La numérisation est-elle encore un problème ?

Réflexions autour de l'OCR du corpus des thèses de l'Université libre de Bruxelles

Journées d'études ***Reconnaissance automatique de caractères et extraction d'information spatiale à partir de sources anciennes numérisées***
10 décembre 2025 - Campus Condorcet, Aubervilliers

Matthieu Pichon

Chercheur postdoctoral (Université libre de Bruxelles - Ratio DH)
matthieu.pichon@ulb.be

Introduction

Une réflexion à trois voix, avec :

- **Sébastien de Valeriola**, chargé de cours, chaire Humanités numériques de l'ULB, responsable du projet.
- **Carmen Carrasco Luján**, chercheuse postdoctorale, projet « *De main de maître* » (transcription automatique des archives de l'historien Henri Pirenne)

Introduction

Une réflexion à trois voix, avec :

- **Sébastien de Valeriola**, chargé de cours, chaire Humanités numériques de l'ULB, responsable du projet.
- **Carmen Carrasco Luján**, chercheuse postdoctorale, projet « *De main de maître* » (transcription automatique des archives de l'historien Henri Pirenne)

→ Article en cours de rédaction, principalement sur l'OCR.

Introduction

Une réflexion à trois voix, avec :

- **Sébastien de Valeriola**, chargé de cours, chaire Humanités numériques de l'ULB, responsable du projet.
- **Carmen Carrasco Luján**, chercheuse postdoctorale, projet « *De main de maître* » (transcription automatique des archives de l'historien Henri Pirenne)

→ Article en cours de rédaction, principalement sur l'OCR. **Double contexte :**

Introduction

Une réflexion à trois voix, avec :

- **Sébastien de Valeriola**, chargé de cours, chaire Humanités numériques de l'ULB, responsable du projet.
- **Carmen Carrasco Luján**, chercheuse postdoctorale, projet « *De main de maître* » (transcription automatique des archives de l'historien Henri Pirenne)

→ Article en cours de rédaction, principalement sur l'OCR. **Double contexte :**

1. Analyses exploratoires dans le cadre du projet sur le corpus des thèses de l'ULB.
2. Plus généralement, plusieurs personnes qui travaillent sur de l'*Automatic Text Recognition* (ATR) + mise en place d'une plateforme dédiée en interne

Introduction

Point de départ : l'OCR n'est plus un problème ?

Introduction

Point de départ : l'OCR n'est plus un problème ?

1. OCR déjà ancienne $\rightarrow$ nombreux outils (Abbyy, Tesseract, etc.)

Introduction

Point de départ : l'OCR n'est plus un problème ?

1. OCR déjà ancienne → nombreux outils (Abbyy, Tesseract, etc.)
2. Qualité désormais suffisante pour la recherche

Introduction

Point de départ : l'OCR n'est plus un problème ?

1. OCR déjà ancienne → nombreux outils (Abbyy, Tesseract, etc.)
2. Qualité désormais suffisante pour la recherche
3. L'HTR est un plus grand défi

Introduction

Point de départ : l'OCR n'est plus un problème ?

1. OCR déjà ancienne → nombreux outils (Abbyy, Tesseract, etc.)
2. Qualité désormais suffisante pour la recherche
3. L'HTR est un plus grand défi
4. L'OCR (et peut-être l'HTR) est un problème réglé par le développement des LLM

Introduction

Point de départ : l'OCR n'est plus un problème ?

1. OCR déjà ancienne → nombreux outils (Abbyy, Tesseract, etc.)
2. Qualité désormais suffisante pour la recherche
3. L'HTR est un plus grand défi
4. L'OCR (et peut-être l'HTR) est un problème réglé par le développement des LLM

Notre avis → **c'est peut-être plus compliqué que ça**

Numériser des archives académiques

(OCR et corpus des thèses de l'ULB)

Numériser des archives académiques

Le projet

Projet « Déchiffrer la recherche à l'ULB », direction : Sébastien de Valeriola et Kenneth Bertrams (ULB)

Numériser des archives académiques

Le projet

Corpus ≈ 10 000 thèses

Numériser des archives académiques

Le projet

Corpus $\approx$ 10 000 thèses

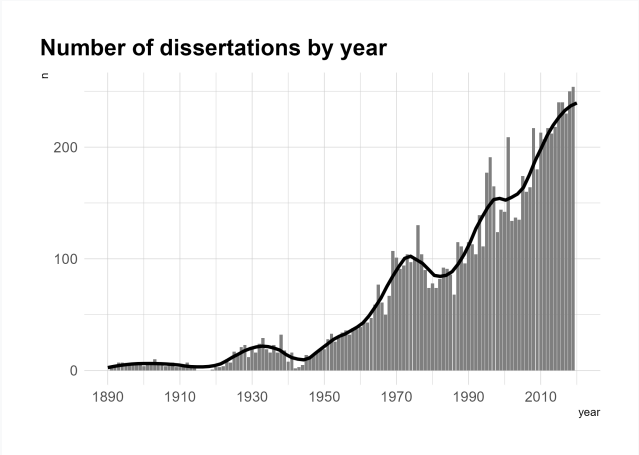
- 8000 numérisées + 2000(+) nativement numériques (à partir de 2009)

Numériser des archives académiques

Le projet

Corpus ≈ 10 000 thèses

- 8000 numérisées + 2000(+) nativement numériques (à partir de 2009)
- Depuis la création de l'ULB (mais surtout 1890) → années 2020

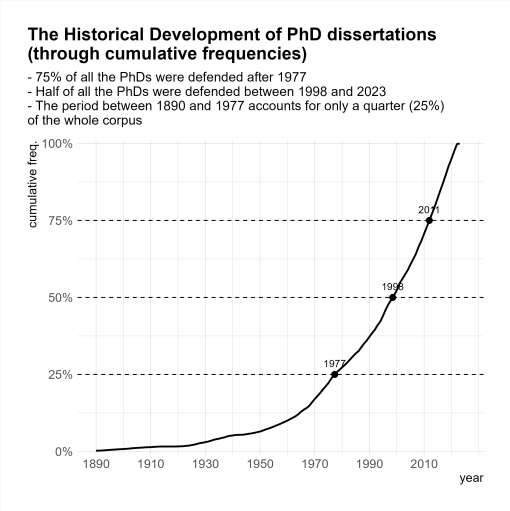


Numériser des archives académiques

Le projet

Corpus ≈ 10 000 thèses

- 8000 numérisées + 2000(+) nativement numériques (à partir de 2009)
- Depuis la création de l'ULB (mais surtout 1890) → années 2020

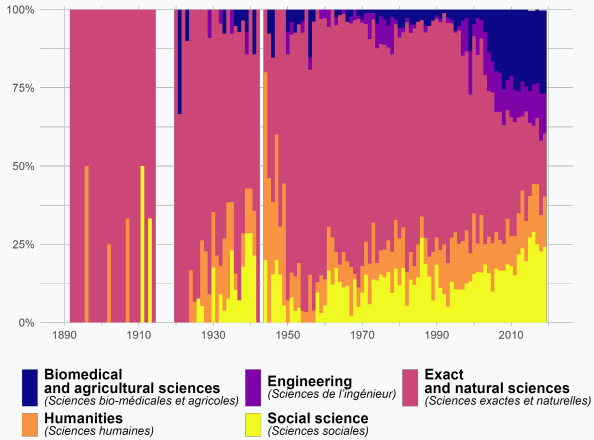


Numériser des archives académiques

Le projet

Corpus ≈ 10 000 thèses

- 8000 numérisées + 2000(+) nativement numériques (à partir de 2009)
- Depuis la création de l'ULB (mais surtout 1890) → années 2020
- Dans toutes les disciplines (de l'ULB)



Numériser des archives académiques

Le projet

Genèse du projet

- Années 2010 : le Département des bibliothèques et de l'information scientifique (DBIS) de l'ULB entrepend la numérisation des thèses archivées
- Service de numérisation en interne (scanners, équipes)
- Objectif : préserver et diffuser en ligne (https://difusion.ulb.ac.be/)

→ Projet très représentatif de la notion de *Collections as Data*

Numériser des archives académiques

Le projet

Numérisation en deux étapes :

1. Abbyy FineReader 8.0 (sorti en 2005)
2. Abbyy FineReader Engine 11 : objectif de réduction du poids des PDF pour la mise en ligne ; seulement une partie du corpus

→ Différents *outputs*, dont des **XML** ressemblant à du **ALTO** (*Analyzed Layout and Text Object*) ; but principal : PDF légers et cherchables

→ Une opération titanesque, mais avec plusieurs problèmes

Numériser des archives académiques

Problèmes

Des modèles d'OCR anciens et propriétaires

- Des gains de qualité depuis Abbyy 8 et 11 ?
- Licences expirées, modèles périmés : comment mettre à jour le corpus au fil des années ?
- Une boîte noire propriétaire

→ Dépendance et contrainte du propriétaire

Numériser des archives académiques

Problèmes

Des *outputs* hétérogènes : le jeu des *n* différences

```
−<document version="1.0" producer="FineReader 8.0" xsi:schemaLocation="http://
www.abbyy.com/FineReader_xml/FineReader8-schema-v2.xml http://www.abbyy.com/
FineReader_xml/FineReader8-schema-v2.xml" pagesCount="26"
mainLanguage="FrenchStandard" languages="FrenchStandard">
  <page width="2365" height="3177" resolution="300" originalCoords="true"> </
  page>
 −<page width="2335" height="3168" resolution="300" originalCoords="true">
  −<block blockType="Text" blockName="" isHidden="true" l="920" t="106" r="1349"
   b="180">
   −<region>
      <rect l="920" t="106" r="1349" b="180"/>
    </region>
   −<text>
    −<par align="Justified">
     −<line baseline="171" l="1316" t="125" r="1342" b="171">
      −<formatting lang="FrenchStandard">
         <charParams l="1316" t="125" r="1342" b="171" characterHeight="46"
         hasUncertainHeight="false" baseLine="0" wordStart="true"
         wordFromDictionary="false" wordNormal="false" wordNumeric="true"
         wordIdentifier="true" wordPenalty="0" meanStrokeWidth="22"
         charConfidence="85" serifProbability="255">I</charParams>
       </formatting>
      </line>
     </par>
    </text>
   </block>
```

Abbyy 8

```
−<document version="1.0" producer="ABBYY FineReader Engine 11" languages=""
xsi:schemaLocation="http://www.abbyy.com/FineReader_xml/FineReader10-schema-
v1.xml http://www.abbyy.com/FineReader_xml/FineReader10-schema-v1.xml">
 −<page width="2315" height="3257" resolution="300" originalCoords="1">
  −<block blockType="Text" blockName="" l="521" t="702" r="1701" b="1063">
   −<region>
      <rect l="1351" t="702" r="1700" b="703"/>
      <rect l="961" t="703" r="1700" b="704"/>
      <rect l="571" t="704" r="1700" b="705"/>
      <rect l="521" t="705" r="1700" b="1042"/>
      <rect l="521" t="1042" r="1701" b="1045"/>
      <rect l="522" t="1045" r="1701" b="1059"/>
      <rect l="522" t="1059" r="1700" b="1060"/>
      <rect l="522" t="1060" r="1354" b="1061"/>
      <rect l="522" t="1061" r="964" b="1062"/>
      <rect l="522" t="1062" r="574" b="1063"/>
    </region>
   −<text>
    −<par lineSpacing="1824">
     −<line baseline="748" l="527" t="708" r="1574" b="756">
        <formatting lang="FrenchStandard">Sur 1'entraînement des
        électrolytes</formatting>
      </line>
     −<line baseline="823" l="527" t="783" r="1694" b="831">
        <formatting lang="FrenchStandard">par les précipités. Influence de
        l'agi-</formatting>
      </line>
     −<line baseline="900" l="528" t="866" r="725" b="900">
        <formatting lang="FrenchStandard">tation .</formatting>
      </line>
     </par>
```

Abbyy 11

Numériser des archives académiques

Problèmes

Des *outputs* hétérogènes : le jeu des *n* différences

- Une segmentation au caractère dans les XML Abbyy 8 ; à la ligne dans les XML Abbyy 11
- Présence d'un indice de qualité (*charConfidence*) uniquement dans les XML Abbyy 8 : problème beaucoup plus embêtant.
- Problème subséquent : pas de certitude sur le calcul de *charConfidence* ($\approx$ probabilité que le caractère soit le bon)

Numériser des archives académiques

Problèmes

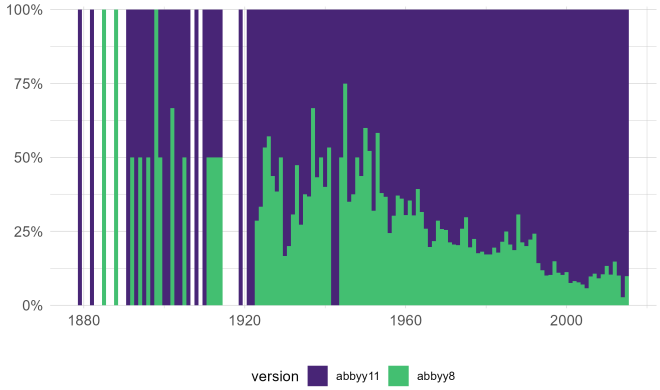
Des *outputs* hétérogènes

Version	*n*	percent
Abbyy 11	6757	82.2 %
Abbyy 8	1458	17.7 %

Distribution des fichiers

Numériser des archives académiques

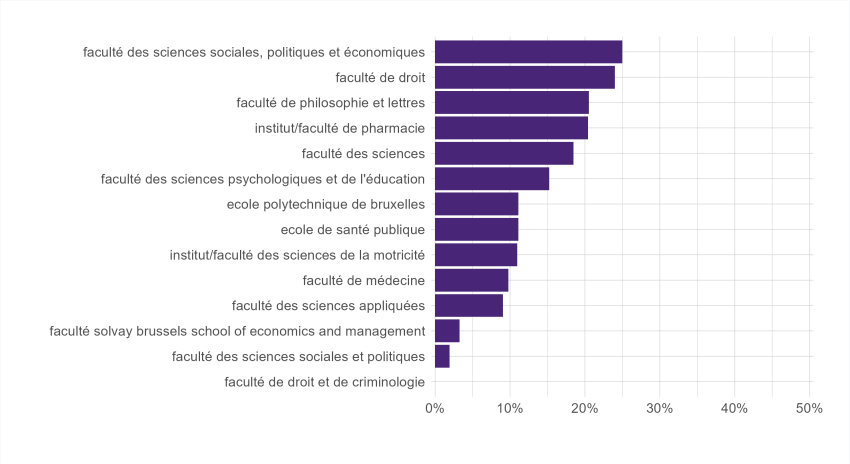
Des *outputs* hétérogènes



Numériser des archives académiques

Problèmes

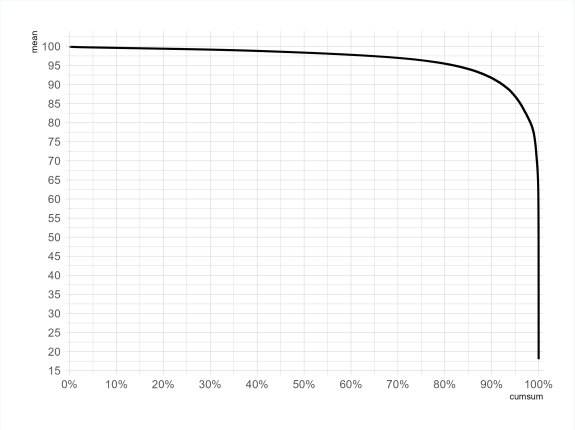
Des *outputs* hétérogènes : le jeu des *n* différences



Numériser des archives académiques

Qualité de l'OCR

- Malgré tout : se baser sur les fichiers Abbyy 8 pour estimer la qualité d'OCR du corpus
- Moyenne de *charConfidence* (de tous les caractères) : 97,7
- 75% des pages ont une *charConfidence* moyenne supérieure ou égale à 97 / 90% des pages ont une *charConfidence* supérieure ou égale à 92.



Numériser des archives académiques

Qualité de l'OCR

→ Moyenne élevée : pas de problème ?

→ Littérature montre que pour une série d'analyses, passé un certain seuil, un gain supplémentaire de qualité d'OCR ne change pas radicalement la donne [1]

→ Une qualité parfaite est illusoire et pas nécessaire

1. Ex : Hill, Hengchen, 2019, « Quantifying the impact of dirty OCR on historical text analysis… », *Digit. Scholarsh. Humanit.*, 34, p. 825-843

Numériser des archives académiques

Qualité de l'OCR

→ Moyenne élevée : pas de problème ?

→ Littérature montre que pour une série d'analyses, passé un certain seuil, un gain supplémentaire de qualité d'OCR ne change pas radicalement la donne [1]

→ Une qualité parfaite est illusoire et pas nécessaire

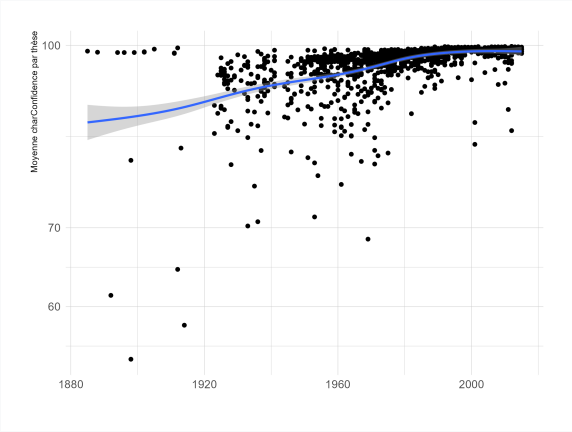
Oui mais… Quels biais ?

1. Ex : Hill, Hengchen, 2019, « Quantifying the impact of dirty OCR on historical text analysis… », *Digit. Scholarsh. Humanit.*, 34, p. 825-843

Numériser des archives académiques

Biais historique

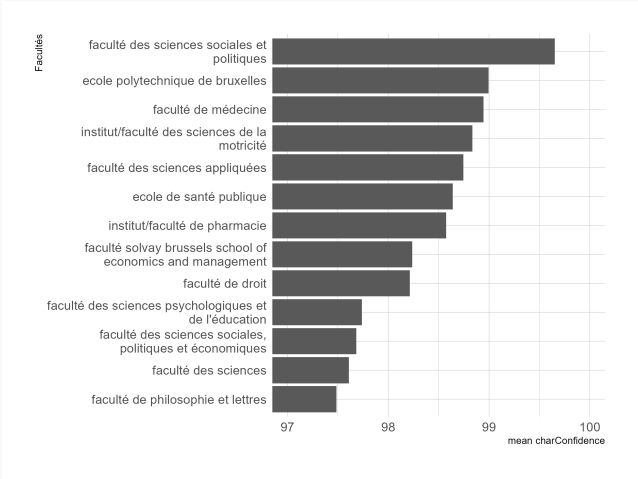
- Qualité d'OCR hétérogène d'un point de vue historique
- Amélioration progressive de la qualité de l'OCR au fil du temps (qualité de la conservation, de l'impression, etc.)
- Moindre dispersion au fil du temps



Numériser des archives académiques

Biais disciplinaire

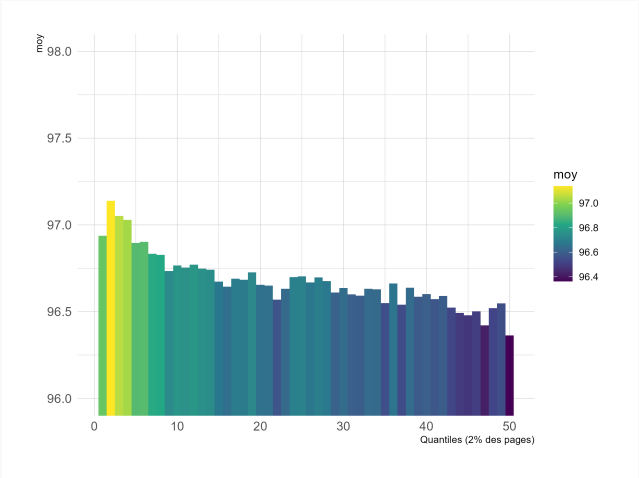
- Qualité d'OCR hétérogène d'un point de vue disciplinaire



Numériser des archives académiques

Biais interne

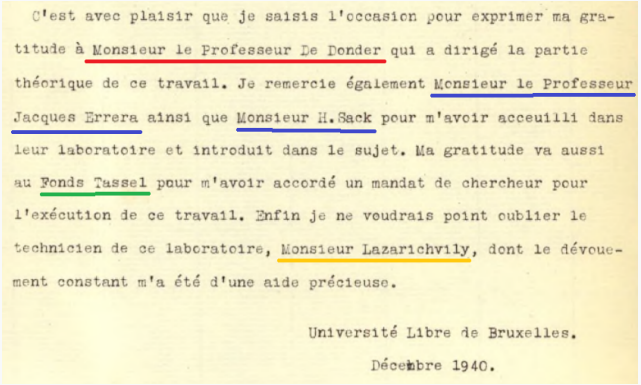
- En moyenne, la qualité d'OCR varie en fonction des pages



Numériser des archives académiques

Biais interne

- En moyenne, la qualité d'OCR varie en fonction des pages
- Que se passe-t-il si certaines pages m'intéressent particulièrement ?



C'est avec plaisir que je saisis l'occasion pour exprimer ma gratitude à Monsieur le Professeur De Donder qui a dirigé la partie théorique de ce travail. Je remercie également Monsieur le Professeur Jacques Errera ainsi que Monsieur H.Sack pour m'avoir acceuilli dans leur laboratoire et introduit dans le sujet. Ma gratitude va aussi au Fonds Tassel pour m'avoir accordé un mandat de chercheur pour l'exécution de ce travail. Enfin je ne voudrais point oublier le technicien de ce laboratoire, Monsieur Lazarichvily, dont le dévouement constant m'a été d'une aide précieuse.

Université Libre de Bruxelles.

Décembre 1940.

Se (ré)approprier la numérisation

(OCR libre et réfléchie)

Se (ré)approprier l'OCR

- Besoin d'un outil d'ATR en interne
- Création il y a environ 1 an d'une plateforme de support en humanités numériques : ***Praesto DH*** et mise à disposition d'outils pour les chercheurs.ses de l'ULB (instance Wikibase, etc.)
- Création de **TAMI** (*Transcription Automatique des Manuscrits et Imprimés*)

Se (ré)approprier l'OCR

TAMI, un outil basé sur :

- **eScriptorium**, développé à l'EPHE : plateforme web pour corriger la segmentation et la reconnaissance et préparer l'entraînement de modèles d'ATR
- **Kraken** :
  - Moteur d'OCR/HTR sur lequel repose eScriptorium
  - Pensé spécifiquement pour les écritures non-latines et les textes anciens/historiques
  - Mobilise des modèles d'OCR/HTR libres et réentraînables (dépôt Zenodo)

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer Abbyy (8 et 11) à eScriptorium-Kraken

- Modèle *CATMuS-Print [Large]* (UNIGE/INRIA), disponible sur Zenodo.org, et décrit comme *Diachronic model for French prints and other West European languages* : entraîné sur des textes depuis le XVIe jusqu'à des textes numériques du XXIe siècle
- Travail en cours :
  - Comparaison des *confidences*
  - Comparaison avec une *vérité terrain*

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer les *confidences*

- Abbyy 8 : *charConfidence* / Kraken : *glyph confidence* ou *GC* (XML-ALTO)
- Test sur une thèse choisie au hasard [a]

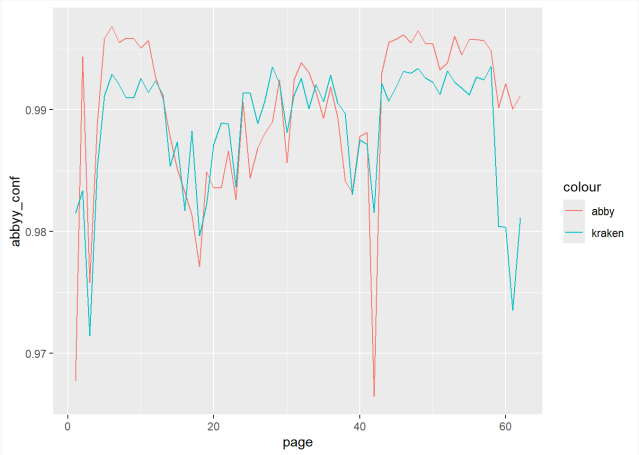
a. Hubert Chantrenne, *Recherches sur le mécanisme de la synthèse des protéines*, 1951

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer les *confidences*

- Pas très clair : sur certaines pages, Kraken fait mieux, sur d'autres non (ou similaire)
- Problème : peut-on comparer deux indices dont l'un est une boîte noire ?



Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer sur la base d'une *vérité terrain*

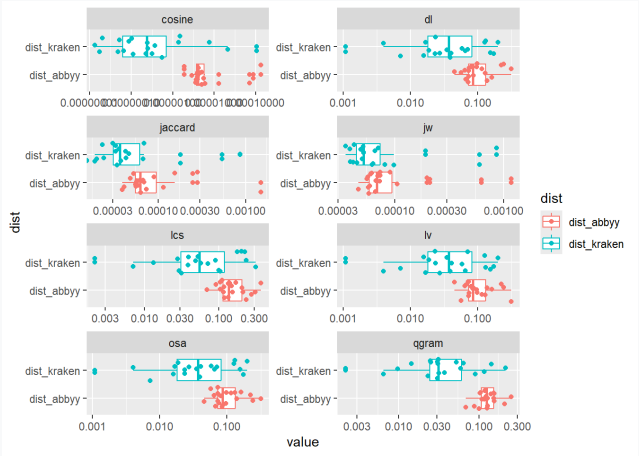
- Même thèse, correction manuelle des 20 premières pages

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer sur la base d'une *vérité terrain*

- ***Character Error Rate*** au niveau de chaque page : plus la distance est plus grande, moins l'OCR est proche de la « réalité »

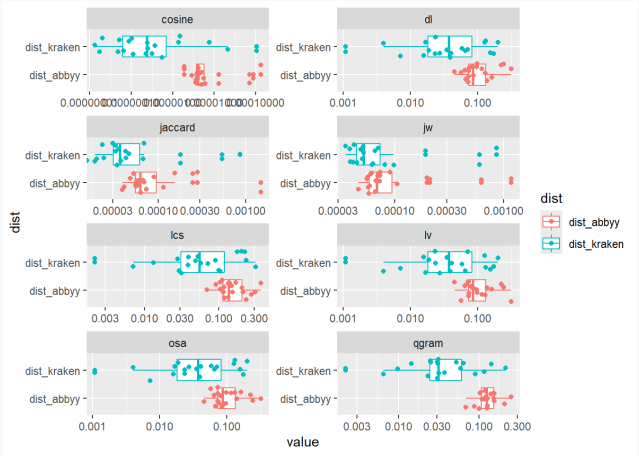


Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer sur la base d'une *vérité terrain*

- ***Character Error Rate*** au niveau de chaque page : plus la distance est plus grande, moins l'OCR est proche de la « réalité »
- Kraken fait mieux !

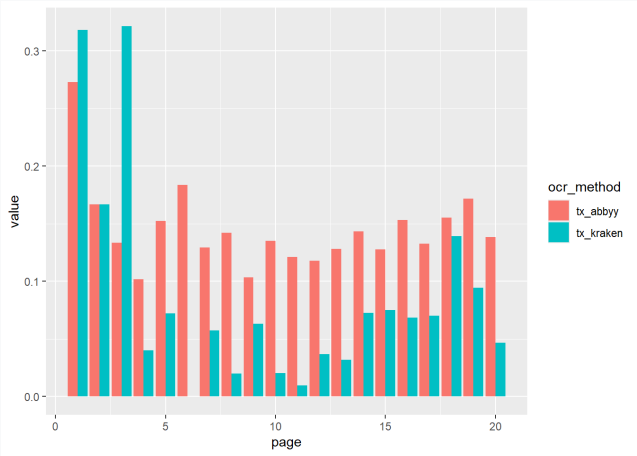


Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer sur la base d'une *vérité terrain*

- ***Word Error Rate***, comparaison page par page : plus la distance est plus grande, moins l'OCR est proche de la « réalité »

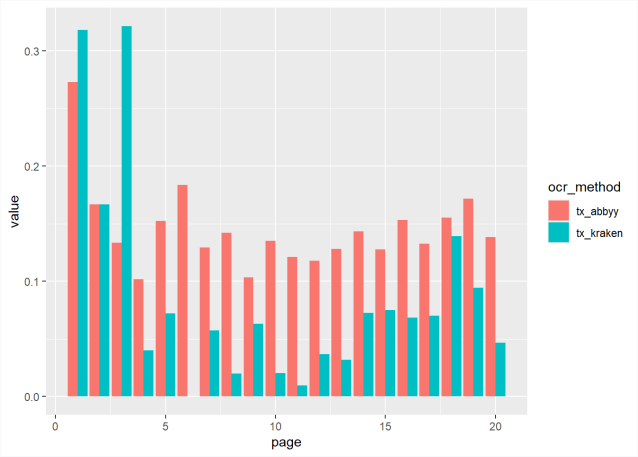


Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Comparer sur la base d'une *vérité terrain*

- ***Word Error Rate***, comparaison page par page : plus la distance est plus grande, moins l'OCR est proche de la « réalité »
- Kraken fait mieux ! Sauf sur les premières pages



Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

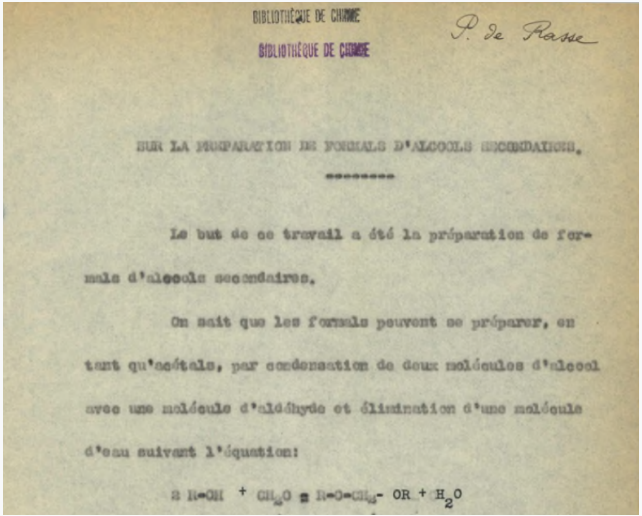
- Fait problablement mieux ?

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

- Fait problablement mieux ?
- Tester sur une thèse particulièrement difficile (encre qui a bavé, papier abîmé…) : P. de Rasse, *Sur la préparation de formals d'alcools secondaires*, s. d.



BIBLIOTHÈQUE DE CHIMIE

BIBLIOTHÈQUE DE CHIMIE

P. de Rasse

SUR LA PREPARATION DE FORMALS D'ALCOOLS SECONDAIRES.

Le but de ce travail a été la préparation de formals d'alcools secondaires.

On sait que les formals peuvent se préparer, en tant qu'acétals, par condensation de deux molécules d'alcool avec une molécule d'aldéhyde et élimination d'une molécule d'eau suivant l'équation:

$$2\ R\text{-}OH\ +\ CH_2O \rightleftarrows R\text{-}O\text{-}CH_2\text{-}\ OR\ +\ H_2O$$

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

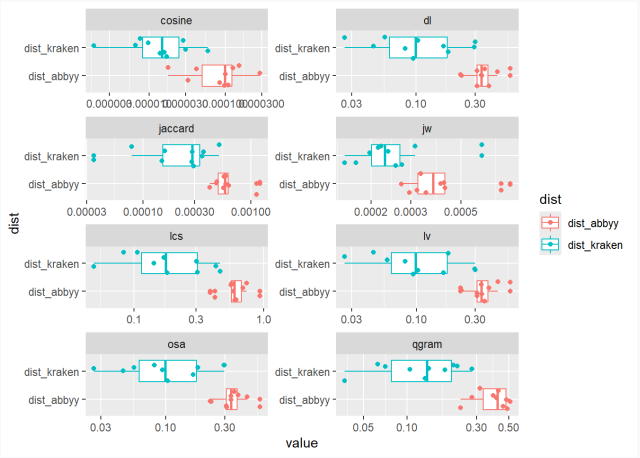
- ***Character Error Rate*** sur les 10 premières pages

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

- Kraken fait mieux !



Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

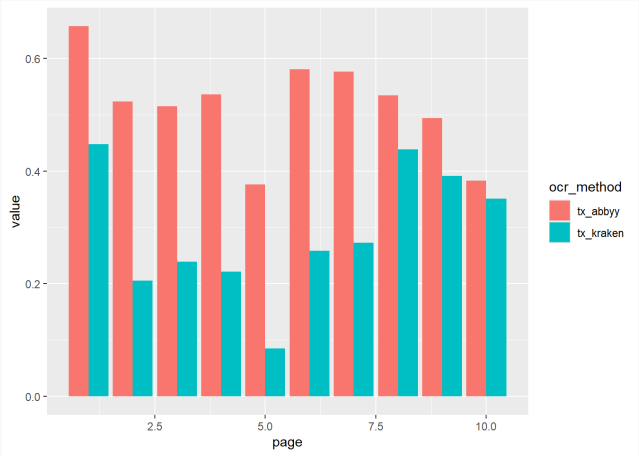
- ***Word Error Rate*** sur les 10 premières pages

Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Super, mais Abbyy 11 ?

- Kraken fait mieux !



Se (ré)approprier l'OCR

Ré-océriser les thèses ?

Bilan provisoire

- Des résultats prometteurs, suffisamment pour lancer une réocérisation des thèses
- La réocérisation va permettre d'analyser la *confidence* de manière plus robuste (échantillons moins hétérogènes, etc.)
- Pour CER et WER, démarche à consolider : partir de textes propres et simuler des dégradations
- *Fine-tuner* le modèle d'OCR, notamment pour les premières et dernières pages (couverture, remerciements, bibliographie…) ?

Se (ré)approprier l'OCR

Bonus : et un LLM ?

L'OCR réglée par les LLM ?

Se (ré)approprier l'OCR

Bonus : et un LLM ?

L'OCR réglée par les LLM ? Anecdote sur un corpus de presse belge numérisé

texte	Kraken	LLM
Nous avons par voie d'Angleterre des nouvelles de Lisbonne	Nous avone par voie d'Anjieterre dca nouvelles de Lr.-"vne	Nous avons par voie d'Angleterre des nouvelles de Lausanne

Se (ré)approprier l'OCR

Bonus : et un LLM ?

texte	Kraken	LLM
Nous avons par voie d'Angleterre des nouvelles de Lisbonne	Nous avone par voie d'Anjieterre dca nouvelles de Lr.-"vne	Nous avons par voie d'Angleterre des nouvelles de Lausanne

Se (ré)approprier l'OCR

Bonus : et un LLM ?

texte	Kraken	LLM
`Nous avons par voie d'Angleterre des nouvelles de Lisbonne`	`Nous avone par voie d'Anjieterre dca nouvelles de Lr.-"vne`	`Nous avons par voie d'Angleterre des nouvelles de Lausanne`

Distance entre `Lr.-"vne` et `Lausanne`/`Lisbonne` = strictement similaire (= 5)

Distance (CER) entre OCR-Kraken/réalité > Distance entre LLM/réalité : **problème!**

Se (ré)approprier l'OCR

Bonus : et un LLM ?

texte	Kraken	LLM
`Nous avons par voie d'Angleterre des nouvelles de Lisbonne`	`Nous avone par voie d'Anjieterre dca nouvelles de Lr.-"vne`	`Nous avons par voie d'Angleterre des nouvelles de Lausanne`

→ Méfiance sur la CER pour estimer la qualité de transcription d'un LLM

→ Vaut-il mieux une information parcellaire ou une contre-vérité ?

Conclusion

- **L'OCR pose toujours des questions**
- Ne pas négliger les biais, surtout pour des textes anciens/des démarches historiennes
- Promouvoir des protocoles basés sur des outils et des modèles libres, réentraînables, se méfier des boîtes noires (dont LLM)
- Pour des documents anciens (*ie* des archives) et la recherche, les solutions propriétaires « grand public » ne sont peut-être pas les meilleures solutions
- Adapter l'OCR à son usage, logique de *fitness for use*

Merci !