La recherche en IA appliquée aux SHS

Gaël Lejeune, gael.lejeune@sorbonne-universite.fr

13 Janvier 2025

Sorbonne Université

Qui suis-je ?

Maître de Conférences HDR en Informatique (Sorbonne Université), formation initiale :

- Licence Sciences du Langage
- Master Traductologie et Sciences Cognitives
- Doctorat en Informatique et Applications

Parlons donc de **pluridisciplinarité** :

Qui suis-je ?

Maître de Conférences HDR en Informatique (Sorbonne Université), formation initiale :

- Licence Sciences du Langage
- Master Traductologie et Sciences Cognitives
- Doctorat en Informatique et Applications

Parlons donc de **pluridisciplinarité** :

L'informatique (ou IA) pour les Humanités

Et inversement !

Plan de la présentation

Informatique en Faculté de Lettres

Mettre des mots sur du bruit

Des méthodes robustes aux variations locales ?

Travail sur Corpus : Liberté d'Expression

CERES : outiller la recherche en SHS

Informatique en Faculté de Lettres

Vérifier/Exemplifier des Théories Linguistiques

- Penser la qualité des Données . . .en liaison avec les besoins

Vérifier/Exemplifier des Théories Linguistiques

- Penser la qualité des Données . . .en liaison avec les besoins
- ! Ne pas surestimer l'importance de la qualité "visible" (ou intrinsèque)

Vérifier/Exemplifier des Théories Linguistiques

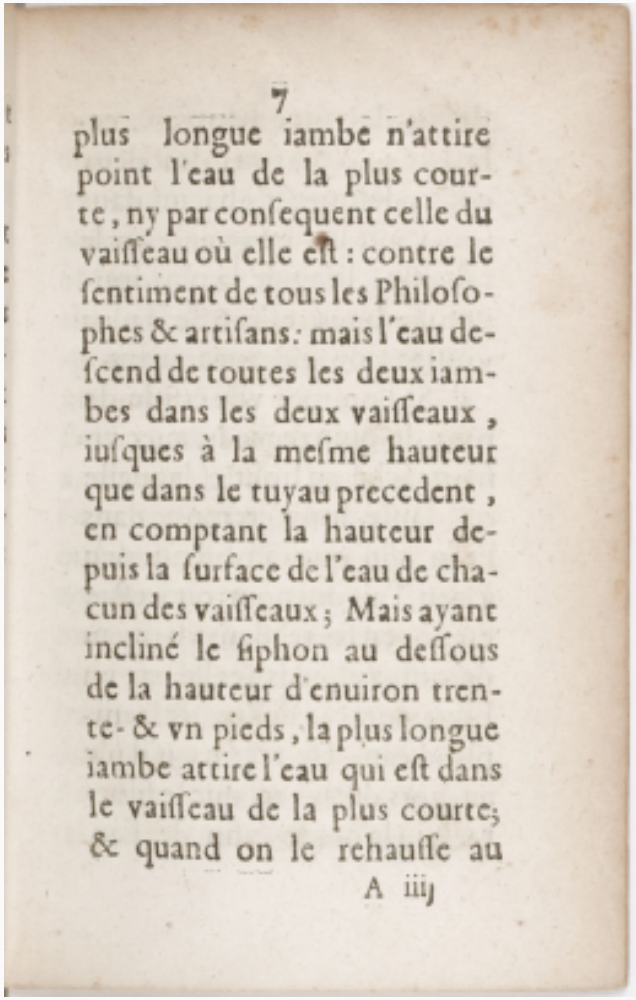
- Penser la qualité des Données . . .en liaison avec les besoins
- ! Ne pas surestimer l'importance de la qualité "visible" (ou intrinsèque)
- Penser des Corpus cohérents, équilibrés . . .et passer à l'échelle

Vérifier/Exemplifier des Théories Linguistiques

- Penser la qualité des Données . . .en liaison avec les besoins
- ! Ne pas surestimer l'importance de la qualité "visible" (ou intrinsèque)
- Penser des Corpus cohérents, équilibrés . . .et passer à l'échelle
- Faisabilité de la Collecte ?

Mettre des mots sur du bruit

Dans l'idéal on a (presque) le texte complet



7

plus longue iambe n'attire point l'eau de la plus courte, ny par conſequent celle du vaiſſeau où elle eſt : contre le ſentiment de tous les Philoſophes & artiſans: mais l'eau deſcend de toutes les deux iambes dans les deux vaiſſeaux, iuſques à la meſme hauteur que dans le tuyau precedent, en comptant la hauteur depuis la ſurface de l'eau de chacun des vaiſſeaux; Mais ayant incliné le ſiphon au deſſous de la hauteur d'enuiron trente- & vn pieds, la plus longue iambe attire l'eau qui eſt dans le vaiſſeau de la plus courte; & quand on le rehauſſe au

A iiij

7
plus longue iambe n'attire
point l'eau de la plus cour¬
te, ny par conſequent celle du
vaiſſeau où elle eſt: contre le
ſentiment de tous les Philoſo¬
phes & artiſans: mais l'eau de¬
ſcend de toutes les deux iam¬
bes dans les deux vaiſſeaux,
iuſques à la meſme hauteur
que dans le tuyau precedent,
en comptant la hauteur de¬
puis la ſurface de l'eau de cha¬
cun des vaiſſeaux; Mais ayant
incliné le ſiphon au deſſous
de la hauteur d'enuiron tren¬
te-& vn pieds, la plus longue
iambe attire l'eau qui eſt dans
le vaiſſeau de la plus courte;
& quand on le rehauſſe au
A iiij

Une page bien "formée", des caractères bien détectés (modèle OCR de Simon Gabay)

Dans les faits

1
l
r
1
3
e
aaaaa
A
pn eer
ees?
ee
e
aae?
ateu
ae
aeoe
ees¬
¬
dai
5
il
ee
aa
aeaaaaul
a
.
leu
:
eee
an)
IVLES a2a C dnsr eiller du y
l en les onseils, Abba des dbbayes de SsNab or enmorraine,
de SMda-d de Soissonis, et de SPiere de Corbie Cc.
DPaletar oneoret ecuuae-pmanl du Bex
, ¬
38

Une page difficile à traiter, du bruit

Que faire avec ces données ?

Faut-il toujours standardiser ?

Données Tout Gallica 1600-1800 [1]

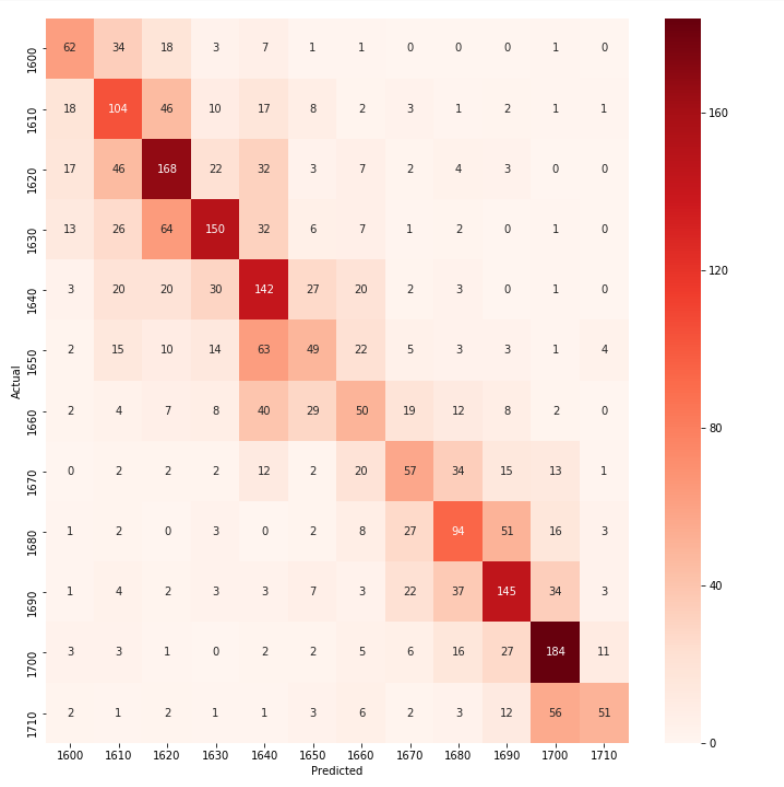
Tâche Datation automatique (en contexte bruité)

1. [Baledent et al., 2020]

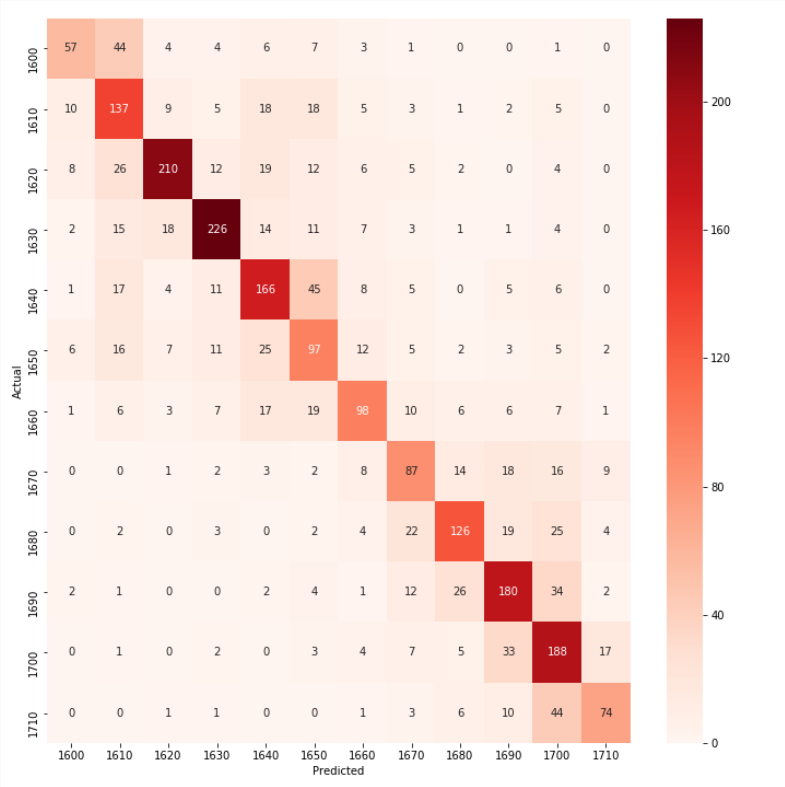
Faut-il toujours standardiser ?

Données Tout Gallica 1600-1800 [1]

Tâche Datation automatique (en contexte bruité)



Mots (F.mes 46.3, Sim. 92.8)



1 – 4 grammes (F-mes 71.43, Sim. 95.0)

1. [Baledent et al., 2020]

Le bruit et la classification

Données Corpus multilingue bruité artificiellement [2]

Tâche Classification et détection d'évènements

2. [Nguyen et al., 2020]

Le bruit et la classification

Données Corpus multilingue bruité artificiellement [2]

Tâche Classification et détection d'évènements

(i) (ii) (iii) (iv) (v) (vi)

Figure 2 – (i) clean image, (ii) *Phantom Character*, (iii) *Character Degradation*, (iv) *Bleed Through*, (v) *Blur*, and (vi) all mixed together.

2. [Nguyen et al., 2020]

Adapter les données ou adapter la méthode ?

- Chercher des mots, des phrases là où il n'y en a pas (ou plus) ?
  - A-t-on des systèmes robustes aux variations locales ?

Adapter les données ou adapter la méthode ?

- Chercher des mots, des phrases là où il n'y en a pas (ou plus) ?
  - A-t-on des systèmes robustes aux variations locales ?
- Comment travailler hors ≪ données de laboratoire ≫ ?

Adapter les données ou adapter la méthode ?

- Chercher des mots, des phrases là où il n'y en a pas (ou plus) ?
  - A-t-on des systèmes robustes aux variations locales ?
- Comment travailler hors ≪ données de laboratoire ≫ ?
- Quelles sont les interférences qui sont vraiment problématiques
  - bruit ressenti VS bruit réel

Adapter les données ou adapter la méthode ?

- Chercher des mots, des phrases là où il n'y en a pas (ou plus) ?
  - A-t-on des systèmes robustes aux variations locales ?
- Comment travailler hors ≪ données de laboratoire ≫ ?
- Quelles sont les interférences qui sont vraiment problématiques
  - bruit ressenti VS bruit réel
- **Quels observables sont pertinents ?**

Des méthodes robustes aux variations locales ?

Transformer un problème "Humanités" en problème HN

Mission : trouver des noms de lieux, personnages dans des romans

Transformer un problème "Humanités" en problème HN

Mission : trouver des noms de lieux, personnages dans des romans

Reformulation : Reconnaissance d'Entités Nommées (NER) après Reconnaissance Optique de Caractères (OCR)[Koudoro-Parfait, 2025].

Transformer un problème "Humanités" en problème HN

Mission : trouver des noms de lieux, personnages dans des romans

Reformulation : Reconnaissance d'Entités Nommées (NER) après Reconnaissance Optique de Caractères (OCR)[Koudoro-Parfait, 2025].

Données Corpus ELTEC (retranscriptions de référence de 10 romans) VS 4 océrisations différentes (Kraken_base, Kraken_17, Tesseract_fr) [Koudoro-Parfait et al., 2021]

Tâche Comparaison des sorties NER (Spacy et Stanza) sur la version propre et les différentes versions OCR

Influence de la qualité de l'image sur la sortie OCR ?

Kraken 3.0	**Tess. fr 0.3.6**
Ses voisines plumaient leurs	Ses voisines plumaient leurs
oioes quatre fois avant de les ven-	vies quatre fois avant de les ven-
LL I II II I I IIF M ii I I II E E g	
Chamnnlhrs de ta mn Mamnnetta	Chaumière de la mè1. Nannette
dre ; mais la mere Nannette	dre ; mais la mère Nannette
disait que eetait une	disait que c'était une
mauvaise m6thode,	mauvaise méthode,
paree qu'ainsi la plume ...	parce qu'ainsi l2 plume ...

Table 1 – Transcriptions* OCR d'une illustration et de sa légende. illustration, légende, contaminations orthographiques.

* Z. Carraud, *La petite Jeanne*, 1884.

Quelles Entités nommées sur corpus bruité

texte = "Ses voisines plumaient leurs vies quatre fois avant de les ven-
Chaumière de la mè1. Nannette dre ; mais la mère Nannette disait que c'était
une mauvaise méthode, parce qu'ainsi l2 plume ..."

Quelles Entités nommées sur corpus bruité

texte = "Ses voisines plumaient leurs vies quatre fois avant de les ven-Chaumière de la mè1. Nannette dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi l2 plume ..."

Passons dans un moteur de REN,

avec le modèle `spacy_small`

- MISC : Chaumière de la mè1
- PER : Nannette
- PER : l2 plume

Quelles Entités nommées sur corpus bruité

texte = "Ses voisines plumaient leurs vies quatre fois avant de les ven-Chaumière de la mè1. Nannette dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi l2 plume ..."

Passons dans un moteur de REN,

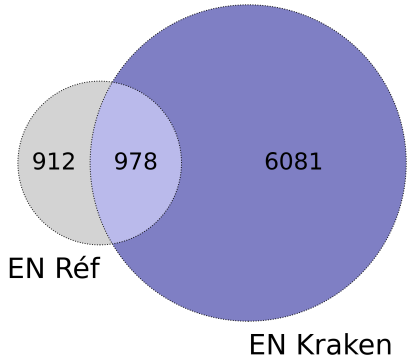
avec le modèle `spacy_small`

- MISC : Chaumière de la mè1
- PER : Nannette
- PER : l2 plume

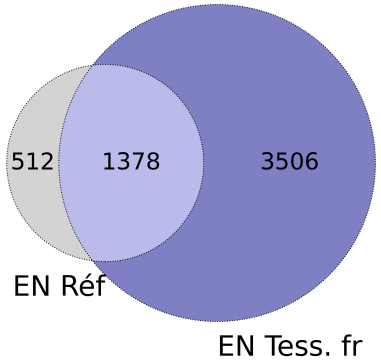
avec le modèle `spacy_medium`

- PER : Chaumière
- LOC : Nannette

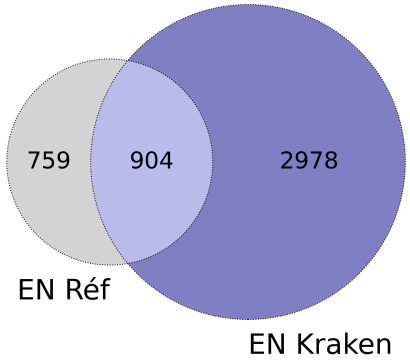
Impact du modèle NER selon la version (ELtec `fr`)



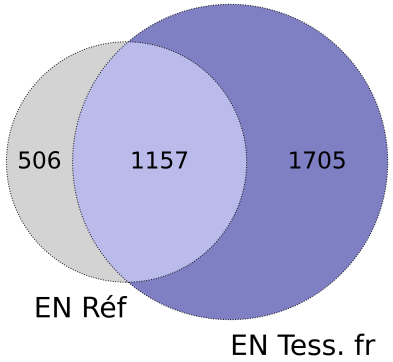
(a) Kraken (Spacy large)



(b) Tesseract (Spacy large)



(c) Kraken-Base (Stanza)



(d) Tesseract (Stanza)

Quelles sont ces différences ?

Filtrer l'output plutôt que corriger l'input

pour "l'Amérique" (version de référence), on trouve dans la version OCR :

- 0.625, "l'Amerique"
- 0.4743416490252569, 'Aerique'
- 0.4330127018922193, 'Amerique'

Quelles sont ces différences ?

Filtrer l'output plutôt que corriger l'input

pour "l'Amérique" (version de référence), on trouve dans la version OCR :

- 0.625, "l'Amerique"
- 0.4743416490252569, 'Aerique'
- 0.4330127018922193, 'Amerique'

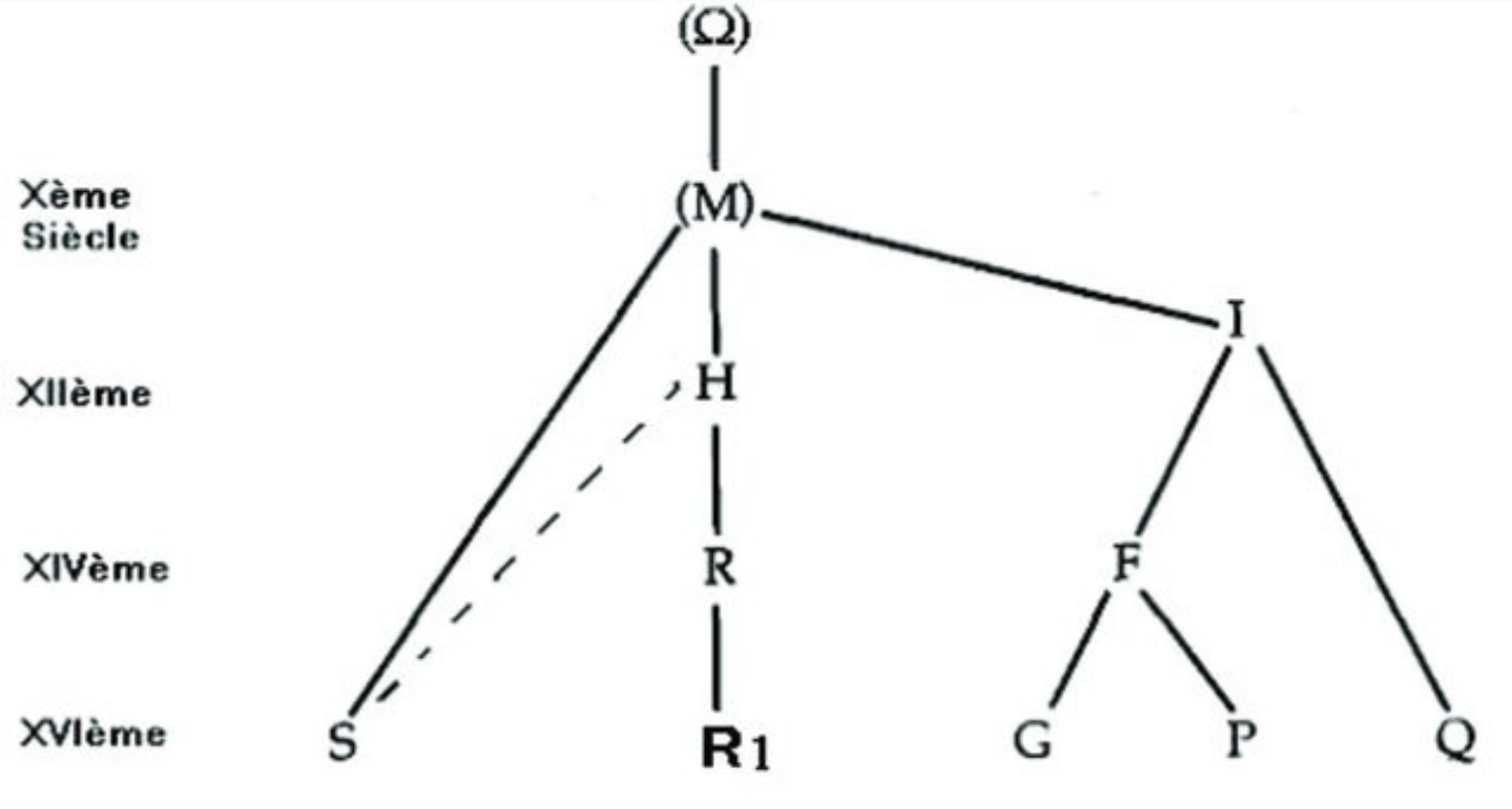
Pour "Canadienne" :

- 1.0, 'Canadienne'
- 0.8660254037844386, 'Canadien'
- 0.5773502691896257, 'Canadisn'

Pour "Comanches" :

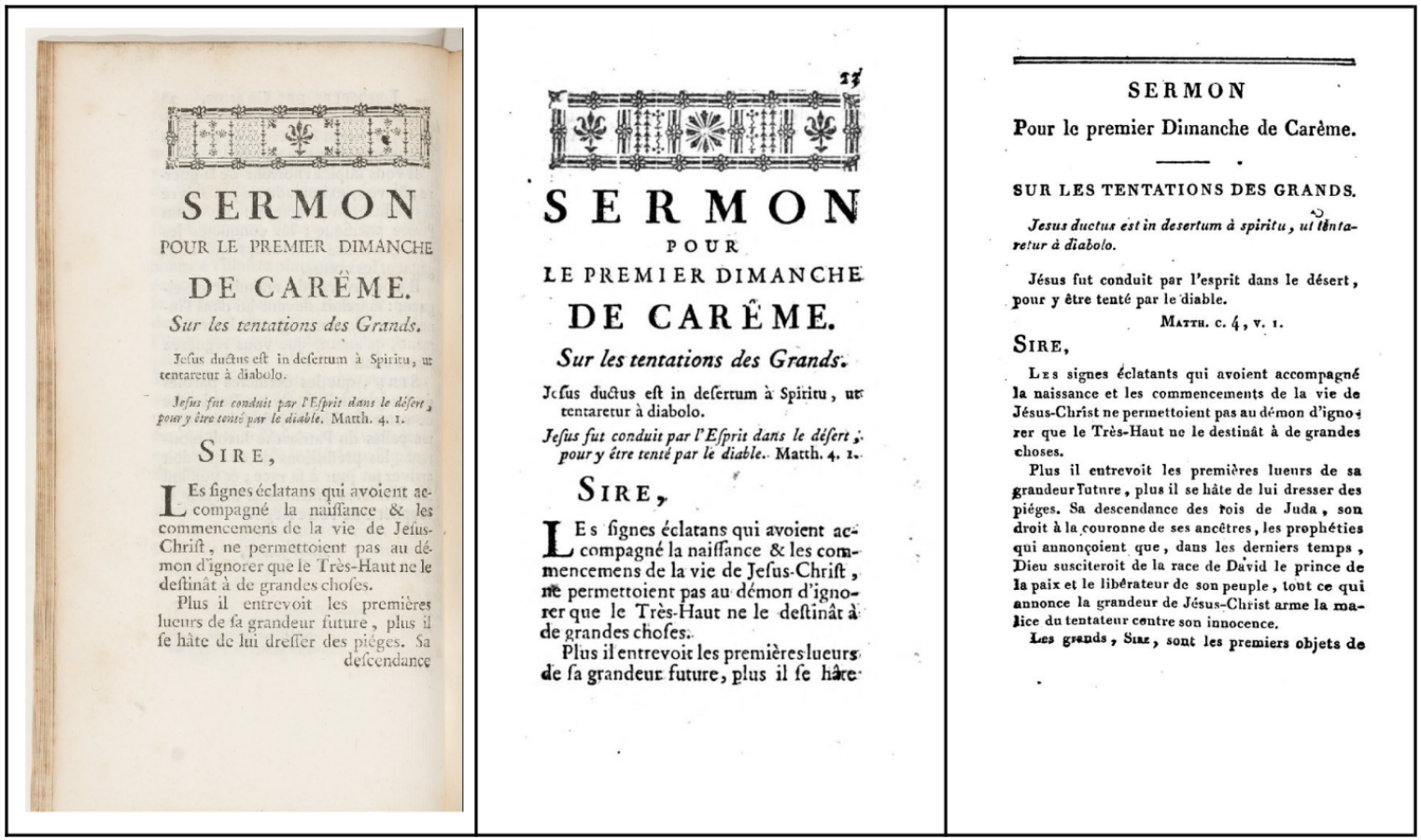
- 0.7462025072446364, 'Comanchessont'
- 0.707106781186547 6, 'le Comanche'
- 0.5345224838248488, 'Comancnes'
- 0.5345224838248488, 'Comancles'

La variation prend différentes formes



Stemma Codicum de manuscrits (Source : Marc le Pouliquen)

La variation est partout



SERMON

POUR LE PREMIER DIMANCHE

DE CARÊME.

Sur les tentations des Grands.

Jesus ductus est in desertum à Spiritu, ut tentaretur à diabolo.

Jesus fut conduit par l'Esprit dans le désert, pour y être tenté par le diable. Matth. 4. 1.

SIRE,

LEs signes éclatans qui avoient accompagné la naissance & les commencemens de la vie de Jesus-Christ, ne permettoient pas au démon d'ignorer que le Très-Haut ne le destinât à de grandes choses.

Plus il entrevoit les premières lueurs de sa grandeur future, plus il se hâte de lui dresser des piéges. Sa descendance

23

SERMON

POUR

LE PREMIER DIMANCHE

DE CARÊME.

Sur les tentations des Grands.

Jesus ductus est in desertum à Spiritu, ut tentaretur à diabolo.

Jesus fut conduit par l'Esprit dans le désert, pour y être tenté par le diable. Matth. 4. 1.

SIRE,

LEs signes éclatans qui avoient accompagné la naissance & les commencemens de la vie de Jesus-Christ, ne permettoient pas au démon d'ignorer que le Très-Haut ne le destinât à de grandes choses.

Plus il entrevoit les premières lueurs de sa grandeur future, plus il se hâte

SERMON

Pour le premier Dimanche de Carême.

SUR LES TENTATIONS DES GRANDS.

Jesus ductus est in desertum à spiritu, ut tentaretur à diabolo.

Jésus fut conduit par l'esprit dans le désert, pour y être tenté par le diable.

MATTH. c. 4, v. 1.

SIRE,

LEs signes éclatants qui avoient accompagné la naissance et les commencements de la vie de Jésus-Christ ne permettoient pas au démon d'ignorer que le Très-Haut ne le destinât à de grandes choses.

Plus il entrevoit les premières lueurs de sa grandeur future, plus il se hâte de lui dresser des piéges. Sa descendance des rois de Juda, son droit à la couronne de ses ancêtres, les prophéties qui annonçoient que, dans les derniers temps, Dieu susciteroit de la race de David le prince de la paix et le libérateur de son peuple, tout ce qui annonce la grandeur de Jésus-Christ arme la malice du tentateur contre son innocence.

Les grands, SIRE, sont les premiers objets de

Différents imprimés des Sermons de Massillon (ANR ECOLE)

La variation est multiple

ChapBooks (Source : University of Oxford)

Variation et observables

Que cherche t-on dans des corpus textuels ?

Que contiennent les données textuelles ?
Des mots ?

- Encore des mots,

Variation et observables

Que cherche t-on dans des corpus textuels ?

Que contiennent les données textuelles ?
Des mots ?

- Encore des mots,
- Toujours des mots

Variation et observables

Que cherche t-on dans des corpus textuels ?

Que contiennent les données textuelles ?
Des mots ?

- Encore des mots,
- Toujours des mots
- Rien que des mots ?

Variation et observables

Que cherche t-on dans des corpus textuels ?

Que contiennent les données textuelles ?
Des mots ?

- Encore des mots,
- Toujours des mots
- Rien que des mots ?
- Les mêmes mots ?

Variation et observables

Que cherche t-on dans des corpus textuels ?

Que contiennent les données textuelles ?
Des mots ?

- Encore des mots,
- Toujours des mots
- Rien que des mots ?
- Les mêmes mots ?

! La variation peut nécessiter des pré-traitements (standardiser la langue) mais sans garantie d'amélioration des résultats [Siino et al., 2024].

Travail sur Corpus : Liberté d'Expression

Exemple sur des corpus nativement numériques

Mission : étudier la liberté d'expression des différents acteurs (sportifs mais pas que) pendant les JO dans différents types de médias.

Exemple sur des corpus nativement numériques

Mission : étudier la liberté d'expression des différents acteurs (sportifs mais pas que) pendant les JO dans différents types de médias.

Problématiques de corpus :

1. accéder aux données X/Twitter ?

Exemple sur des corpus nativement numériques

Mission : étudier la liberté d'expression des différents acteurs (sportifs mais pas que) pendant les JO dans différents types de médias.

Problématiques de corpus :

1. accéder aux données X/Twitter ?
2. avoir une taille significative ?

Exemple sur des corpus nativement numériques

Mission : étudier la liberté d'expression des différents acteurs (sportifs mais pas que) pendant les JO dans différents types de médias.

Problématiques de corpus :

1. accéder aux données X/Twitter ?
2. avoir une taille significative ?
3. Pénibilité ?

Reformulation :

- → limitation à la Presse
- → collecte (semi-)automatique
- → Europresse
- → Europarser

Obtenir des corpus structurés depuis Europresse (I)

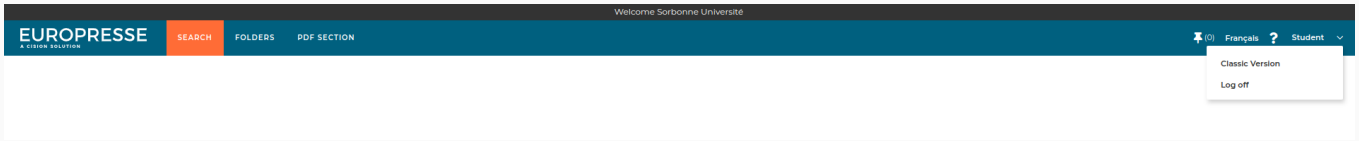


Figure 4 – La version classique d'Europresse

Obtenir des corpus structurés depuis Europresse (II)

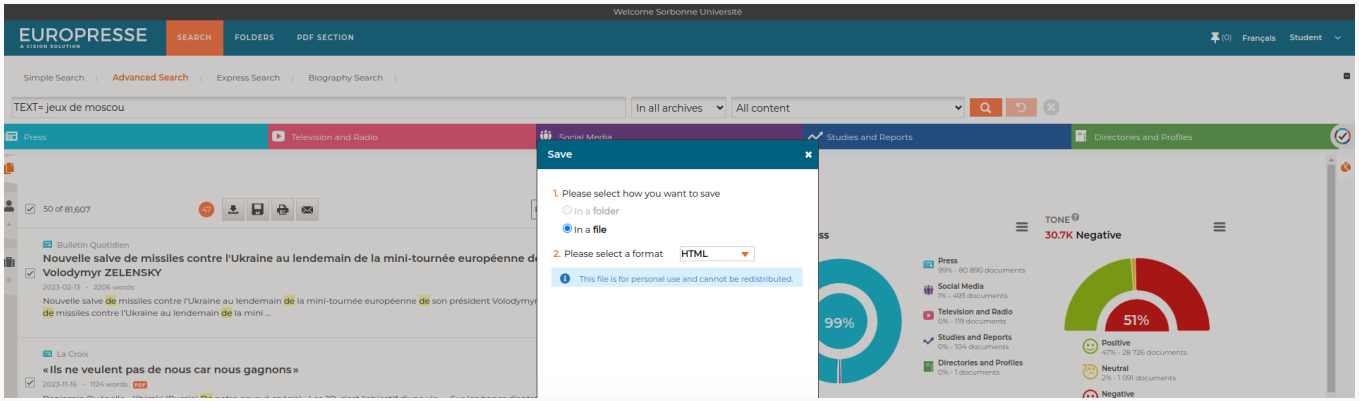


Figure 5 – Exporter en un mode "lisible par la machine" : HTML

Obtenir des corpus structurés depuis Europresse (III)

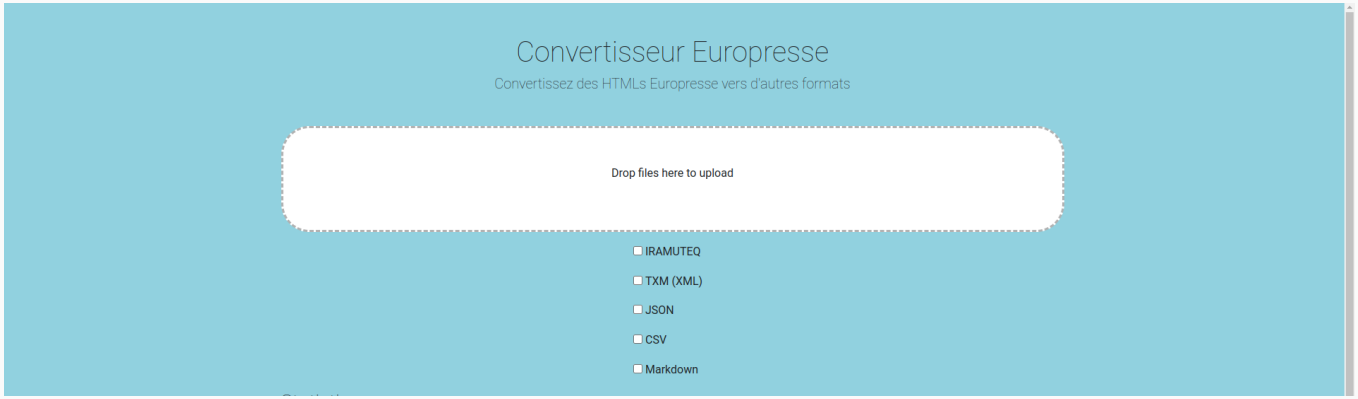
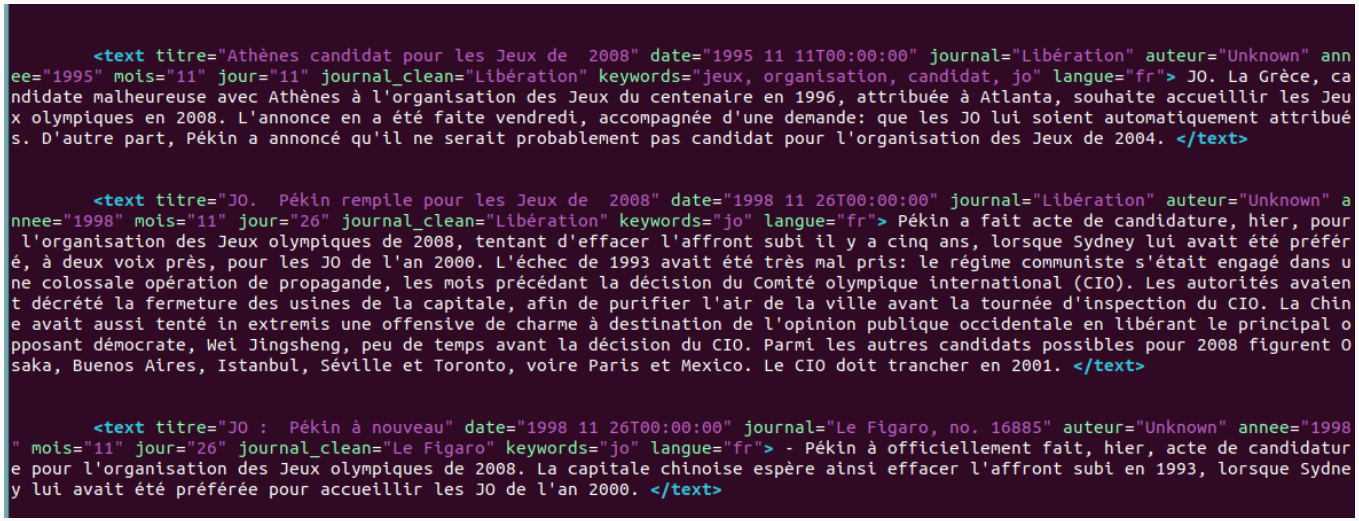


Figure 6 – Transformer avec EUROPARSER

`https://ceres.huma-num.fr/europarser/`

Obtenir des corpus structurés depuis Europresse (III)



```
<text titre="Athènes candidat pour les Jeux de  2008" date="1995 11 11T00:00:00" journal="Libération" auteur="Unknown" annee="1995" mois="11" jour="11" journal_clean="Libération" keywords="jeux, organisation, candidat, jo" langue="fr"> JO. La Grèce, candidate malheureuse avec Athènes à l'organisation des Jeux du centenaire en 1996, attribuée à Atlanta, souhaite accueillir les Jeux olympiques en 2008. L'annonce en a été faite vendredi, accompagnée d'une demande: que les JO lui soient automatiquement attribués. D'autre part, Pékin a annoncé qu'il ne serait probablement pas candidat pour l'organisation des Jeux de 2004. </text>

<text titre="JO.  Pékin rempile pour les Jeux de  2008" date="1998 11 26T00:00:00" journal="Libération" auteur="Unknown" annee="1998" mois="11" jour="26" journal_clean="Libération" keywords="jo" langue="fr"> Pékin a fait acte de candidature, hier, pour l'organisation des Jeux olympiques de 2008, tentant d'effacer l'affront subi il y a cinq ans, lorsque Sydney lui avait été préféré, à deux voix près, pour les JO de l'an 2000. L'échec de 1993 avait été très mal pris: le régime communiste s'était engagé dans une colossale opération de propagande, les mois précédant la décision du Comité olympique international (CIO). Les autorités avaient décrété la fermeture des usines de la capitale, afin de purifier l'air de la ville avant la tournée d'inspection du CIO. La Chine avait aussi tenté in extremis une offensive de charme à destination de l'opinion publique occidentale en libérant le principal opposant démocrate, Wei Jingsheng, peu de temps avant la décision du CIO. Parmi les autres candidats possibles pour 2008 figurent Osaka, Buenos Aires, Istanbul, Séville et Toronto, voire Paris et Mexico. Le CIO doit trancher en 2001. </text>

<text titre="JO :  Pékin à nouveau" date="1998 11 26T00:00:00" journal="Le Figaro, no. 16885" auteur="Unknown" annee="1998" mois="11" jour="26" journal_clean="Le Figaro" keywords="jo" langue="fr"> - Pékin à officiellement fait, hier, acte de candidature pour l'organisation des Jeux olympiques de 2008. La capitale chinoise espère ainsi effacer l'affront subi en 1993, lorsque Sydney lui avait été préférée pour accueillir les JO de l'an 2000. </text>
```

Figure 7 – Résultat d'EUROPARSER : un document structuré

Collecte de Corpus en Ligne ?

À retenir :

- disponibilité VS utilisabilité ?

Collecte de Corpus en Ligne ?

À retenir :

- disponibilité VS utilisabilité ?
- pour l'humain VS pour la machine ?

Collecte de Corpus en Ligne ?

À retenir :

- disponibilité VS utilisabilité ?
- pour l'humain VS pour la machine ?
- complétude ? exhaustivité ?

CERES : outiller la recherche en SHS

Des outils, des usages

Besoins identifiées par CERES

- Collecter ET Formater des données (ex Europarser)

Des outils, des usages

Besoins identifiées par CERES

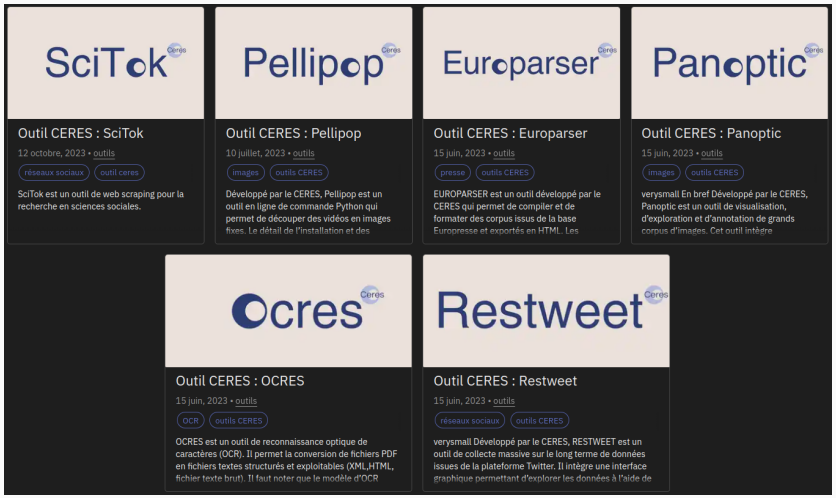
- Collecter ET Formater des données (ex Europarser)
- Évaluer la Qualité des Données (Waddle, Tecquel et Mordor)

Des outils, des usages

Besoins identifiées par CERES

- Collecter ET Formater des données (ex Europarser)
- Évaluer la Qualité des Données (Waddle, Tecquel et Mordor)

Travailler sur d'**autres types de données**



La méthode CRS

Les conditions pour automatiser

- **C**asse-pieds : pas intéressant à faire
- **R**épétitif : impression de predre son temps

La méthode CRS

Les conditions pour automatiser

- **C**asse-pieds : pas intéressant à faire
- **R**épétitif : impression de predre son temps
- **S**imple : simplifier pour pouvoir le coder

Détecter des dé-figements linguistiques

Mission : : Chercher les variantes de "Que la force soit avec toi" et les comparer grammaticalement

3. 1M de Tweets ressemblant à des figements collectés via feu l'API Twitter

Détecter des dé-figements linguistiques

Mission : : Chercher les variantes de "Que la force soit avec toi" et les comparer grammaticalement

Casse-pieds et Répétitif ?

3. 1M de Tweets ressemblant à des figements collectés via feu l'API Twitter

Détecter des dé-figements linguistiques

Mission : : Chercher les variantes de "Que la force soit avec toi" et les comparer grammaticalement

Casse-pieds et Répétitif ?

Reformulation : collecte via API [3], Similarité + Alignement

que	*la*	*force*	-	*soit*	*avec*	*toi*
que	*la*	*force*	**ouvrière**	*soit*	*avec*	*toi*
que	*la*	*force*	**de la Ligue 1**	*soit*	*avec*	*toi*

3. 1M de Tweets ressemblant à des figements collectés via feu l'API Twitter

Autres exemples d'outils

- OCRes : Transformer en texte des séries de PDF

Autres exemples d'outils

- OCRes : Transformer en texte des séries de PDF
- Pellipop : Transformer des séquences vidéos en série d'images

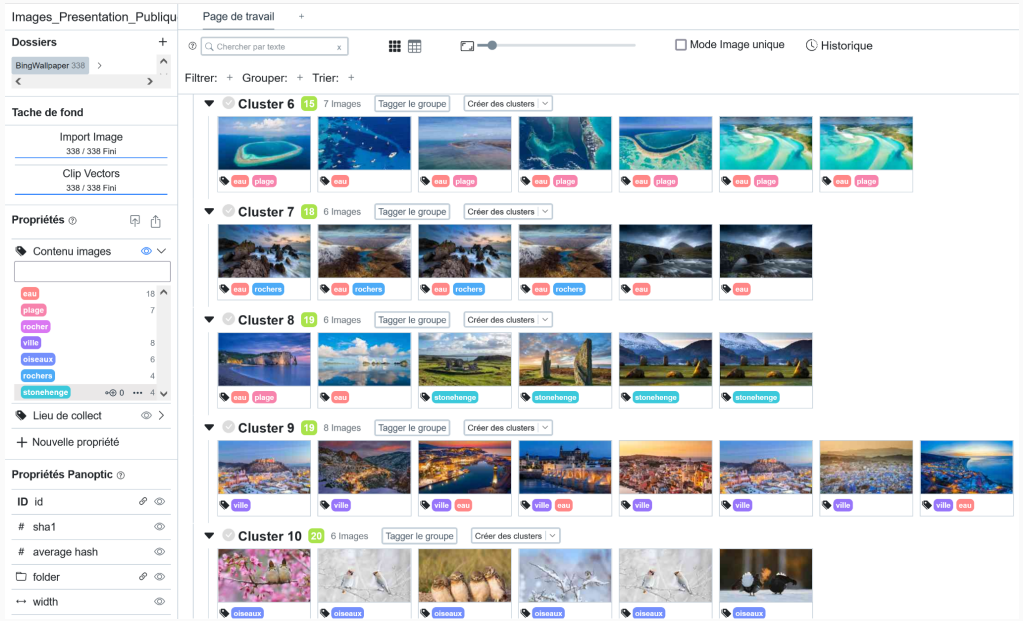
Autres exemples d'outils

- OCRes : Transformer en texte des séries de PDF
- Pellipop : Transformer des séquences vidéos en série d'images
- SciTok : Collecter sur SciTok avec les méta-données

Autres exemples d'outils

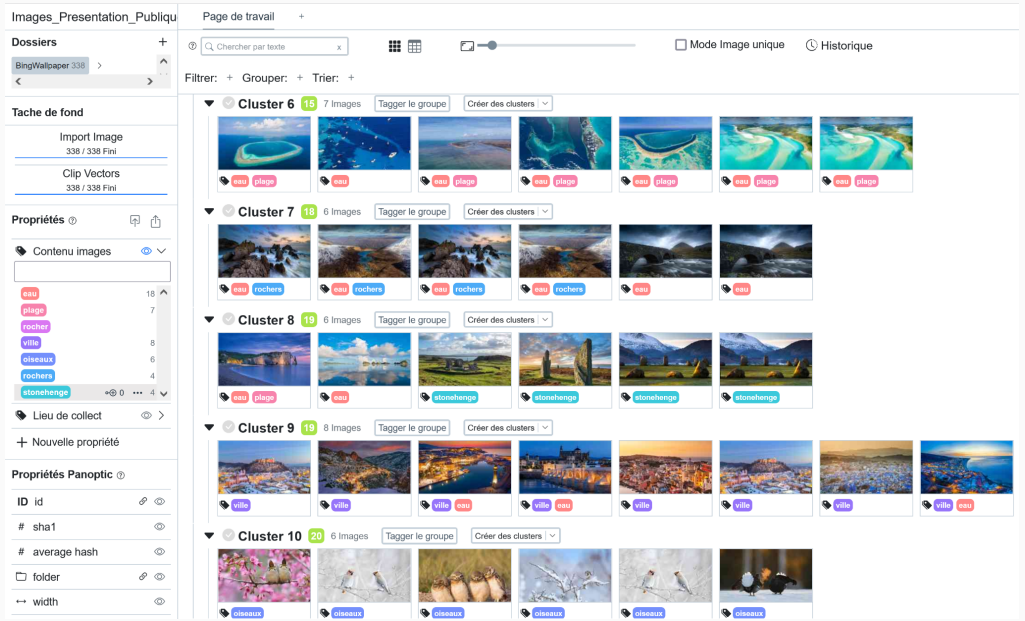
- OCRes : Transformer en texte des séries de PDF
- Pellipop : Transformer des séquences vidéos en série d'images
- SciTok : Collecter sur SciTok avec les méta-données
- Panoptic : **Organiser (manuellement et automatiquement) des collections d'images**[Bouté et al., 2024]

Panoptic



4. https://ceres.sorbonne-universite.fr/4f4013c8-f350-48c9-89cd-37e02dbe5c8a/

Panoptic



PANOPTIC est conçu pour la manipulation d'images pour la recherche en Info-Com/Sociologie, . . .détourné pour la recherche sur la désinformation en histoire (projet VIRAPIC)[4]

4. https://ceres.sorbonne-universite.fr/4f4013c8-f350-48c9-89cd-37e02dbe5c8a/

En conclusion

- Les questions des HN, les réponses des IA ?

En conclusion

- Les questions des HN, les réponses des IA ?
- ET inversement :
- IA pour les Humanités Numériques
- IA par les Humanités Numériques

Recommandations pratiques :

- Ne pas imaginer que tout est soluble dans l'info/IA/ChatGPT/Whatever

En conclusion

- Les questions des HN, les réponses des IA ?
- ET inversement :
- IA pour les Humanités Numériques
- IA par les Humanités Numériques

Recommandations pratiques :

- Ne pas imaginer que tout est soluble dans l'info/IA/ChatGPT/Whatever
- Automatiser ce qui peut l'être
- En en acceptant/connaissant les limites

Soyez curieux/curieuses :

décomposez, reformulez, détournez

Références i

Baledent, A., Hiebel, N., and Lejeune, G. (2020).
Dating Ancient texts : an Approach for Noisy French Documents.
In *Language Tech. for Historical and Ancient Languages*, .

Bouté, É., Julliard, V., Alié, F., Gödicke, D., Pailler, F., and Ecrement, V. (2024).
Panoptic, un outil d'exploration par similarité de vastes corpus d'images.
In *Humanistica 2024*.

Koudoro-Parfait, C. (2025).
Des IA au service de l'espace littéraire du XIX^{e} siècle : évaluation et analyse des outils de reconnaissance d'entités nommées spatiales.
PhD thesis, Thèse, Sorbonne Université.

Références ii


Koudoro-Parfait, C., Lejeune, G., Alrahabi, M., and Roe, G. (2021).
Discovering Spatial Relations in Litterature : what is the influence of OCR noise ?
In *NewsEye's international conference*, Paris, France.


Nguyen, N. K., Boros, E., Lejeune, G., and Doucet, A. (2020).
Impact analysis of document digitization on event extraction.
In *Natural Language for Artificial Intelligence (NL4AI)*, . -.


Siino, M., Tinnirello, I., and La Cascia, M. (2024).
Is text preprocessing still worth the time ? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers.
Information Systems, 121 :102342.