École Pratique
des Hautes Études

PSL

aoroc

eScriptorium

P.A. Stokes
peter.stokes@ephe.psl.eu

Kraken et eScriptorium

- Conçus principalement pour la transcription automatique des livres imprimés et manuscrits (OCR/HTR/ATR)
- **Kraken** : Un moteur ATR modulaire (cli, en python), développé principalement par Ben Kiessling (EPHE/AOROC)
- **eScriptorium** : Interface web pour Kraken (entre autres), développé principalement par une équipe EPHE/AOROC
- Libres, gratuits, ouverts (le logiciel **et les modèles**)
  - eScriptorium doit être utile et facile d'utiliser
  - Il faut des bons algorithmes, mais aussi une bonne ingénierie

Useful Links

The Kraken Website
kraken @ PyPI
kraken @ github
Issue Tracker

Navigation

Advanced Usage
Training
API Tutorial
API Reference
Models

kraken

kraken is a turn-key OCR system optimized for historical and non-Latin script material.

Features

kraken's main features are:

- Fully trainable layout analysis and character recognition
- Right-to-Left, BiDi, and Top-to-Bottom script support
- ALTO, PageXML, abbyXML, and hO
- Word bounding boxes and character
- Multi-script recognition support
- Public repository of model files
- Lightweight model files
- Variable recognition network archite

Pull requests and code contributions are alw

Installation

Kraken can be run on Linux or Mac OS X (bo
on-board *pip* utility and the anaconda scient

Installation using Pip

http://kraken.re

ESCRIPTORIUM DOCUMENTATION

Search docs

Home
Contribute to the documentation
About this documentation

QUICK-START
Quick-start
Terminology
FAQ

WALKTHROUGH
Import data
Automatic prediction
Manual segmentation
Manual transcription
Manual annotation
Virtual keyboard
Train models

» 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 🚀

The **Quick-start** is intended as a tutorial to introduce completely new users to the basics of using eScriptorium. It should be envisioned as a gateway to the walkthrough section which is much more detailed.

https://escriptorium.readthedocs.io

Quelques types d'utilisateurs

- Ceux qui veulent accéder à tous les paramètres
- Ceux qui veulent des solutions clés en main
- Ayant des documents « simples » (lignes droites horizontales, écriture latine, pas d'annotations, ...)
- Ayant des documents complexes (lignes courbes dans différentes directions, écritures « rares » non alphabétiques, plusieurs mains/écritures différents, ...)
- Ceux qui veulent/doivent suivre les principes FAIR et Open Science/Open Data (pour les données d'entraînement, les modèles et les résultats)

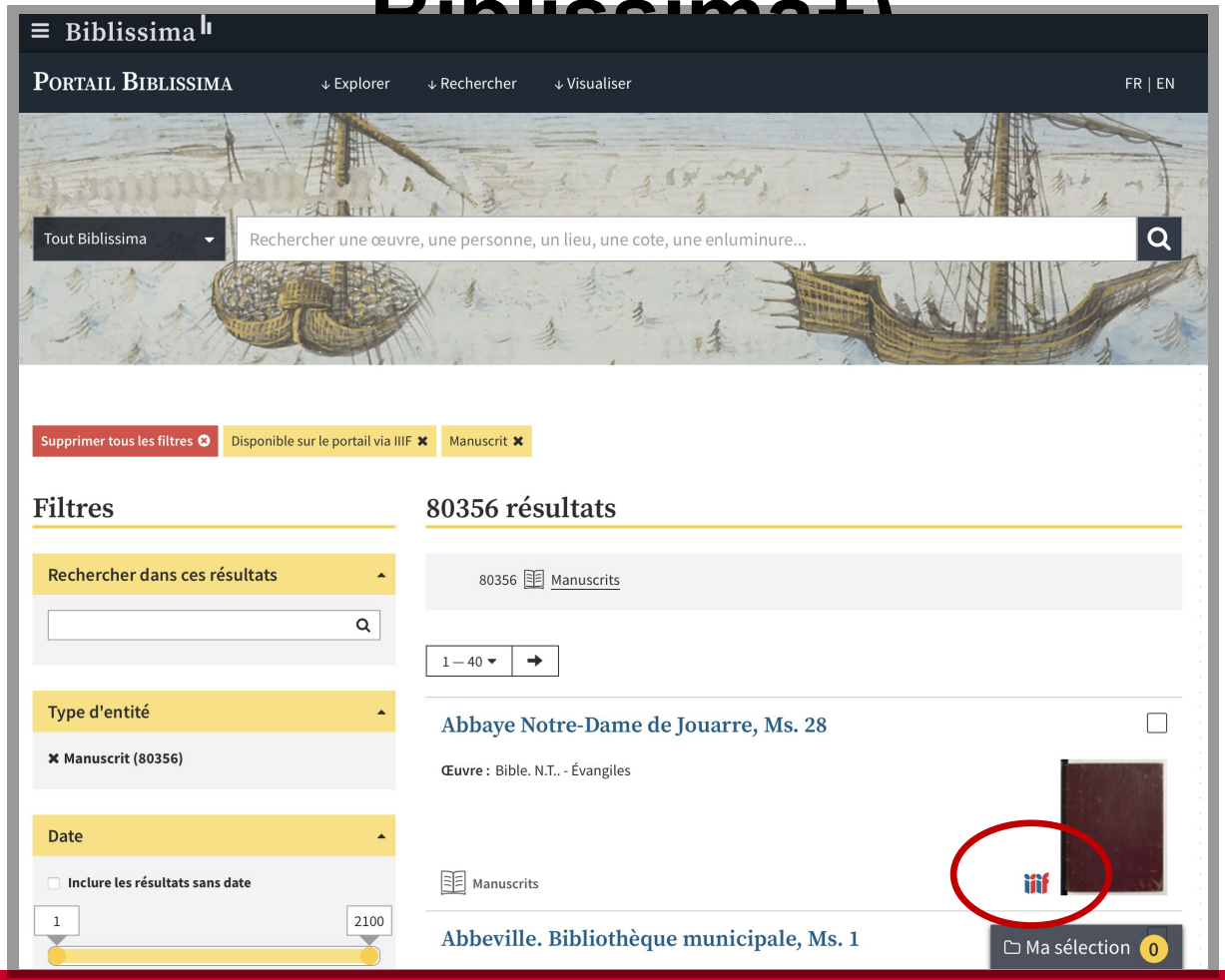
Des Langues de Scripta-PSL (sélection)

- Araméen ancien
- Arménien médiéval
- Chinois pré-imperial
- Égyptien ptolémaïque
- Élamite
- Moyen-iranien
- Japonais médiéval
- Vieux javanais
- Vieux khmer
- Méroitique
- Pali
- Soghdien
- Sumerien
- Tamil classique
- Tai-lue
- Tokharien
- Ougaritique
- Ombrie
- Vietnamien ancien
- …

Cas d’utilisation fréquents

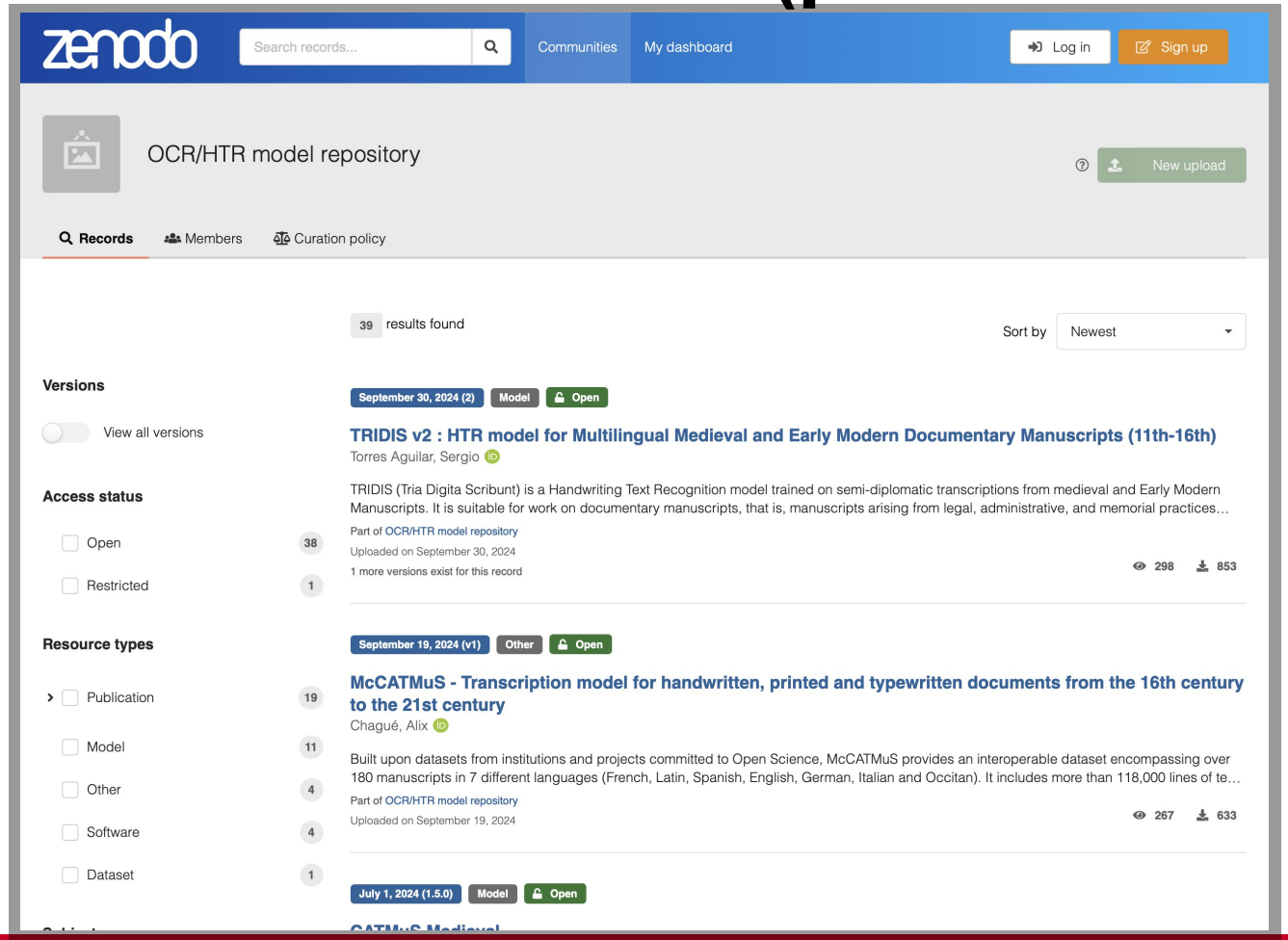
- Un (long) livre manuscrit
- Une grande collection de dizaines de milliers de documents
- Des carnets de recherche (mise en page très irrégulière, diagrammes, etc.)
- Journaux personnels (d’une seule main, mais pouvant varier dans le temps, format souvent différent, ...)
- Des archives des documents courts à plusieurs mains.
- « Tout »

Importer les images via IIIF (par ex. Biblissima+)



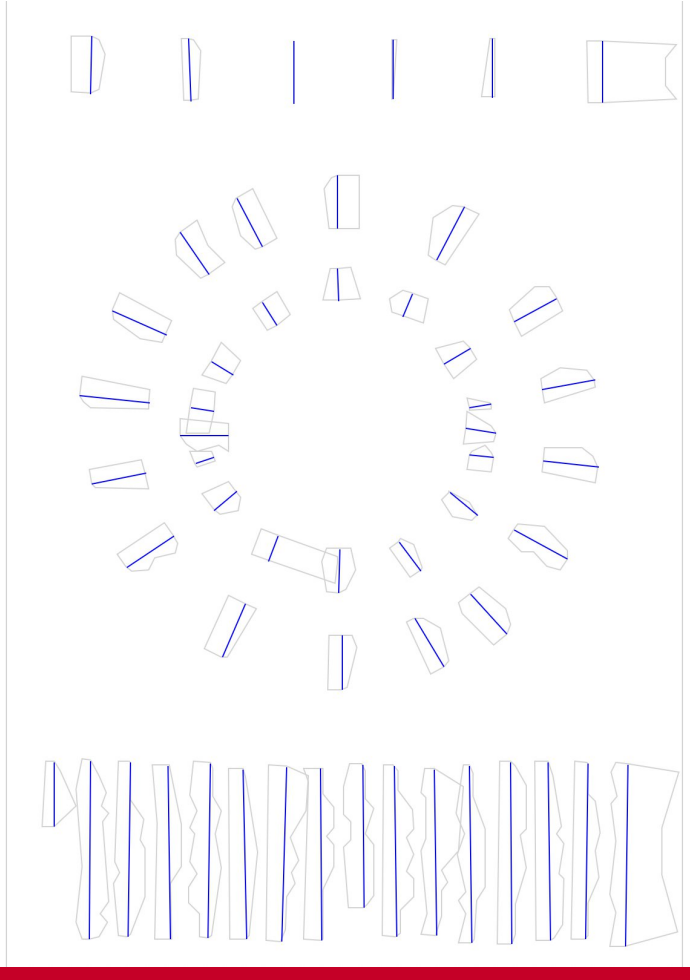
https://portail.biblissima.fr/

Importer des modèles (par ex. Zenodo)



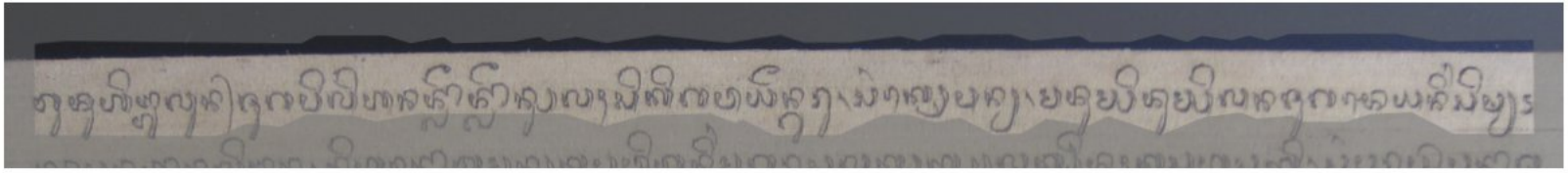
https://zenodo.org/communities/ocr_models/

Segmentation des zones et des lignes



Transcription (ou correction) à la main

Line #1



tuduhiṅ hulun·, juga pilihana dlə:-dlə:n salaḥ siki gavayən guru, saṁkṣepanya, madum pidum pilana juga deyaniṁ siṣya §

by marine.s (eScriptorium) on Thu May 27 2021 10:39:19 GMT+0200

-Toggle history

tuduhiṅg hulun , juga pilihana dlə:-dlə:n salaḥ siki gavayən guru, saṁkṣepanya, madum pidum pilana juga deyaniṁ siṣya §	marine.s (eScriptorium)
tuduhiṅgulun , juga saṁkṣepanya, madum pidum	marine.s (eScriptorium)

Prédiction (transcription) automatique

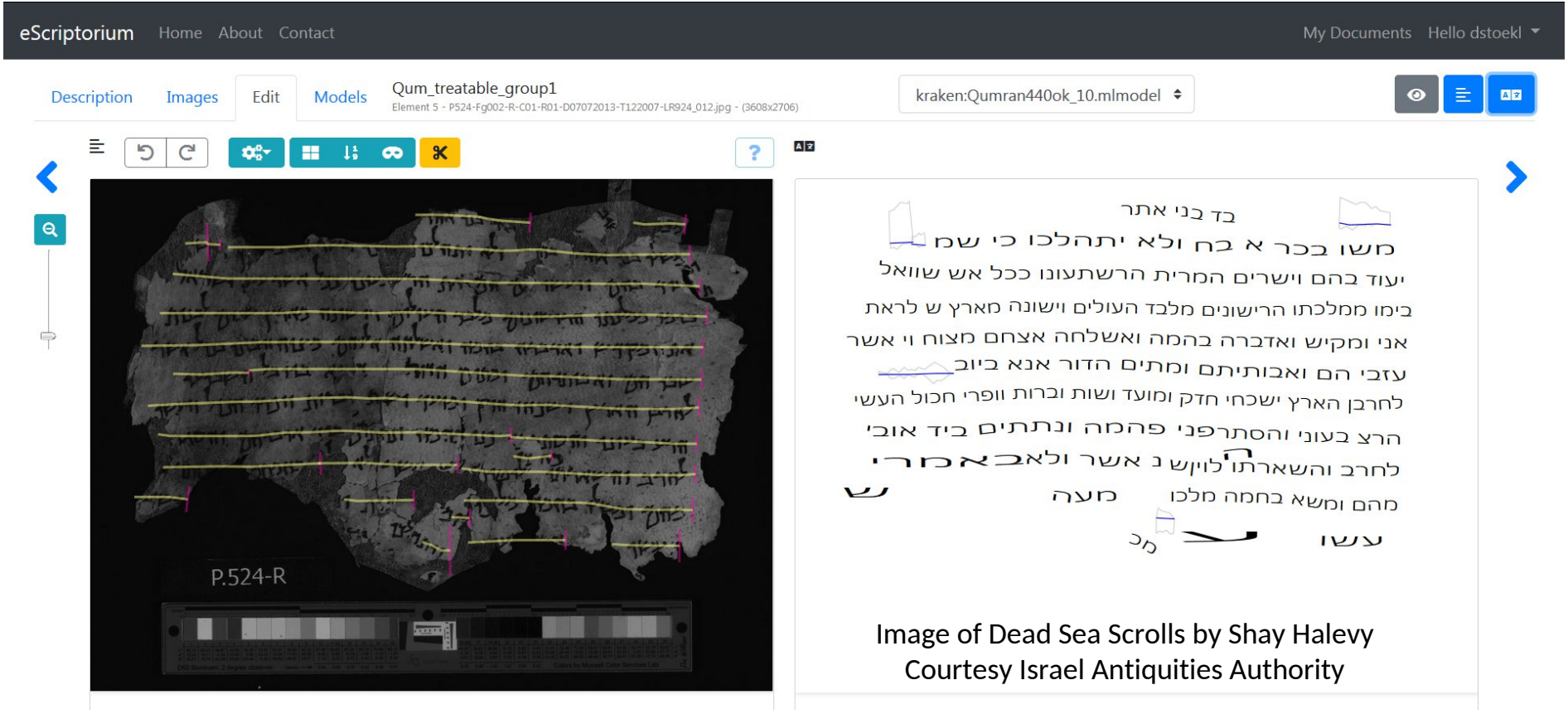
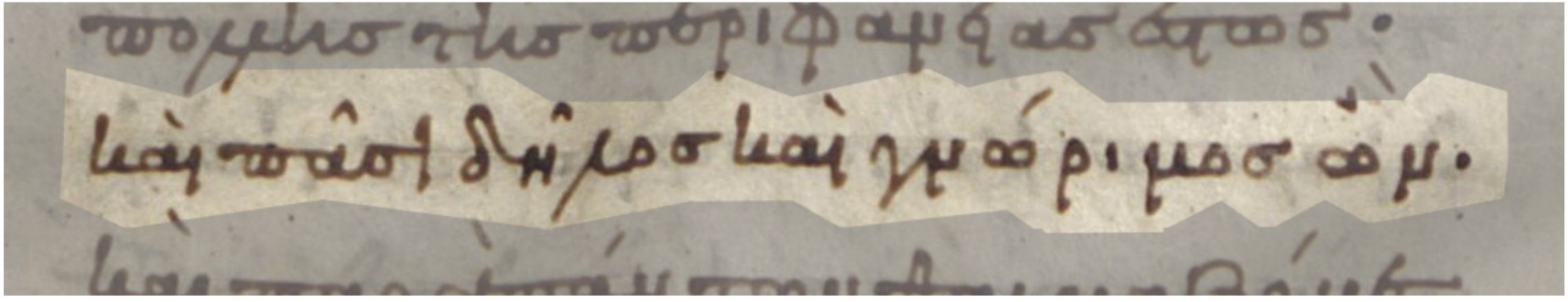


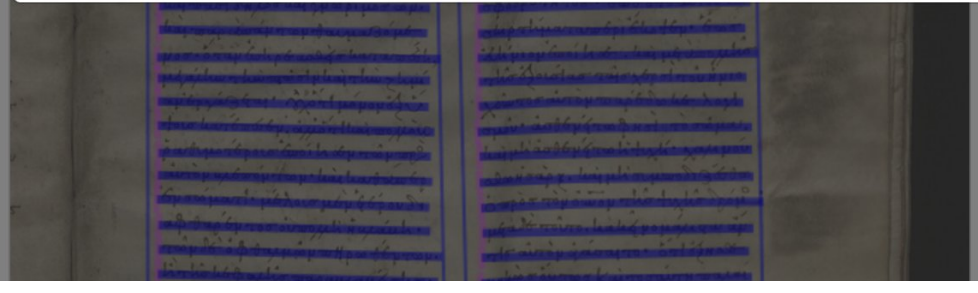
Image of Dead Sea Scrolls by Shay Halevy
Courtesy Israel Antiquities Authority

Line #7



καὶ πᾶσι δῆλος καὶ γνώριμος ὢν·

by pstokes (kraken:GreekMin_5mss_humanum_56) on Tue Jan 16 2024 10:11:57 GMT+0100



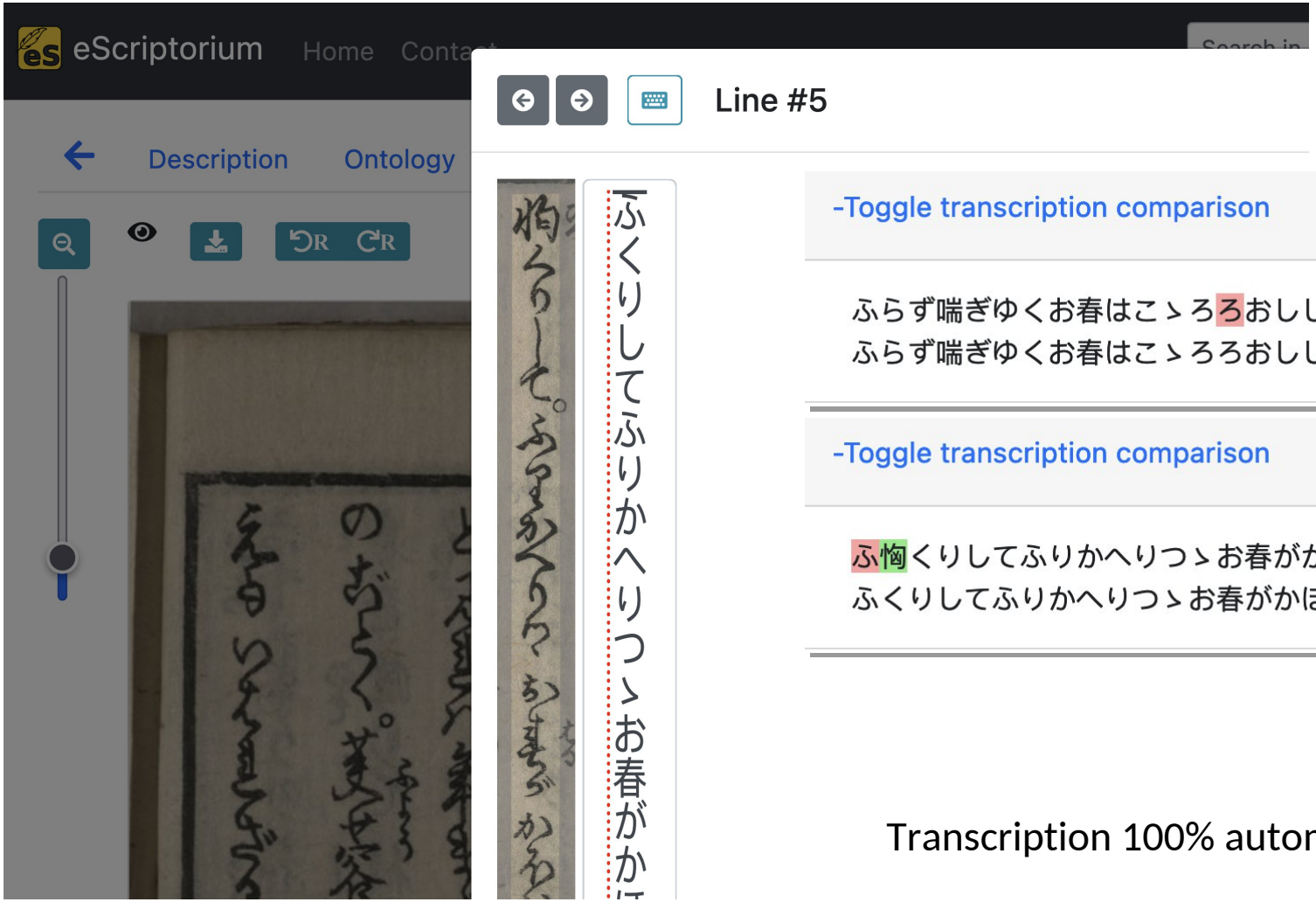
καὶ παρὰ πάντων θαυμαζό
νος· ὁ τὰν ἐπηρεασθεὶς κατα
μεγάλην τὴν πτῶσιν καὶ τὴ
αν ἐργάζεται· οὐχότι μόνον
ψους κατέπεσεν, ἀλλ' ὅτι κ
ῥαθυμοτέρους ἐποίησεν τῶ
αὐτὸν βλεπόντων· καὶ καθαπ
ἐν σώματι· μέλους μὲν ἐτέ
αφθαρέντος οὐ πολλῆ ἡ βλ
τῶν δὲ ὀφθαλμῶν πειρωθέν
ἢ τῆς κεφαλῆς παραβλαβε

σηιρτήματα περιέκοψεν· ἐ
εὐήνιον ἐποίησε. καὶ μετὰ πο
τῆς ἐξούσιας ταῖς χέρσι τοῦ ἡ
χοῦντος αὐτὸν παρέδωκε λ
σμου· ἀσθενείτω φησὶ τοσῶμ
καὶ μὴ ἀσθενείτω ἡ ψυχή·
σθωησὰρξ, καὶ μὴ συμποδιξέ
ὅπρος τὸν (οὐνον) τῆς ψυχῆ
μετα δὲ τοῦτο, κακεινομάλισ
τις αὐτὸν ἀγάσαιτο· ὅτι ἔξνσθ

Transcription 100% automatique. Merci à Annette von Stockhausen (BBAW)

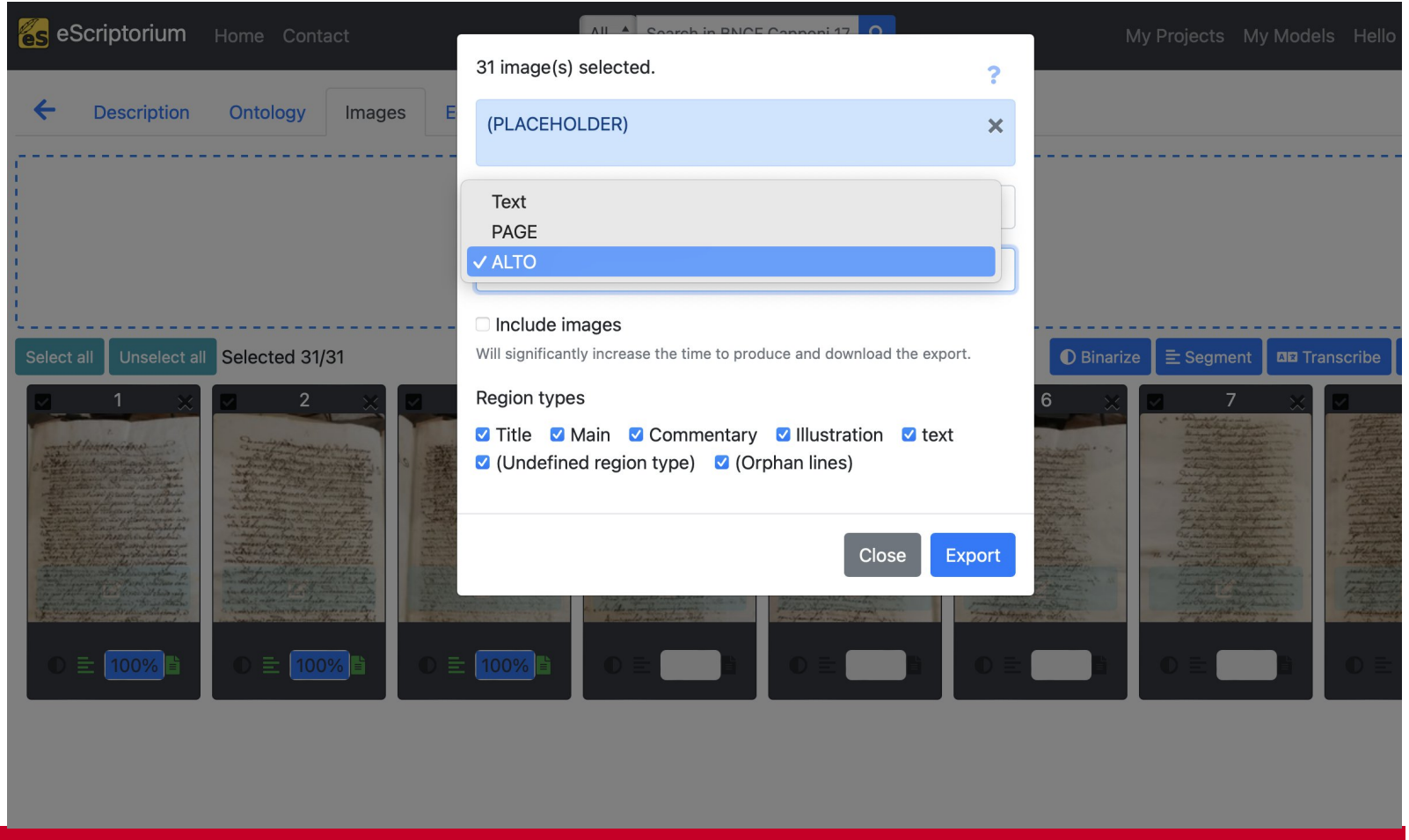
Line #26

by (eScriptorium) on Sat Jan 23 2021 23:44:51 GMT+0100

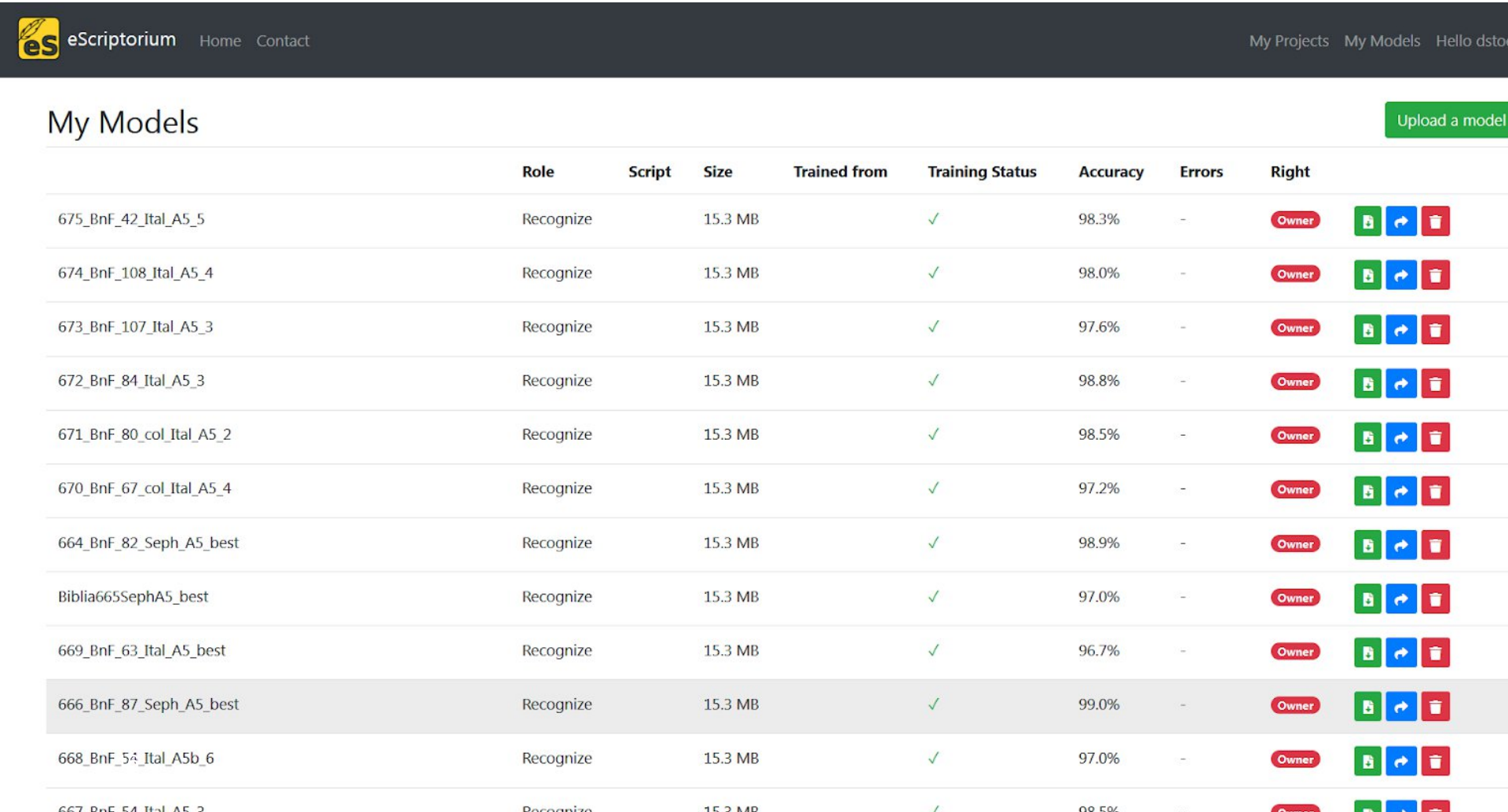


Transcription 100% automatique

Export des données (txt, ALTO, PAGE)



Export des modèles



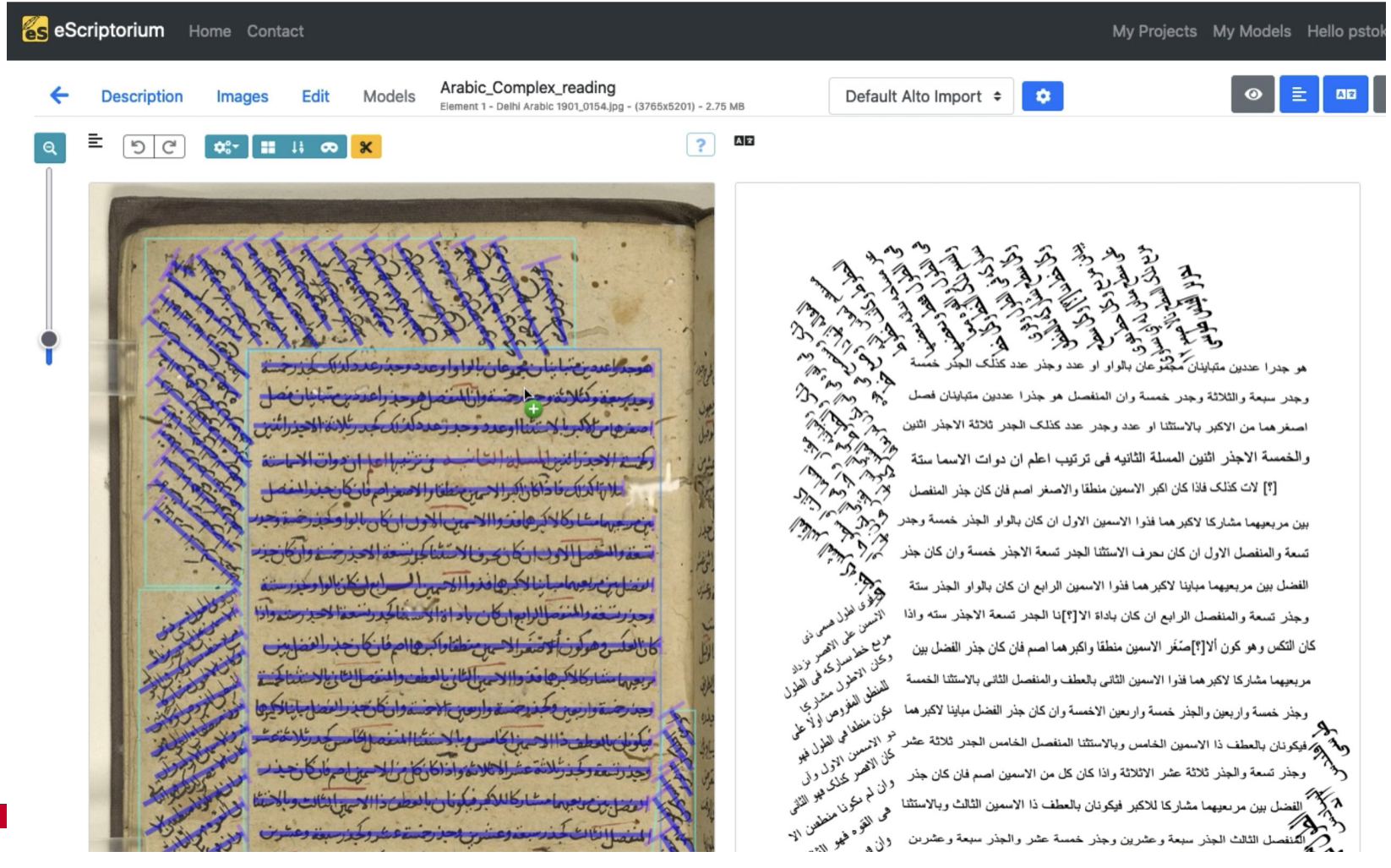
	Role	Script	Size	Trained from	Training Status	Accuracy	Errors	Right
675_BnF_42_Ital_A5_5	Recognize		15.3 MB		✓	98.3%	-	Owner
674_BnF_108_Ital_A5_4	Recognize		15.3 MB		✓	98.0%	-	Owner
673_BnF_107_Ital_A5_3	Recognize		15.3 MB		✓	97.6%	-	Owner
672_BnF_84_Ital_A5_3	Recognize		15.3 MB		✓	98.8%	-	Owner
671_BnF_80_col_Ital_A5_2	Recognize		15.3 MB		✓	98.5%	-	Owner
670_BnF_67_col_Ital_A5_4	Recognize		15.3 MB		✓	97.2%	-	Owner
664_BnF_82_Seph_A5_best	Recognize		15.3 MB		✓	98.9%	-	Owner
Biblia665SephA5_best	Recognize		15.3 MB		✓	97.0%	-	Owner
669_BnF_63_Ital_A5_best	Recognize		15.3 MB		✓	96.7%	-	Owner
666_BnF_87_Seph_A5_best	Recognize		15.3 MB		✓	99.0%	-	Owner
668_BnF_54_Ital_A5b_6	Recognize		15.3 MB		✓	97.0%	-	Owner

Défis d'écritures « rares » et diverses

- Différents types d'écriture, polices, blocs Unicode (peut-être !)
  - Alphabets, logogrammes, hiéroglyphes...
- Différentes directions
  - Gauche-droite, droite-gauche, haut-bas (puis D à G ou G à D), bas-haut, boustrophédon, diagonale, non linéaire, mixte, ...
  - Écrire sur la ligne de base, sur la ligne supérieure ou verticale, dans la grille, ...
- Différentes conventions pour la transcription et la présentation académique
- Un corpus souvent limité
  - Il y a peut-être peu d'exemples pour entrainer une machine
  - Souvent pas de modèles préexistants pour la langue, la mise en page …

Il nous faut des bons algorithmes, mais aussi une bonne ingénierie

Un exemple « facile » : la ligne d’écriture



Utilisation de kraken/eScriptorium

- kraken et eScriptorium sont entièrement gratuits et ouverts
  - Nous ne voulons pas que vous soyez enfermés dans nos serveurs.
  - Vous pouvez les télécharger et les utiliser dès maintenant.
- De manière réaliste, on a besoin d'accès aux GPU
  - Nous n'avons pas les ressources pour fournir un serveur public.
  - Il existe plus d'une douzaine d'installations différentes d'eScriptorium (et probablement beaucoup plus que nous ne connaissons pas !).
  - Comment gérer les besoins de l'IA de plus en plus gourmande et les utilisateurs qui veulent un système simple sur leur machine ?
  - Qui devrait héberger et soutenir le(s) serveur(s) ?

Il y a toujours un coût quelque part : ici, c'est le serveur

Un élément dans un flux de travail

- Segmentation via « YALTAI » (Thibault Clérice)
- Alignement texte-image avec PASSIM (OpenITI, …)
- Reconnaissance de l'écriture chinoise (Colin Brisson)
- Post-processing via modèles de langue (Thibault Clérice, …)
- Structuration semi-automatique de texte via régions, signes dans le texte (Simon Gabay et al., Daniel Stökl Ben Ezra et al.)
- Conversion en TEI via ALTO + XSLT ou API + python (Alix Chagué, Caroline Plumel, …)
- Publication via TEI Publisher (D. Stökl, …)
- …

D’autres possibilités pour l’ATR

- Analyse de mise en page à grande échelle
  - Variation par texte/date/lieu de production (décoration, no. Colones, lignes par page, lettres par ligne, …)
  - Est-ce que certaines parties d’un texte sont plus glosés que d’autres?
- Détection de la langue, du style d’écriture à grande échelle
  - Proportion de latin *versus* vernaculaire, d’un style d’écriture, …
- Identification automatique de texte (œuvre)
  - ATR+detection de la réutilisation de texte (PASSIM, TRACER, …)
- Alignement texte-image
  - Pour entrainer, mais aussi (par ex.) pour analyse paléographique, …
- *« Distant Reading »*, stylométrie, etc. (mais avec prudence !)
- Et *beaucoup* de plus*!*

La « Philosophie » d'eScriptorium

- Il n'est pas faisable (même possible ?) de répondre à tous
  - L'équipe d'eScriptorium ne peut pas tout faire
- Le plus important est de permettre la souplesse, la facilité d'importation et d'exportation, la facilité de créer vos propres outils externes.
- Il est aussi important de permettre et d'encourager le partage et la réutilisation
  - De logiciels, d'outils, de données, de modèles, de bonnes pratiques, ...
- Si nous travaillons ensemble, un petit effort peut faire beaucoup !

Pour aller plus loin…

- **escriptorium.readthedocs.io**
- **gitter.im/escripta/escriptorium**
- gitlab.inria.fr/scripta/escriptorium
- github.com/mittagessen/kraken
- zenodo.org/communities/ocr_models/
- ephenum.hypotheses.org/1412
- Voir aussi des publications de S.Gabay, A.Pinche, A. Chagué, ...
  - « From eScriptorium to TEI Publisher » hal.inria.fr/hal-03538115/
  - « Mutualisons la VT » hal.archives-ouvertes.fr/hal-03398740/
  - …
- peter.stokes@ephe.psl.eu

Remerciements

Ce projet a bénéficié de l'aide de l'Université PSL, de la Mellon Foundation, de l'État français gérée par l'Agence Nationale de la Recherche, et de l'Union Européenne, portant les références suivants (entre d'autres) :

- Université PSL programme IRIS : Scripta-PSL
- EU Horizon 2020 Research and Innovation Programme : Grant Agreement N° 871127 (RESILIENCE) et 101071829 (MIDRASH)
- Programme d'Investissements d'Avenir : ref. ANR-21-ESRE-0005 (Biblissima+)
- The Mellon Foundation (OpenITI)

PSL | SCRIPTA histoire et pratiques de l'écrit

DIM Sciences du texte et connaissances nouvelles — Région îledeFrance

aoroc

École Pratique des Hautes Études

RESILIENCE

Biblissima

INVESTIR L'AVENIR

Funded by the European Union

THE ANDREW W. MELLON FOUNDATION