

De la donnée avant toute chose ?

De la donnée avant toute chose ? Retour d'expérience de l'utilisation de l'HTR dans des projets d'édition et d'étude des textes médiévaux

Journées d'Études « De la transcription manuelle participative des textes à la reconnaissance automatique de texte (ATR/HTR) : outils, théorie, pratiques. »

Matthias GILLE LEVENSON

École nationale des chartes – Centre Jean Mabillon & École Normale Supérieure de Lyon – CIHAM UMR 5648

Nancy, 10 septembre 2024

2024-09-10

Matthias GILLE LEVENSON

De la donnée avant toute chose ? Retour d'expérience de l'utilisation de l'HTR dans des projets d'édition et d'étude des textes médiévaux

Journées d'Études « De la transcription manuelle participative des textes à la reconnaissance automatique de texte (ATR/HTR) : outils, théorie, pratiques. »

Matthias GILLE LEVENSON

École nationale des chartes – Centre Jean Mabillon & École Normale Supérieure de Lyon – CIHAM UMR 5648

Nancy, 10 septembre 2024

2024-09-10

# Plan

## 1 Introduction

## 2 Phase de production des données

- Penser la production en amont
- CATMuS, un projet de production collaboratif de données d'ATR

## 3 Après l'HTR : tout change / rien ne change

- Quand s'arrête la correction ?
- Structurer les documents
- Identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

## 4 Conclusions

De la donnée avant toute chose ?

2024-09-10

└ Plan

Plan

- Introduction
- Phase de production des données
  - Penser la production en amont
  - CATMuS, un projet de production collaboratif de données d'ATR
- Après l'HTR : tout change / rien ne change
  - Quand s'arrête la correction ?
  - Structurer les documents
  - Identifier la césure à la ligne
  - Résoudre les abréviations
  - Et l'édition critique ?
- Conclusions

## 1 Introduction

## 2 Phase de production des données

## 3 Après l'HTR : tout change / rien ne change

## 4 Conclusions

## Introduction

2024-09-10

- Introduction
- Phase de production des données
- Après l'HTR : tout change / rien ne change
- Conclusions

# Expériences personnelles avec l'HTR

- Pour de l'édition
- Pour l'étude proprement dite du texte
- Actuellement, pour des expériences de collation multilingue

## Idée de la communication

- Proposer un retour d'expérience
- Donner des pistes et mes idées sur les bonnes pratiques en ATR dans une perspective plus globale d'étude et de traitement du texte patrimonial
- De la donnée avant toute chose? Changement de paradigme avec l'apparition de l'apprentissage supervisé : il faut assumer ces changements et intégrer la nouvelle donne (ou la nouvelle donnée). Avant toute chose est donc à prendre au sens chronologique et non pas logique : la production du savoir et du texte reste prédominante.

# Édition critique

E deuen fazer otra puerta de fierro que llaman puerta de traycion con grandes fierros ayuso. E deue ser foradada la torre por que se pueda sobir e desçender con cadenas de fierro e deuese fazer ante de las puertas principales<sup>[R: fol. 279v]</sup> por que non las pueda quemar. [A, III-3-20, traduction, fol. 266r, éd. p. 647]

2 desçender *ABRQJZ* | defender *G* 2 deuese *ABGRJZ* | dévense *Q* 2 deuese *BAQJZ* | [dévese *BGRJAZ* | dévense *Q*] [ø *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om. Q* 3 pueda *AZ* | puedan *BGRQJ*

Édition critique du *Regimiento de los Príncipes*, issue de collation automatisée. Les témoins A et Z sont issus d'HTR : (GILLE LEVENSON 2023a) et (GILLE LEVENSON 2023b)

2024-09-10  
De la donnée avant toute chose?  
└ Introduction  
└ Édition critique

E deuen fazer otra puerta de fierro que llaman puerta de traycion con grandes fierros ayuso. E deue ser foradada la torre por que se pueda sobir e desçender con cadenas de fierro e deuese fazer ante de las puertas principales<sup>[R: fol. 279v]</sup> por que non las pueda quemar. [A, III-3-20, traduction, fol. 266r, éd. p. 647]

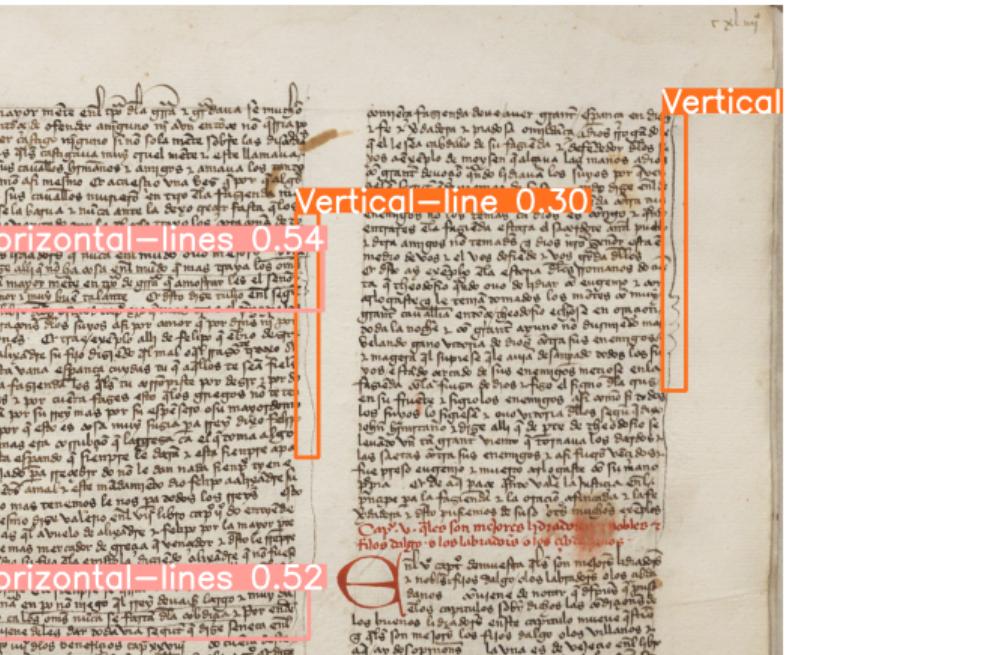
2 desçender *ABRQJZ* | defender *G* 2 deuese *ABGRJZ* | dévense *Q* 2 deuese *BAQJZ* | [dévese *BGRJAZ* | dévense *Q*] [ø *BAQJZ* | de *GR*] 3 de *BAGRJZ* | *om. Q* 3 pueda *AZ* | puedan *BGRQJ*

Édition critique du *Regimiento de los Príncipes*, issue de collation automatisée. Les témoins A et Z sont issus d'HTR : (GILLE LEVENSON 2023a) et (GILLE LEVENSON 2023b)

Idée de la communication

- Travail sur une traduction castillane du *De Regimine Principum* de Gilles de Rome.
- Je parle depuis le point de vue d'un médiéviste, d'un philologue et d'un éditeur de textes patrimoniaux

# Études de la réception d'un manuscrit par ses marques de lecture



Identification automatisée avec YOLO v5 (REDMON 2016) de zones de texte marquées par un lecteur. Escorial  
Ms. K.I.5, fol. 144r

## └ Introduction

### └ Études de la réception d'un manuscrit par ses marques de lecture

2024-09-10



Identification automatisée avec YOLO v5 (REDMON 2016) de zones de texte marquées par un lecteur. Escorial  
Ms. K.I.5, fol. 144r

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, transcription

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité

# Rappels sur le fonctionnement (actuel) de l'HTR

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité
- Avec Kraken (KIESSLING 2019), pas de transcription globale du texte mais ligne par ligne

- Trois phases distinctes : segmentation en **zones**, segmentation en **lignes**, **transcription**
- Respecter les phases et bien découper le travail permet de gagner du temps *in fine* et de produire des données de meilleure qualité
- Avec Kraken (KIESSLING 2019), pas de transcription globale du texte mais ligne par ligne

## 1 Introduction

## 2 Phase de production des données

- Penser la production en amont
- CATMuS, un projet de production collaboratif de données d'ATR

## 3 Après l'HTR : tout change / rien ne change

## 4 Conclusions

2024-09-10

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds

- └ Phase de production des données
- └ Penser la production en amont
- └ Penser la production en amont

2024-09-10

- Question de pérennité : les données restent, les modèles disparaissent.

## Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :

- Phase de production des données
  - Penser la production en amont
    - Penser la production en amont

2024-09-10

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :

## Penser la production en amont

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?

- └ Phase de production des données
  - └ Penser la production en amont
  - └ Penser la production en amont

2024-09-10

- Différencier et **hiérarchiser données et modèles** et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?

- Question de pérennité : les données restent, les modèles disparaissent.

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?

De la donnée avant toute chose ?

Phase de production des données

Penser la production en amont

Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?

De la donnée avant toute chose ?

Phase de production des données

Penser la production en amont

Penser la production en amont

2024-09-10

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

De la donnée avant toute chose ?

Phase de production des données

Penser la production en amont

Penser la production en amont

2024-09-10

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

- Question de pérennité : les données restent, les modèles disparaissent.

# Penser la production en amont

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

1-23 fechura no deue paran mjero  
 1-24 ala color \$da q qere\$ a los  
 1-25 fralcons q soy cntrados o  
 1-26 f Fuara a manallos.  
 1-27 oq<ue> torna contra umneio prriua

Un cas de modèle produit à partir de données hétérogènes, identifié par J.M. Fradejas (PINCHE et al. 2024)

## De la donnée avant toute chose ?

- Phase de production des données
- Penser la production en amont
- Penser la production en amont

2024-09-10

- Différencier et hiérarchiser données et modèles et privilégier les premières sur les seconds
- Se mettre d'accord en amont de la production sur des normes d'annotation :
  - Quelle typologie des zones et des lignes utiliser ? Identifier les titres de section ? Identifier toutes les zones ?
  - Comment transcrire ?
  - Que faire des abréviations ?
- Éviter à tout prix les modèles « boule de neige » :

fechura no deue paran mjero  
 ala color \$da q qere\$ a los  
 fralcons q soy cntrados o  
 f Fuara a manallos.  
 oq<ue> torna contra umneio prriua

Un cas de modèle produit à partir de données hétérogènes, identifié par J.M. Fradejas (PINCHE et al. 2024)

Sur la fin que fu il en hoot ier estut vierge devant le festendà delos oños es ca mejor prodere  
a la p[re]mier p[er]sonne des m[es]me que et si nora el p[ro]p[ri]et[é] onferme del signe io el fisiò en a  
m[on]teur du bien que au temps se pert en b[ea]tent idem. Pour ce q[ue] p[ro]toit h[ab]it. Eala vero  
disciplina gheince sive le fais Dieudale vero. potestuenter, la  
cune cause elle r[es]t. Imē god tu te puise ioue q[ue] la f[or]g[iv]e[re]z voint eu a fait co  
q[ue] tellos fauallorat q[ue] l'oreste t[ra]nscrit. ne n[on] es gheen v[e]lue si hi p[ro]me h[ab]e[re] mo  
un j[ou]g an h[ab]ileste. en t[ra]xeler li oreste d'alle sit ale p[ro]leone p[er] crudelitate  
serie turum ne r[es]p. dier belles de la sereur qui deute faire supplicatio[n]e. et deute est en



**CATMuS  
Medieval**

classe et cramece sive data foraleja que i en que top[er]. Et d[icit] la soliloq[ue] de la re[al]le d[omi]ne in. Il  
ensamble onostre aymant de lu la serial acord. v[er]at du heu ma manca es  
iv En attie u[er] dat hochste les r[es] de bone o conlas manos por sus se contre top[er]  
urz par hoye. Sicut et emam ou p[ro]p[ri]et[é] et gherem[us] vider Edicte lo quae fu illib[er]a  
scordia. Et stando on[us] remunerat amicis q[ue] ghe[n]te omni eris. e[st] pro p[ro]p[ri]et[é] mas aqu[us] ro  
faemlyde ghe[n]te q[ue] h[ab]ent deu[el]lo. No p[er]fessio. In senecte d[omi]ni magis. I tercero d[omi]ni alma d[omi]n[us]  
re dicha q[ue] sive ad p[re]dictum. et luanu et filio. m[on]tes qui creuerunt monachos q[ue] los principes en e  
tence et bone. p[er] d[omi]n[u]m. L'omnia abucais quis iprest a my desen te comene. v[er]at

- Volonté autour de 2022 de réunir des producteur.ices de données (philologues) venant d'horizons distincts
- Naissance du projet CATMuS pour « *Consistent Approaches for Transcribing Manuscripts* »
- Le corpus est récemment publié (CLÉRICE et al. 2024)

## De la donnée avant toute chose?

### Phase de production des données

#### CATMuS, un projet de production collaboratif de données d'ATR

2024-09-10



- Volonté autour de 2022 de réunir des producteur.ices de données (philologues) venant d'horizons distincts
- Naissance du projet CATMuS pour « *Consistent Approaches for Transcribing Manuscripts* »
- Le corpus est récemment publié (CLÉRICE et al. 2024)



# Statistiques sur le corpus CATMuS

- Publié sur HuggingFace : <https://huggingface.co/datasets/CATMuS/medieval>
- Recueil de 17 dépôts différents **homogénéisés** selon les normes du projet
- Environ 175.000 lignes transcrites et intégralement corrigées (*corpus gold*)
- 245 documents différents en 9 langues
- Du IX<sup>e</sup> au XV<sup>e</sup> siècle, avec une prédominance du bas Moyen Âge (XIII<sup>e</sup>-XV<sup>e</sup> siècles)
- Corpus encore biaisé en raison de l'histoire du projet

## De la donnée avant toute chose ?

- └ Phase de production des données
  - └ CATMuS, un projet de production collaboratif de données d'ATR
    - └ Statistiques sur le corpus CATMuS

2024-09-10

- Publié sur HuggingFace : <https://huggingface.co/datasets/CATMuS/medieval>
- Recueil de 17 dépôts différents **homogénéisés** selon les normes du projet
- Environ 175.000 lignes transcrites et intégralement corrigées (*corpus gold*)
- 245 documents différents en 9 langues
- Du IX<sup>e</sup> au XV<sup>e</sup> siècle, avec une prédominance du bas Moyen Âge (XIII<sup>e</sup>-XV<sup>e</sup> siècles)
- Corpus encore biaisé en raison de l'histoire du projet

# Statistiques sur le corpus CATMuS

- Publié sur HuggingFace : <https://huggingface.co/datasets/CATMuS/medieval>
- Recueil de 17 dépôts différents homogénéisés selon les normes du projet
- Environ 175.000 lignes transcrites et intégralement corrigées (corpus *gold*)
- 245 documents différents en 9 langues
- Du IX<sup>e</sup> au XV<sup>e</sup> siècle, avec une prédominance du bas Moyen Âge (XIII<sup>e</sup>-XV<sup>e</sup> siècles)
- Corpus encore biaisé en raison de l'histoire du projet

Split	Software	Training Time	Character Error Rate (%)		
			Validation	Test	Space-related
General	Kraken	2112 min ± 163	5.7 ± 0.07	4.7 ± 0.06	1.0 ± 0.02
Feature	Kraken	1464 min ± 238	6.8 ± 0.16	13.1 ± 0.24	2.7 ± 0.06
General	Pylaia	308 min ± 047	9.1 ± 0.63	8.4 ± 0.73	1.8 ± 0.11
Feature	Pylaia	295 min ± 078	11.3 ± 0.24	21.2 ± 0.92	3.8 ± 0.06

Modèles produits à partir des données de CATMuS. (CLÉRICE et al. 2024, p. 16)

## De la donnée avant toute chose ?

- └ Phase de production des données
- └ CATMuS, un projet de production collaboratif de données d'ATR
  - └ Statistiques sur le corpus CATMuS

2024-09-10

## Statistiques sur le corpus CATMuS

Split	Software	Training Time	Character Error Rate (%)		
			Validation	Test	Space-related
General	Kraken	2112 min ± 163	5.7 ± 0.07	4.7 ± 0.06	1.0 ± 0.02
Feature	Kraken	1464 min ± 238	6.8 ± 0.16	13.1 ± 0.24	2.7 ± 0.06
General	Pylaia	308 min ± 047	9.1 ± 0.63	8.4 ± 0.73	1.8 ± 0.11
Feature	Pylaia	295 min ± 078	11.3 ± 0.24	21.2 ± 0.92	3.8 ± 0.06

Modèles produits à partir des données de CATMuS. (Clérice et al. 2024, p. 16)

# Une pomme de discorde : les abréviations

Notre point de vue est celui de philologues :

- Nous considérons la résolution des abréviations comme une tâche de TAL plutôt que de vision assistée par ordinateur (CLÉRICE et al. 2024)
- La résolution via ATR pose des problèmes de **généralisation** et d'**adaptation**.
  - Généralisation : le développement des abréviations peut poser problème dans le cadre de corpus multilingue ; les résultats sont moins bons en développant les abréviations
  - Adaptation : le développement des abréviations est dépendant du contexte linguistique et historique du document.

2024-09-10

De la donnée avant toute chose ?

└ Phase de production des données

└ CATMuS, un projet de production collaboratif de données d'ATR

└ Une pomme de discorde : les abréviations

Notre point de vue est celui de philologues :

- Nous considérons la résolution des abréviations comme une tâche de TAL, plutôt que de vision assistée par ordinateur (CLÉRICE et al. 2024)
- La résolution via ATR pose des problèmes de généralisation et d'adaptation.

# Le choix de la conservation des abréviations

Une norme de transcription **graphématisque** (STUTZMANN 2010) :

- Réduction des allographes au graphème
- Réduction des allographes *<i>/<j>* et *<u>/<v>* à *<i>* et à *<u>*
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>

De la donnée avant toute chose ?

- └ Phase de production des données
- └ CATMuS, un projet de production collaboratif de données d'ATR
  - └ Le choix de la conservation des abréviations

2024-09-10

Le choix de la conservation des abréviations

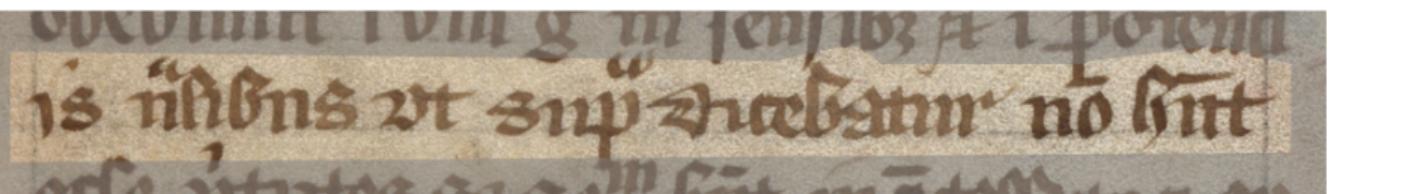
Une norme de transcription **graphématisique** (STUTZMANN 2010) :

- Réduction des allographes au graphème
- Réduction des allographes *<i>/<j>* et *<u>/<v>* à *<i>* et à *<u>*
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>

# Le choix de la conservation des abréviations

Une norme de transcription **graphématisque** (STUTZMANN 2010) :

- Réduction des allographes au graphème
- Réduction des allographes *<i>/<j>* et *<u>/<v>* à *<i>* et à *<u>*
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>



is <sup>a</sup>n̄libus ut sup̄ dicebatur nō h̄nt

De la donnée avant toute chose ?

- └ Phase de production des données
- └ CATMuS, un projet de production collaboratif de données d'ATR
- └ Le choix de la conservation des abréviations

2024-09-10

- Réduction des allographes au graphème
- Réduction des allographes *<i>/<j>* et *<u>/<v>* à *<i>* et à *<u>*
- Non développement des abréviations
- Alignement sur les caractères proposés par la MUFI (HAUGEN 2013) :  
<https://mufi.info/>

is n̄libus ut sup̄ dicebatur nō h̄nt  
is n̄libus ut sup̄ dicebatur nō h̄nt

# Concilier l'intérêt particulier et les besoins généraux

- Cette solution nous semble être le plus à même de **concilier besoins généraux et particuliers**
- La question de l'**homogénéité des données** est fondamentale : manuel d'annotation et outils de contrôle (<https://github.com/PonteIneptique/choco-mufin> et <https://github.com/HTR-United/HTRVX>)
- La contrepartie est la nécessité de **travailler en aval** de l'acquisition du texte pour normaliser le **texte**
  - **Transition** : produire des données, c'est bien, mais comment faire après ? Le travail n'est pas du tout fini. Comment corriger le texte ? Comment gérer la segmentation ? Les abréviations ? Qu'est-ce qui change pour l'édition ?
- Le manuel en ligne est disponible : <https://catmus-guidelines.github.io/> (en cours de rédaction)

2024-09-10

- Cette solution nous semble être le plus à même de concilier besoins généraux et particuliers
- La question de l'**homogénéité des données** est fondamentale : manuel d'annotation et outils de contrôle (<https://github.com/PonteIneptique/choco-mufin> et <https://github.com/HTR-United/HTRVX>)
- La contrepartie est la nécessité de travailler en aval de l'acquisition du texte pour normaliser le texte
- Le manuel en ligne est disponible : <https://catmus-guidelines.github.io/> (en cours de rédaction)

## 1 Introduction

## 2 Phase de production des données

## 3 Après l'HTR : tout change / rien ne change

- Quand s'arrête la correction ?
- Structurer les documents
- Identifier la césure à la ligne
- Résoudre les abréviations
- Et l'édition critique ?

## 4 Conclusions

2024-09-10

# Quand s'arrête la correction ?

- La phase suivant la transcription automatisée sera généralement celle de la transformation en TEI
- Plusieurs outils permettent de réaliser cette transformation :
  - <https://github.com/Jean-Baptiste-Camps/ALTEI>,
  - <https://github.com/chartes/alto2tei>,
  - [https://github.com/matgille/alto\\_to\\_teii/](https://github.com/matgille/alto_to_teii/)
- Il restera des erreurs dans les données
- Faut-il intégrer les corrections faites dans les données d'entraînement ?
- Si oui, cela suppose de penser en amont une modélisation en TEI qui soit **rétroconvertible**
- En d'autres termes, il faudra conserver un premier état de TEI pseudo-diplomatique (conservation des tei:lb)

- └ Après l'HTR : tout change / rien ne change
  - └ Quand s'arrête la correction ?
    - └ Quand s'arrête la correction ?

2024-09-10

- La phase suivant la transcription automatisée sera généralement celle de la transformation en TEI
- Plusieurs outils permettent de réaliser cette transformation :
  - <https://github.com/Jean-Baptiste-Camps/ALTEI>,
  - <https://github.com/chartes/alto2tei>,
  - [https://github.com/matgille/alto\\_to\\_teii/](https://github.com/matgille/alto_to_teii/)
- Il restera des erreurs dans les données
- Faut-il intégrer les corrections faites dans les données d'entraînement ?
- Si oui, cela suppose de penser en amont une modélisation en TEI qui soit **rétroconvertible**
- En d'autres termes, il faudra conserver un premier état de TEI pseudo-diplomatique (conservation des tei:lb)

# Assurer la rétroconvertibilité

```

<lb break="yes" facs="#facs_eSc_line_92b25bc7" xml:id="elem_eSc_line_92b25bc7">cū uegetablib- γ plātis ut pō
<lb break="?" facs="#facs_eSc_line_4fc0c0e" xml:id="elem_eSc_line_4fc0c0e"/>nūttia augm̄tia. ḡnatia γ
<lb break="?" facs="#facs_eSc_line_1e32f48c" xml:id="elem_eSc_line_1e32f48c"/>tlia q γ ip̄s arborib- c̄petūt.
<lb break="?" facs="#facs_eSc_line_2fddd609" xml:id="elem_eSc_line_2fddd609"/>po γ cognit̄e sfit̄e sūt uis
<lb break="yes" facs="#facs_eSc_line_b1fb0fe4" xml:id="elem_eSc_line_b1fb0fe4"/>gust⁹ γ tactus. γ tlia in qb- q
<lb break="?" facs="#facs_eSc_line_63be87f7" xml:id="elem_eSc_line_63be87f7"/>cam⁹ cū brutis. appetitie ū di
<lb break="?" facs="#facs_eSc_line_2304726d" xml:id="elem_eSc_line_2304726d"/>stīgūt. nā qdam ē appetit⁹ i
<lb break="?" facs="#facs_eSc_line_9e3287cc" xml:id="elem_eSc_line_9e3287cc"/>hoie i q n̄ qcat cū brutis ut
<lb break="?" facs="#facs_eSc_line_4f2eeff0c" xml:id="elem_eSc_line_4f2eeff0c"/>appetit⁹ seq̄s itllem. Qdā ū
<lb break="yes" facs="#facs_eSc_line_887a268e" xml:id="elem_eSc_line_887a268e"/>i q qcat cū eis appetit⁹ seq̄s se
<lb break="?" facs="#facs_eSc_line_3c8f19df" xml:id="elem_eSc_line_3c8f19df"/>sū. Appetit⁹ āt seq̄s sfm pt
<lb break="?" facs="#facs_eSc_line_cd2f421c" xml:id="elem_eSc_line_cd2f421c"/>noīari sfualitas. seq̄s itlē
<lb break="?" facs="#facs_eSc_line_5d351c82" xml:id="elem_eSc_line_5d351c82"/>ctū noīē uolūtas. f q̄mod lo-
<lb break="?" facs="#facs_eSc_line_fc03f518" xml:id="elem_eSc_line_fc03f518"/>q̄di bruta h̄- sfualitate γ appen-
<lb break="?" facs="#facs_eSc_line_3ecf893f" xml:id="elem_eSc_line_3ecf893f"/>titū sfit̄im. s- n̄ uolūtate h̄-

```

Le document TEI avec des identifiants présents dans le fichier ALTO original

## De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Quand s'arrête la correction ?
- └ Assurer la rétroconvertibilité



Le document TEI avec des identifiants présents dans le fichier ALTO original

## • PASSER

# Assurer la rétroconvertibilité

```

-<TextLine ID="eSc_line_4fc80c0e" TAGREFS="LT6426" BASELINE="1581 1374 2153 1366" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0">
+<Shape></Shape>
<String CONTENT="nuttia augm̄tia. ghatia γ" HPOS="1579.0" VPOS="1322.0" WIDTH="574.0" HEIGHT="81.0"/>
</TextLine>
-<TextLine ID="eSc_line_1e32f48c" TAGREFS="LT6426" BASELINE="1573 1434 2164 1421" HPOS="1571.0" VPOS="1375.0" WIDTH="593.0" HEIGHT="89.0">
+<Shape></Shape>
<String CONTENT="tlia ̄ q ̄ ip̄is arborib̄ cōpetūt." HPOS="1571.0" VPOS="1375.0" WIDTH="593.0" HEIGHT="89.0"/>
</TextLine>
-<TextLine ID="eSc_line_2fddd609" TAGREFS="LT6426" BASELINE="1579 1487 2155 1479" HPOS="1578.0" VPOS="1438.0" WIDTH="577.0" HEIGHT="81.0">
+<Shape></Shape>
<String CONTENT="po' v cognit̄e sfit̄e sūt uis̄" HPOS="1578.0" VPOS="1438.0" WIDTH="577.0" HEIGHT="81.0"/>
</TextLine>
-<TextLine ID="eSc_line_b1fb0fe4" TAGREFS="LT6426" BASELINE="1575 1545 2159 1534" HPOS="1573.0" VPOS="1490.0" WIDTH="586.0" HEIGHT="84.0">
+<Shape></Shape>
<String CONTENT="gust̄ ̄ tactus. ̄ tlia in qb̄ ̄ g" HPOS="1573.0" VPOS="1490.0" WIDTH="586.0" HEIGHT="84.0"/>
</TextLine>
-<TextLine ID="eSc_line_63be87f7" TAGREFS="LT6426" BASELINE="1576 1602 2164 1589" HPOS="1575.0" VPOS="1545.0" WIDTH="589.0" HEIGHT="78.0">
+<Shape></Shape>
<String CONTENT="cam̄ cū brutis. appetitie u di" HPOS="1575.0" VPOS="1545.0" WIDTH="589.0" HEIGHT="78.0"/>
</TextLine>
-<TextLine ID="eSc_line_2304726d" TAGREFS="LT6426" BASELINE="1573 1660 2150 1648" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0">
+<Shape></Shape>
<String CONTENT="st̄igūf. nā qdam ē appetit̄ i" HPOS="1571.0" VPOS="1598.0" WIDTH="579.0" HEIGHT="91.0"/>
</TextLine>

```

Le fichier ALTO d'origine

- └ Après l'HTR : tout change / rien ne change
- └ Quand s'arrête la correction ?
- └ Assurer la rétroconvertibilité

2024-09-10

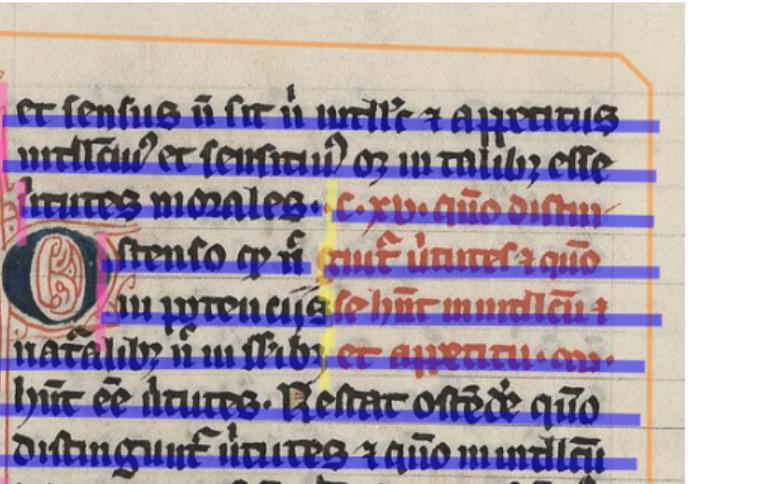
Assurer la rétroconvertibilité

Le fichier ALTO d'origine

## • PASSER

# Structurer les documents

- Classifier les zones et les lignes lors de la phase d'ATR peut permettre de faciliter la structuration (semi-)automatisée du document :



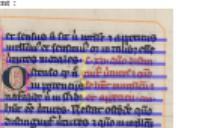
Classification des lignes suivant le vocabulaire contrôlé SegmOnto (GABAY et al. 2021). En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric. Vat. Borg. 360, fol 190v.

2024-09-10

De la donnée avant toute chose ?

- Après l'HTR : tout change / rien ne change
  - Structurer les documents
    - Structurer les documents

- Classifier les zones et les lignes lors de la phase d'ATR peut permettre de faciliter la structuration (semi-)automatisée du document :



Classification des lignes suivant le vocabulaire contrôlé SegmOnto (GABAY et al. 2021). En fuchsia : les lignes de type DefaultLine ; en jaune, les lignes de type Headingline:rubric. Vat. Borg. 360, fol 190v.

- La classification des zones de textes permet de ne s'intéresser qu'à des zones précises (la colonne, ou au contraire les titres courants par exemple).
- On utilisera la première ligne de la rubrique (dans la zone de texte principal) pour identifier une borne de division (un chapitre par exemple), et ainsi de suite pour tout le document.
- L'exemple donné ici est doublement problématique : le numéro de chapitre est erronné, ce qui peut poser problème si l'on utilise le texte pour numérotter les divisions ; en second lieu, l'ordre des lignes est évident pour l'humain mais plus difficile à identifier pour la machine (mais on y arrive !), ce qui peut de même poser problème pour la structuration du texte.

# Identifier la césure à la ligne

- Dans les manuscrits médiévaux la césure à la ligne n'est pas systématiquement indiquée

*podies sensituos. Qd assi como nñgu  
ome nñ es alabado nñ es tenido por bue  
no por q muelle bie su uata nñ por q cre  
sar bie assi nñ es alabado por q bie agu  
da mete o eye solit mete. Saluo ende*

*rōne ptiapant qma p se a sñq si no  
obedimt rōm g m sensibz a i potenti  
is nñbs ut sup dñebar no hñt  
esse vñtes sñ so hñt m i tellez a ap  
petitu in nobis an dñpser est appeti*

*v appetitu intelléctu. dñtates g de qb  
loqui intendim q sñt qdam hic lan  
tabiles ut eunt i potentiis alibus  
ul m ipis sñb ul m appetitu. Vñtu  
ul m appetitu intelléctuo ul m ipo idle  
tu ul m omibz hñs ul m aliqbz hñs*

*C'el asturiano atodo esto alen  
trase z fielo de manaz concis  
del Rey con fernando z dela Reyna  
dona maria Su madre quele en  
biana a pedr por meqnd al papa sñ  
bre esta pason mas con pedro  
que esa obiso de bnygos a esa sa  
son E referencatio del papa ua*

*S'ann si hñ est. media è intellé  
ctus morales. intelléctu. pñ m  
spñta qñ unibz moralibz. Pñ pñ  
den nñ hñ hñs leis si nñ vñtes  
moralis. pñ e hñs si pñt e hñ  
entes. alibi. nñspells. m pñtēs*

*sophia. geometria. metaphi  
sica. i. talia uirtutes uero si  
phicæ morales. sñt ille q sñt  
in appetitu. siue appetitus il  
le sit sensitivæ. siue intellé  
ctuus. cuiusmodi aut sunt  
iustitia. temperancia. fortitudo*

Valladolid, 251 fol. 17r ; Cambridge, Corpus Christi College, MS 283, fol 14r ; Vatican, Borg. 360, fol. 190r ;  
BNF Esp 36, fol 1v ; BNE MSS/958, fol. 14r ; Valencia BH Ms 0594, fol. 23r

## De la donnée avant toute chose ?

- Après l'HTR : tout change / rien ne change
  - Identifier la césure à la ligne
  - Identifier la césure à la ligne

2024-09-10

Dans les manuscrits médiévaux la césure à la ligne n'est pas systématiquement indiquée

*Qd assi como nñgu  
ome nñ es alabado nñ es tenido por bue  
no por q muelle bie su uata nñ por q cre  
sar bie assi nñ es alabado por q bie agu  
da mete o eye solit mete. Saluo ende*

*rōne ptiapant qma p se a sñq si no  
obedimt rōm g m sensibz a i potenti  
is nñbs ut sup dñebar no hñt  
esse vñtes sñ so hñt m i tellez a ap  
petitu in nobis an dñpser est appeti*

*v appetitu intelléctu. dñtates g de qb  
loqui intendim q sñt qdam hic lan  
tabiles ut eunt i potentiis alibus  
ul m ipis sñb ul m appetitu. Vñtu  
ul m appetitu intelléctuo ul m ipo idle  
tu ul m omibz hñs ul m aliqbz hñs*

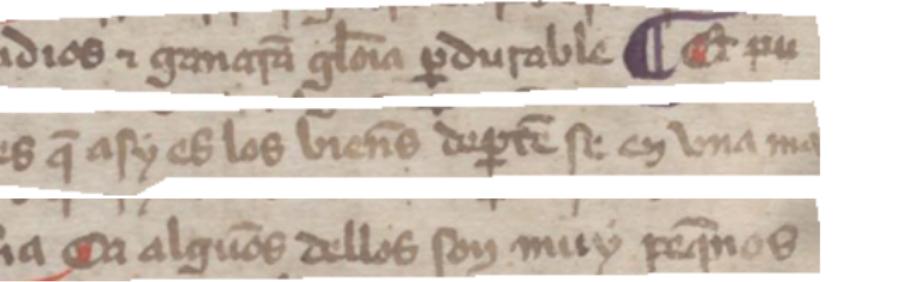
*C'el asturiano atodo esto alen  
trase z fielo de manaz concis  
del Rey con fernando z dela Reyna  
dona maria Su madre quele en  
biana a pedr por meqnd al papa sñ  
bre esta pason mas con pedro  
que esa obiso de bnygos a esa sa  
son E referencatio del papa ua*

*S'ann si hñ est. media è intellé  
ctus morales. intelléctu. pñ m  
spñta qñ unibz moralibz. Pñ pñ  
den nñ hñ hñs leis si nñ vñtes  
moralis. pñ e hñs si pñt e hñ  
entes. alibi. nñspells. m pñtēs*

*sophia. geometria. metaphi  
sica. i. talia uirtutes uero si  
phicæ morales. sñt ille q sñt  
in appetitu. siue appetitus il  
le sit sensitivæ. siue intellé  
ctuus. cuiusmodi aut sunt  
iustitia. temperancia. fortitudo*

Valladolid, 251 fol. 17r ; Cambridge, Corpus Christi College, MS 283, fol 14r ; Vatican, Borg. 360, fol. 190r ;  
BNF Esp 36, fol 1v ; BNE MSS/958, fol. 14r ; Valencia BH Ms 0594, fol. 23r

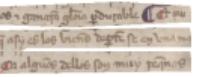
## Identifier la césure



adios a granapa gloria perdurable Cest pu  
es q asy es los bienes dupert se en una ma  
na con algunos de los soy muy perdon

- └ Après l'HTR : tout change / rien ne change
  - └ Identifier la césure à la ligne
  - └ Identifier la césure

2024-09-10



Dios a granapa gloria perdurable Cest pu  
es q asy es los bienes dupert se en una ma  
na con algunos de los soy muy perdon

Il suffit donc d'associer deux à deux les lignes (faire des bigrammes de lignes) afin d'identifier si la chaîne de caractère de fin de ligne correspond aussi à une fin de mot.

## Identifier la césure

adios ⁊ ganarā gloia pdurable ¶ Et pu  
es q̄ asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos

adios ⁊ ganarā gloia pdurable ¶ Et pu  
es q̄ asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos



adios ⁊ ganarā gloia pdurable ¶ Et pues es q̄ asy es los bieñs deptē se en una maña Ca alguōs dellos son  
muy peqnos

- └ Après l'HTR : tout change / rien ne change
  - └ Identifier la césure à la ligne
  - └ Identifier la césure

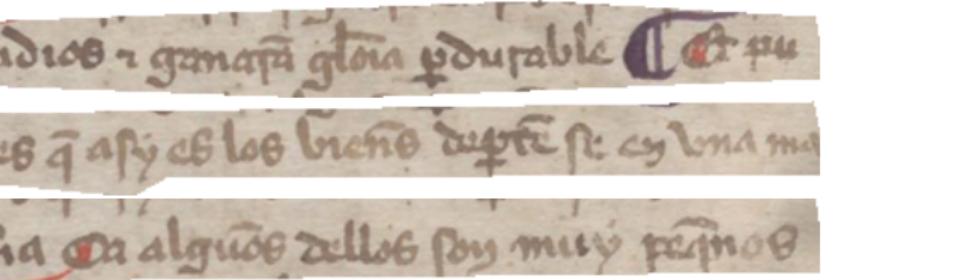
2024-09-10

adios ⁊ ganarā gloia pdurable ¶ Et pu  
es q̄ asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos

adios ⁊ ganarā gloia pdurable ¶ Et pues es q̄ asy es los bieñs deptē se en una maña Ca alguōs dellos son  
muy peqnos

Il suffit donc d'associer deux à deux les lignes (faire des bigrammes de lignes) afin d'identifier si la chaîne de caractère de fin de ligne correspond aussi à une fin de mot.

## Identifier la césure



adios j ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqños

adios j ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos

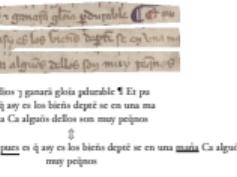
↔

adios j ganarā gloia pdurable ¶ Et pues es q asy es los bieñs deptē se en una maña Ca alguōs dellos son  
muy peqnos

Voir (CLÉRICE 2020) et [https://github.com/matgille/boudams\\_like\\_tokenizer](https://github.com/matgille/boudams_like_tokenizer)

- └ Après l'HTR : tout change / rien ne change
- └ Identifier la césure à la ligne
- └ Identifier la césure

2024-09-10



adios j ganarā gloia pdurable ¶ Et pu  
es q asy es los bieñs deptē se en una ma  
ña Ca alguōs dellos son muy peqnos

Voir (CLÉRICE 2020) et [https://github.com/matgille/boudams\\_like\\_tokenizer](https://github.com/matgille/boudams_like_tokenizer)

Il suffit donc d'associer deux à deux les lignes (faire des bigrammes de lignes) afin d'identifier si la chaîne de caractère de fin de ligne correspond aussi à une fin de mot.

# Résoudre les abréviations

- Le plus « simple » est d'utiliser une table de conversion à base de règles
- Suppose de connaître en amont toutes les abréviations et cas possibles **en contexte**
- Un peu laborieux quand le volume de documents différents est important

```

243 felicit e-l      felicitatem
244 cit e-l      citatem
245 uot e-l      uoluntatem
246 <SOT>cai t e<EOT> ~caritate~
247 <SOT>p<EOT>      ~prout~
248 <SOT>ad<EOT>      ~aliquid~
249 gn̄          gener
  
```

Table d'abréviation. <SOT> et <EOT> viennent marquer un début et/ou une fin de mot

## De la donnée avant toute chose ?

- Après l'HTR : tout change / rien ne change
  - Résoudre les abréviations
    - Résoudre les abréviations

2024-09-10

- Le plus « simple » est d'utiliser une table de conversion à base de règles
  - Suppose de connaître en amont toutes les abréviations et cas possibles **en contexte**
  - Un peu laborieux quand le volume de documents différents est important
- ```

243 felicit e-l      felicitatem
244 cit e-l      citatem
245 uot e-l      uoluntatem
246 <SOT>cai t e<EOT> ~caritate-
247 <SOT>p<EOT>      ~prout-
248 <SOT>ad<EOT>      ~aliquid-
249 gn̄          gener
  
```

Table d'abréviation. <SOT> et <EOT> viennent marquer un début et/ou une fin de mot

- Un travail en cours à préciser ; encore des irrégularités dans la table
- les choix

# Et l'édition critique ?

L'ATR appelle assez naturellement des méthodes de collation automatisée du texte

De la donnée avant toute chose ?

- └ Après l'HTR : tout change / rien ne change
- └ Et l'édition critique ?
- └ Et l'édition critique ?

2024-09-10

# Et l'édition critique ?

|                         | ab                     | ab       | ab      | ab                  | ab             | ab              | ab       | ab      | ab |
|-------------------------|------------------------|----------|---------|---------------------|----------------|-----------------|----------|---------|----|
| homéotéleute            | lex                    | ~        | lex.    | om./lex.            | om./lex.       | om./lex.        | om./lex. | ~       | ~  |
| Rome_1607               | quaedam                |          | mediae  |                     | inter          | intellectuales  | et       | morales |    |
| Rome_1556               | quædon                 |          | mediæ   |                     | inter          | intellectuales  | et       | morales |    |
| Planck                  | quedam                 | medie    |         |                     | icirer         | iitellecituales | et       | morales |    |
| Bevilaqua_1498          | quadam                 |          | mediae  |                     | inter          | intellectuales  | et       | morales |    |
| Vat_Lat_590             | quedam                 | die      | me      |                     | inter          | intellectuales  | et       | morales |    |
| CCC_MSS_283             | qued                   |          | mē      |                     | inter          | intellectuales  | et       | morales |    |
| Borgh_360               | et                     | uirtutes | morales |                     |                | intellectuales  | et       | morales |    |
| Geneve_Ms_Lat_92        | et                     | medie    | quedum  |                     | inter          |                 |          | morales |    |
| Metz_Mediatheque_1234   | quedam                 | medie    |         | interintellectuales |                |                 | et       | morales |    |
| Beinecke_Marston_MS_139 | quadam                 | medie    |         |                     | inter          | intellectumales | et       | morales |    |
| BNE_MSS_958             | quad <small>RC</small> | medie    |         |                     | inter          | intellectuales  | et       | morales |    |
| BNF_Lat_6477            | quedam                 | medie    |         |                     |                | intellectuales  | et       | morales |    |
| BNE_9236                | quedam                 | medie    |         |                     | inter          | intellectumales | et       | morales |    |
| BNF_Lat_1234            | quedam                 | medie    |         | interintellectuales |                |                 |          |         |    |
| Valencia_BH_Ms_0594     | quedam                 | medie    |         | inter               | intellectuales |                 | et       | morales |    |

Table d'alignement avant la phase de correction, après segmentation et résolution des abréviations. Gilles de Rome,  
*De Regimine Principum*, chapitre 1.2.2

- └ Après l'HTR : tout change / rien ne change
- └ Et l'édition critique ?
- └ Et l'édition critique ?

2024-09-10

| ab        | ab      | ab     | ab    | ab             | ab | ab      | ab |
|-----------|---------|--------|-------|----------------|----|---------|----|
| Rome_1607 | quaedam | mediae | inter | intellectuales | et | morales |    |

Table d'alignement avant la phase de correction, après segmentation et résolution des abréviations. Gilles de Rome,  
*De Regimine Principum*, chapitre 1.2.2

# Et l'édition critique ?

| ab                      | ab    | ab       | ab         | ab        | ab     | ab       | ab     | ab       | homéotéleute | ab     | ab    | ab    |
|-------------------------|-------|----------|------------|-----------|--------|----------|--------|----------|--------------|--------|-------|-------|
| lex.                    | norm. | om./lex. | lex.       | graph.    | lex.   | om./lex. | om.    | om./lex. | om.          | lex.   | ~     |       |
| Rome_1607               |       | e        | sensitiuae | sunt      | visus  |          | gustus | auditus  |              | et     | talia |       |
| Rome_1556               |       | e        | sensitiuae | sunt      | visus  |          | gustus | auditus  |              | et     | talia |       |
| Planck                  |       |          | sensitiue  | sunt      | uisus  |          | gustus | auditus  |              | et     | talia |       |
| Bevilacqua_1498         |       |          | sensitiua  | sunt      | ui-sus |          | gustus | auditus  |              | et     | talia |       |
| Vat Lat 590             |       |          | sensitiem  | sunt      | uisus  |          | gustus | et       | tactus       | et     | talia |       |
| CCC_MSS_283             |       |          | senfitiue  | sunt      | uisus  | gus-tus  |        |          | et           | tactus | et    | talia |
| Borgh_360               |       |          | sensituue  | sunt      | ui sus |          | gustus |          | tactus       | et     | talia |       |
| Geneve Ms_Lat_92        |       |          | sensitiue  | sunt      | uisus  | auditus  | gustus |          |              | et     | talia |       |
| Metz_Mediatheque_1234   |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| Beinecke_Marston_MS_139 |       |          | sensitiue  | sunt      | uisus  |          | gustus | et       | tactus       | et     | talia |       |
| BNE_MSS_958             |       |          | sensitiue  | suntuisus |        |          | gustus |          | tactus       | et     | talia |       |
| BNF_Lat_6477            |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | e      | talia |       |
| BNE_9236                |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| BNF_Lat_1234            |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| Valencia_BH_Ms_0594     |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |

Table d'alignement après la phase de correction. Gilles de Rome, *De Regimine Principum*, chapitre 1.2.1

De la donnée avant toute chose ?

- Après l'HTR : tout change / rien ne change
- Et l'édition critique ?
- Et l'édition critique ?

2024-09-10

| ab                      | ab    | ab       | ab         | ab        | ab     | ab       | ab     | ab       | homéotéleute | ab     | ab    | ab    |
|-------------------------|-------|----------|------------|-----------|--------|----------|--------|----------|--------------|--------|-------|-------|
| lex.                    | norm. | om./lex. | lex.       | graph.    | lex.   | om./lex. | om.    | om./lex. | om.          | lex.   | ~     |       |
| Rome_1607               |       | e        | sensitiuae | sunt      | visus  |          | gustus | auditus  |              | et     | talia |       |
| Rome_1556               |       | e        | sensitiuae | sunt      | visus  |          | gustus | auditus  |              | et     | talia |       |
| Planck                  |       |          | sensitiue  | sunt      | uisus  |          | gustus | auditus  |              | et     | talia |       |
| Bevilacqua_1498         |       |          | sensitiua  | sunt      | ui-sus |          | gustus | auditus  |              | et     | talia |       |
| Vat Lat 590             |       |          | sensitiem  | sunt      | uisus  |          | gustus | et       | tactus       | et     | talia |       |
| CCC_MSS_283             |       |          | senfitiue  | sunt      | uisus  | gus-tus  |        |          | et           | tactus | et    | talia |
| Borgh_360               |       |          | sensituue  | sunt      | ui sus |          | gustus |          | tactus       | et     | talia |       |
| Geneve Ms_Lat_92        |       |          | sensitiue  | sunt      | uisus  | auditus  | gustus |          |              | et     | talia |       |
| Metz_Mediatheque_1234   |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| Beinecke_Marston_MS_139 |       |          | sensitiue  | sunt      | uisus  |          | gustus | et       | tactus       | et     | talia |       |
| BNE_MSS_958             |       |          | sensitiue  | suntuisus |        |          | gustus |          | tactus       | et     | talia |       |
| BNF_Lat_6477            |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | e      | talia |       |
| BNE_9236                |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| BNF_Lat_1234            |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |
| Valencia_BH_Ms_0594     |       |          | sensitiue  | sunt      | uisus  |          | gustus |          | tactus       | et     | talia |       |

Et l'édition critique ?

Table d'alignement après la phase de correction. Gilles de Rome, *De Regimine Principum*, chapitre 1.2.1

1 Introduction

2 Phase de production des données

3 Après l'HTR : tout change / rien ne change

4 Conclusions

De la donnée avant toute chose ?  
└ Conclusions

2024-09-10

- Introduction
- Phase de production des données
- Après l'HTR : tout change / rien ne change
- Conclusions

## ■ Penser en amont les principes d'annotation est fondamental

- Identifier les intérêts et les inconvénients de chaque choix en terme de balance intérêt particulier / intérêt général
- Lancer de courtes campagnes test pour mettre à l'épreuve les choix d'annotation
- Identifier le type d'information nécessaire aux étapes ultérieures de traitement du texte manuscrit ou imprimé : penser l'ensemble de la production des données et pas seulement l'HTR pour l'HTR
- Se mettre en conformité (ou pas) avec les normes existantes et le documenter clairement

## ■ Du travail reste à mener

- Sur tout la chaîne post-ATR afin d'arriver à un texte final plus propre
- Et plus particulièrement pour la collation
- Quid de l'accumulation du bruit avec la multiplication des étapes ?

## De la donnée avant toute chose ? └ Conclusions

2024-09-10

### De la donnée avant toute chose ?

#### └ Conclusions

- Penser en amont les principes d'annotation est fondamental
  - Identifier les intérêts et les inconvénients de chaque choix en terme de balance intérêt particulier / intérêt général
  - Lancer de courtes campagnes test pour mettre à l'épreuve les choix d'annotation
  - Identifier le type d'information nécessaire aux étapes ultérieures de traitement du texte manuscrit ou imprimé : penser l'ensemble de la production des données et pas seulement l'HTR pour l'HTR
  - Se mettre en conformité (ou pas) avec les normes existantes et le documenter clairement
- Du travail reste à mener
  - Sur tout la chaîne post-ATR afin d'arriver à un texte final plus propre
  - Et plus particulièrement pour la collation
  - Quid de l'accumulation du bruit avec la multiplication des étapes ?

# Merci!

Merci de votre attention !

- Diapos : [https://github.com/matgille/Comm\\_Nancy\\_sept\\_2024](https://github.com/matgille/Comm_Nancy_sept_2024)

└ Conclusions

└ Merci !

2024-09-10

Merci de votre attention !

■ Diapos : [https://github.com/matgille/Comm\\_Nancy\\_sept\\_2024](https://github.com/matgille/Comm_Nancy_sept_2024)

# Références I

- [1] Thibault CLÉRICE. « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin ». *Journal of Data Mining & Digital Humanities* 2020 (7 avr. 2020). URL : <https://jdmdh.episciences.org/6264/pdf>.
- [2] Thibault CLÉRICE, Ariane PINCHE, Malamenia VLACHOU-EFSTATHIOU, Alix CHAGUÉ, Jean-Baptiste CAMPS, Matthias Gilles LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Patricia O'CONNOR, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Benjamin KISSLING. « CATMuS Medieval : A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond ». *Document Analysis and Recognition - ICDAR 2024*. Sous la dir. d'Elisa H. BARNEY SMITH, Marcus LIWICKI et Liangrui PENG. Cham : Springer Nature Switzerland, 2024, p. 174-194. ISBN : 978-3-031-70543-4. DOI : 10.1007/978-3-031-70543-4\_11.
- [3] Simon GABAY, Jean-Baptiste CAMPS, Ariane PINCHE et Claire JAHAN. « SegmOnto : Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More) ». *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*. 2021. ISBN : 978-3030865481. URL : <https://segmonto.github.io/>.
- [4] Matthias GILLE LEVENSON. « Le Regimiento de Los Príncipes et sa glose : étude et édition numérique de la partie sur le Gouvernement de la cité en temps de guerre (III, 3) ». Codir. Carlos HEUSCH et Jesús R. VELASCO. École Normale Supérieure de Lyon, 2023. URL : <https://theses.hal.science/tel-04337406>.
- [5] Matthias GILLE LEVENSON. « Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR) ». *Journal of Data Mining and Digital Humanities* (2023). DOI : 10.46298/jdmdh.10416. URL : <https://zenodo.org/records/8340483>.
- [6] Odd Einar HAUGEN. « Dealing with Glyphs and Characters : Challenges in Encoding Medieval Scripts ». *Document numérique* 16.3 (2013), p. 97-111. URL : <https://www.cairn.info/revue-document-numerique-2013-3-page-97.htm>.
- [7] Benjamin KISSLING. « Kraken - an Universal Text Recognizer for the Humanities ». DH2019 : Complexity. Utrecht, 2019. URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html>.
- [8] Ariane PINCHE, Thibault CLÉRICE, Alix CHAGUÉ, Jean-Baptiste CAMPS, Malamenia VLACHOU-EFSTATHIOU, Matthias Gilles LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Patricia O'CONNOR. « CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts ». *Digital Humanities - DH2024*, ADHO. Washington, D.C., 2024.

2024-09-10

- [1] Thibault CLÉRICE. « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Text in Old French and Latin ». *Journal of Data Mining & Digital Humanities* 2020 (7 avr. 2020). URL : <https://jdmdh.episciences.org/6264/pdf>.
- [2] Thibault CLÉRICE, Ariane PINCHE, Malamenia VLACHOU-EFSTATHIOU, Alix CHAGUÉ, Jean-Baptiste CAMPS, Matthias Gilles LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Patricia O'CONNOR, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Benjamin KISSLING. « CATMuS Medieval : A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond ». *Document Analysis and Recognition - ICDAR 2024*. Sous la dir. d'Elisa H. BARNEY SMITH, Marcus LIWICKI et Liangrui PENG. Cham : Springer Nature Switzerland, 2024, p. 174-194. ISBN : 978-3-031-70543-4. DOI : 10.1007/978-3-031-70543-4\_11.
- [3] Simon GABAY, Jean-Baptiste CAMPS, Ariane PINCHE et Claire JAHAN. « SegmOnto : Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More) ». *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*. 2021. ISBN : 978-3030865481. URL : <https://segmonto.github.io/>.
- [4] Matthias GILLE LEVENSON. « Le Regimiento de Los Príncipes et sa glose : étude et édition numérique de la partie sur le Gouvernement de la cité en temps de guerre (III, 3) ». Codir. Carlos HEUSCH et Jesús R. VELASCO. École Normale Supérieure de Lyon, 2023. URL : <https://theses.hal.science/tel-04337406>.
- [5] Matthias GILLE LEVENSON. « Towards a General Open Dataset and Model for Late Medieval Castilian Text Recognition (HTR/OCR) ». *Journal of Data Mining and Digital Humanities* (2023). DOI : 10.46298/jdmdh.10416. URL : <https://zenodo.org/records/8340483>.
- [6] Odd Einar HAUGEN. « Dealing with Glyphs and Characters : Challenges in Encoding Medieval Scripts ». *Document numérique* 16.3 (2013), p. 97-111. URL : <https://www.cairn.info/revue-document-numerique-2013-3-page-97.htm>.
- [7] Benjamin KISSLING. « Kraken - an Universal Text Recognizer for the Humanities ». DH2019 : Complexity. Utrecht, 2019. URL : <https://dev.clariah.nl/files/dh2019/boa/0673.html>.
- [8] Ariane PINCHE, Thibault CLÉRICE, Alix CHAGUÉ, Jean-Baptiste CAMPS, Malamenia VLACHOU-EFSTATHIOU, Matthias Gilles LEVENSON, Olivier BRISVILLE-FERTIN, Federico BOSCHETTI, Franz FISCHER, Michael GERVERS, Agnès BOUTREUX, Avery MANTON, Simon GABAY, Wouter HAVERALS, Mike KESTEMONT, Caroline VANDYCK et Patricia O'CONNOR. « CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts ». *Digital Humanities - DH2024*, ADHO. Washington, D.C., 2024.

# Références II

- [9] J. REDMON. « You Only Look Once : Unified, Real-Time Object Detection ». *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. url : [https://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016/html/Redmon\\_You\\_Only\\_Look\\_CVPR\\_2016\\_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html).
- [10] Dominique STUTZMANN. « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? » *Codicology and Palaeography in the Digital Age*. T. 2. 2010, p. 34.