Reconnaissance optique des caractères et des écritures manuscrites

Projet E-NDP

Jean-Baptiste Camps[1] Nicolas Perreaux[2]

[1] : CJM, École nationale des chartes | Paris, Sciences & Lettres

[2]LAMOP | Université Paris 1 Panthéon-Sorbonne

16 février 2021

Des robots paléographes ?

- Prédiction d'un contenu texte
- à partir d'une image de la source
- par une intelligence artificielle
- entraînée par un humain
- dans un processus alternant
  - phases d'intervention humaines ;
  - phases de calcul.

Collaboration active entre humain et machine.

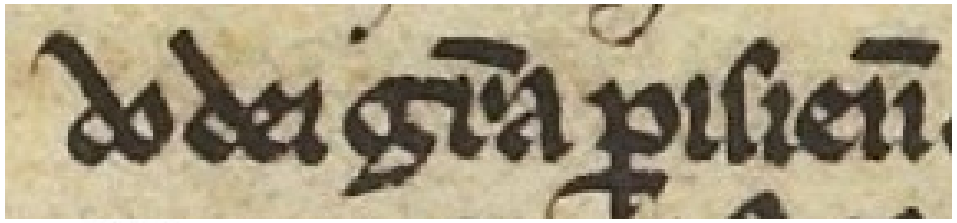
Objectifs

Comprendre comment ces technologies peuvent s'appliquer aux registres de Notre-Dame de Paris.

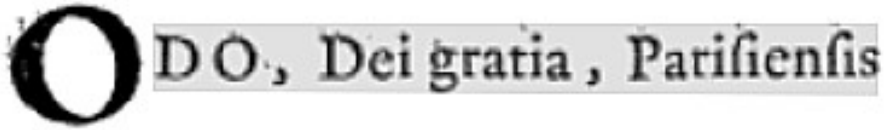
OCR/HTR-isation - méthodes et histoire(s)

Qu'est-ce qu'un OCR et le processus d'OCRisation ?

- **OCR** (parfois ROC en français) : « Optical Character Recognition » ou « Reconnaissance optique de caractère. » (Wikipedia)
- **OCRiser** : « Transformer automatiquement un fichier contenant l'image d'un document en fichier texte, grâce à un logiciel OCR. » (Wikipedia)
- Processus qui consiste à convertir un ensemble de signes graphiques, le plus souvent alphanumériques (mais aussi les ponctuations, espacements...), encodés sous la forme d'une image, en mode texte.
- L'OCR désigne à la fois un **processus** (d'OCR) et un **logiciel** (d'OCR).



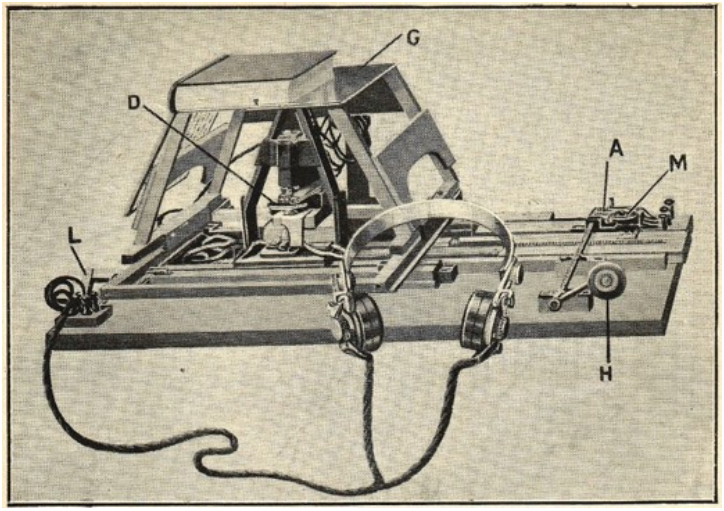
[O]do dei gra pisien
Paris, BnF, ms. 5526

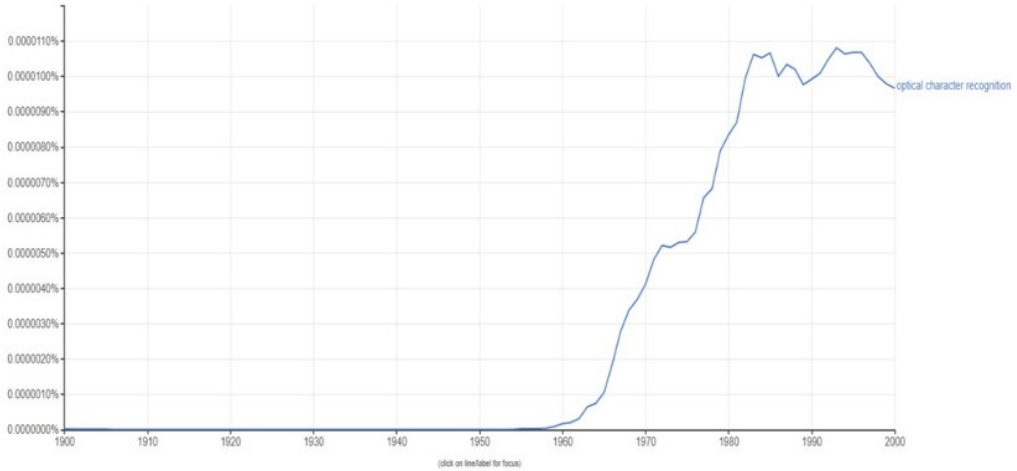


Odo, Dei gratia, Parisiensis

Quelques repères chronologiques

- Prémisses **dès 1870-1930** : transmission d'images, « scanner » pour aveugles (Charles Carey, Paul Nipkow, Edmund Fournier d'Albe...).
- 1929, **Gustav Tauschek** : machine qui compare le résultat d'un scan à des modèles mémorisés.
- Développements très nets **entre 1940 et 1990 :**
  - cryptanalyse ; prix ; passeports ; codes postaux ; rapports en cartes ; contrôle des populations.
  - production de machines de plus en plus portables et rapides (MIT, IBM...).
  - compagnies dédiées à la création de machines puis de logiciels (*Kurzweil Computer* en 1978).
- **Généralisation au XXIe siècle...** et rétro-conversion de l'héritage culturel.





OCR vs. HTR

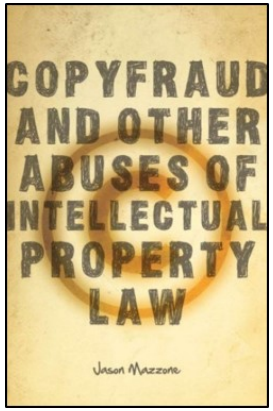
- L'**HTR** (*Handwritten Text Recognition* | Reconnaissance de l'écriture manuscrite) n'est pas fondamentalement différente de l'OCR, mais plus récente.
- La **difficulté est plus grande**, car les données plus complexes/variables (travaux de Shelia Guberman).
- Différents **essais industriels existent dès les années 1980** (*pen computing*).
- De **nombreux programmes scientifiques depuis 2010**, grâce aux développements de l'**IA** :
  - disponibilité de nouveaux algorithmes (réseaux de neurones), notamment `OCRopy` (Thomas Breuel) ;
  - Velum, Himanis, Horae, Home, Camps-Clérice-Pinche...
- **Les logiciels** employés ne sont donc pas différents de ceux pour l'OCR ;
- Modèles spécifiques à entraîner (pour chaque main, écriture,...), mais se généralise peu à peu → enjeu de disponibilité des données.

Quelles éditions et manuscrits peut-on OCRiser ?

- Les droits patrimoniaux sur une création expirent 70 ans après la mort de son auteur (100 ans si « mort(s) pour la France). Cf. le Code de la propriété intellectuelle et la Convention de Berne (1886-1979).
- **L'édition d'un texte n'est pas une création** (jugement Droz-Garnier).
- **Tous les textes médiévaux** peuvent donc être OCRisés et distribués !

Quelques références bibliographiques

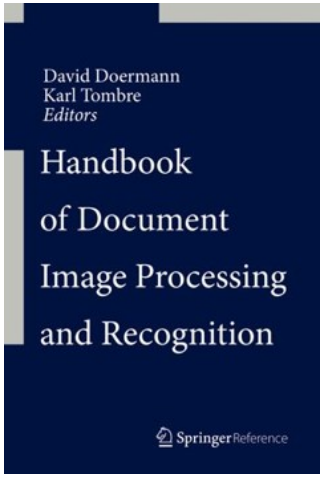
- J. Mazzone, Copyfraud and Other Abuses of Intellectual Property Law (2006)
- A. Guerreau, « L'avenir de la philologie. Textes anciens et domaine public » (2015, ID-HAL : 01112213).
- Le jugement Droz-Garnier et les réactions qu'il a entraîné.

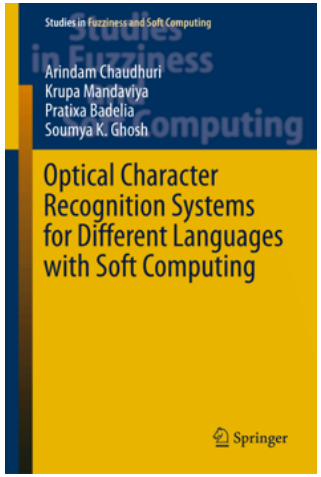


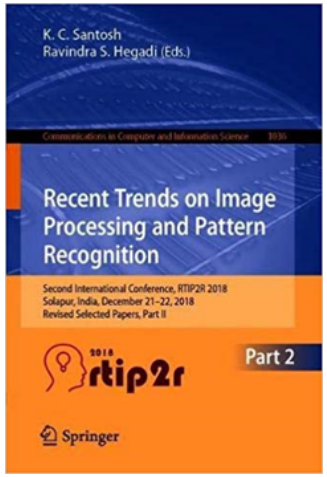
2006 et 2015

Quelques références bibliographiques

- Schantz, *The history of OCR, optical character recognition* (1982).
- *Optical character recognition in the historical discipline* (1993).
- Doermann et al., *Handbook of Document Image Processing and Recognition* (2014).
- Chaudhuri et al., *Optical Character Recognition Systems for Different Languages with Soft Computing* (2017).
- Santosch et al. (éd.), *Recent Trends in Image Processing and Pattern Recognition*, 3 volumes (2017-2019).







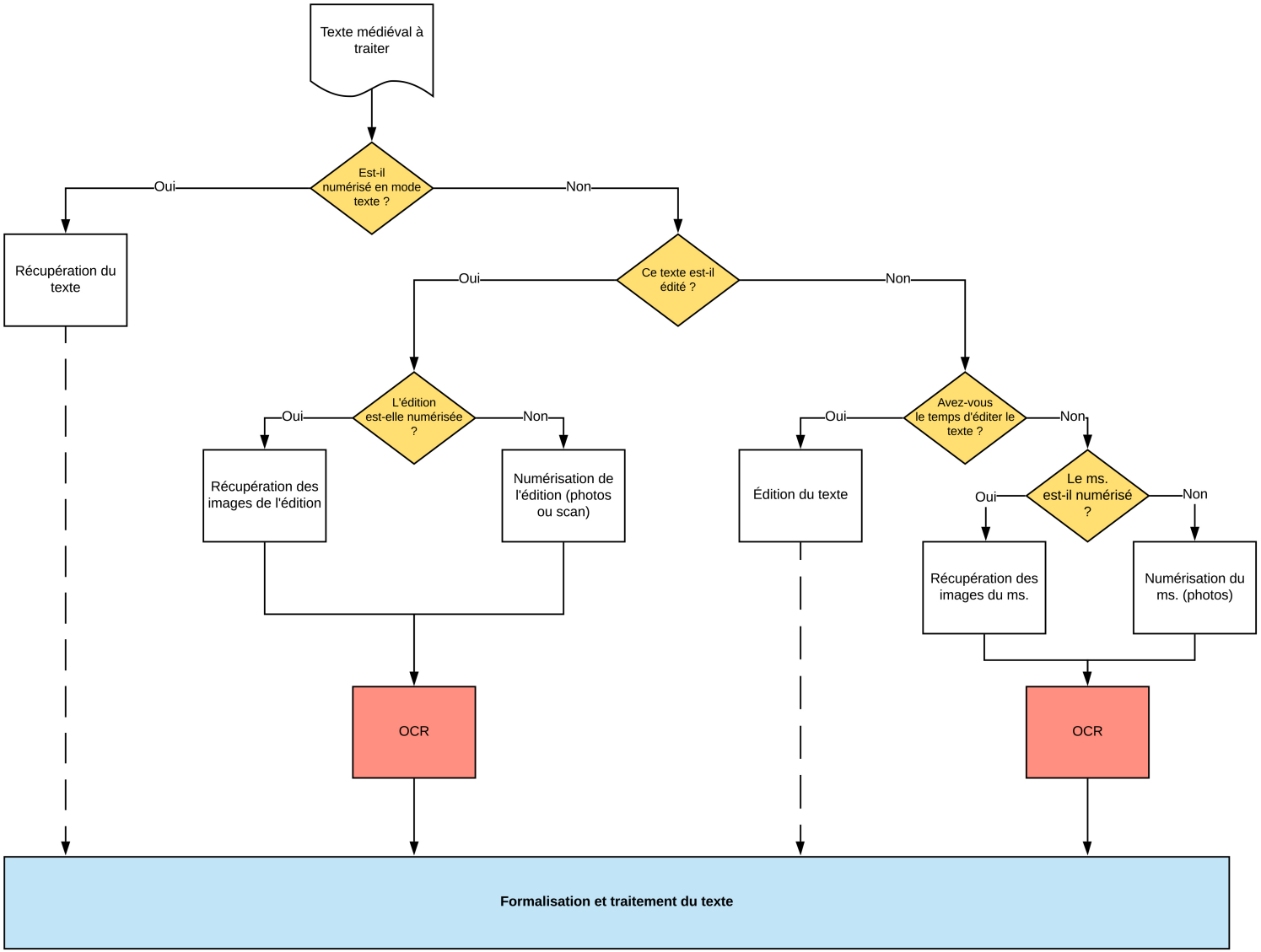
Plan

Plan

Quatre situations pour l'acquisition

- **L'intérêt de l'OCR/HTR dépend des objectifs**, du temps dont on dispose et de la situation historiographique / numérique antérieure.
- Se poser une série de questions :
  - Est-ce que le texte dont je souhaite réaliser l'analyse est numérisé ?
  - Si non, est-il édité ?
  - Si l'édition existe, est-elle facilement disponible ? Est-elle numérisée ? Est-elle de bonne qualité ?
  - Sinon encore, existe-t-il un ms. numérisé / numérisable ?
  - ...
  - Faut-il éditer ce texte inédit ?

Quatre situations pour l'acquisition

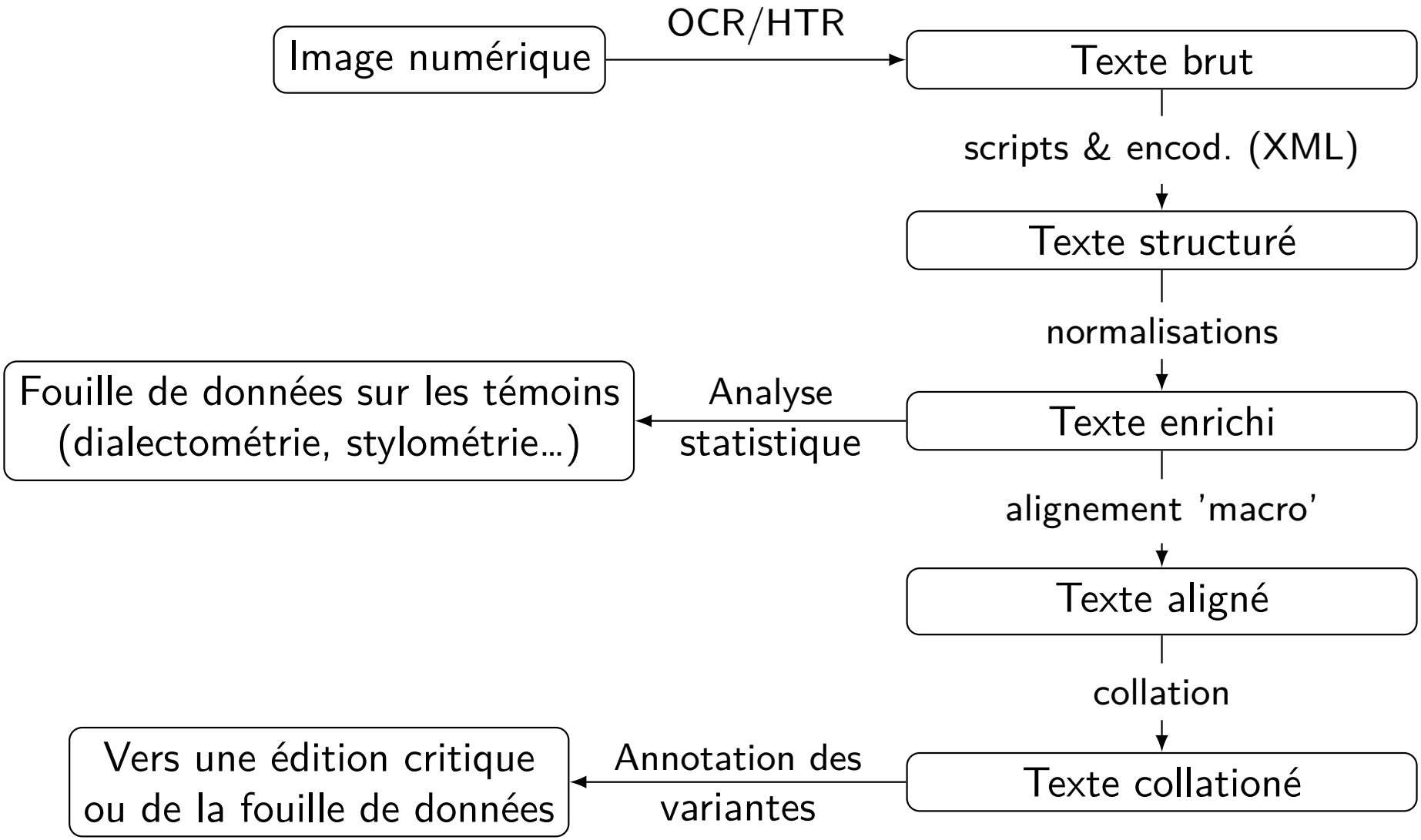


Quatre situations pour l'acquisition

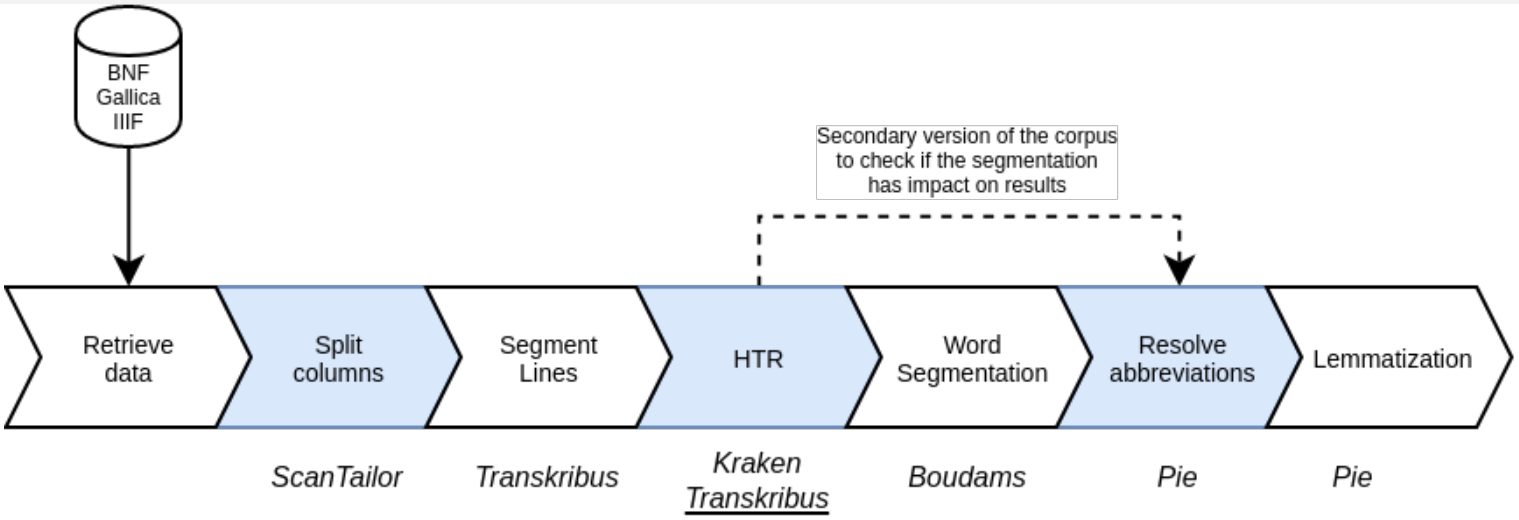
- **L'intérêt de l'OCR/HTR dépend des objectifs**, du temps dont on dispose et de la situation historiographique / numérique antérieure.
- Se poser une série de questions :
  - Est-ce que le texte dont je souhaite réaliser l'analyse est numérisé ?
  - Si non, est-il édité ?
  - Si l'édition existe, est-elle facilement disponible ? Est-elle numérisée ? Est-elle de bonne qualité ?
  - Sinon encore, existe-t-il un ms. numérisé / numérisable ?
  - ...
  - Faut-il éditer ce texte inédit ?
- Il est important de **bien évaluer la situation** pour prendre les bonnes décisions. Les processus techniques engagés varient en fonction des étapes nécessaires.
- Dans le **cas d'e-NDP, c'est le nombre des registres qui oriente vers l'HTR**.

Plan

Une chaîne de traitement théorique



Légendier français (BnF, fr. 412)



Data

- 961 columns
- 46 lines / columns
- 64 texts

Input

Output

qaz	car
il	il
nest	naistre
pas	pas
drois	droit
qe	que4

J.B. Camps, Th. Clérice & A. Pinche, *Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis*, `https://arxiv.org/abs/2012.03845`.

Les étapes du processus d'OCR/HTR

1. Pré-traitement des images ;
2. Analyse de la mise en page et identification des lignes (processus central dans le cas de l'HTR, appelé la segmentation) ;
3. Reconnaissance des caractères ;
4. Éventuels post-traitements, visant à améliorer les résultats et à enrichir le texte.

Plan

Interfaces : Transkribus

- développé par un consortium de recherche européen (projet READ ; Univ. Innsbruck et al.) ;
- financé par la Commission Européenne (Horizon 2020) ;
- permet de charger des images, analyser la mise en page, segmenter…
- opérations réalisées sur les serveurs de Transkribus.

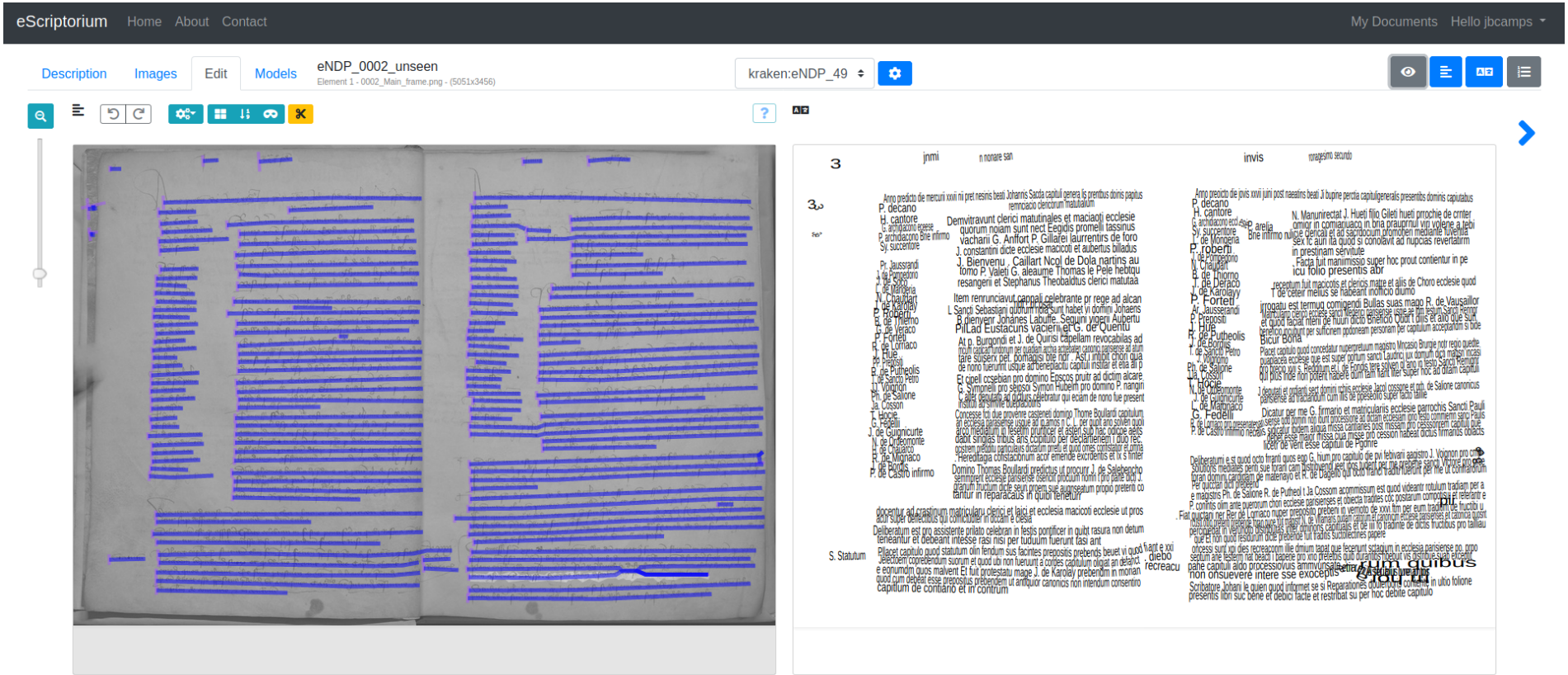
- Transkribus est devenu payant le 19 oct. 2020
- Achats de crédits pour le traitement automatique (HTR) de pages ;

Interfaces : eScriptorium

- logiciel libre, développé dans le cadre du projet Scripta (PSL) ;
- se branche sur `Kraken` pour l'analyse de mise en page et HTR ;
- nécessite d'être déployé sur un serveur par une institution ;
- Code : `https://gitlab.inria.fr/scripta/escriptorium` ;
- démos vidéos : `https://escripta.hypotheses.org/escriptorium-video-gallery`.

Démo.

eScriptorium



Algorithmes et bibliothèques logicielles : `Kraken`

- outil d'analyse de mise en page et d'HTR ;
- fondé sur de l'apprentissage profond (RNN LSTM) ;
- développé par Ben Kiessling dans le projet Scripta (PSL) ;
- Module Python, `https://github.com/mittagessen/kraken` ;
- Doc : `kraken.re`.

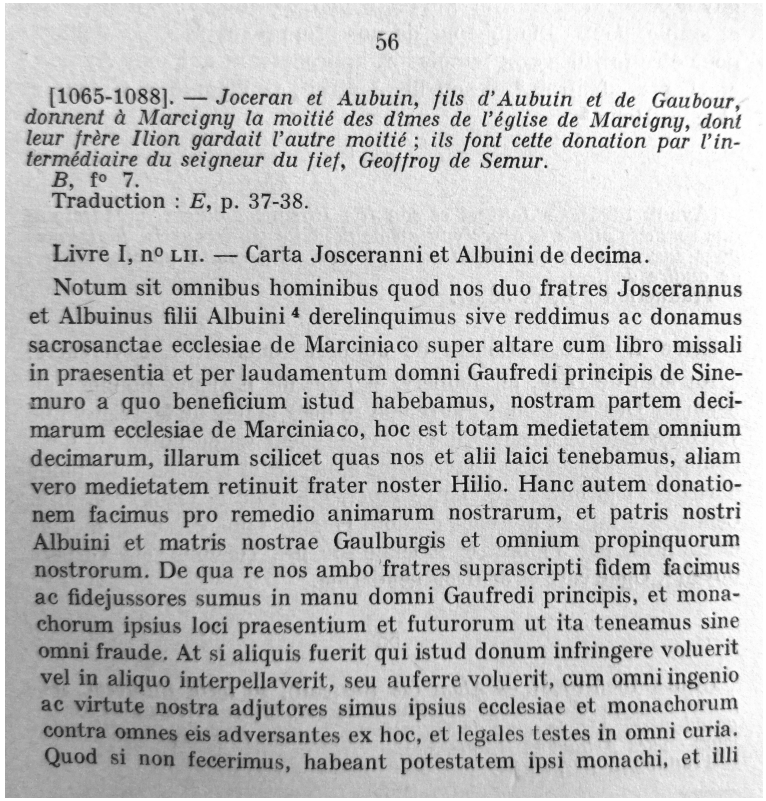
Pour en savoir plus, éléments techniques en annexe, 59

Plan

Plan

Processus d'acquisition et traitement

- Que peut **l'OCR** pour des éditions de **textes dits "de la pratique"** ?
- Journées *Du parchemin à la fouille de données. Nouveaux outils pour la création, la formalisation et l'analyse des corpus médiévaux* (2019).
- Smartphone sur les chartes de Marcigny-sur-Loire (éd. J. Richard, 1957).



56

[1065-1088]. — *Joceran et Aubuin, fils d'Aubuin et de Gaubour, donnent à Marcigny la moitié des dîmes de l'église de Marcigny, dont leur frère Ilion gardait l'autre moitié ; ils font cette donation par l'intermédiaire du seigneur du fief, Geoffroy de Semur.*
B, f° 7.
Traduction : *E*, p. 37-38.

Livre I, n° LII. — Carta Josceranni et Albuini de decima.

Notum sit omnibus hominibus quod nos duo fratres Joscerannus et Albuinus filii Albuini [4] derelinquimus sive reddimus ac donamus sacrosanctae ecclesiae de Marciniaco super altare cum libro missali in praesentia et per laudamentum domni Gaufredi principis de Sine-muro a quo beneficium istud habebamus, nostram partem decimarum ecclesiae de Marciniaco, hoc est totam medietatem omnium decimarum, illarum scilicet quas nos et alii laici tenebamus, aliam vero medietatem retinuit frater noster Hilio. Hanc autem donationem facimus pro remedio animarum nostrarum, et patris nostri Albuini et matris nostrae Gaulburgis et omnium propinquorum nostrorum. De qua re nos ambo fratres suprascripti fidem facimus ac fidejussores sumus in manu domni Gaufredi principis, et monachorum ipsius loci praesentium et futurorum ut ita teneamus sine omni fraude. At si aliquis fuerit qui istud donum infringere voluerit vel in aliquo interpellaverit, seu auferre voluerit, cum omni ingenio ac virtute nostra adjutores simus ipsius ecclesiae et monachorum contra omnes eis adversantes ex hoc, et legales testes in omni curia. Quod si non fecerimus, habeant potestatem ipsi monachi, et illi

Notum sit omnibus hominibus quod nos duo fratres Joscerannus || et Albuinus filii Albuini4 derelinquimus sive reddimus ac donamus || sacrosanctae ecclesiae de Marciniaco super altare cum libro missali || in praesentia et per laudamentum domni Gaufredi principis de Sine-||-muro a quo beneficium istud habebamus, || nostram partem deci-||-marum ecclesiae de Marciniaco, hoc est totam medietatem omnium || decimarum, illarum scilicet quas nos et alii laici tenebamus, aliam || vero medietatem retinuit frater noster Hilio. Hanc autem donatio-||-nem facimus pro remedio animarum nostrarum, et patris nostri || Albuini et matris nostrae Gaulburgis et omnium propinquorum || nostrorum. [...]

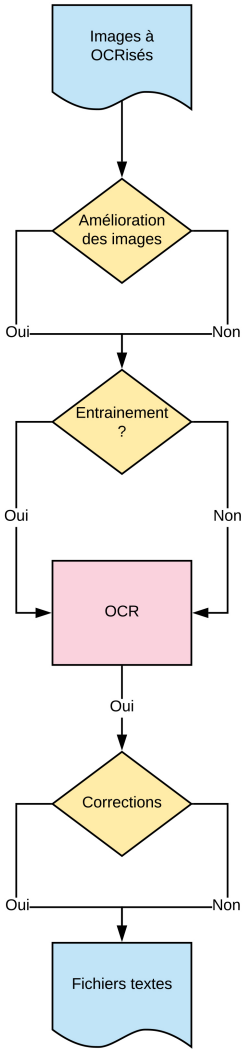
Processus d'acquisition et traitement

- **Chartes de Notre-Dame** : disponibles dans de nombreux formats sur Internet.
- Gallica = téléchargement direct (PDF) ou via Pyllica (Pierre-Carl Langlais).
- Internet Archive = téléchargement direct (PDF-JPEG | traités ou non).
- Comment déterminer **quelles images conviennent le mieux pour l'OCR** ?

- Résultats de la requête « titre = Cartulaire de Notre-Dame de Paris » sur Internet Archive.
- 22 volumes en tout, alors que l'édition ne compte que 4 volumes.
  - 5 versions du tome 1 ;
  - 6 versions du tome 2 ;
  - 4 versions du tome 3 ;
  - 7 versions du tome 4.
- Pour chaque volume, plusieurs versions (brute ou pré-traitée), etc.

Plan

Un processus itératif (1)



- L'OCR/HTR se composent donc de **nombreuses étapes**.
- **Comparer les résultats** est essentiel.
- Nécessite de nombreux allers-retours = **chronophage**.
- La plupart des OCR/HTR incluent des paramètres (=modèles) de base, qui permettent de transcrire sans entraîner...
- ... mais l'**entraînement est à privilégier** si l'on fait un long travail, comme dans le cas d'e-NDP.

Un processus itératif (2)

- Autre possibilité : fournir au logiciel un **dictionnaire**.
- Méthode intéressante, mais **risques de sur-corrections**.
- Pour les chartes de Notre-Dame, des essais sur **différents logiciels** :
  - Abbyy 14, Transkribus, Kraken, Tesseract, etc.

- Très chronophage : « images » (6 versions) * « traitements » (rien / dic / entr / dic + entr = 4) x « logiciels » (4) = 96 possibilités !
- **Quelle combinaison va donner les meilleurs résultats ?** C'est difficilement prévisible...

- **L'évaluation des résultats d'OCR/HTR est donc délicate** :
  - Erreurs sur les caractères vs. sur les mots
  - Sur-corrections
  - Problèmes de mise en page, etc.

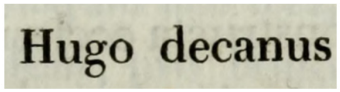
Difficultés de l'OCR d'une édition (diplomatique) (1)

(a) Essai sur la résolution

Cart. ND, n° 104

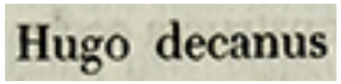
Hugo decanus et magister Theobaldus, canonicus Parisiensis, omnibus presentes litteras inspecturis salutem in Domino. Universitati vestre notum facimus quod, cum esset contentio inter venerabilem patrem P.[3], Parisiensem episcopum, ex una parte, et W.[4] Marmerel, militem, ex altera, idem W. dicebat se, auctoritate sua, posse recipere redditus sibi debitos ab hospitibus de terra de Prunoi; episcopus autem ex adverso dicebat quod dominium et justitia tota illius terre ad se pertinebat , quicumque eam teneret, ita quod predictus W. predictos redditus non debebat recipere, nisi per manus servientium episcopi; dicens quod in possessione hujus juris fuerat Mauricius, bone memorie episcopus, et Odo, antecessor suus, usque ad quamdam compositionem que facta fuerat inter predictum Odonem et predictum W., que dicitur facta fuisse per Odonem de Sancto Mederico et R. de Lineriis, quam compositionem nolebat episcopus sibi in aliquo prejudicium facere. Tandem compromissum est in nos hoc modo, ut inquireremus a testibus ex utraque parte productis, ut quem inveniremus in possessione hujus juris, tempore Mauricii et Odonis usque ad compositionem predictam , illi possessionem illius juris adjudicaremus , salvo jure proprietatis utrique parti.

2600 x 2000



Hugo decanus et magister 'fheobaldus, canonicus Parisiensis, omnibus presentes litteras inspecturis salutem in Domino. Universitati vcstre notum facimus quod, cum esset contentio inter venerabilem patrem P.', Parisiensem episcopum,'ex. una parte, et W.[4] Marmorei, militem, ex altera, idcmW. dicebat se, auctoritate sua, posse recipere redditus sibi debitos ab hospitibus de terra de Prunoi; episcopus autem ex adverso dicebat quod dominium et justitia tota illius terre ad se perlinebat, quicumque eam teneret, ita quod prediclus W. prcdictos redditus non debebat recipere, nisi per manus servientium episcopi; dicens quod in possessione hujus juris fuerat Mauricius, bone memorie episcopus, et Odo, antecessor suus, usque ad quamdam 'compositionem que facta fuerat inter predictum Odonem et prodictum W., que dicitur facta fuisse per Odonem de Sancto Mederico et R. de Lineriis, quam compositionem nolebat episcopus sibi in aliquo prejudicium facere. Tandem compromissum est in nos hoc modo, ut inquireremus a testibus ex utraque parte productis, ut quem inveniremus in possessione hujus juris, tempore Mauricii et Odonis usque ad compositionem predictam , illi possessionem illius juris adjudicaremus, salvo jure proprietatis utrique parti.

400 x 310

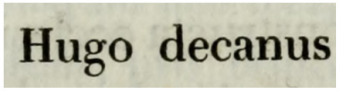


Difficultés de l'OCR d'une édition (diplomatique) (2)

(b) Essai sur la netteté
Cart. ND, n° 104

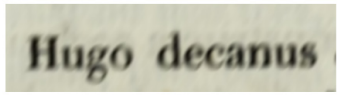
Hugo decanus et magister Theobaldus, canonicus Parisiensis, omnibus presentes litteras inspecturis salutem in Domino. Universitati vestre notum facimus quod, cum esset contentio inter venerabilem patrem P.[3], Parisiensem episcopum, ex una parte, et W.[4] Marmerel, militem, ex altera, idem W. dicebat se, auctoritate sua, posse recipere redditus sibi debitos ab hospitibus de terra de Prunoi; episcopus autem ex adverso dicebat quod dominium et justitia tota illius terre ad se pertinebat , quicumque eam teneret, ita quod predictus W. predictos redditus non debebat recipere, nisi per manus servientium episcopi; dicens quod in possessione hujus juris fuerat Mauricius, bone memorie episcopus, et Odo, antecessor suus, usque ad quamdam compositionem que facta fuerat inter predictum Odonem et predictum W., que dicitur facta fuisse per Odonem de Sancto Mederico et R. de Lineriis, quam compositionem nolebat episcopus sibi in aliquo prejudicium facere. Tandem compromissum est in nos hoc modo, ut inquireremus a testibus ex utraque parte productis, ut quem inveniremus in possessione hujus juris, tempore Mauricii et Odonis usque ad compositionem predictam , illi possessionem illius juris adjudicaremus , salvo jure proprietatis utrique parti.

Photographie nette

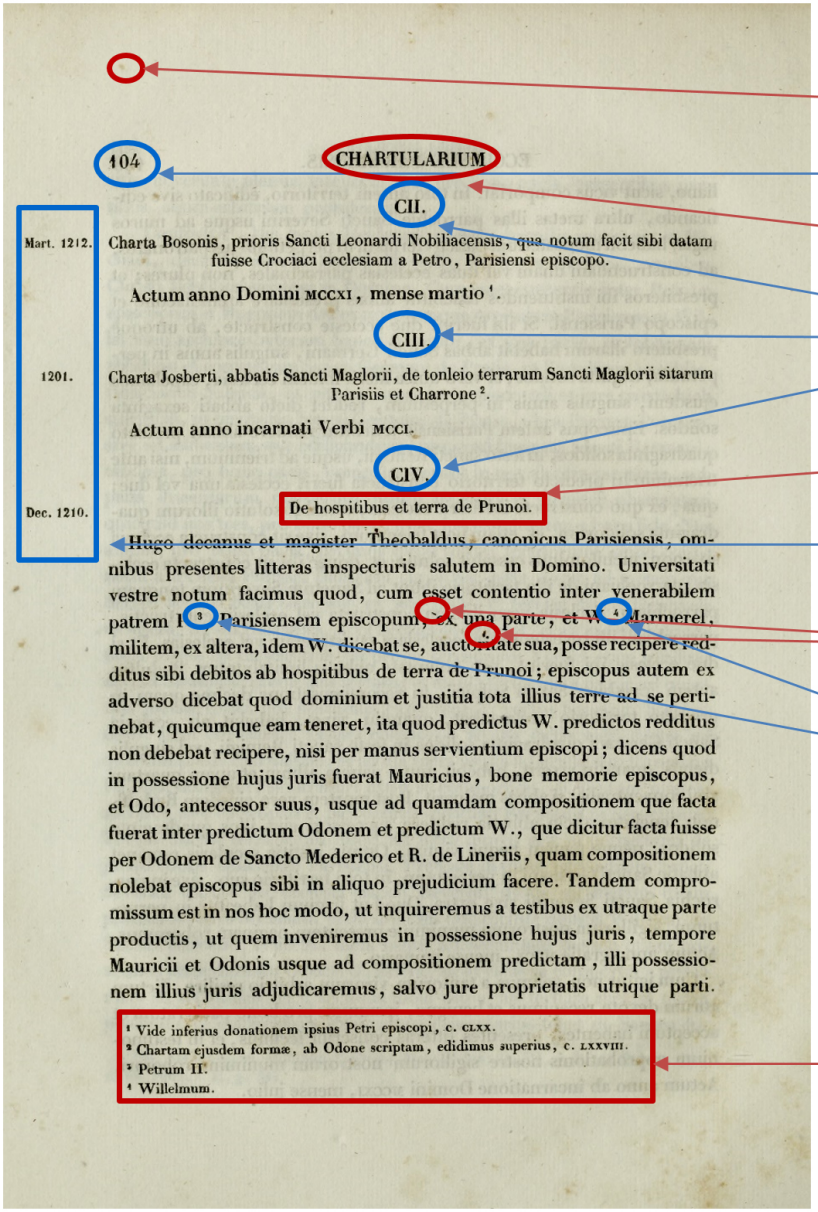


Ilugo decanus et magister fheobaldus, canonicus Parisiensis, omnibus presentes litteras inspecturi» salutem in Dornino, Universitati vestre notum facimus quod, cum esset contentio inter venerabilem *patrem* I*.', Parisiensem episcopum, ex, ima *parte* ,, et W/ Marmerel, militem, ex altera, idern W, <1 icebat se, auctoritatcsua,posscrecipere redditus sibi debitos ab hospitibus de terra de Prunoi; *epiwtpu»* autem ex adverso dicebat quod dominium et justitia tota illius terre ad s<† pertinebat , quicumque eam teneret, ita quod predictus W. predictos redditus non debebat recipere, nisi per manus servientium episcopi; dicens quod in possessione hujus juris fuerat Mauricius, Ixme memorie episcopus, et Odo, antecessor suus, usque ad quamdam compositionem que facta fuerat inter predictum Odonem et predictum W,, que dicitur facta fuisse per <Jdonern de Sancto Mederico et It de Lineriis, quam compositionem nolebat episcopus sibi in aliquo prejudiciurn *facere. Tandem compro-* missum est in nos hoc modo, ut inquireremus a testibus ex utraque parte productis, ut quem inveniremus in possessione hujus juris, tempore Mauricii et Odonis usque ad compositionem predictam , illi possessionem illius juris adjudicaremus, salvo jure proprietatis utrique parti,

Photographie floue



Difficultés de l'OCR d'une édition (diplomatique) (3)



104 CHARTULARIUM

CII.

Mart. 1212. Charta Bosonis, prioris Sancti Leonardi Nobiliacensis, qua notum facit sibi datam fuisse Crociaci ecclesiam a Petro, Parisiensi episcopo.

Actum anno Domini MCCXI, mense martio [1].

CIII.

1201. Charta Josberti, abbatis Sancti Maglorii, de tonleio terrarum Sancti Maglorii sitarum Parisiis et Charrone [2].

Actum anno incarnati Verbi MCCI.

CIV.

Dec. 1210. De hospitibus et terra de Prunoi.

Hugo decanus et magister Theobaldus, canonicus Parisiensis, omnibus presentes litteras inspecturis salutem in Domino. Universitati vestre notum facimus quod, cum esset contentio inter venerabilem patrem I. [3] Parisiensem episcopum, ex una parte, et W. [4] Marmerel, militem, ex altera, idem W. dicebat se, auctoritate sua, posse recipere redditus sibi debitos ab hospitibus de terra de Prunoi; episcopus autem ex adverso dicebat quod dominium et justitia tota illius terre ad se pertinebat, quicumque eam teneret, ita quod predictus W. predictos redditus non debebat recipere, nisi per manus servientium episcopi; dicens quod in possessione hujus juris fuerat Mauricius, bone memorie episcopus, et Odo, antecessor suus, usque ad quamdam compositionem que facta fuerat inter predictum Odonem et predictum W., que dicitur facta fuisse per Odonem de Sancto Mederico et R. de Lineriis, quam compositionem nolebat episcopus sibi in aliquo prejudicium facere. Tandem compromissum est in nos hoc modo, ut inquireremus a testibus ex utraque parte productis, ut quem inveniremus in possessione hujus juris, tempore Mauricii et Odonis usque ad compositionem predictam, illi possessionem illius juris adjudicaremus, salvo jure proprietatis utrique parti.

[1] Vide inferius donationem ipsius Petri episcopi, c. CLXX.
[2] Chartam ejusdem formæ, ab Odone scriptam, edidimus superius, c. LXXVIII.
[3] Petrum II.
[4] Willelmum.

taches sur le papier

numéros de page

en-tête

numéros des actes

analyses (en latin !)

dates

taches d'encre

appels de notes

(?) caractères mal formés (justitia)

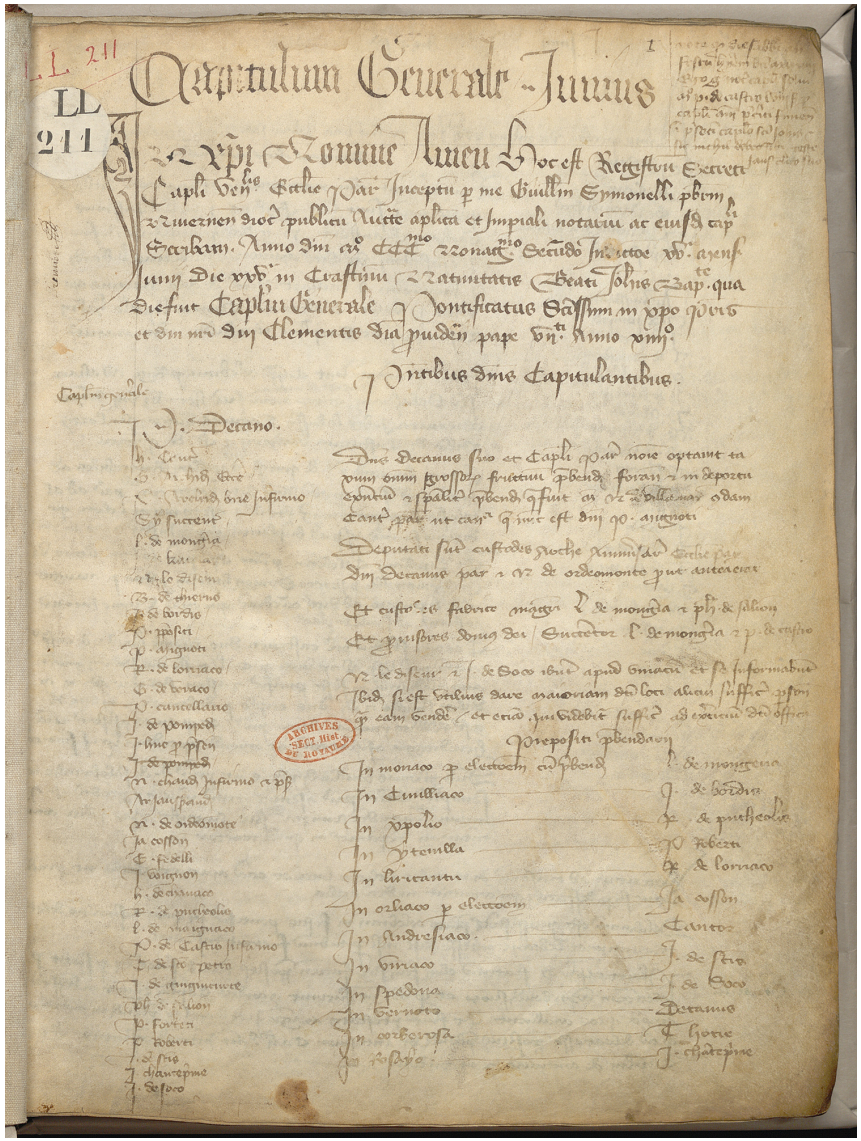
notes

Transition : de l'OCR à l'HTR

- **De l'OCR à l'HTR, les difficultés sont multipliées :**
  - Graphies très variables
  - Multiples mains dans les mss.
  - Irrégularités des lignes
  - Mises en page complexes (colonnes multiples)
  - Salissures, lettres effacées, textes cancellés, etc.
  - Numéros de folios, tampons, écritures modernes, etc.

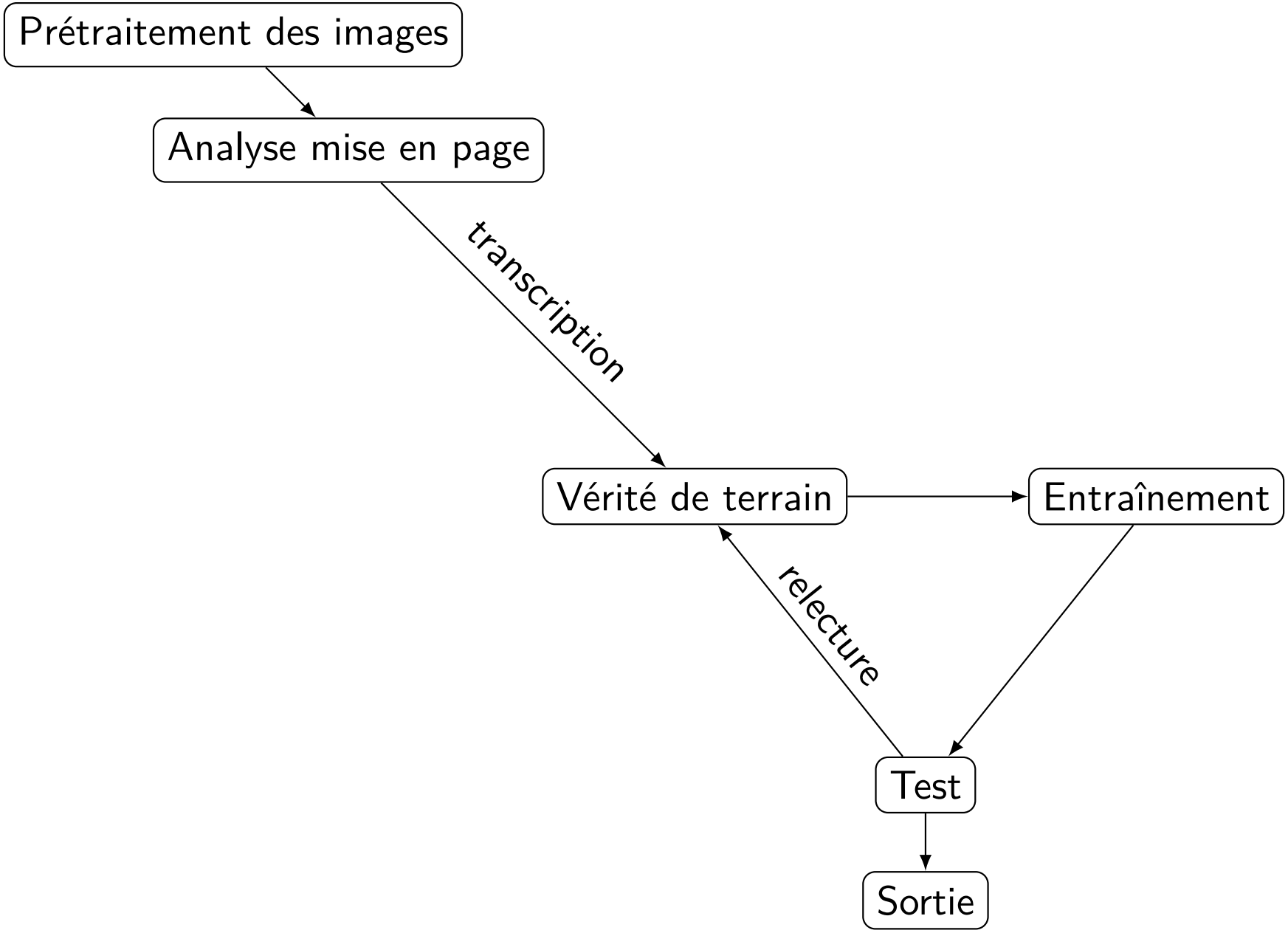
- Ce sont **des "évidences" pour les médiévistes, mais pas pour les machines**...

Ci-contre : Registre LL108, fol. 1r.



Plan

Une collaboration homme machine



Plan

Mise en page

Délimiter

surface des zones ou région (rectangles, polygones,...)

tracé des lignes d'écriture

Typer

zones texte principale, glose, initiale, signature,...

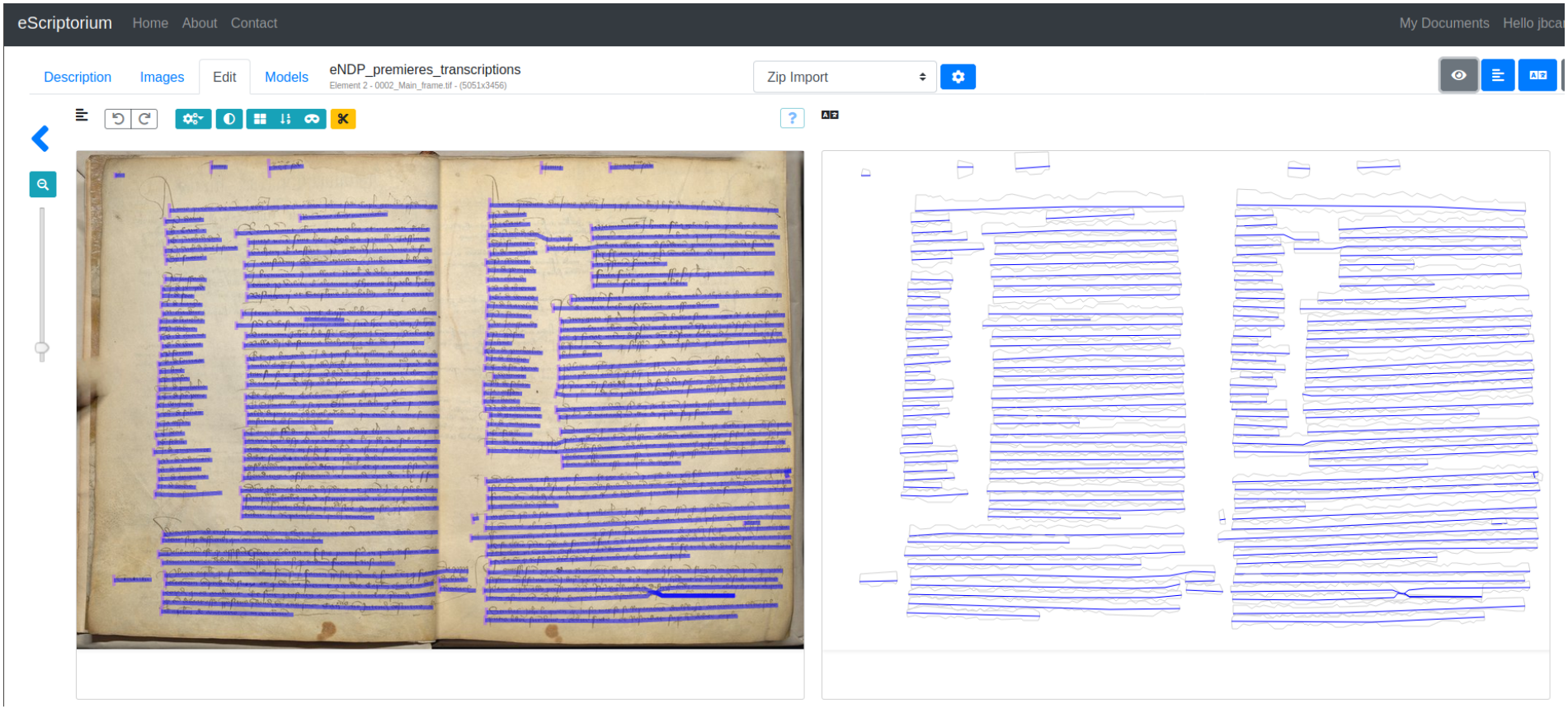
lignes défaut, ajout interlinéaire,...

Quelle ontologie ?

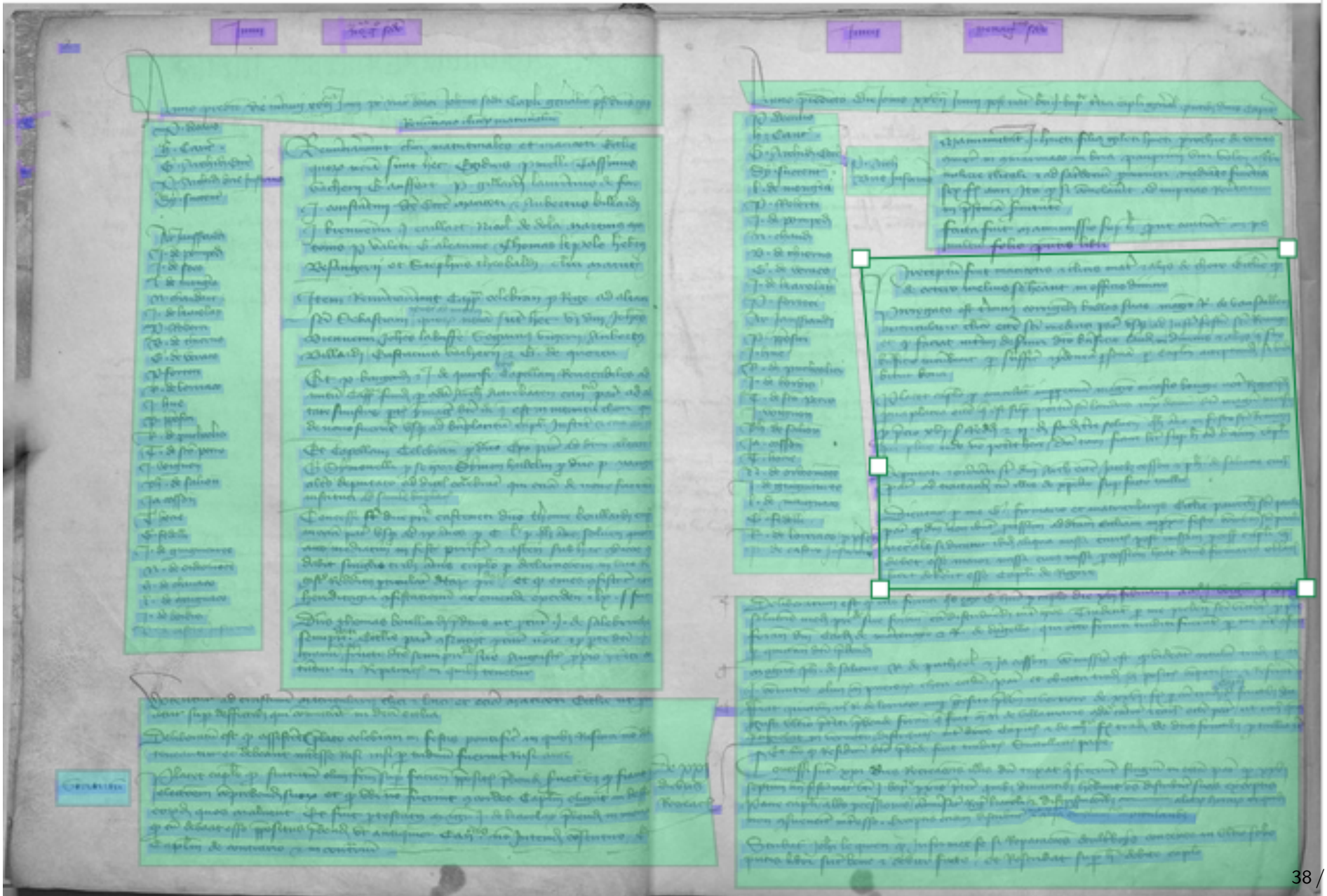
Projet SegmOnto (lancé en janvier dernier), `https://github.com/SegmOnto`.
→ recenser les cas, enregistrer des exemples et proposer une ontologie commune.

Détection des lignes

(Modèle CBAD par défaut)

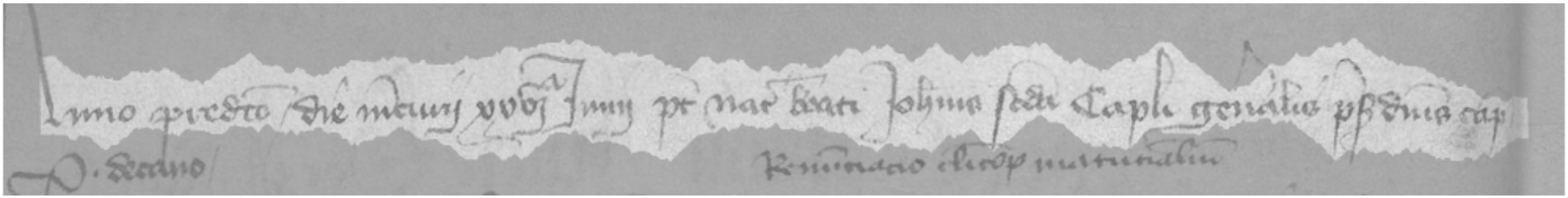


Zones : quelles délimitations et quels types ?

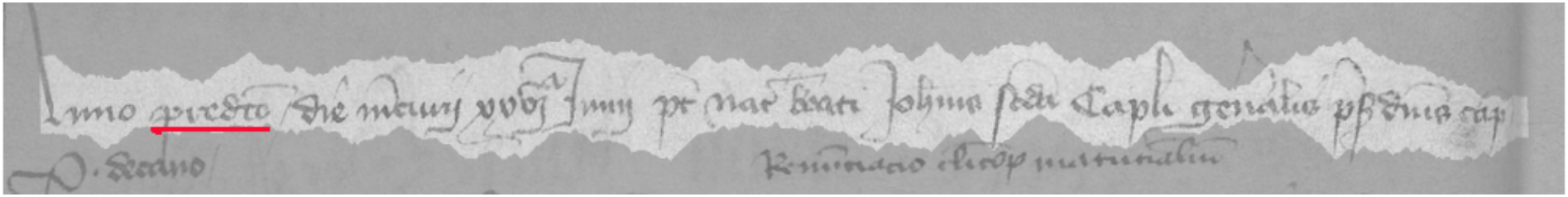


Plan

Que choisir ?



Que choisir ?



"Contenu" textuel

Sans bijection

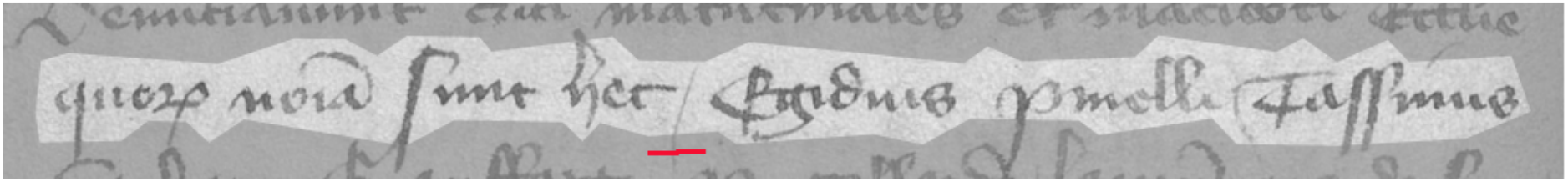
predicto contenu normalisé

Avec bijection (i.e., à un seul signe de la source correspond un seul signe de la transcription et réciproquement)

predtõ graphèmes (simplifiés)

preȣtõ allographes

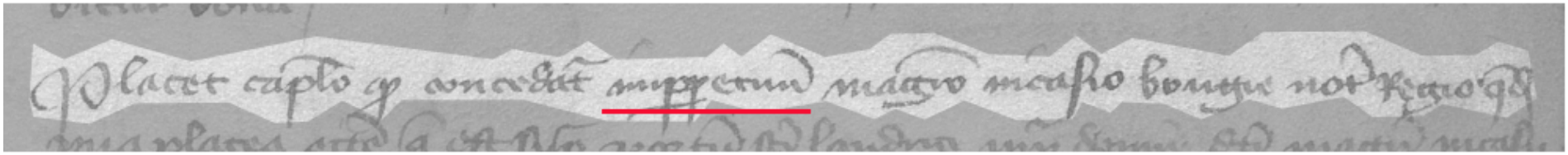
Que choisir ?



Ponctuation

- : modernisée ?
- , graphématique ?
- / allographétique ?

Que choisir ?



Segmentation

in perpetuum	modernisée ?
inperpetuum	présence/absence ?
inpp etuum	proportionnelle ?

Sur quels niveaux se place une transcription ?

Robinson et Solopova 1993

- **regularized** normalisation des graphies (i.e. Charles) ;
- **graphemic** graphies conservées (i.e. Carles) ;
- **graphetic** allographes conservées (i.e. carl.) ;
- **graphic** « every mark in the manuscript, every space, is represented in the transcription, even to the point of decomposition of letter forms into discrete marks ».

Choix terminologiques

graphemic ↦ graphématique

graphetic ↦ allographétique (cf. Stutzmann 2014).

Problématiques dans un projet d'OCR / HTR

Choix de transcription

1. Coller à la source ? (conserver les abréviations, la segmentation, la ponctuation...)
→ plus simple pour la reconnaissance
2. Se rapprocher d'un résultat normalisé ? Transcrire en résolvant les abréviations, etc.
→ besoin de plus de données pour des résultats (mais, cf. projet Himanis).
3. Distinguer les différentes étapes de ce traitement (prédiction, segmentation, normalisation, avec des outils distincts ((cf. Boudams, Pie,...).

Garantir l'homogénéïté des données

Très difficile dès que le nombre de transcripteurs > 1 (voire ≥ 1). Encore plus difficile dès qu'on s'écarte d'une transcription modernisée.

Plan

Une expérience avec des étudiants (ms. fr. 412 ; légendier en prose du XIIIe siècle)

Données d'entraînement

- 271 colonnes retranscrites par A. Pinche
  - Entraînement : 244 colonnes
  - Évaluation : (Kraken) 27 colonnes *Inconnu sur Transkribus*
  - Test : 27 colonnes
- 39 colonnes par des étudiantes
  - Variance parmi les transcriptions (caractères, abrev., segm.)
  - 8 transcriptrices
  - Non-Spécialistes

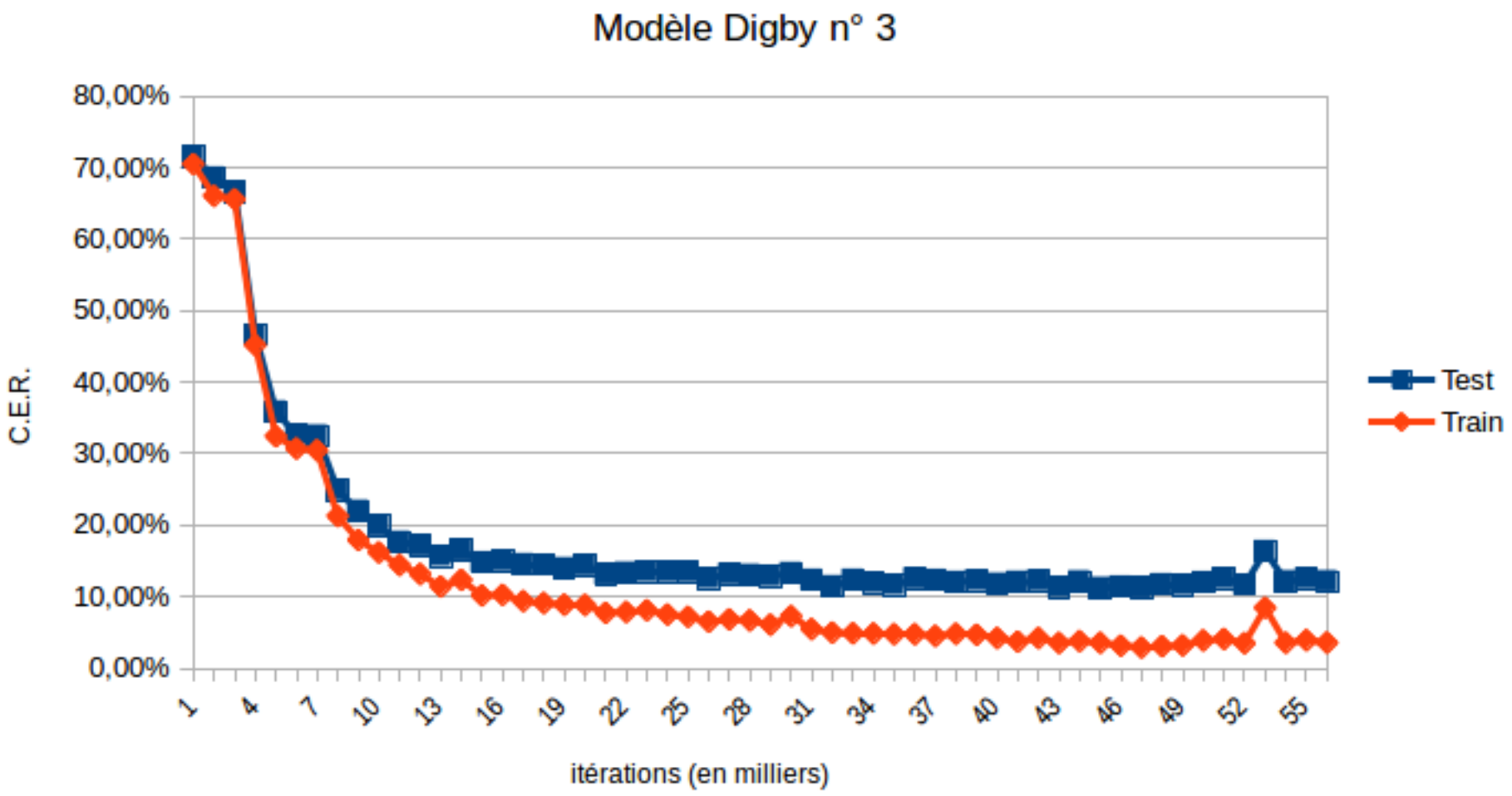
CER

	Test	*Étudiants*
Kraken	4.87	31.16
K. NoSpace	3.37	27.16
Transk HTR+	**2.29**	**8.07**

J.B. Camps, Th. Clérice & A. Pinche, *Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis*, `https://arxiv.org/abs/2012.03845`.

Entraînement

Entraînement sur un ms. (ici Roland d'Oxford)



Résultats

Facteurs

- quantité des données disponibles ;
- qualité des données disponibles ;
- difficulté de la tâche (nombre de mains, régularité de l'écriture, etc.)
- choix de transcription ;

Estimations très empiriques du taux d'erreur attendu

Pour une quantité substantielle de pages d'entraînement,

graphèmes et signes abréviatifs, une seule main $\rightarrow$ $<5\%$;

graphèmes et signes abréviatifs, plusieurs mains similaires $\rightarrow$ 15-5% ;

normalisé (pas de bijection) $\rightarrow$ $>15\%$.

Évaluation et confusions fréquentes

```
Loading model ../models-best/HTR/eNDP_49.mlmodel
Evaluating ../models-best/HTR/eNDP_49.mlmodel
=== report  ===
11142   Characters
1922    Errors
82.75\%  Accuracy
888     Insertions
282     Deletions
752     Substitutions

Errors  Correct-Generated
111     { i } - {  }
82      { e } - {  }
64      { t } - {  }
64      { n } - {  }
62      { s } - {  }
60      { m } - {  }
46      { SPACE } - {  }
```

Plan

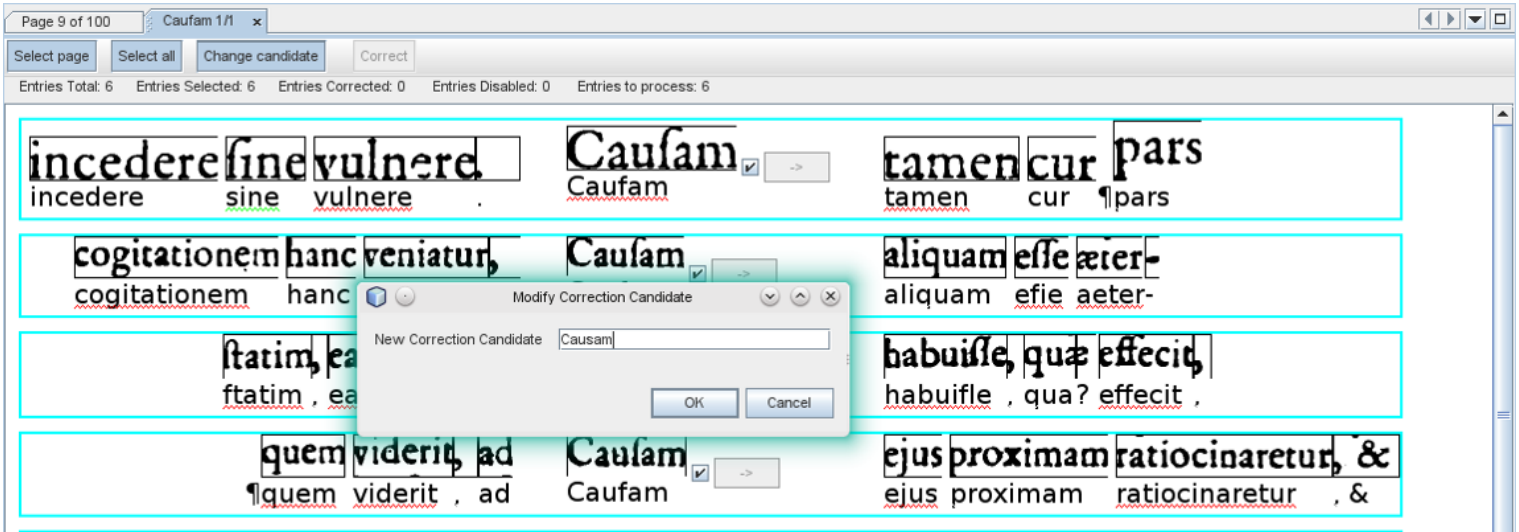
Le post-traitement de l'HTR

- De **nombreux traitements sont ensuite envisageables** sur les sorties d'HTR.
- Il faut distinguer entre :
  - Les traitement permettant d'améliorer le texte.
  - L'identification des différentes zones de texte (=dimension sémantique).
  - L'enrichissement du corpus.
- Ces post-traitements doivent être pensés en amont, et **ils dépendent de ce que l'on souhaite faire avec les "données"** (qui ne sont de fait, à ce stade, plus des "données", mais bien un résultat) :
  - Analyse chronologique, cartographique, sémantique, etc.

Améliorer le texte

- Il existe **différentes façons d'améliorer le texte** de l'HTR.
- On distingue en générale entre **corrections supervisées** et **corrections non-supervisées**.

- **L'approche supervisée** consiste à repérer les erreurs induites par l'OCR, et à les corriger, soit une par une, soit en « lots » :
  - Les logiciels d'OCR affichent la probabilité qu'un caractère ait été bien reconnu.
  - Le logiciel PoCoTo (Thorsten Vobl) permet de contrôler si une erreur est présente de multiples fois dans la sortie d'OCR et de les corriger simultanément (par ex. : « ct » => « et » ; « noturn » => « notum », etc.).
- **L'approche non-supervisée** consiste à utiliser des règles pour corriger.



L'identification (et le traitement) des zones

- **Les différentes zones** dans la mise en forme d'un manuscrit **sont souvent significatives**.
- Dans les registres de Notre-Dame : listes de chanoines, éléments de datation, commentaires, textes au sens strict, etc.

- On peut détecter ces différentes zones soit par leurs **caractères spatiaux**, soit par leur **contenu sémantique** :
  - Une zone haute et peu large a une forte probabilité d'être une liste de chanoines.
  - Une zone qui ne contient que des éléments de datation... à une forte probabilité d'être une date !
- **En combinant ces méthodes**, on pourrait mieux détecter le sens de la mise en page des registres, et ensuite **attribuer des métadonnées aux différentes zones** (par ex. tel texte se situant à droite d'une zone de datation, possède la métadonnée `date` en question, etc.).

L'enrichissement du corpus

- **Nombreuses possibilités** :
  - Détection des **entités nommées** (et attribution de coordonnées géographiques pour les toponymes, par une articulation à différents référentiels).
  - **Lemmatisation** des textes (via différents paramètres, développés dans plusieurs projets : ANR OMNIA et ANR VELUM, travaux de l'EDC, ERC LILLA, etc.)
  - **Structuration XML** reprenant l'ensemble des ces éléments, depuis la datation à la lemmatisation.

- Liées aux **traitements envisagés**. Trois exemples :
  - 1. la lemmatisation pour l'analyse sémantique et stylométrique ;
  - 2. l'identification d'entités nommées (toponymes) pour les SIG ;
  - 3. l'identification d'entités nommées (anthroponymes) et la formalisation pour l'analyse de réseaux (Gephi), etc.

- Enfin, il ne faut pas oublier que les données e-NDP pourront **enrichir en retour de futurs projets** :
  - modèles pour l'HTR ; lemmatisation du latin tardo-médiéval ; analyse des zones dans les manuscrits, etc.

Plan

Plan

Exporter et convertir ses données de Transkribus

Objectif

- Récupérer la vérité de terrain produite avec Transkribus ;
- pour la réutiliser avec une autre interface ;
- ou pour s'en servir de matériaux d'entraînement.

Les alternatives

- eScriptorium ;
- Kraken (Scripta | PSL).

Le format pivot

`Alto` (mais avec des spécificités).

Aspyre GT

Outil de conversion d'Alto-2 (Transkribus) à Alto-4 (eScriptorium/Kraken),

- dév. par Alix Chagué (INRIA)
- **Code Python** : `https://gitlab.inria.fr/dh-projects/aspyre-gt` ;
- **GUI web** : `https://aspyre-gui.herokuapp.com/`.

Aspyre GT

```
# Pour l'installation, consulter le README
# Conversion des Alto
python3 aspyre/run.py -i export_transkr -o output_Aspyre -t
# Si plusieurs pages, ziper ensuite en un fichier unique
zip output_Aspyre.zip output_Aspyre/* -j
```

Plan

Le moteur : différentes solutions techniques

- approches segmentées (avec *template matching*) ou non segmentées ;
- mesures de distance ; méthodes statistiques (chaînes de Markov, CRF) ou d'intelligence artificielle (réseaux de neurones convolutifs ou récurrents, LSTM 1D, LSTM 2D, etc.) ;
- outils directement opérationnels ou nécessitant un entraînement.

Ocropy, CLSTM et Kraken

`OCRopy` et CLSTM développés par Thomas M. Breuel ; `Kraken`, *fork* d'`OCRopy` développé par Ben Kiessling (PSL).

- approche non segmentées ;
- réseaux de neurones récurrents (LSTM) ;
- *open source* et nécessitant l'entraînement d'un modèle.

Exemple d'approche fondée sur l'apprentissage machine

- Les lignes sont toutes normalisées à une hauteur donnée ;
- une fenêtre de largeur 1 px, et de la hauteur normalisée parcourt l'image et les valeurs de ses pixels servent d'entrée au réseau de neurones ;
- qui doit produire en sortie, via une couche de "décodage" (CTC), la transcription de la ligne.

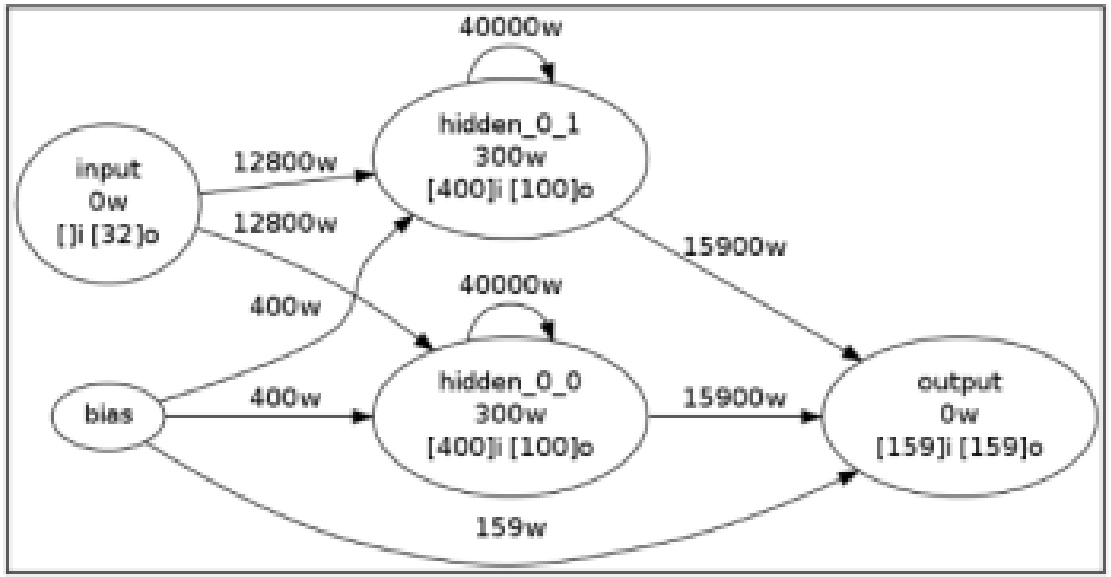


FIGURE – OCRopy BLSTM network (Breul et al. 2013)

Plan

Références

- Transkribus, `http://transkribus.eu/` ;
- Kraken (*fork* d'OCRopy, développé par Ben Kiessling ; dév. du projet Scripta PSL), `http://kraken.re/` et `https://github.com/mittagessen/kraken/` ;
- `eScriptorium` : `https://gitlab.inria.fr/scripta/escriptorium`.