



École Pratique
des Hautes Études

PSL 

 doroc

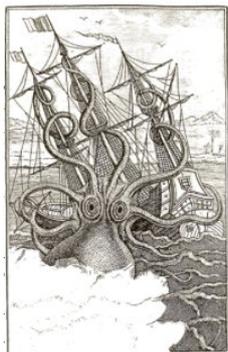
eScriptorium

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





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 [hOCR](#)
- Word bounding boxes and character bounding boxes
- Multi-script recognition support
- [Public repository](#) of model files
- [Lightweight model files](#)
- [Variable recognition network architecture](#)

Pull requests and code contributions are always welcome.

Installation

Kraken can be run on Linux or Mac OS X (both on-board [pip](#) utility and the [anaconda](#) scientific Python environment).

Installation using Pip

Useful Links

[The Kraken Website](#)
[kraken @ PyPI](#)
[kraken @ github](#)
[Issue Tracker](#)

Navigation

[Advanced Usage](#)
[Training](#)
[API Tutorial](#)
[API Reference](#)
[Models](#)

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.

<http://kraken.re>

<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
- Méroïtique
- 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 IIF (par ex. Biblissima)

The screenshot shows the Biblissima portal interface. At the top, there is a navigation bar with the logo and menu items: "PORTAIL BIBLISSIMA", "Explorer", "Rechercher", "Visualiser", and language options "FR | EN". Below the navigation bar is a search bar with the placeholder text "Rechercher une œuvre, une personne, un lieu, une cote, une enluminure...".

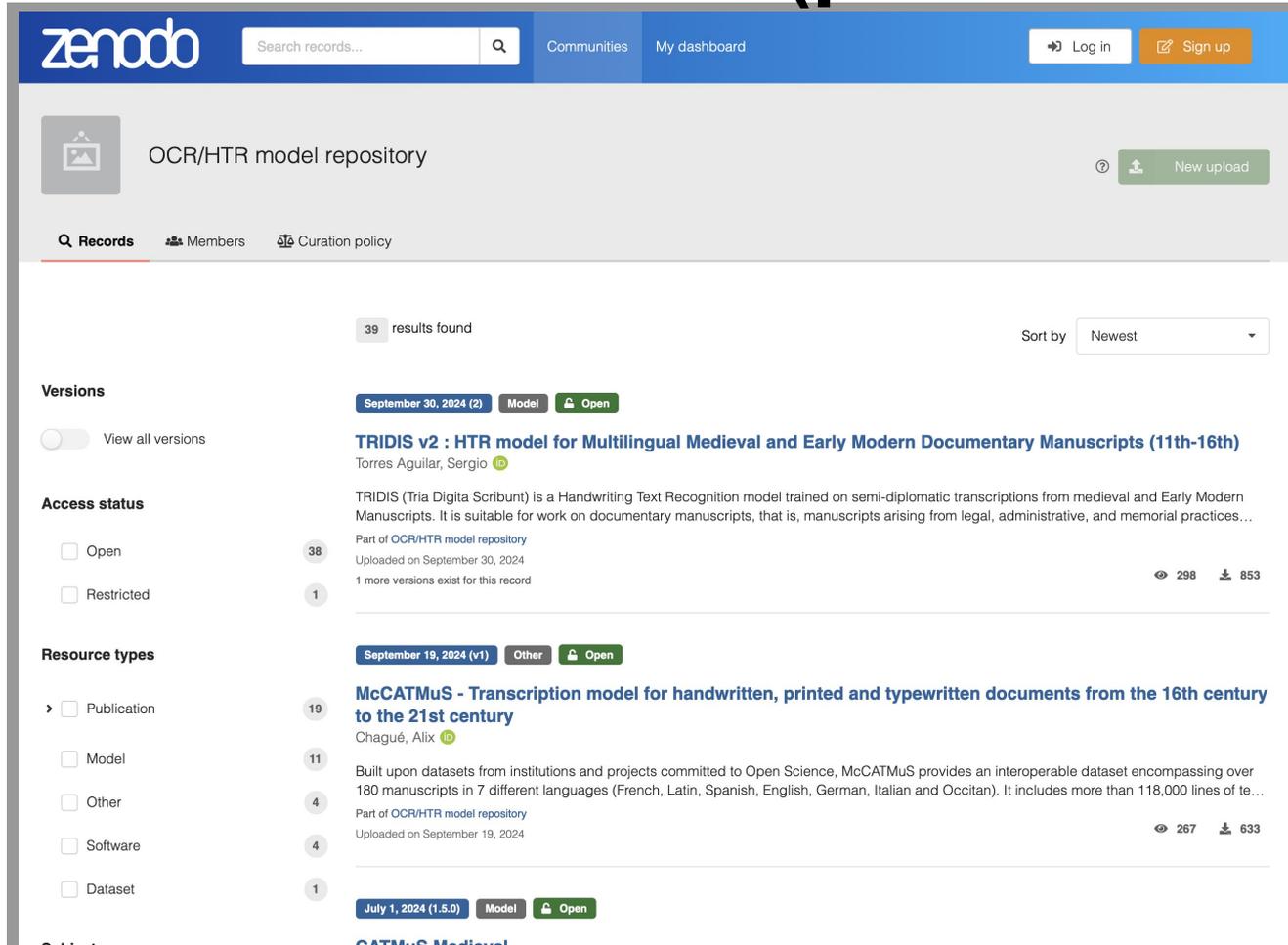
Below the search bar, there are filter buttons: "Supprimer tous les filtres", "Disponible sur le portail via IIF", and "Manuscrit".

The main content area is divided into two columns. The left column contains filter sections: "Filtres", "Rechercher dans ces résultats", "Type d'entité" (with "Manuscrit (80356)" selected), and "Date" (with "Inclure les résultats sans date" unchecked and a date range from 1 to 2100). The right column shows "80356 résultats" and a list of search results. The first result is "Abbaye Notre-Dame de Jouarre, Ms. 28" with the description "Œuvre : Bible. N.T. - Évangiles". Below this result, there is a small thumbnail image of a book cover and a red circle highlighting the IIF icon. The second result is "Abbeville. Bibliothèque municipale, Ms. 1". At the bottom right, there is a button labeled "Ma sélection" with a counter showing "0".

<https://portail.biblissima.fr/>



Importer des modèles (par ex. Zenodo)



The screenshot shows the Zenodo website interface for the 'OCR/HTR model repository'. The top navigation bar includes the Zenodo logo, a search bar, and links for 'Communities', 'My dashboard', 'Log in', and 'Sign up'. The repository title 'OCR/HTR model repository' is displayed with a 'New upload' button. Below the title, there are tabs for 'Records', 'Members', and 'Curation policy'. The main content area shows 39 results found, sorted by 'Newest'. On the left, there are filters for 'Versions', 'Access status', and 'Resource types'. The first result is 'TRIDIS v2 : HTR model for Multilingual Medieval and Early Modern Documentary Manuscripts (11th-16th)' by Torres Aguilar, Sergio, with 298 views and 853 downloads. The second result is 'McCATMuS - Transcription model for handwritten, printed and typewritten documents from the 16th century to the 21st century' by Chagué, Alix, with 267 views and 633 downloads.

zenodo Search records... Communities My dashboard Log in Sign up

OCR/HTR model repository New upload

Records Members Curation policy

39 results found Sort by Newest

Versions
 View all versions

Access status
 Open
 Restricted

Resource types
 Publication
 Model
 Other
 Software
 Dataset

September 30, 2024 (2) Model Open

TRIDIS v2 : HTR model for Multilingual Medieval and Early Modern Documentary Manuscripts (11th-16th)
Torres Aguilar, Sergio

TRIDIS (Tria Digita Scribunt) is a Handwriting Text Recognition model trained on semi-diplomatic transcriptions from medieval and Early Modern Manuscripts. It is suitable for work on documentary manuscripts, that is, manuscripts arising from legal, administrative, and memorial practices...
Part of OCR/HTR model repository
Uploaded on September 30, 2024
1 more versions exist for this record

298 853

September 19, 2024 (v1) Other Open

McCATMuS - Transcription model for handwritten, printed and typewritten documents from the 16th century to the 21st century
Chagué, Alix

Built upon datasets from institutions and projects committed to Open Science, McCATMuS provides an interoperable dataset encompassing over 180 manuscripts in 7 different languages (French, Latin, Spanish, English, German, Italian and Occitan). It includes more than 118,000 lines of te...
Part of OCR/HTR model repository
Uploaded on September 19, 2024

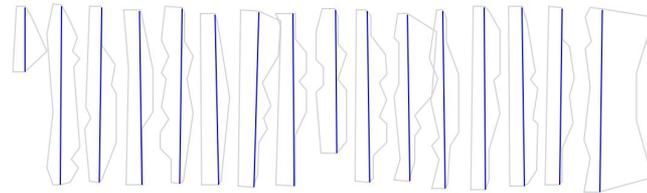
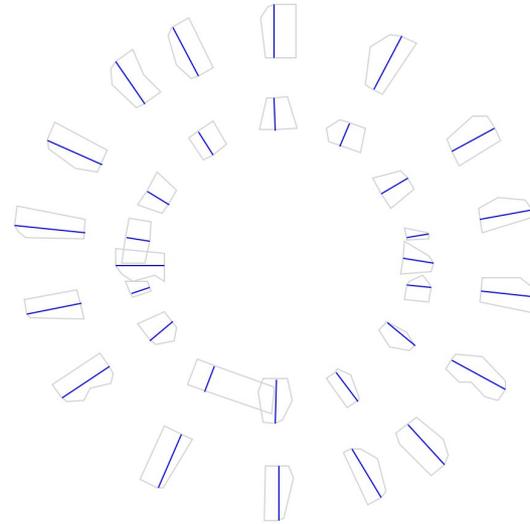
267 633

July 1, 2024 (1.5.0) Model Open

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

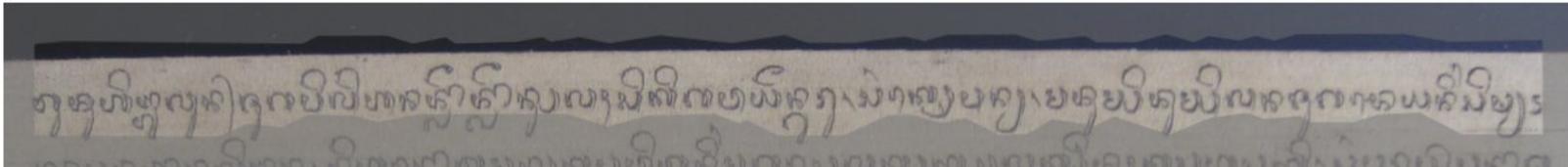


Segmentation des zones et des lignes



Transcription (ou correction) à la main

← → Line #1 ×



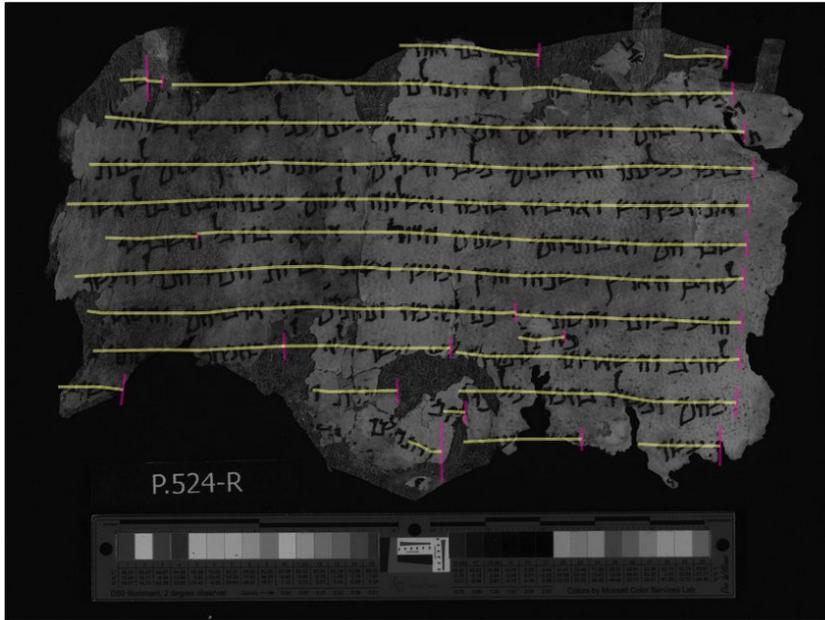
tuduhin hulun, juga pilihana dlə:-dlə:n salah siki gavayən guru, samkšepanya, madum pidum pilana juga deyanim sişya §

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

[-Toggle history](#) ?

tuduhin hulun, juga pilihana dlə:-dlə:n salah siki gavayən guru, samkšepanya, madum pidum pilana juga deyanim sişya §	marine.s (eScriptorium)	05/2
tuduhingulun , juga samkšepanya, madum pidum	marine.s (eScriptorium)	05/2

Prédiction (transcription) automatique



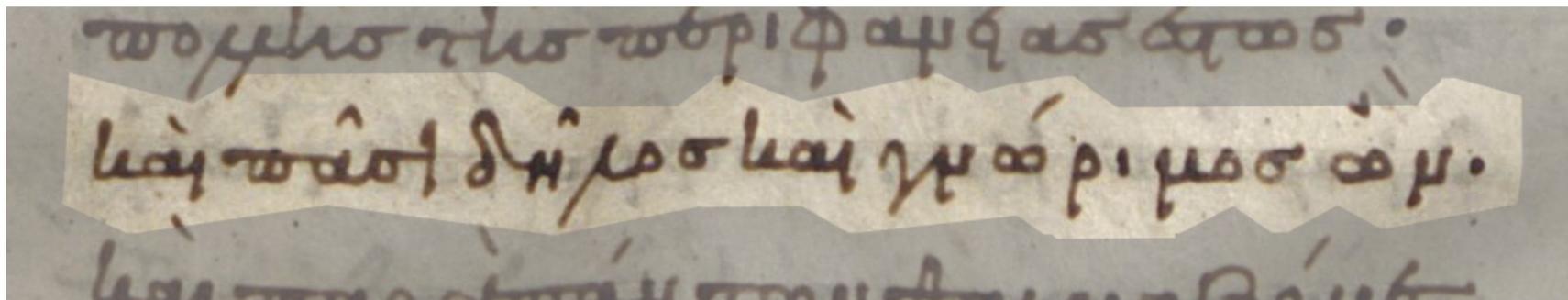
בד בני אתר
משו בכר א בח ולא יתהלכו כי שמ
יעוד בהם וישרים המרית הרשתעונו ככל אש שוואל
בימו ממלכתו הרישונים מלבד העולים וישונה מארץ ש לראת
אני ומקיש ואדברה בהמה ואשלחה אצחם מצוח וי אשר
עזבי הם ואבותיתם ומתים הדור אנא ביוכ
לחרבן הארץ ישכחי חזק ומועד ושות וברות וופרי חכול העשי
הרצ בעוני והסתרפני פהמה ונתתים ביד אובי
לחרב והשארתי לויקש נ אשר ולא באמרי
מהם ומשא בחמה מלכו מעה
עשו

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



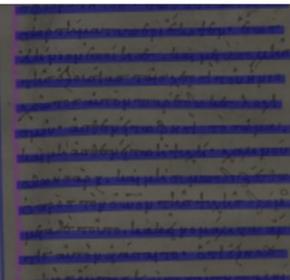
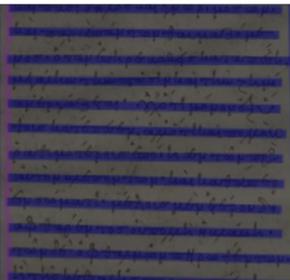


Line #7



καὶ πᾶσι δῆλος καὶ γνῶριμος ὤν.

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



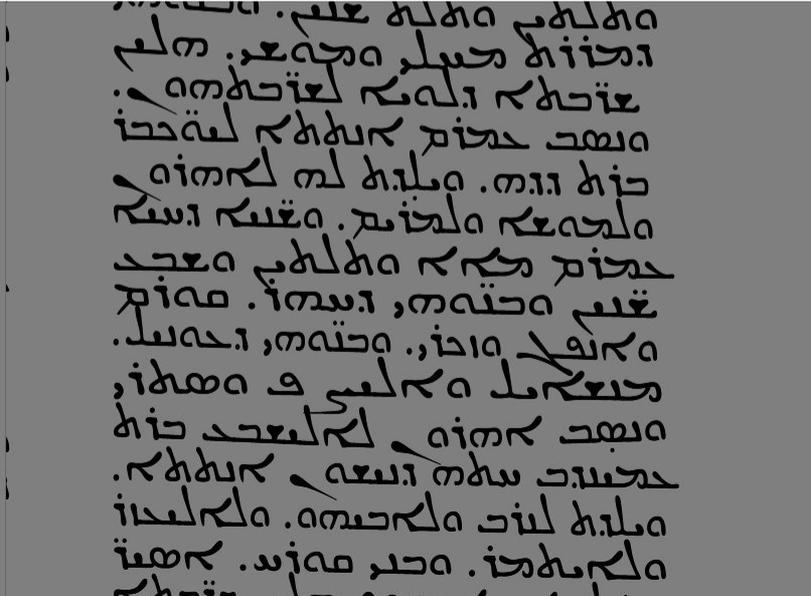
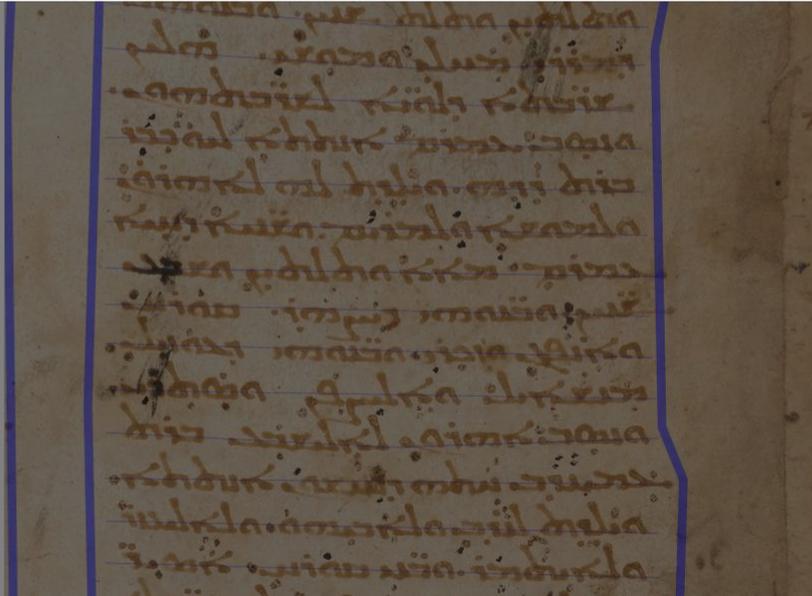
καὶ πᾶσι δῆλος καὶ γνῶριμος
καὶ παρὰ πάντων θαυματο-
νος· ὁ τὰν ἐπηρεασθεὶς κατα-
μεγάλην τὴν πτώσιν καὶ τῆ-
αν ἐργάζεται· οὐχότι μόνον
ψους κατέπεσεν, ἀλλ' ὅτι κ-
ραθυμοτέρους ἐποίησεν τῷ
αὐτὸν βλέπόντων· καὶ καθαπ-
έν σώματι· μέλους μὲν ἐτέ-
αφθαρέντος οὐ πολλῆ ἢ βλ-
τῶν δὲ ὀφθαλμῶν πειρωθέν
ἢ τῆς κεφαλῆς παραβλαβε-

σφοδρῆτος· εὐς ὅτι ἐπι-
σητήματα περιέκοψεν· ἐ-
εὐήνιον ἐποίησε. καὶ μετὰ τοῦ
τῆς ἐξούσιας ταῖς χέρσι τοῦ ἢ
χοῦντος αὐτὸν παρέδωκε λ-
σμου· ἀσθeneίτω φησὶ τοσῶμ
καὶ μὴ ἀσθeneίτω ἢ ψυχῆ·
σθωησάρξ, καὶ μὴ συμποδιξέ-
ἔπρος τὸν (οὔνον) τῆς ψυχῆ
μετα δὲ τοῦτο, κακεινομάλισ-
τις αὐτὸν ἀγάσαιτο· ὅτι ἐξ-
σθω οὔτως καὶ τοσαύτη παλα-

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



הַמִּשְׁבַּח אֲמִירָה לְלִמְעַבְדֵי כֹהֵן



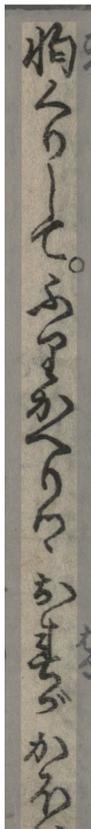
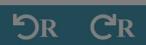


Line #5



Description

Ontology



ふくりしてふりかへりつゝお春がかほ

-Toggle transcription comparison

ふらず喘ぎゆくお春はこゝろおししづめ衝とかけ
ふらず喘ぎゆくお春はこゝろおししづめ衝とかけ

-Toggle transcription comparison

ふ愧くりしてふりかへりつゝお春がかほを朧月夜に侘
ふくりしてふりかへりつゝお春がかほを朧月夜に侘

Transcription 100% automatique



Export des données (txt, ALTO, PAGE)

The screenshot shows the eScriptorium web interface with a modal dialog open. The dialog is titled "31 image(s) selected." and contains the following elements:

- A search bar with the text "(PLACEHOLDER)".
- A list of export options: "Text", "PAGE", and "✓ ALTO" (which is highlighted in blue).
- An unchecked checkbox for "Include images" with the note "Will significantly increase the time to produce and download the export."
- A section for "Region types" with the following checked options: "Title", "Main", "Commentary", "Illustration", "text", "(Undefined region type)", and "(Orphan lines)".
- Buttons for "Close" and "Export".

The background interface shows a grid of document images, with the first two labeled "1" and "2", and a status bar at the bottom indicating "Selected 31/31".



Export des modèles



eScriptorium

[Home](#)

[Contact](#)

[My Projects](#)

[My Models](#)

[Hello dsto](#)

My Models

[Upload a model](#)

	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	  
667_BnF_54_Ital_A5_3	Recognize		15.3 MB		✓	98.5%	-	Owner	  

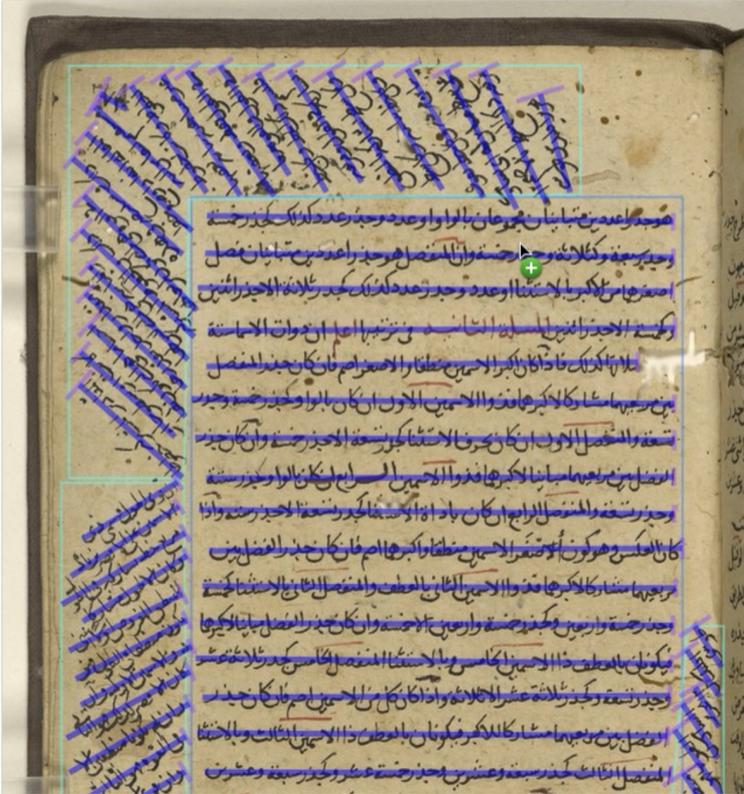
Défis d'écritures « rares » et diverses

- 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 entraîner 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



هو جذرا عددين متباينان مجموعان بالواو او عدد وجذر عدد كذلك الجذر خمسة
وجذر سبعة والثلاثة وجذر خمسة وان المنفصل هو جذرا عددين متباينان فصل
اصغرها من الاكبر بالاستتعا او عدد وجذر عدد كذلك الجذر ثلاثة الاجذر اثنين
والخمسة الاجذر اثنين المسئلة الثانيه في ترتيب اعلم ان دوات الاسماء ستة
[?] لات كذلك فاذا اكبر الاسمين منطلقا والاصغر اصم فان كان جذر المنفصل
بين مربعيهما مشاركا لاكبرهما فذوا الاسمين الاول ان كان بالواو الجذر خمسة وجذر
تسعة والمنفصل الاول ان كان بحرف الاستتعا الجذر تسعة الاجذر خمسة وان كان جذر
الفضل بين مربعيهما مباينا لاكبرهما فذوا الاسمين الرابع ان كان بالواو الجذر ستة
وجذر تسعة والمنفصل الرابع ان كان باداء الا[?]نا الجذر تسعة الاجذر ستة واذا
كان التمس وهو كون [?]صغر الاسمين منطلقا واكبرهما اصم فان كان جذر الفضل بين
مربعيهما مشاركا لاكبرهما فذوا الاسمين الثاني بالعطف والمنفصل الثاني بالاستتعا الخمسة
وجذر خمسة واربعين والجذر خمسة واربعين الاخمسة وان كان جذر الفضل مباينا لاكبرهما
فيكونان بالعطف ذا الاسمين الخامس وبالاتتعا المنفصل الخامس الجذر ثلاثة عشر
وجذر تسعة والجذر ثلاثة عشر الاثلاثة واذا كان كل من الاسمين اصم فان كان جذر
الفضل بين مربعيهما مشاركا لاكبر فيكونان بالعطف ذا الاسمين الثالث وبالاتتعا
المنفصل الثالث الجذر سبعة وعشرين وجذر خمسة عشر والجذر سبعة وعشرين
وان ه



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+détection 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](https://github.com/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)



École Pratique des Hautes Études



Biblissima II



Funded by the European Union

