

# Création auto-supervisée de vérité de terrain pour l'entraînement de modèles de transcription automatique de documents anciens hébraïques.

**Présentation par :** Matthieu Freyder

**Encadrant :** Prof. Daniel Stoeckl Ben Ezra

**Date :** Lundi 24 juin 2024

Stage effectué du 12 février au 16 juin au sein du laboratoire Archéologie & Philologie d'Orient et d'Occident (AOROC )



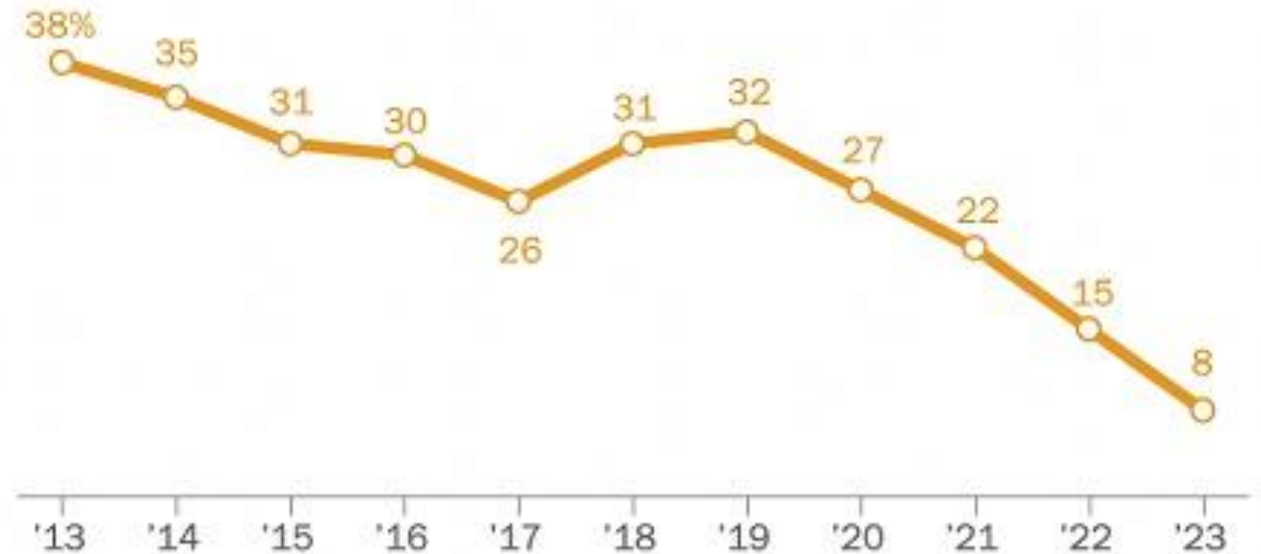
## Internet perd la mémoire...



Une mémoire fragile... mais vive !

## 38% of webpages from 2013 are no longer accessible

*% of links from each year that are no longer accessible as of October 2023*



Source: Pew Research Center analysis of a random selection of URLs collected by the Common Crawl web repository (n=999,989) and checked using page and DNS response codes. Web pages defined as inaccessible if they returned a status code of 204, 400, 404, 410, 500, 501, 502, 503, 523 or did not return a valid status code.

"When Online Content Disappears"

**PEW RESEARCH CENTER**

# Le livre doit raviver ses souvenirs



Une mémoire robuste... mais endormie !

! La transcription des documents anciens: un enjeu majeur pour réactiver cette mémoire de l'humanité !

תפלה על אבעבועות  
ישראל גנף בנשת בני ישראל אל הקדש: גפ  
אתה הוא יי אלהינו שהקטירו אבותינו את קטרת  
הסמים בזמן שבית המקדש קיים כגאשר  
צוית אותם על יד נביאך ככתוב בתורתך: ויאמר יי  
אל משה קח לך סמים נטף ושחלת וחלבנה סמים  
ולבונה זכרה בד בבד יהיה: ועשית אותה קטרת  
רוקח מעשה רוקח ממלח טהור קדש: ושחקת ממנה  
הדק ונתתה ממנה לפני העדות באהל מועד אשר  
אועד לך שמה קדש קדשים תהיה לכם: והקטרת  
אשר תעשה במתכנתה לא תעשו לכם קדש תהיה  
לך ליי: ונאמר והקטיר עליו אהרן קטרת סמים בבקר  
בבקר בהטיבו את הנרות יקטירנה: ובחלות אהרן  
את הנרות בין הערבים יקטירנה קטרת תמיד לפני  
יי לדורותיכם: פטום הקטרת כיצד שלש מאות וששים  
ושמנה בנים היו בה שלש מאות וששים וחמשה  
כמנן ימות החמה מנה בכל יום ויום מחציתה בבקר  
ומחציתה בערב: ושלשה מנים יתרים שמהם היה  
מבנים כהן גדול מלא חפניו ביום הכפורים לפניכם  
ומחזירין אותם למכתשת מאתמול לקיים מצות דקה  
מן הדקה ואחד עשר סמנים היו בה ואלו הן: הצרי  
והצפורן: החלבנה: והלבונה משקל שבעים שבעים  
מנה: מור וקציפה: שבולת נרד: וכרכום משקל ששה  
עשר



# Si cette mémoire était ravivée...



accès universel, moteurs de recherche, analyse de texte, comparaison de texte, traduction automatique, etc.



meilleures connaissances de l'histoire des manuscrits et des auteurs



découvertes (ou redécouvertes) d'informations inédites



utilisation du big data pour des analyses statistiques et des corrélations inédites

# Le projet MiDRASH

## Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script

(Migrations des traditions textuelles et sribales via l'analyse computationnelle à grande échelle des manuscrits médiévaux en écriture hébraïque)

**Projet** : Projet de six ans en sciences humaines computationnelles sur les manuscrits hébraïques médiévaux. Lancement en 2023.

**Financement** : Subvention ERC Synergy de plus de 10 millions d'euros accordée par le Conseil européen de la recherche.

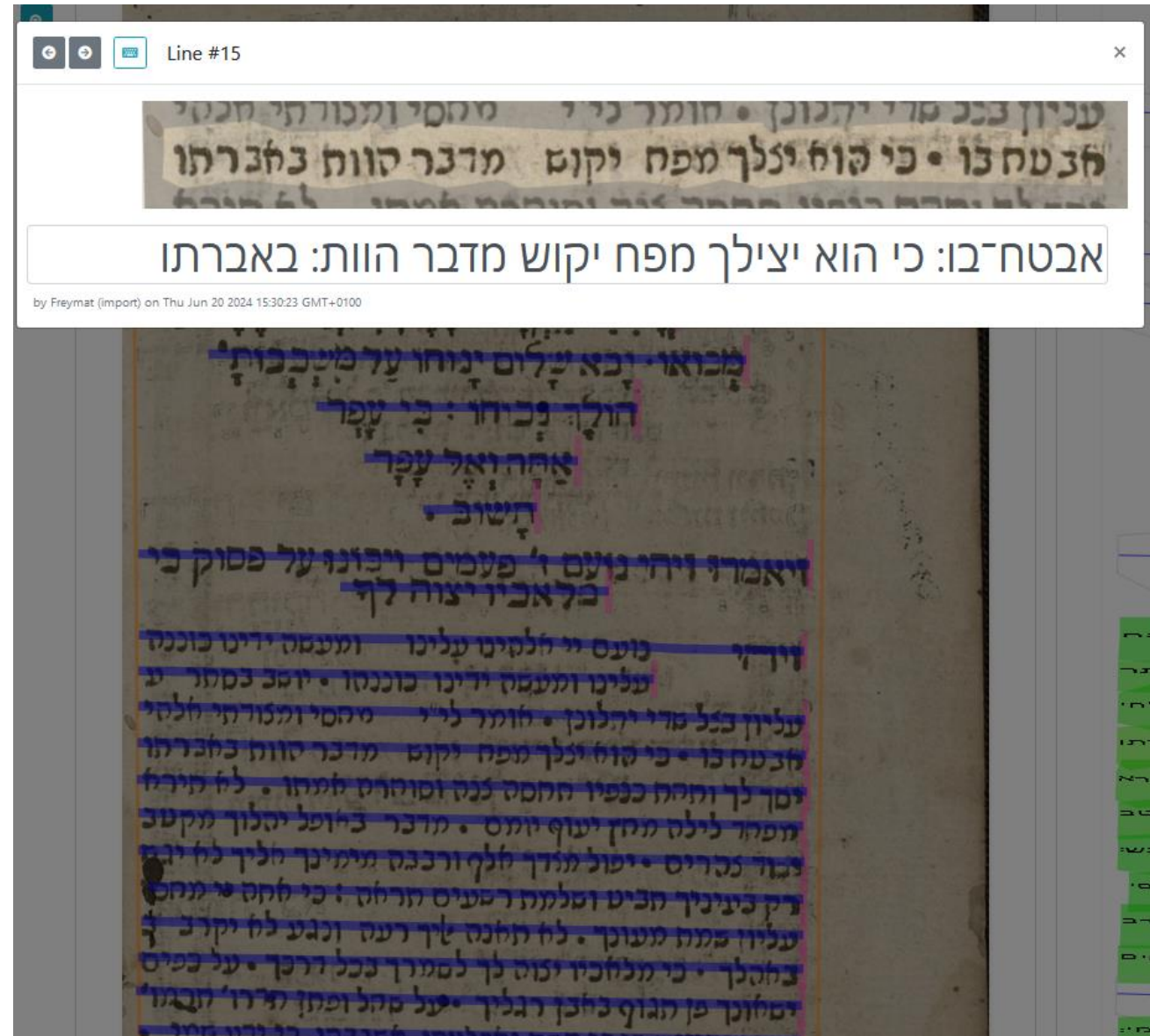
**Institutions** : EPHE et ses partenaires en Israël (Université de Tel-Aviv, Bar Ilan, Bibliothèque nationale d'Israël, Université de Haïfa).

✓ Todo:

- 🔄 Transcription automatique des manuscrits
- 🔄 Analyse intertextuelle
- 🔄 Analyse paléographique
- 🔄 Analyse linguistique
- 🔄 Études de cas
- 🔄 Publier les résultats en open source

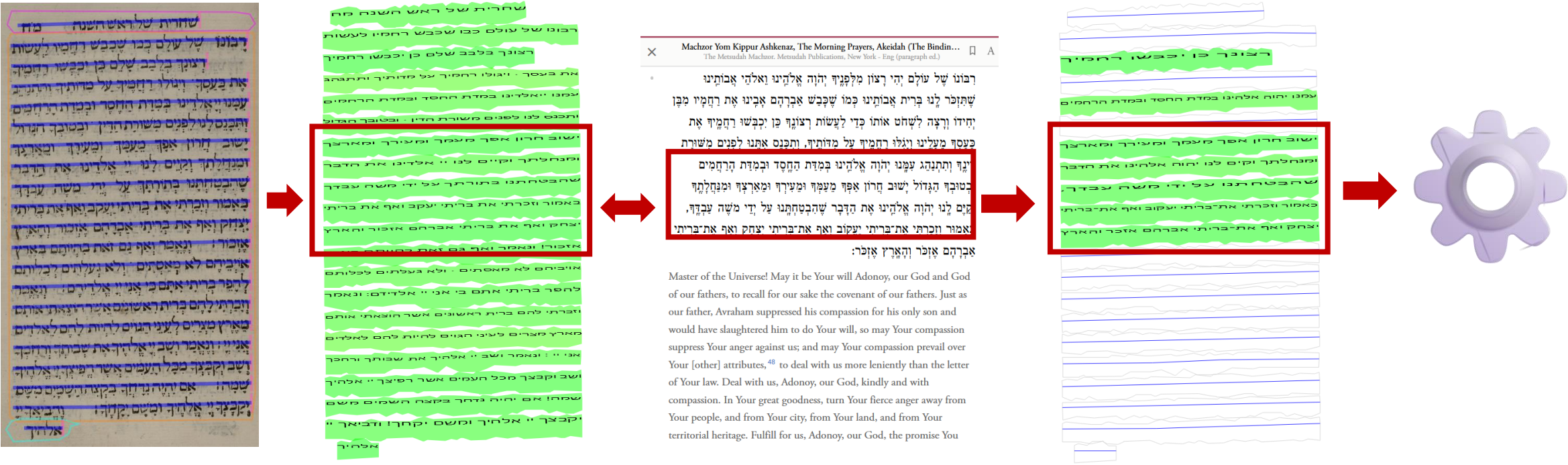
# Ma mission

- Le projet MiDRASH souhaite **transcrire** des manuscrits hébraïques médiévaux en texte interrogeable.
- Outils: des **modèles de transcription** qui nécessitent d'être **entraînés**.
- **Pour entraîner** un modèle de transcription, il est nécessaire de disposer d'un corpus de **données annotées**. L'annotation manuelle de ces données **est une tâche longue, fastidieuse et coûteuse**.
- Ma mission consiste à **développer un outil d'annotation auto-supervisée**, afin de **créer automatiquement des corpus de données annotées** pour l'entraînement des modèles de transcription.





# Principe de l'annotation auto-supervisée



1. Transcription avec un modèle générique.  
Résultat « bruité »
2. Recherche de correspondances avec un témoin numérique (ici sefaria.org)
3. Si correspondance, alors sélection du passage
4. Pour entrainer un meilleur modèle de transcription.

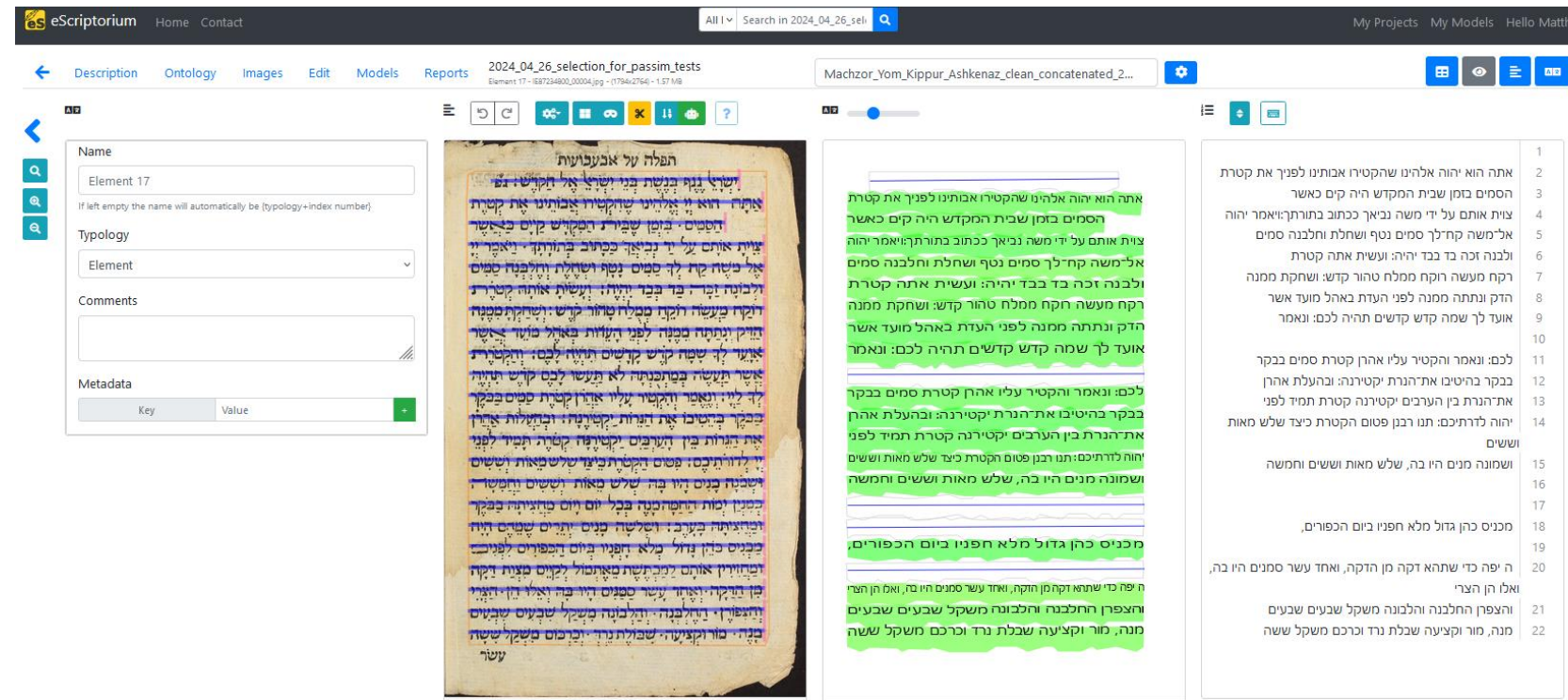
# Les outils: 1. eScriptorium

**eScriptorium** : Une plateforme avec interface visuelle **open source**, pour l'analyse de documents historiques.

**Objectif** : Combiner outils informatiques et numériques pour la transcription et l'annotation des textes.

## Fonctionnalités :

- Gestion de projets et utilisateurs
- Création et gestion de documents et métadonnées
- Importation et exportation d'images et de textes
- Segmentation et transcription
- Entraînement de modèles de segmentation et de transcription

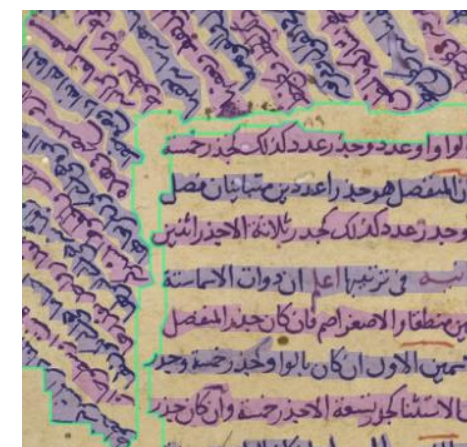
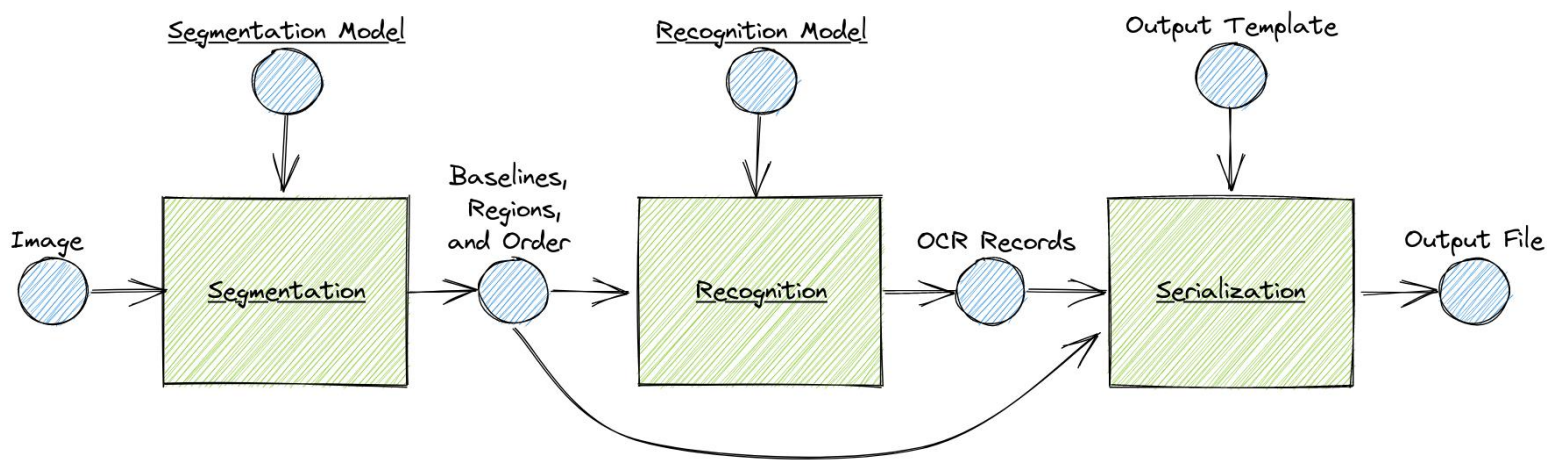




# Les outils: 2. Kraken

**Kraken: moteur d'HTR/OCR open source** utilisé par eScriptorium, optimisé pour les **documents historiques** et les **scripts non latins**.

- **Segmentation**: détection des régions dans la page, détection des lignes de base, des polygones
- **Transcription**: conversion des caractères en texte brut



# Les outils: 3. Passim

- **Passim** : Outil open source de détection automatique de **réutilisation de texte**.
- **Fonctionnement** : Comparaison de textes par fenêtre de caractères (n-grams) et alignements raffinés par l'algorithme de Smith-Waterman.
- **Utilisations** :
  - Détection de rééditions multiples
  - Identification de citations
  - Étude de la viralité des textes
  - Filtrage des documents dupliqués

```
{
  "uid": 6446647668810100226,
  "id": "eSc_textblock_11cd7539_IE103402206_00014",
  "ref": "0",
  "series": "OCR",
  "text": "במספר שמות לגלגלמתמחולתולדתם למשפחתם לבית אבתם פקדיחולאלף וחמש מאות: פ וכבן לבני שמעון|  
ושלש מאות: פ כדחולפקדיהם למטה שמעון תשעה וחמשים אלףחול(עשרים שנה ומעלה כל יצא צבא: וכנחולכל--זכר מבן  
כה) פקדיהם למטה גדחולמבן עשרים שנה ומעלה כל יצא צבאחוללמשפחתם לבית אבתם במספר שמותחוללבני גד תולדתם  
חמשה וארבעים",
  "lines": [
    {
      "begin": 0,
      "text": "חולאלף וחמש מאות: פ וכבן לבני שמעון|",
      "wits": [
        {
          "ref": "1",
          "id": "MT_NoVoc_concatenated.txt",
          "series": "GT",
          "begin": 243870,
          "alg": "אלף וחמש מאות: ----- לבני שמעון|",
          "alg2": "חולאלף וחמש מאות: פ וכבן לבני שמעון|",
          "matches": 26,
          "text": "אלף וחמש מאות : לבני שמעון|"
        }
      ]
    }
  ]
}
```

# Les outils: 3. Passim

Exemple de résultats: comparaison entre OCR d'une page et des corpus de témoins numériques (.txt).

Bloc de texte (OCR) ←

Lignes du bloc (OCR) ←


Alignements trouvés pour une ligne ←

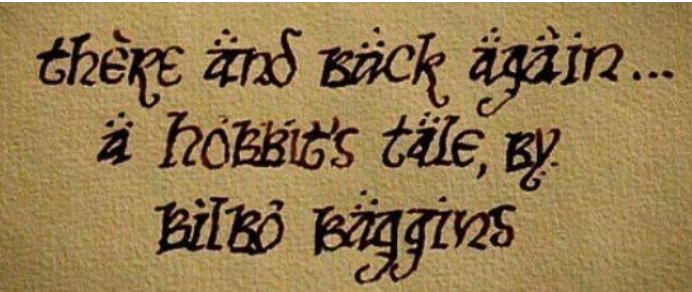
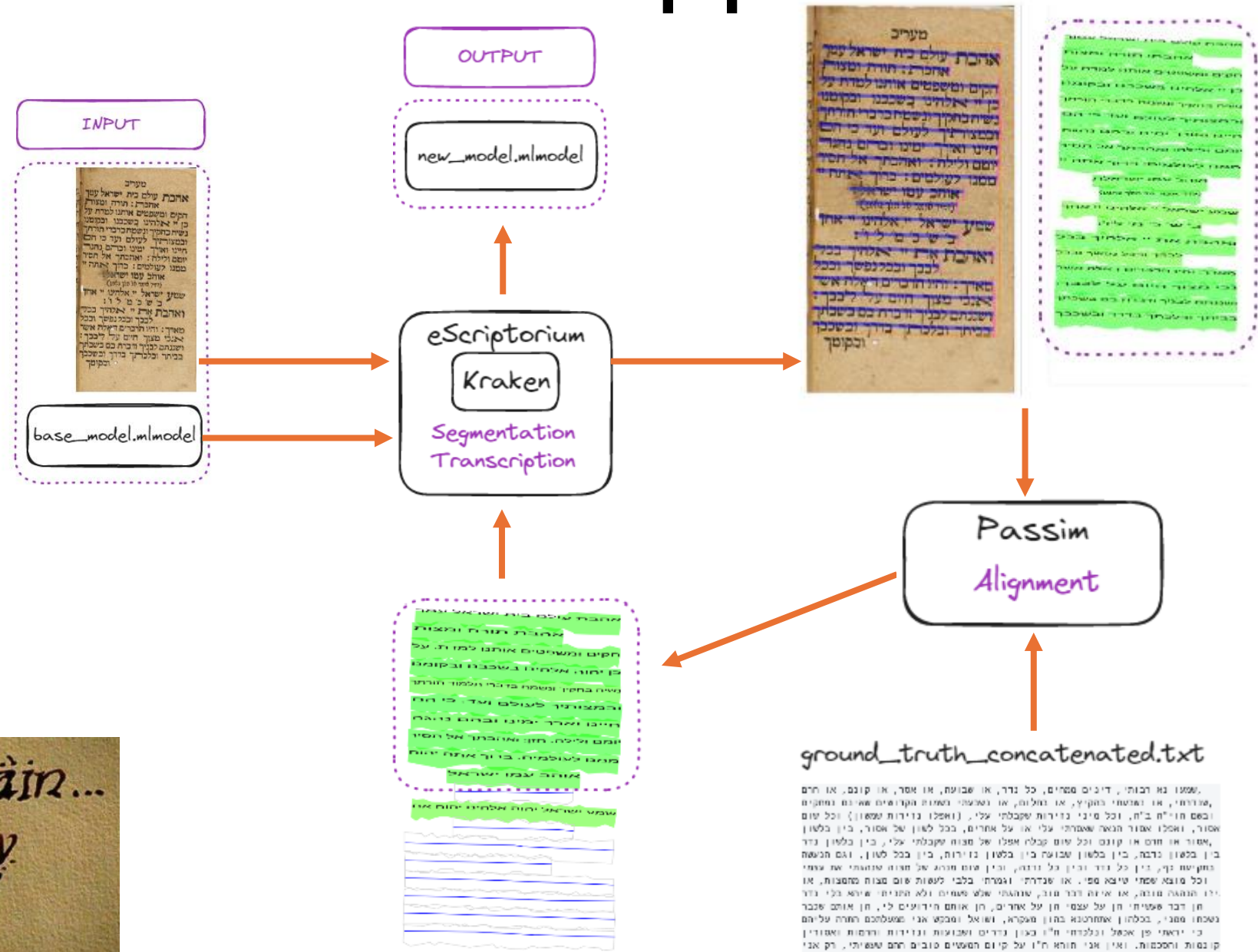
Texte historique détecté (GT) dans l'OCR ←

```
{
  "uid": 6446647668810100226,
  "id": "eSc_textblock_11cd7539_IE103402206_00014",
  "ref": "0",
  "series": "OCR",
  "text": "במספר שמות\תולדתם למשפחתם לבית אבתם פקדי\חאלף וחמש מאות: פ וכבן לבני שמעון|
פקדיהם למטה שמעון תשעה וחמיסח\עשרים שנה ומעלה כל יצא צבא: וכנח\לגלגלתם כל--זכר מבן
מבן עשרים שנה ומעלה כלח\למשפחתם לבית אבתם במספר שמות\ושלש מאות: פ כד לבני גד תולדתסח\אלף
כ"ה) פקדיהם למטה גד חמשה וארבעיסח\יצא צבא",
  "lines": [
    {
      "begin": 0,
      "text": "אלף וחמש מאות: פ וכבן לבני שמעון|",
      "wits": [
        {
          "ref": "1",
          "id": "MT_NoVoc_concatenated.txt",
          "series": "GT",
          "begin": 243870,
          "alg": "אלף וחמש מאות: ----- לבני שמעון|",
          "alg2": "אלף וחמש מאות: פ וכבן לבני שמעון|",
          "matches": 26,
          "text": "אלף וחמש מאות : לבני שמעון|"
        }
      ]
    }
  ]
},
```



# Principe de fonctionnement du pipeline TABA

-  **Rappel de notre mission:**
  - Produire de la **vérité de terrain**
  - en générant de façon **auto-supervisée** un corpus labélisé
  - pour l'**entraînement** de modèles de transcription.



# Collecter les textes numériques existants (GT)

**Sources de textes numériques :** Sefaria  
(<https://www.sefaria.org/>), récupération via API

**Extraction et concaténation des textes :** produire des textes concaténés (.txt)

**Création d'indexes :** Permet d'identifier les alignements entre les documents et les textes numériques.

**Code et fichiers disponibles sur GitHub:**  
<https://github.com/Freymat>



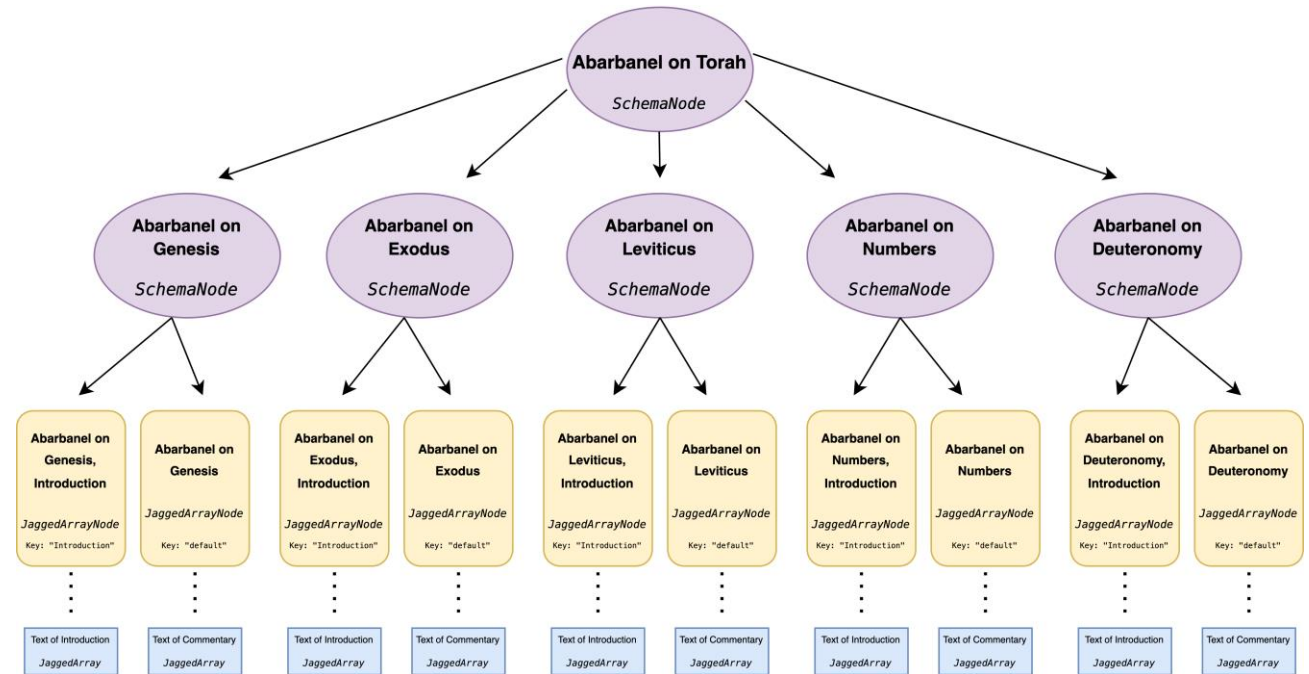
**Naviguer dans la structure des données, un défi !**

- Fichiers json

- **Objets JaggedArray:** listes de listes, qui contiennent le texte

- **Objets JaggedArrayNode:** dictionnaires contenant les métadonnées: profondeur du niveau de texte, niveau et nom de sections.

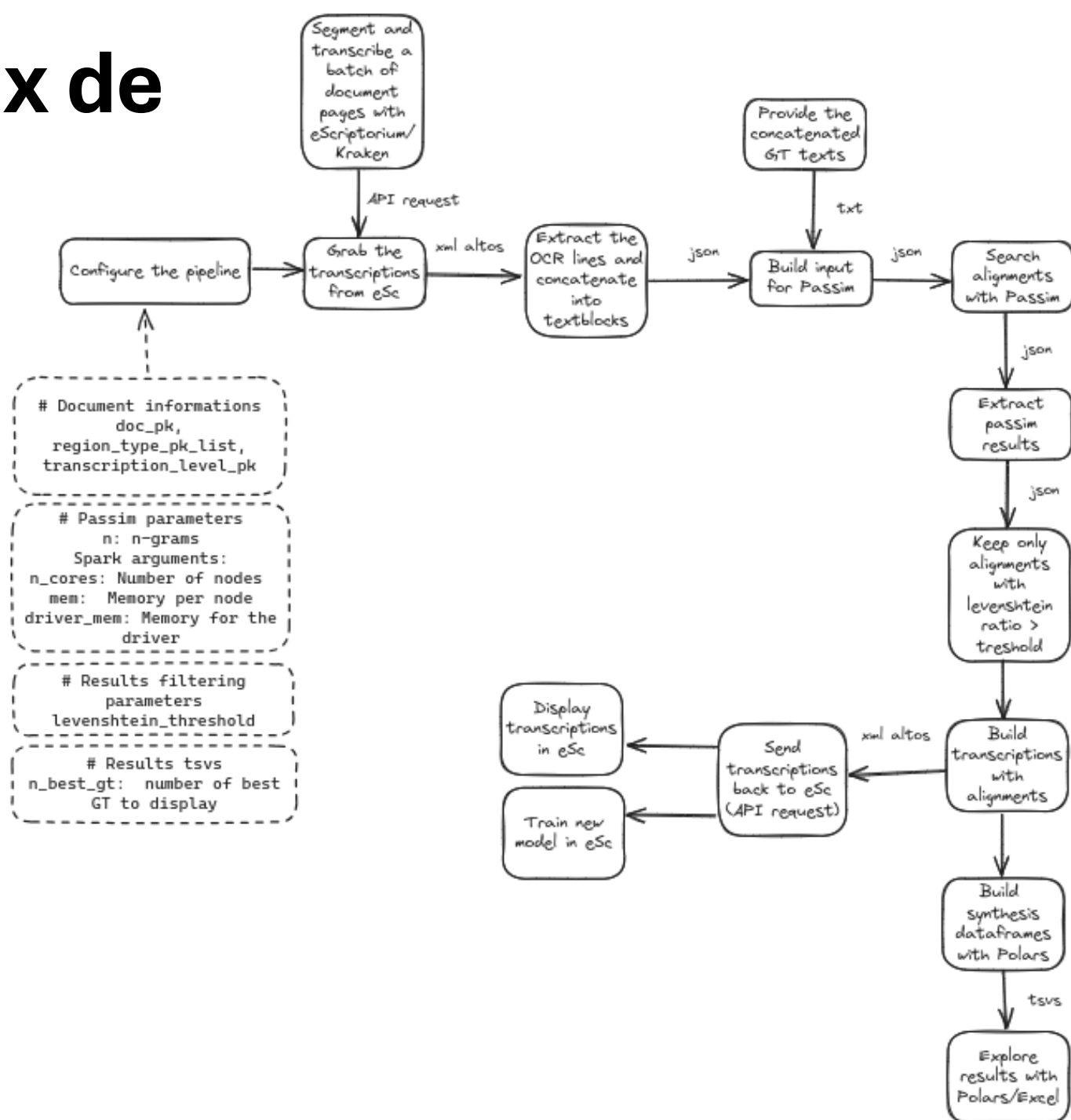
- Permet **de gérer des structures** de documents complexes, incomplets...



# Diagramme de flux de TABA



Code en python disponible sur:  
<https://github.com/Freymat>





# Sortie du pipeline

Dans eScriptorium: visualisation des alignements, entraînement des modèles.

Les données brutes sont stockées par le pipeline (structure des alignements, contenu textuel des lignes d'OCR et des alignements...)



```
{
  "filename": "IE87581919_00018.xml",
  "part_pk": 734771,
  "part_title": "Element 37",
  "levenshtein_threshold": 0.8,
  "total_aligned_lines_count": 10,
  "aligned_clusters_size": [
    4,
    1,
    1,
    1,
    3
  ],
  "GT_id": "Machzor_Yom_Kippur_Ashkenaz_clean_concatenated.txt"
}
```

# Utilisation de Polars pour l'analyse des résultats

Le pipeline produit des résultats de synthèse:

```
df = pl.read_csv('Results_based_on_total_aligned_lines_doc_pk_4381_n7_lev_0.8.tsv', separator='\t')
df.head()
```

shape: (5, 11)								
id	filename	doc_pk	part_pk	title	ocr_lines_in_part	Machzor_Rosh_Hashanah_Ashkenaz_clean_concatenated.txt	MT_NoVoc_concatenated.txt	Machzor
i64	str	i64	i64	str	i64	i64	i64	
1	"IE103336373_00043.xml"	4381	775261	"Element 65"	48	16	16	
2	"IE103402206_00010.xml"	4381	734735	"Element 1"	11	null	11	
3	"IE103402206_00014.xml"	4381	734736	"Element 2"	19	null	9	
4	"IE103409244_00025.xml"	4381	734737	"Element 3"	28	null	27	
5	"IE108110062_00015.xml"	4381	775262	"Element 66"	84	null	6	

2 tableaux : 1) le nombre total de lignes alignées pour chaque GT, et 2) la taille du plus gros groupe d'alignements successifs.

```
df = pl.read_csv('Top_5_GT_max_cluster_length_doc_pk_4381_n7_lev_0.8.tsv', separator='\t')
df.head()
```

shape: (5, 16)										
id	filename	doc_pk	part_pk	title	ocr_lines_in_part	Top1_GT_id	max_aligned_clusters_size_1	Top2_GT_id	max_aligned_clusters_size_2	
i64	str	i64	i64	str	i64	str	i64	str	i64	
1	"IE87744435_00039.xml"	4381	734794	"Element 60"	29	"Siddur_Ashkenaz_novoc_no_lbs_D..."	19	"Siddur_Ashkenaz_clean_concaten..."	18	
2	"IE36628376_00046.xml"	4381	775264	"Element 68"	40	"MT_NoVoc_concatenated.txt"	14	null	null	
3	"IE87234800_00004.xml"	4381	734751	"Element 17"	22	"Siddur_Ashkenaz_novoc_no_lbs_D..."	8	"Machzor_Yom_Kippur_Ashkenaz_cl..."	8	
4	"IE87733114_00008.xml"	4381	734791	"Element 57"	24	"Siddur_Ashkenaz_novoc_no_lbs_D..."	23	"Siddur_Ashkenaz_clean_concaten..."	23	
5	"IE87580382_00041.xml"	4381	734768	"Element 34"	22	"Siddur_Ashkenaz_novoc_no_lbs_D..."	13	"Siddur_Ashkenaz_clean_concaten..."	13	

2 tableaux des tops GT: 1) Top en terme de nombre de lignes alignées, 2) Top en terme de taille du plus gros groupe d'alignements successifs.

```
df = pl.read_csv('Overall_results_doc_pk_4381_n7_lev_0.8.tsv', separator='\t')
df.head()
```

shape: (5, 10)										
id	filename	doc_pk	part_pk	title	GT_id	ocr_lines_in_part	total_aligned_lines	aligned_lines_ratio	max_aligned_clusters_size	
i64	str	i64	i64	str	str	i64	i64	f64	i64	
1	"IE103336373_00043.xml"	4381	775261	"Element 65"	"Siddur_Ashkenaz_novoc_no_lbs_D..."	48	16	33.3	6	
2	"IE103336373_00043.xml"	4381	775261	"Element 65"	"Siddur_Ashkenaz_clean_concaten..."	48	16	33.3	6	
3	"IE103336373_00043.xml"	4381	775261	"Element 65"	"MT_NoVoc_concatenated.txt"	48	16	33.3	6	
4	"IE103336373_00043.xml"	4381	775261	"Element 65"	"Machzor_Rosh_Hashanah_Ashkenaz..."	48	16	33.3	4	
5	"IE103336373_00043.xml"	4381	775261	"Element 65"	"Machzor_Yom_Kippur_Ashkenaz_cl..."	48	16	33.3	4	

1 tableau général de synthèse recense pour chaque image et chaque GT le nombre total de lignes alignées, le ratio lignes alignées / nombre de lignes d'OCR, et la taille du plus gros groupe d'alignements successifs.

# Difficultés rencontrées

😞 'Read the doc, ... if there is one.'

💡 La solution: la communication avec les développeurs et la communauté d'utilisateurs.

😞 le passage à l'échelle (traitement de très grands lots d'images)

💡 La solution: l'optimisation du code.



# Passage à l'échelle

🧪 **Notre code a été testé sur différents lots de documents :**

📄 68 images et 5 textes numériques (tests du code et vérification des résultats)

📄 📄 1000 images et 5 textes numériques (dataset intermédiaire)

📄 📄 📄 45.000 images et 151 textes numériques (dataset de conditions réelles sur cluster de calcul)

## 💪 3 optimisations majeures

📄 Remplacement du parsing de fichiers XML (avec lxml) par l'utilisation d'expressions régulières (regex) pour reconstituer les xmls contenant les transcriptions d'alignement.  
=> 11s au lieu de 12h

🐼 Parallélisation avec concurrent.futures. Gains majeurs sur le module de reconstruction des xmls (pour chaque image, un fichier xml est créé avec les transcriptions alignées de chaque GT).

🦅 utilisation de Polars pour la création des dataframes de synthèse

```
1 Current date: 2024-06-12 17:45:20
2
3 doc_pk: None
4 Passim n-grams: 7
5 Spark parameters: n_cores=40, mem=90 GB, driver_mem=30 GB
6 Levenshtein ratio threshold: 0.7
7
8 Number of xml files processed (OCR): 45334
9 Number of txt files processed (Ground truth texts): 151
10
11 Step 2 (prepare OCR lines for Passim): 0:15:33.043257
12 Step 3 (Passim computation): 0:10:01.576732
13 Step 4 (xmls update with alignments from Passim): 0:34:33.870268
14 Step 5 (Tsv with results creation): 16:39:46.711925
```

```
1 Current date: 2024-06-16 08:31:11
2
3 doc_pk: False
4 Passim n-grams: 7
5 Spark parameters: n_cores=40, mem=90 GB, driver_mem=30 GB
6 Levenshtein ratio threshold: 0.7
7
8 Number of xml files processed (OCR): 45334
9 Number of txt files processed (Ground truth texts): 151
10
11 Step 2 (prepare OCR lines for Passim): 0:16:32.713447
12 Step 3 (Passim computation): 0:10:48.607443
13 Step 4 (xmls update with alignments from Passim): 0:43:42.903069
14 Step 5 (Tsv with results creation): 0:01:06.859488
```

Thank you my bear !



# Améliorations à apporter:



Améliorer la qualité du nettoyage des textes numériques de vérité de terrain (GT)



Ajouter de nouveaux GT, regrouper certains GT par famille de documents.



Améliorer la gestion des erreurs et des exceptions (TABA)

# Exemples de résultats :

Batch: TABA\_2024\_06\_17\_bav\_1col4\_n7\_results

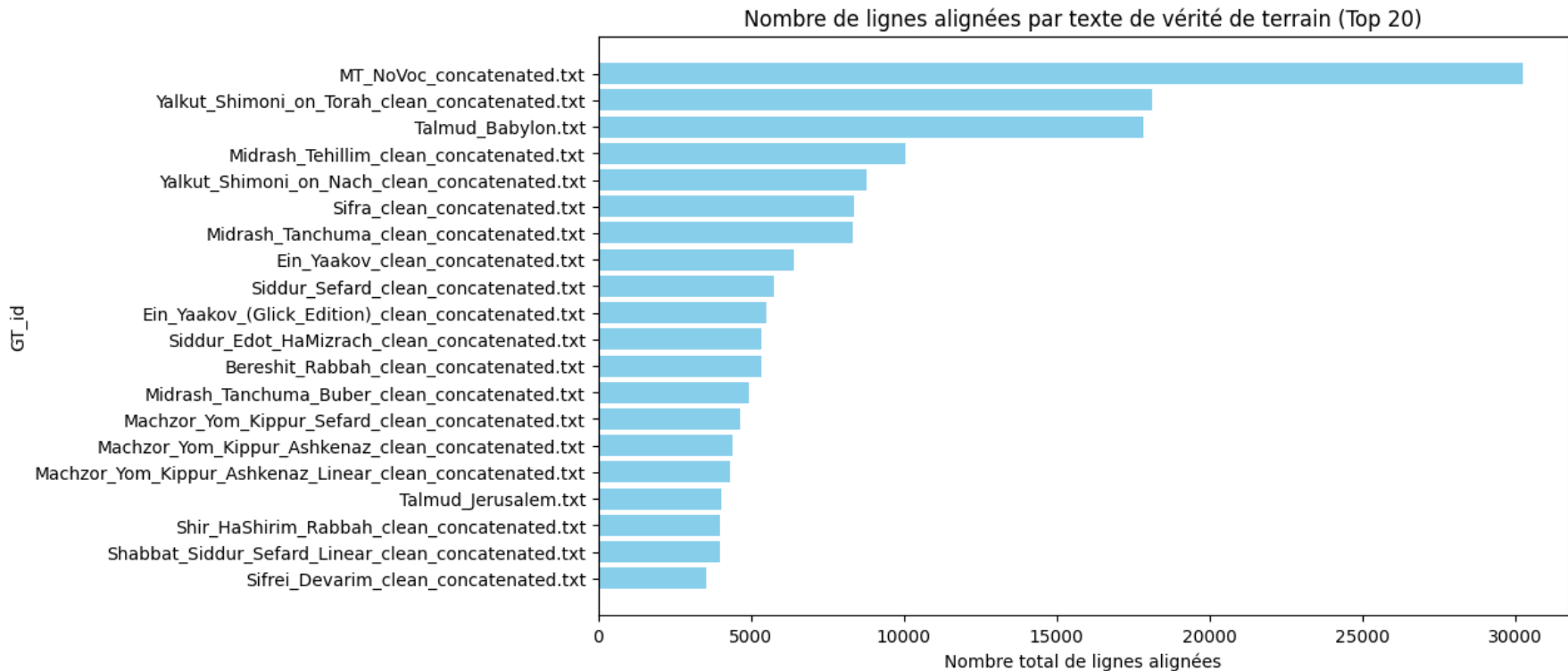
Nombre de fichiers uniques : 45332

Nombre de textes de vérité de terrain (GT) : 151

Nombre total de lignes OCR : 1.276.391

Nombre de fichiers avec alignements trouvés : 14.217

Pourcentage de fichiers avec alignements trouvés : 31,4 %





# La suite pour TABA, au sein du projet MiDRASH...

## A transcrire:

30.000 manuscrits

300.000 fragments

### **Des millions de page !**

En écriture hébraïque, mais en langues diverses (hébreux, araméen, judéo-arabe, mais aussi yiddish, ladino, etc.)

## Élaboration du protocole d'entraînement :

**Classification** des documents pour adapter les modèles de transcription.

**Critères** de classification envisagés :

- **Zone géographique** : ashkénaze, séfarade, italien, byzantin, oriental, yéménite, etc.
- **Type de script** : livresque, semi-cursif, cursif, etc.
- **Milieu de production** : chrétien, samaritain, etc.
- **Époque**.

## Résultats préliminaires :

- Utilisation de TABA sur des documents de la bibliothèque du Vatican (lots de 45.000 pages)
- Le catalogue **confirme les alignements identifiés par TABA**.
- **Identification** de documents dont le contenu n'était pas détaillé dans le catalogue, comme certains recueils (קובץ).

# Remerciements ! 🧡

**Professeur Daniel Stoekl Ben Ezra** pour sa confiance, le partage de connaissance et sa disponibilité.  
**L'équipe du laboratoire Archéologie & Philologie d'Orient et d'Occident** (EPHE - AOROC), pour leur expertise et leur collaboration.  
**L'Icam et la Région Grand Est** pour l'opportunité exceptionnelle de me former en Data.

Retrouvez:

**Le code du pipeline TABA :**

[https://github.com/Freymat/from\\_eScriptorium\\_to\\_Passim\\_and\\_back](https://github.com/Freymat/from_eScriptorium_to_Passim_and_back)



**Le code de constitution du corpus de textes numériques (GT)**

[https://github.com/Freymat/from\\_Sefaria\\_to\\_Passim](https://github.com/Freymat/from_Sefaria_to_Passim)



**Le texte de cette présentation:**

[https://github.com/Freymat/Soutenance\\_contenu/blob/main/soutenance\\_texte.md](https://github.com/Freymat/Soutenance_contenu/blob/main/soutenance_texte.md)

