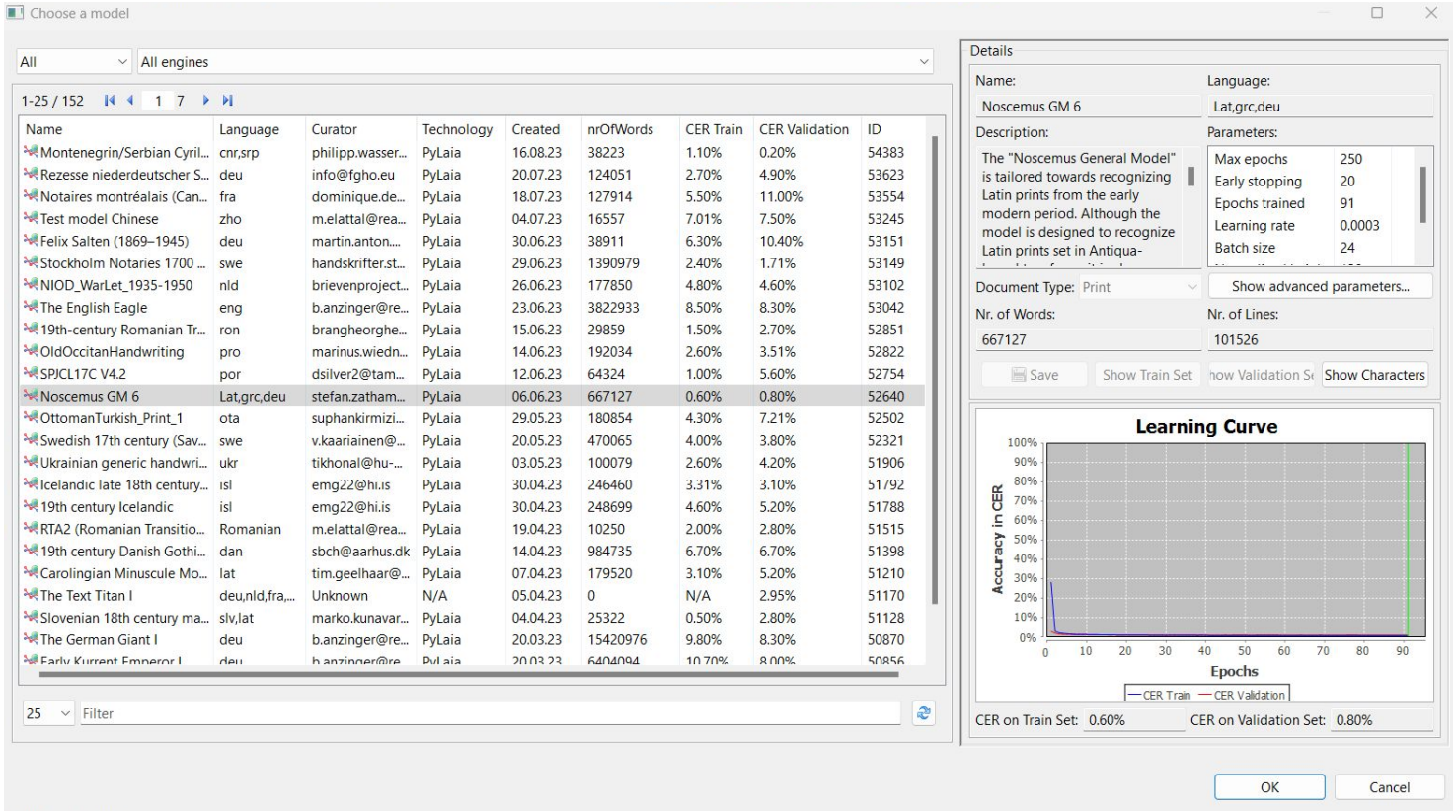
Fantastic models (and where to find them)

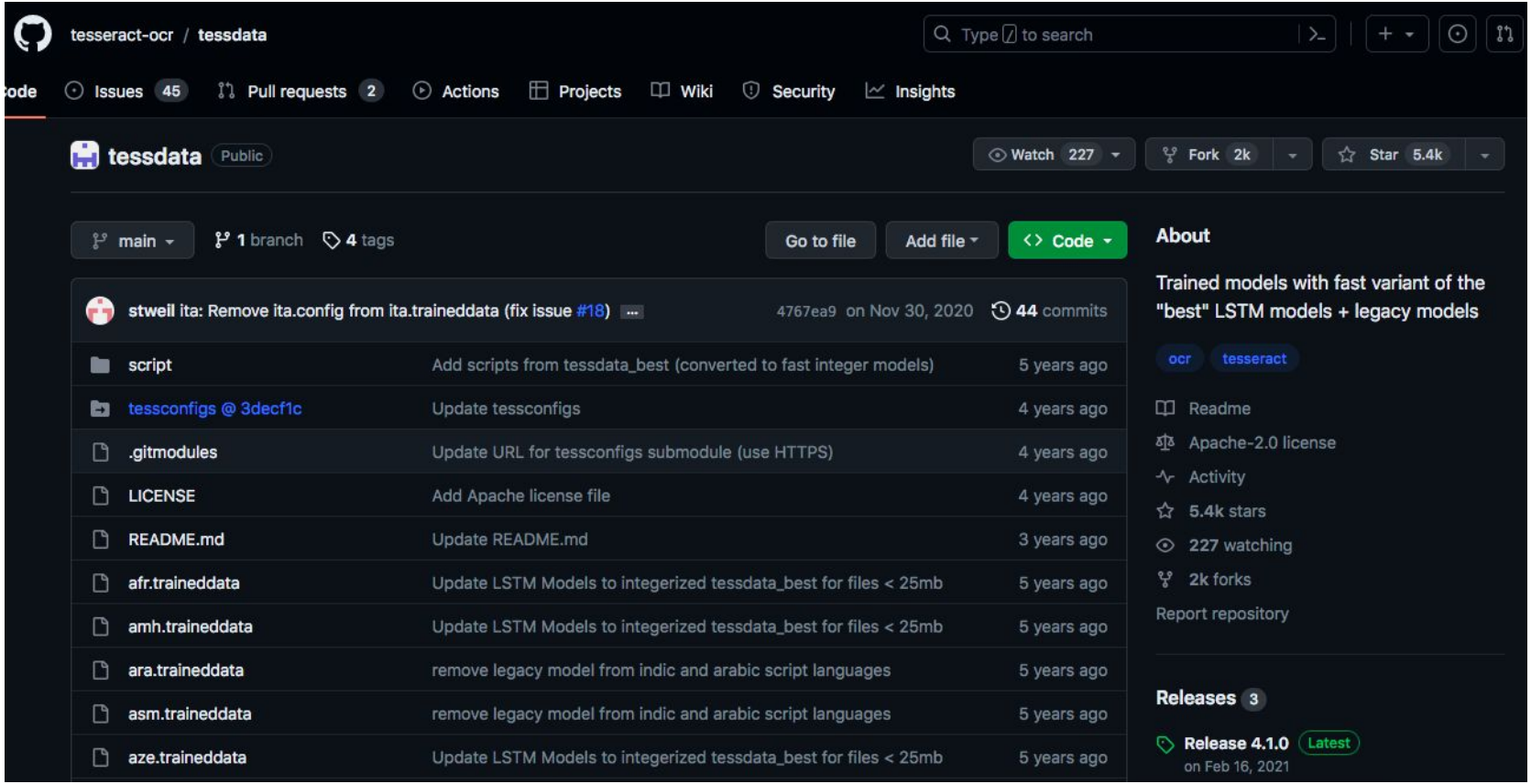
Where to find them ?

- For Kraken and eScriptorium
  - In the tab “My models” on eScriptorium → it contains public models, available for every user and the models the user uploaded/trained themself
  - In the repository “OCR/HTR model” created on Zenodo
- For other software
  - Inside software, but not downloadable (Transkribus, FineReader, etc.) (most common case)
  - Along with software source code (Tesseract)


eScriptorium
Home
Contact
My Projects
My Models
Hello Demo
My Models
Upload a model
Role
Script
Size
Trained from
Training Status
Accuracy
Errors
Right
generic_lectau_nfd_rs_26
Recognize
15.4 MB
82.5%
-
User
Generic CREMMA Model for Medieval Manuscripts (DOI 10.5281/zenodo.7631619)
Recognize
Latin
21.7 MB
91.3%
-
Public
GalliCorpora+ (French Early Modern Print) (DOI 10.5281/zenodo.7410359)
Recognize
Latin
15.5 MB
96.6%
-
Public
HTR-United-Manu McFrench V1 (DOI 10.5281/zenodo.6657808)
Recognize
Latin
15.4 MB
90.6%
-
Public
HTR-United-Manu McFrench V3 (DOI 10.5281/zenodo.6657808)
Recognize
Latin
21.7 MB
90.3%
-
Public
finetune_modelpec_9360_NFC
Recognize
16.6 MB
93.6%
-
Owner
modelpec_9378_NFC
Recognize
16.6 MB
93.8%
-
Owner


zenodo
Search
Upload
Communities
Log in
Sign up
OCR/HTR model repository
Recent uploads
Search OCR/HTR model repository
July 28, 2023 (V2)
Other
Open Access
View
Transcription model for Lucien Peraire's handwriting (French, 20th century)
Alix Chagué;
This model was trained on the peraire-ground-truth dataset (v.2.0.0) and using Manu McFrench as a base model. The peraire-ground-truth dataset contains documents written in French by Lucien Peraire in the late 1920s and in the late 1960s. For more details on the dataset and the transcription guideli
Uploaded on July 28, 2023
May 13, 2023 (v1)
Other
Open Access
View
HTR model for German manuscripts trained from several datasets
Stefan Weil;
This German handwriting recognition model has been trained with ground truth from different sources. It is based on a
New upload
OCR/HTR model repository
A repository of OCR/HTR models with metadata assisting in model selection.
Curated by:
bkiessling
Curation policy:
We're accepting any inclusion request with a valid metadata file. This will eventually be automated when the Zenodo REST API supports community curation.
Created:
February 26, 2019
Harvesting API:
OAI-PMH Interface


Choose a model
All
All engines
1-25 / 152
1 7
Name Language Curator Technology Created nrOfWords CER Train CER Validation ID
Montenegrin/Serbian Cyril... cnr,srp philipp.wasser... PyLaia 16.08.23 38223 1.10% 0.20% 54383
Rezesse niederdeutscher S... deu info@fgho.eu PyLaia 20.07.23 124051 2.70% 4.90% 53623
Notaires montréalais (Can... fra dominique.de... PyLaia 18.07.23 127914 5.50% 11.00% 53554
Test model Chinese zho m.elattal@rea... PyLaia 04.07.23 16557 7.01% 7.50% 53245
Felix Salten (1869–1945) deu martin.anton.... PyLaia 30.06.23 38911 6.30% 10.40% 53151
Stockholm Notaries 1700 ... swe handskrifter.st... PyLaia 29.06.23 1390979 2.40% 1.71% 53149
NIOD_WarLet_1935-1950 nld brievenproject... PyLaia 26.06.23 177850 4.80% 4.60% 53102
The English Eagle eng b.anzinger@re... PyLaia 23.06.23 3822933 8.50% 8.30% 53042
19th-century Romanian Tr... ron brangheorghe... PyLaia 15.06.23 29859 1.50% 2.70% 52851
OldOccitanHandwriting pro marinus.wiedn... PyLaia 14.06.23 192034 2.60% 3.51% 52822
SPJCL17C V4.2 por dsilver2@tam... PyLaia 12.06.23 64324 1.00% 5.60% 52754
Noscemus GM 6 Lat,grc,deu stefan.zatham... PyLaia 06.06.23 667127 0.60% 0.80% 52640
OttomanTurkish_Print_1 ota suphankirmizi... PyLaia 29.05.23 180854 4.30% 7.21% 52502
Swedish 17th century (Sav... swe v.kaariainen@... PyLaia 20.05.23 470065 4.00% 3.80% 52321
Ukrainian generic handwri... ukr tikhonal@hu-... PyLaia 03.05.23 100079 2.60% 4.20% 51906
Icelandic late 18th century... isl emg22@hi.is PyLaia 30.04.23 246460 3.31% 3.10% 51792
19th century Icelandic isl emg22@hi.is PyLaia 30.04.23 248699 4.60% 5.20% 51788
RTA2 (Romanian Transitio... Romanian m.elattal@rea... PyLaia 19.04.23 10250 2.00% 2.80% 51515
19th century Danish Gothi... dan sbch@aarhus.dk PyLaia 14.04.23 984735 6.70% 6.70% 51398
Carolingian Minuscule Mo... lat tim.geelhaar@... PyLaia 07.04.23 179520 3.10% 5.20% 51210
The Text Titan I deu,nld,fra,... Unknown N/A 05.04.23 0 N/A 2.95% 51170
Slovenian 18th century ma... slv,lat marko.kunavar... PyLaia 04.04.23 25322 0.50% 2.80% 51128
The German Giant I deu b.anzinger@re... PyLaia 20.03.23 15420976 9.80% 8.30% 50870
Early Kurrent Emperor I deu b.anzinger@re... PyLaia 20.03.23 6404094 10.70% 8.00% 50856
25
Filter
Details
Name:
Noscemus GM 6
Language:
Lat,grc,deu
Description:
The "Noscemus General Model" is tailored towards recognizing Latin prints from the early modern period. Although the model is designed to recognize Latin prints set in Antiqua-
Parameters:
Max epochs 250
Early stopping 20
Epochs trained 91
Learning rate 0.0003
Batch size 24
Document Type: Print
Show advanced parameters...
Nr. of Words:
667127
Nr. of Lines:
101526
Save
Show Train Set
how Validation Se
Show Characters
Learning Curve
Accuracy in CER
100% 90% 80% 70% 60% 50% 40% 30% 20% 10% 0%
0 10 20 30 40 50 60 70 80 90
Epochs
CER Train
CER Validation
CER on Train Set: 0.60%
CER on Validation Set: 0.80%
OK
Cancel



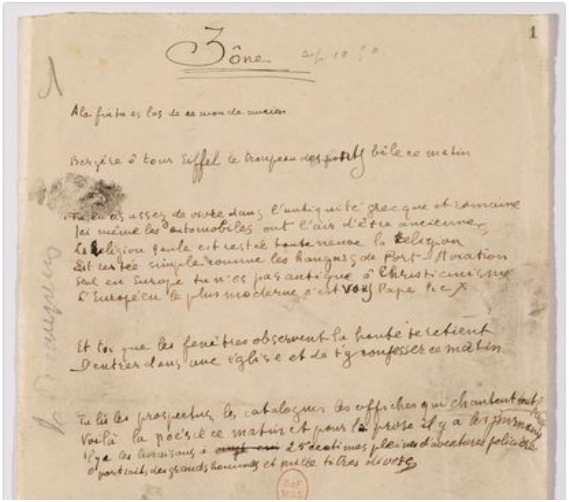
Models in the Tesseract GitHub

Predict and assess: the cornerstone of ATR

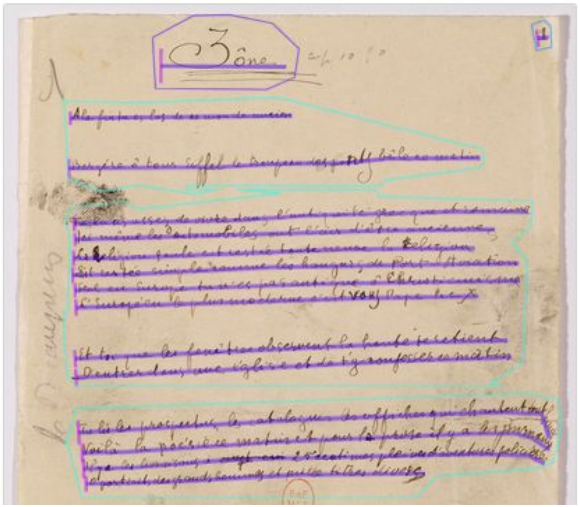
Predict the transcription

- A prediction is the act of using a model to generate (or predict) a layout recognition or a text recognition using an image and a segmentation/transcription model.

Image



Layout



Text

Zône

1

A la fin tu es las de ce monde ancien

Bergère ô tour Eiffel le troupeau des ponts bêle ce matin

Tu en a assez de vivre dans l'antiquité grecque et romaine
Ici même les automobiles ont l'air d'être anciennes
La Religion seule est restée toute neuve la Religion
Est restée simple comme les langues de Port-Aviation
Seul en Europe tu n'es pas antique ô Christianisme
L'Européen le plus moderne c'est vous Pape Pie X

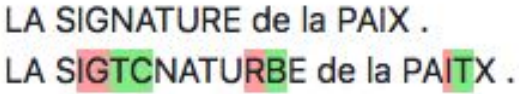
Et toi que les fenêtres observent la honte te retient
D'entrer dans une église et de t'y confesser ce matin

Tu lis les prospectus, les catalogues les affiches qui chantent tout h aut
Voilà la poésie ce matin et pour la prose il y a les journaux
Il y a les livraisons à vingt cin 25 centimes pleines d'aventures policières
O portraits des grands hommes et mille titres divers

Assess the transcription

- How to assess?
  - Eyeballing the result (Can I read it? Does it look gibberish?)
  - Using metrics
    - CER : Character Error Rate
    - WER : Word Error Rate
- How to go deeper?
  - KaMI-Lib or CERberus : ways to nuance the evaluation (caps, digits, diacritics, etc.)

- How to obtain your metrics (from the example below)
  - CER → 25 characters, 4 incorrect characters, CER = 16% (4/25 * 100)
  - WER → 5 words, 2 incorrect words, WER = 40% (2/5 * 100)



Exercise: Getting acquainted with eScriptorium

Exercise: Getting acquainted with eScriptorium

1. Go to the website: escriptorium.inria.fr/
2. Login in eScriptorium:
   - Username: apinche_formation
   - Password: training1234
3. Go to the project: Pinche_FormationDHI
4. Create a new document
   - Name: “workshopdhi_” + your last name → example: “workshopdhi_chiffoleau”
   - Parameters: “Latin”, “Baseline” and “Left to right”
5. Take a look at the models already present on the account
   - “My models”

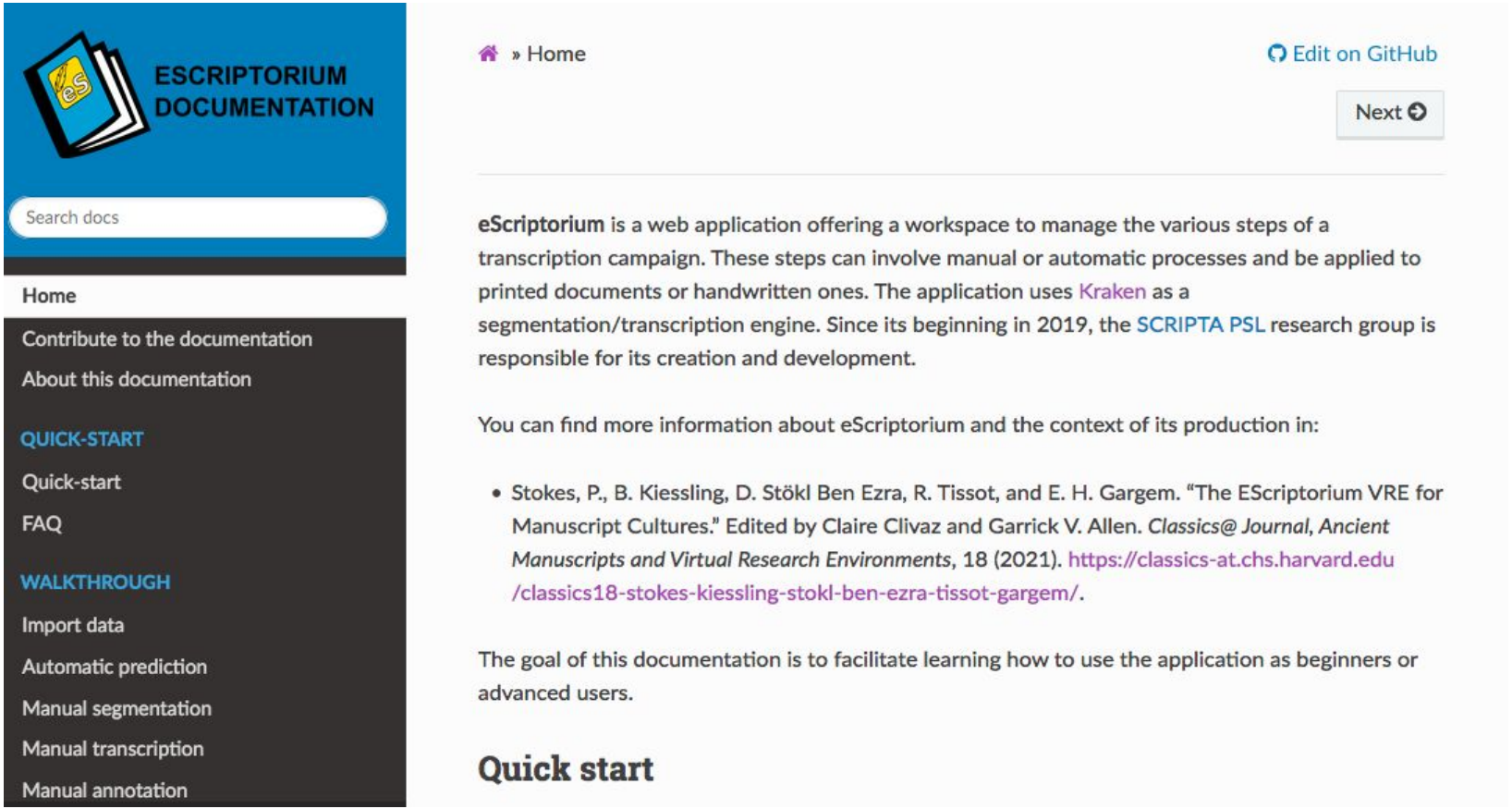
The eScriptorium documentation: a helping hand for users

The eScriptorium documentation: a helping hand for users

- The tutorial aims to provide a **set of (written) walkthroughs detailing eScriptorium's features**, for example:
  - Managing projects and collections of documents,
  - Learning how to segment, annotate and transcribe a text image,
  - Instructions for training models.

The tutorial is available at https://escriptorium-tutorial.readthedocs.io

All collaborations are welcomed: https://github.com/alix-tz/escriptorium-tutorial


ESCRIPTORIUM DOCUMENTATION
Search docs
Home
Contribute to the documentation
About this documentation
QUICK-START
Quick-start
FAQ
WALKTHROUGH
Import data
Automatic prediction
Manual segmentation
Manual transcription
Manual annotation
» Home
Edit on GitHub
Next
eScriptorium is a web application offering a workspace to manage the various steps of a transcription campaign. These steps can involve manual or automatic processes and be applied to printed documents or handwritten ones. The application uses Kraken as a segmentation/transcription engine. Since its beginning in 2019, the SCRIPTA PSL research group is responsible for its creation and development.
You can find more information about eScriptorium and the context of its production in:
• Stokes, P., B. Kiessling, D. Stökl Ben Ezra, R. Tissot, and E. H. Gargem. "The EScriptorium VRE for Manuscript Cultures." Edited by Claire Clivaz and Garrick V. Allen. Classics@ Journal, Ancient Manuscripts and Virtual Research Environments, 18 (2021). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.
The goal of this documentation is to facilitate learning how to use the application as beginners or advanced users.
Quick start

Resources

Resources

- eScriptorium: https://escriptorium.inria.fr/
- eScriptorium documentation: https://escriptorium.readthedocs.io
- Kraken: https://kraken.re/main/index.html
- OCR/HTR models: https://zenodo.org/communities/ocr_models
- CERberus: https://github.com/WHaverals/CERberus
- HTR-United: https://htr-united.github.io/
- KaMILib: https://huggingface.co/spaces/lterriel/kami-app
- PyLaia: https://github.com/jpuigcerver/PyLaia
- Tesseract: https://github.com/tesseract-ocr/tesseract
- Transkribus: https://app.transkribus.eu/

Thank you for your attention

Any questions ?

Contact: *alix.chague[at]inria.fr*
floriane.chiffoleau[at]inria.fr