



**HAL**  
open science

## Gallic(orpor)a: Traitement des sources textuelles en diachronie longue de Gallica

Kelly Christensen, Ariane Pinche, Simon Gabay

► **To cite this version:**

Kelly Christensen, Ariane Pinche, Simon Gabay. Gallic(orpor)a: Traitement des sources textuelles en diachronie longue de Gallica. DataLab de la BnF, Jun 2022, Paris, France. hal-03716534

**HAL Id: hal-03716534**

**<https://hal.science/hal-03716534>**

Submitted on 7 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Gallic(orpor)a

Traitement des sources textuelles en diachronie longue de  
*Gallica*

Kelly Christensen <sup>1</sup>, Ariane Pinche <sup>2</sup>, Simon Gabay <sup>3</sup>

<sup>1</sup>Inria

<sup>2</sup>Ecole nationale des chartes — PSL

<sup>3</sup>Université de Genève

## Avant le projet

### Problématiques de l'apprentissage machine

- Gérer l'hétérogénéité
- Partager les données
- Premiers résultats
- Limites et Gains

### De l'extraction du texte à sa publication

1. Métadonnées
  - Contenu
  - Extraction
  - Mapping
2. Données issues de la  
segment. et de la HTR
  - Mapping
3. Texte pré-éditorialisé
  - Mapping
4. Texte annoté
  - Mapping

## Conclusion

# Avant le projet

- Deucalion pour le traitement des documents médiévaux (J.B. Camps)
- CREMMALab pour le traitement des documents médiévaux (A. Pinche)
- E-ditiones pour le traitement du français d'Ancien Régime (S. Gabay)
- FreEM pour le traitement automatique de la langue (XVI<sup>e</sup>-XVIII<sup>e</sup> s., B. Sagot, P. Ortiz, R. Bawden, Ph. Gambette, A. Bartz, S. Gabay)

**Un enjeu: le moyen français**, compris dans son acception la plus large (XV<sup>e</sup>-XVI<sup>e</sup> s.)

## Avant le projet

### Problématiques de l'apprentissage machine

Gérer l'hétérogénéité  
Partager les données  
Premiers résultats  
Limites et Gains

### De l'extraction du texte à sa publication

1. Métadonnées  
Contenu  
Extraction  
Mapping
2. Données issues de la segment. et de la HTR  
Mapping
3. Texte pré-éditorialisé  
Mapping
4. Texte annoté  
Mapping

### Conclusion

- Données d'apprentissage pour la reconnaissance automatique de caractères
  - Les manuscrits médiévaux en gothique
  - Les imprimés anciens en *antiqua*
  - Premiers tests pour l'analyse de mise en page
  - Données pour la lemmatisation de l'ancien français et du français (pré)classique
- Scripts disponibles
  - Récupération d'images avec IIIF
  - transformation alto → TEI
  - Injection de l'annotation linguistique dans la balise `<standoff>` de l'annotation linguistique

## Avant le projet

### Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

### De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

## Conclusion

## Concevoir une chaîne de traitement

- permettant l'automatisation des différentes tâches standards en humanités numériques
- pour construire un corpus richement annoté
- qui permette de valoriser et d'exploiter les collections numériques de la BnF

# Problématiques de l'apprentissage machine

# Gérer l'hétérogénéité

Galic(orpor)a

K. Christensen et al.

## Comment dépasser l'hétérogénéité des sources sur le temps long ?

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la

segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion



Figure: BnF, Réserve des livres rares, velin, 15<sup>e</sup> s.

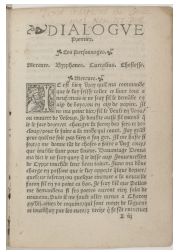


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16<sup>e</sup> s.

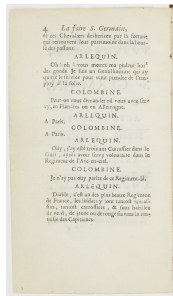


Figure: BnF, Arts du spectacle, réserve 8-RO-1702, 17<sup>e</sup> s.

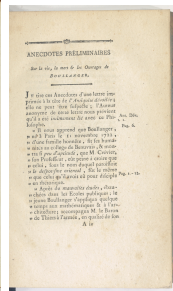


Figure: BnF, Droit, économie, politique, 2012-39571, 18<sup>e</sup> s.

Solution : Développer des systèmes d'annotation compatible avec l'ensemble des documents (mise en page, lemmatisation, entités nommées, etc.)



# Concevoir des pratiques communes : lemmatisation

Harmoniser l'annotation des données : lemme, POS, entités nommées

- Utiliser des étiquettes communes pour assurer la compatibilité de nos données.
- Normalisation linguistique
- La gestion de l'ancien français devient difficile, car les modèles de langue ne sont plus compatibles.

Token	Lemme	POS	Morphologie	Entités
Michel	Michel	NOMP <sub>ro</sub>	NOMB.=s—GENRE=m	PER-B
annotate	annoter	VERc <sub>jg</sub>	MODE=ind—TEMPS=pst—PERS.=3—NOMB.=s	0
les	le	DET <sub>def</sub>	NOMB.=p—GENRE=m	0
textes	texte	NOM <sub>com</sub>	NOMB.=p—GENRE=m	0

Table: Exemple d'annotation

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

# Concevoir des pratiques communes : segmentation

Gallic(orpor)a

K. Christensen et  
al.

Des normes pour décrire les sources :

- **Projet SegmOnto (Gabay et al., 2021)**

*Sélection de zones SegmOnto*      *Lignes SegmOnto*

DropCapitalZone

DefaultLine

GraphicZone

DropCapitalLine

MainZone

HeadingLine

MarginTextZone

InterlinearLine

MusicZone

MusicLine

NumberingZone

QuireMarksZone

RunningTitleZone

TableZone

TitlePageZone

Avant le projet

Problématiques de  
l'apprentissage  
machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du  
texte à sa  
publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la  
segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

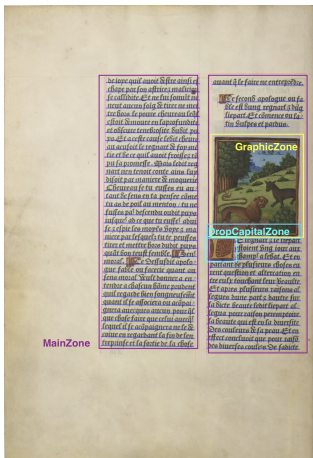


Figure: BnF, Réserve des livres rares, vélin 611, 15<sup>e</sup> s.

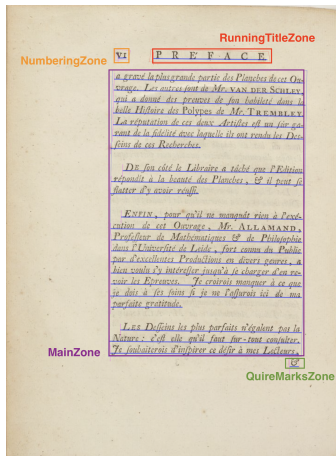


Figure: BnF, Arsenal, 4-S-1534, 18<sup>e</sup> s.

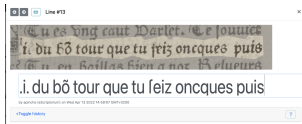
# Concevoir des pratiques communes : HTR

## Harmoniser les transcriptions :

- Définir le degré de précision recherché dans la transcription (abréviations, caractères spéciaux, variation des glyphes)
- Utiliser un set de caractères prédéfini
- Limites : les incunables et les imprimés du 15<sup>e</sup> siècle



**Figure:** BnF, Réserve des livres rares, vélin 611, 15<sup>e</sup> s.



**Figure:** BnF, Réserve des livres rares, RES-Z-2442, 16<sup>e</sup> s.



**Figure:** BnF, Arts du spectacle, réserve 8-RO-1702, 17<sup>e</sup> s.



**Figure:** BnF, Droit, économie, politique, 2012-39571, 18<sup>e</sup> s.

Les modèles créés par des IA ont besoin de données. Si on veut les améliorer, on doit augmenter les corpus d'entraînement. On a donc besoin de :

- Créer des données
- Partager des données
  - Partager les vérités de terrain (GitLab, GitHub, etc.), voir le dépôt des données Gallicorpora
  - S'appuyer sur des catalogues de données d'entraînement, voir HTR-United

# Premiers résultats : Segmentation

Gallic(orpor)a

K. Christensen et al.

## Entraînement des modèles de segmentation - Premiers échecs

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion



Figure: BnF, RésERVE des livres rares, vélin 611, 15<sup>e</sup> s.

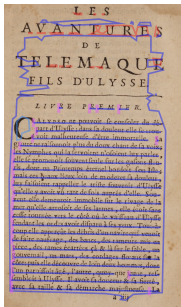


Figure: BnF, RésERVE des livres rares, RES-Z-2442, 16<sup>e</sup> s.

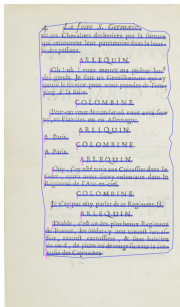


Figure: BnF, Arts du spectacle, RésERVE 8-RO-1702, 17<sup>e</sup> s.

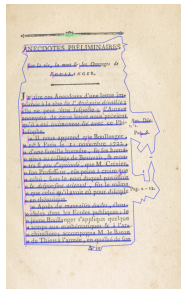


Figure: BnF, département Droit, économie, politique, 2012-39571, 18<sup>e</sup> s.

## Quelques résultats :

Modèles	Baseline			Gallicorpora+		
Test "in-domain"	98,61%			98,66%		
Test set (nb. car.)	114016			146765		
Erreurs fréquentes	NB	Pred.	Vérité	NB	Pred.	Vérité
	293	&#39;	'	327	&#39;	'
	73	SPACE		117	SPACE	
	57	'	&#39;	85		SPACE
	47	.	,	79	'	&#39;
Modèles	GallicorporaAntiqua			GallicorporaAntiquaGothique		
Test "in-domain"	91,10%			96,74%		
Test set (nb. car.)	32749			32749		
Erreurs fréquentes	NB	Pred.	Vérité	NB	Pred.	Vérité
	245	&#39;	'	65	SPACE	
	213	SPACE		64		SPACE
	177	'	&#39;	37	&#39;	'
	66	.	,	33	v	u

# Premiers résultats *Handwritten Text Recognition*

Galic(orpor)a

K. Christensen et al.

Exemples de prédiction sur des documents "out-of-domain" :  
BnF, Arts du spectacle, réserve 8-RO-1702, 17<sup>e</sup> s.



Figure: Prédiction issue du modèle baseline

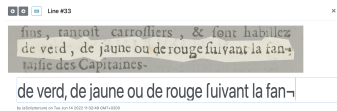


Figure: Prédiction issue du modèle Gallicorpora+

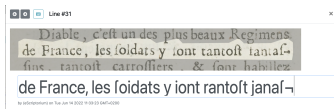


Figure: Prédiction issue du modèle GallicorporaAntiqua

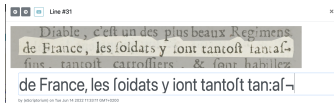


Figure: Prédiction issue du modèle GallicorporaAntiquaGothique

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la

segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion



# Premiers résultats *Handwritten Text Recognition*

Galic(orpor)a

K. Christensen et al.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la

segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

## Exemples de prédiction sur des documents "out-of-domain" :

BnF, département Droit, économie,  
politique, 2012-39571, 18<sup>e</sup> s.

Line #6

Jv tire ces Anecdotes d'une lettre im-

By INFO@BNUFI on Thu, Jul 19, 2012 11:21:41 GMT+0200

**Figure:** Prédiction issue du modèle baseline

Line #6

JE tire ces Anecdotes d'une lettre im-

By INFO@BNUFI on Thu, Jul 19, 2012 11:22:41 GMT+0200

**Figure:** Prédiction issue du modèle Galicorpora+

BnF, Réserve des livres rares, RES-Z-2442, 16<sup>e</sup> s.

Line #7

E est bien vay quit ma commanoe

By INFO@BNUFI on Thu, Jul 19, 2012 11:21:41 GMT+0200

**Figure:** Prédiction issue du modèle baseline

Line #7

L est bien uray quil ma commande

By INFO@BNUFI on Thu, Jul 19, 2012 11:22:41 GMT+0200

**Figure:** Prédiction issue du modèle Galicorpora+

# Premiers résultats *Handwritten Text Recognition*

Gallic(orpor)a

K. Christensen et al.

Exemples de prédiction sur des documents "out-of-domain" :  
La bonne surprise des incunables, BnF, Réserve des livres rares, vélin 611, 15<sup>e</sup> s.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion



Figure: Prédiction issue du modèle Gallicorpora+



Figure: Prédiction issue du modèle cremma-medievalGallicorpora15

# Premiers résultats *Handwritten Text Recognition*

Gallic(orpor)a

K. Christensen et al.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

Cas limites, les manuscrits : BnF, Manuscrits, Fr. 122, 14<sup>e</sup> s.



**Figure:** Prédiction issue du modèle Gallicorpora+



**Figure:** Prédiction issue du modèle cremma-medieval



**Figure:** Prédiction issue du modèle cremma-medievalGallicorpora15

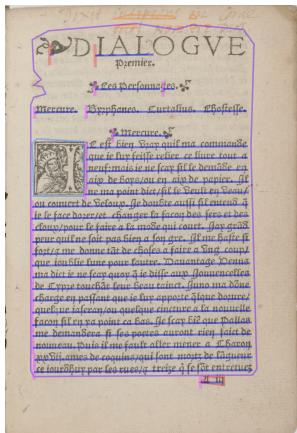


Figure: BnF, département Réserve des livres rares, RES-Z-2442, 16<sup>e</sup> s.

## ► Limites

- Problèmes de segmentation
  - Reconnaissance des headingLines
  - Reconnaissance des zones peu courantes
- Automatisation de l'étape d'éditorialisation du texte

## ► Gains

- Sets d'entraînement pour l'HTR
- Modèle HTR Gallicorpora+

# De l'extraction du texte à sa publication

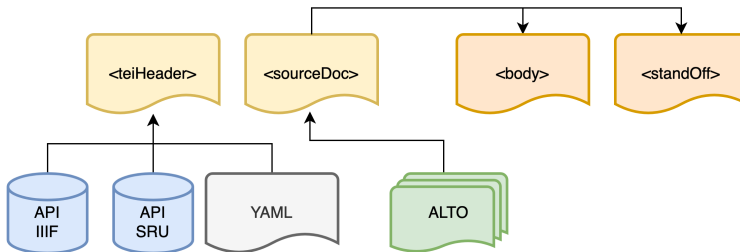
## Générer un TEI avec quatre composants

- 1 Métadonnées → `<teiHeader>`
- 2 Stockage des données issues de la segmentation et de la HTR → `<sourceDoc>`
- 3 Texte pré-éditorialisé → `<body>`
- 4 Texte annoté → `<standOff>`

# Construction automatisée de l'arbre TEI

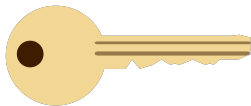
1. Métadonnées  
Contenu  
Extraction  
Mapping
2. Données issues de la segment. et de la HTR  
Mapping
3. Texte pré-éditorialisé  
Mapping
4. Texte annoté  
Mapping

Dérivé des sources externes		Dérivé de la <sourceDoc>	
<teiHeader>	<sourceDoc>	<body>	<standOff>
Métadonnées	Données issues de la seg/htr	Texte pré-éditorialisé	Texte annoté



Trois sources des métadonnées :

- IIIF Image API du document numérisé sur Gallica
- SRU API Catalogue général de la BnF
- Fichier de configuration (à personnaliser selon le projet)



La clef aux APIs = **ARK** (Archival Resource Key)



# Qu'est-ce qu'on veut récupérer du document ?

- Le titre du document
  - Le Romant comique [1re partie], par Mr Scarron
- La responsabilité du document
  - prénom d'auteur : Paul
  - nom d'auteur : Scarron
  - isni d'auteur : 0000000120990126
- La publication du document
  - date : 1655
  - lieu : Leiden
  - éditeur : J. Sambix

# Qu'est-ce qu'on veut récupérer du document ?

- Description de l'objet
  - langue : français
  - type : texte imprimé
- Conservation
  - pays : FR
  - ville : Paris
  - entrepôt : Bibliothèque nationale de France
  - côte (document physique) : 8-Y2-55998
  - côte (numérisation) : bpt6k6424218b

Galic(orpor)a

K. Christensen et al.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

**Contenu**

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

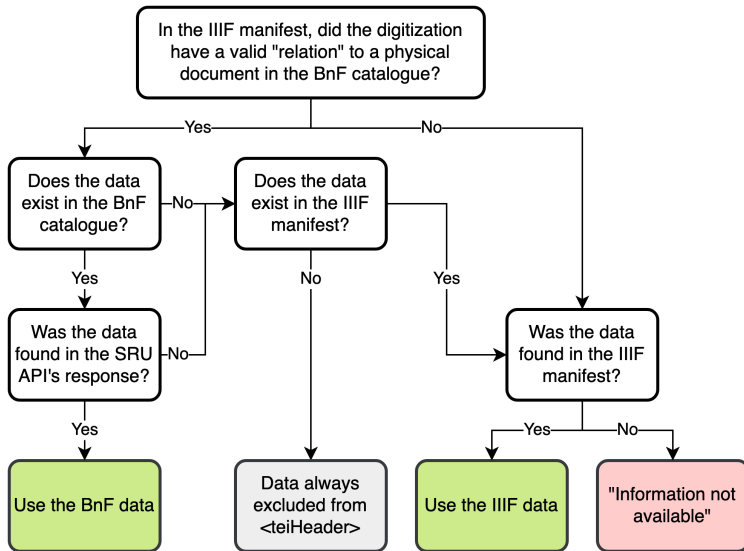
# Qu'est-ce qu'on veut intégrer à propos de notre édition numérique ?

- La responsabilité de l'édition numérique
  - auteur : Kelly Christensen
  - orcid d'auteur : 000000027236874X
- La publication de l'édition numérique
  - éditeur : Gallic(orpor)a
  - responsabilité : BnF DataLab
  - droits : <https://creativecommons.org/licenses/by/4.0/>
  - date : 2022-06-10
- Description de l'objet
  - mesure : 20 pages

# Où peut-on trouver ces infos ?

- APIs → Métadonnées du document
  - titre
  - responsabilité
  - publication
  - description de l'objet
  - conservation
- Fichier de config. → Métadonnées de l'édition
  - responsabilité
  - publication
  - description de l'objet

# Stratégie d'atténuation des risques



# Résultat du mapping des métadonnées

## Avant le projet

## Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

## De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la

segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

## Conclusion

```
1 <?xml version='1.0' encoding='UTF-8'?>
2 <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="ark:12148/bpt6k6424218b">
3 <teiHeader>
4 <fileDesc>
5 <titleStmt>
6 <title>Le Romant comique [Ire partie], par Mr Scarron</title>
7 <author xml:id="SCL">
8 <persName>
9 <forename>Paul</forename>
10 <surname>Scarron</surname>
11 <ptr type="isni" target="0000000120990126"/>
12 </persName>
13 </author>
14 <respStmt>
15 <resp>Transformation from ALTO4 to TEI by</resp>
16 <persName>
17 <forename>Kelly</forename>
18 <surname>Christensen</surname>
19 <ptr type="orcid" target="000000027236874X"/>
20 </persName>
21 <persName>
22 <forename>Simon</forename>
23 <surname>Gabay</surname>
24 <ptr type="orcid" target="0000000190944475"/>
25 </persName>
26 <persName>
27 <forename>Ariane</forename>
28 <surname>Pinche</surname>
29 <ptr type="orcid" target="0000000278435050"/>
30 </persName>
31 </respStmt>
32 </titleStmt>
33 <extent>
34 <measure unit="images" n="20"/>
35 </extent>
36 <publicationStmt>
37 <publisher>Gallic(orpor)a</publisher>
38 <authority>BnF DATALab</authority>
39 <availability status="restricted" n="cc-by">
40 <licence target="https://creativecommons.org/licenses/by/4.0"/>
41 </availability>
42 <date when="2022-06-10"/>
43 </publicationStmt>
44 <sourceDesc>
45 <bibl>
46 <publicationStmt>
47 <sourceDesc>
48 <ptr target="http://catalogue.bnf.fr/ark:/12148/cb31308524b/">
49 <author ref="#SCL">
50 <persName>
51 <forename>Paul</forename>
52 <surname>Scarron</surname>
53 <ptr type="isni" target="0000000120990126"/>
54 </persName>
55 </author>
56 <title>Le Romant comique [Ire partie], par Mr Scarron</title>
57 <pubPlace key="NL">Leiden</pubPlace>
58 <publisher>J. Sambix</publisher>
59 <date when="1655" cert="high" resp="BNF">1655</date>
60 </bibl>
61 <msDesc>
62 <msIdentifier>
63 <country key="FR"/>
64 <settlement>Information not available.</settlement>
65 <repository>Bibliothèque nationale de France</repository>
66 <idno>B-72-55998</idno>
67 <altIdentifier>
68 <idno type="ark">bpt6k6424218b</idno>
69 </altIdentifier>
70 </msIdentifier>
71 <physDesc>
72 <objectDesc>
73 <ptr type="text" target="#text-impriné"/>
74 </objectDesc>
75 </physDesc>
76 </msDesc>
77 </sourceDesc>
78 </fileDesc>
79 <profileDesc>
80 <langUsage>
81 <language id="fre">français</language>
82 </langUsage>
83 </profileDesc>
84 <teiHeader>
85 <sourceDoc>
86 <surfaceDesc>
87 <surface xml:id="f15" n="0" ulx="0" uly="0" lrx="1189" lry="2146">
88 <graphic url="https://gallica.bnf.fr/ark:/12148/bpt6k6424218b/f15/fu">
89 <zone xml:id="f15_z1" type="MainZone" subtype="none" n="none" points="59,1
```

## 2. Données issues de la segment. et de la HTR



Galic(orpor)a

K. Christensen et al.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

# ALTO → <sourceDoc>

## ALTO

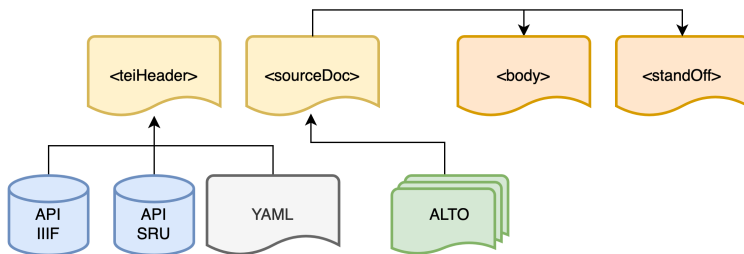
```
<TextLine ID="line_3" TAGREFS="LT832"  
  BASELINE="277 985 734 990" HPOS...>  
<Shape>  
  <Polygon POINTS="277 985 275 940..." /> </Shape>  
<String CONTENT="CHAPITRE I." HPOS="275"  
  VPOS="929" WIDTH="460" HEIGHT="70" ></String>...
```

## TEI

```
<zone xml:id="f15_z1_l1" type="HeadingLine"  
  subtype="none" n="1"  
  points="277,985 275,940..."  
  source="https://.../f15/275,929,460,70...jpg">  
<path xml:id="f15_z1_l1_p"  
  points="277,985 734,990"/>  
<line xml:id="f15_z1_l1_t">CHAPITRE I.</line> ...
```



### 3. Texte pré-éditorialisé (<body>)



Galic(orpor)a

K. Christensen et al.

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité  
Partager les données  
Premiers résultats  
Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées  
Contenu  
Extraction  
Mapping
2. Données issues de la segment. et de la HTR  
Mapping
3. Texte pré-éditorialisé  
Mapping
4. Texte annoté  
Mapping

Conclusion

## Mapping &lt;sourceDoc&gt; → &lt;body&gt;

&lt;sourceDoc&gt;

```

<zone xml:id="f15_z1_l1" type="HeadingLine"
      subtype="none" n="1"
      points="277,985 275,940...>
  <path xml:id="f15_z1_l1_p"
        points="277,985 734,990"/>
  <line xml:id="f15_z1_l1_t">CHAPITRE I.</line>
</zone>

```

&lt;body&gt;

```

<pb corresp="#f15"/>
<ab corresp="#f15_z1" type="MainZone">
  <hi rend="HeadingLine">
    <lb corresp="#f15_z1_l1"/>CHAPITRE I.
  </hi>

```

...

## 4. Texte annoté

- Lemmatisation
- Normalisation
- Reconnaissance des entités nommées

# Mapping <sourceDoc> → annotation

A MON SEIGNEVR DE LANGEI,

Humble salut et recongoissance de sa liberalité enuers luy.

segment

ESTIENNE DOLET, A MON SEIGNEVR DE LANGEI,  
Humble salut et recongoissance de sa  
liberalité enuers luy.

Annotation du texte (plusieurs modèles TAL)

token	Humble	salut	et	recongoissance
lemma	Humble	salut	et	recongoissance
pos	ADJqua	NOMcom	CONcoo	ADJqua
norm	Humble	salut	et	reconnaissance

Avant le projet

Problématiques de l'apprentissage machine

Gérer l'hétérogénéité

Partager les données

Premiers résultats

Limites et Gains

De l'extraction du texte à sa publication

1. Métadonnées

Contenu

Extraction

Mapping

2. Données issues de la segment. et de la HTR

Mapping

3. Texte pré-éditorialisé

Mapping

4. Texte annoté

Mapping

Conclusion

- Des progrès :
  - Augmentation du nombre de vérités de terrain pour l'HTR
  - Amélioration des modèles pour l'HTR et la segmentation
  - Scripts de conversion Alto to TEI
- Des difficultés :
  - Production des données très longue
  - Nécessité de produire une documentation plus détaillée pour la production des corpus (segmentation + HTR)
  - Mise à disposition et la valorisation des données très chronophage
  - Modèles linguistiques en retrait
- Ouverture :
  - Visualisation des corpus à l'aide de TEIPublisher