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

Kelly Christensen, Ariane Pinche, Simon Gabay

Gallic(orpor)a

Traitment des sources textuelles en diachronie longue de *Gallica*

Kelly Christensen [1], Ariane Pinche [2], Simon Gabay [3]

[1]Inria

[2]Ecole nationale des chartes — PSL

[3]Université de Genève

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

Avant le projet

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

Projets préexistants

- 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^{e}-XVIII^{e} 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^{e}-XVI^{e} 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

Éléments disponibles

- 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 $\rightarrow$ 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

Objectif

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

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

Problématiques de l'apprentissage machine

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

Gérer l'hétérogénéité

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

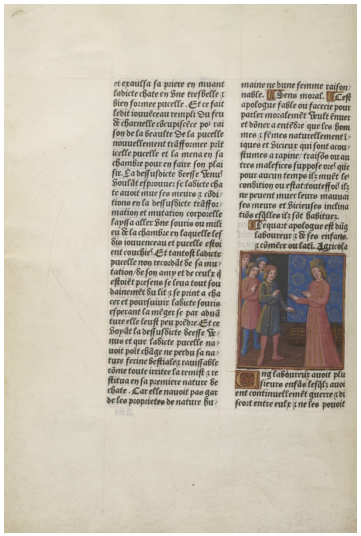


Figure: BnF, Réserve des livres rares, velin, 15e s.

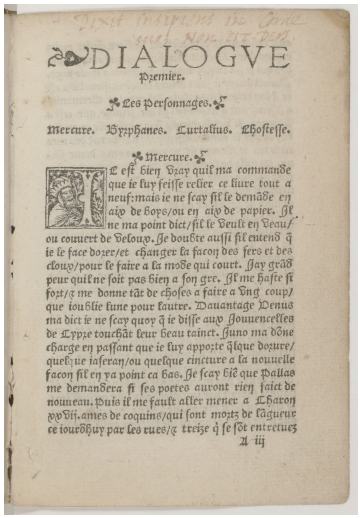


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16e s.

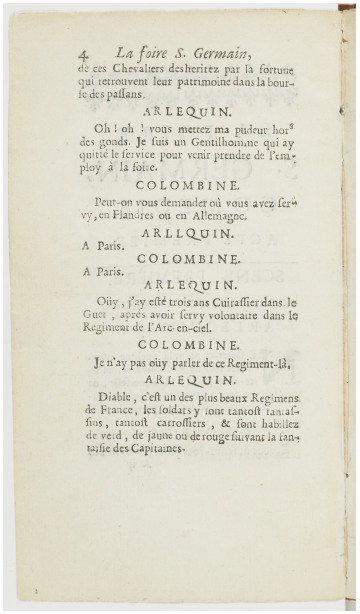


Figure: BnF, Arts du spectacle, réserve 8-RO-1702, 17e s.

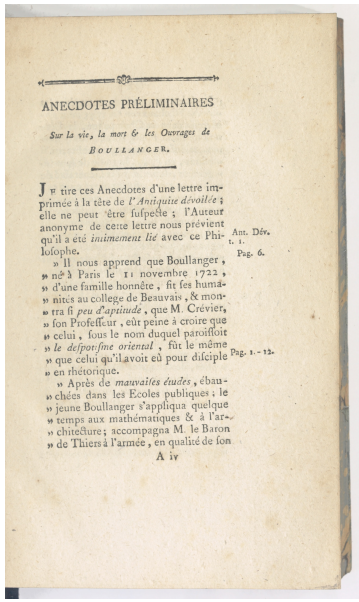


Figure: BnF, Droit, économie, politique, 2012-39571, 18e s.

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

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 : 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	NOMPro	NOMB.=s—GENRE=m	PER-B
annote	annoter	VERcjg	MODE=ind—TEMPS=pst—PERS.=3—NOMB.=s	0
les	le	DETdef	NOMB.=p—GENRE=m	0
textes	texte	NOMcom	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

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

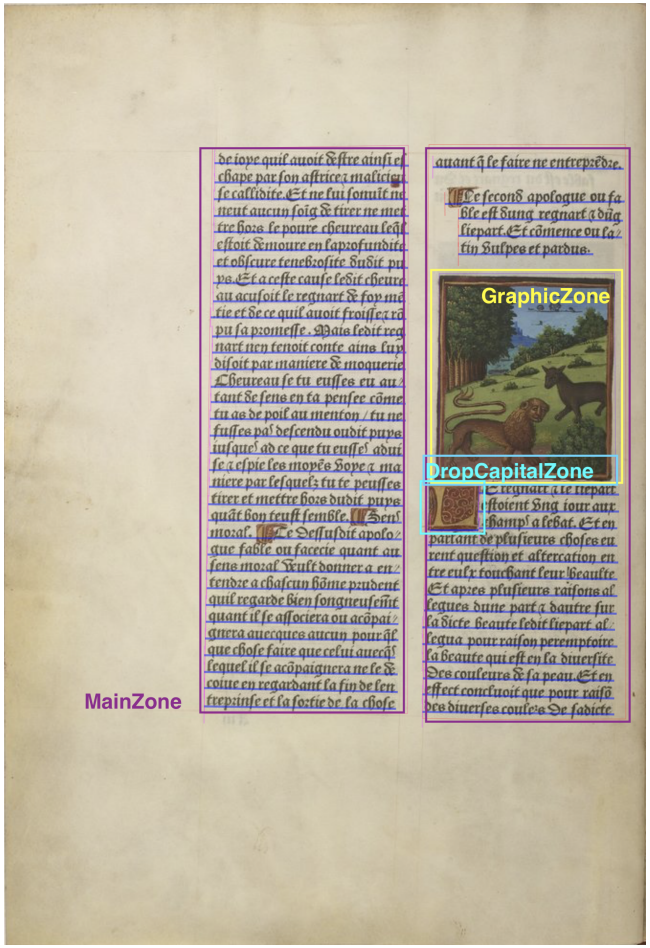


Figure: BnF, Réserve des livres rares, vélins 611, 15e s.

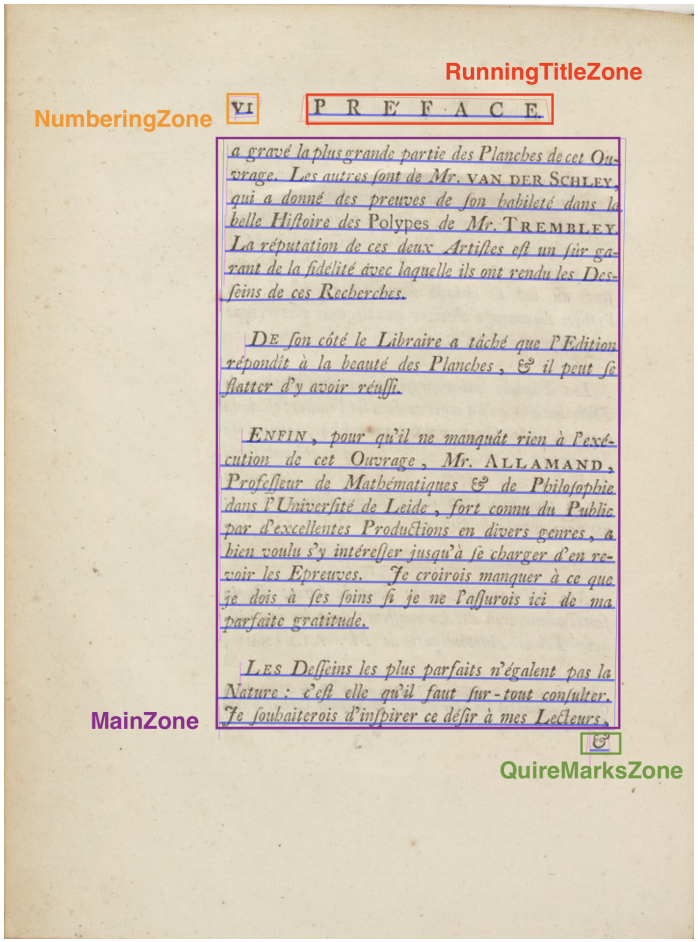


Figure: BnF, Arsenal, 4-S-1534, 18e s.

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

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 15e siècle

Figure: BnF, Réserve des livres rares, vélins 611, 15e s.

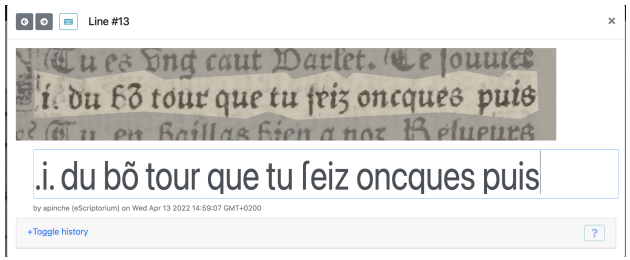


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16e s.

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

Figure: BnF, Droit, économie, politique, 2012-39571, 18e 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

Partager les données

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

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 : Segmentation

Entraînement des modèles de segmentation - Premiers échecs

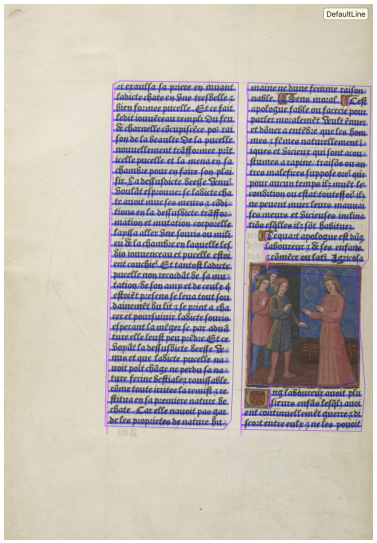


Figure: BnF, Réserve des livres rares, vélins 611, 15e s.

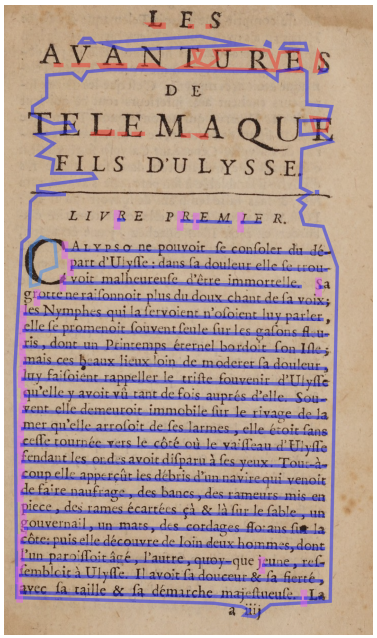


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16e s.

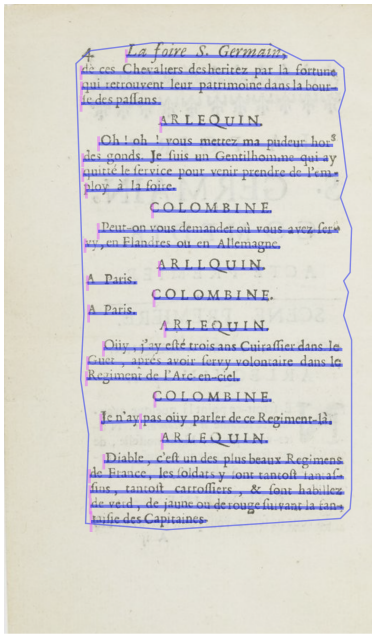


Figure: BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

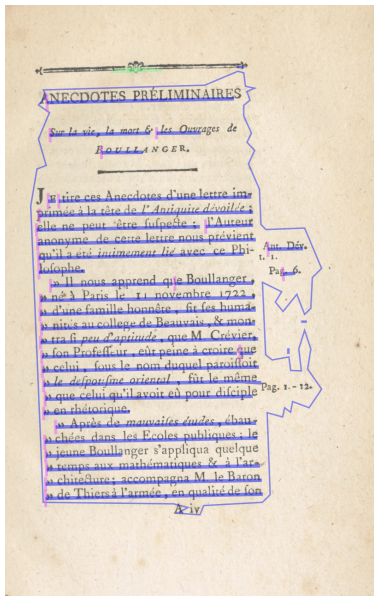


Figure: BnF, département Droit, économie, politique, 2012-39571, 18e s.

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

Premiers résultats *Handrwitten Text Recognition*

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

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 *Handrwitten Text Recognition*

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

Figure: Prédiction issue du modèle baseline

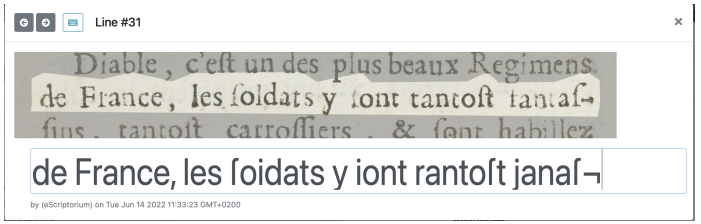


Figure: Prédiction issue du modèle GallicorporaAntiqua

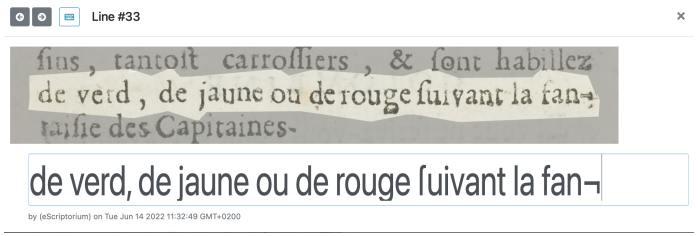


Figure: Prédiction issue du modèle Gallicorpora+

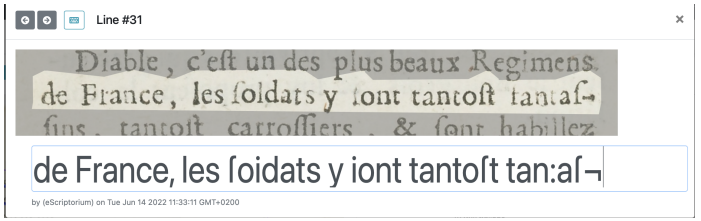


Figure: Prédiction issue du modèle GallicorporaAntiquaGothique

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 *Handrwitten Text Recognition*

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

BnF, département Droit, économie, politique, 2012-39571, 18e s.

Figure: Prédiction issue du modèle baseline

Figure: Prédiction issue du modèle Gallicorpora+

BnF, Réserve des livres rares, RES-Z-2442, 16e s.

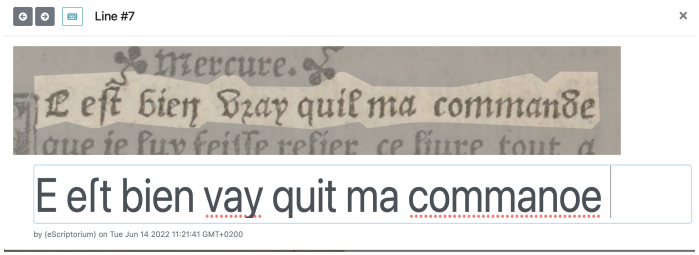


Figure: Prédiction issue du modèle baseline

Figure: Prédiction issue du modèle Gallicorpora+

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 *Handrwitten Text Recognition*

Exemples de prédiction sur des documents "out-of-domain" : La bonne surprise des incunables, BnF, Réserve des livres rares, vélins 611, 15^{e} s.

Figure: Prédiction issue du modèle Gallicorpora+

Figure: Prédiction issue du modèle cremma-medievalGallicorpora15

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 *Handrwitten Text Recognition*

Cas limites, les manuscrits : BnF, Manuscrits, Fr. 122, 14e s.

Figure: Prédiction issue du modèle Gallicorpora+

Figure: Prédiction issue du modèle cremma-medieval

Figure: Prédiction issue du modèle cremma-medievalGallicorpora15

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

Limites et Gains

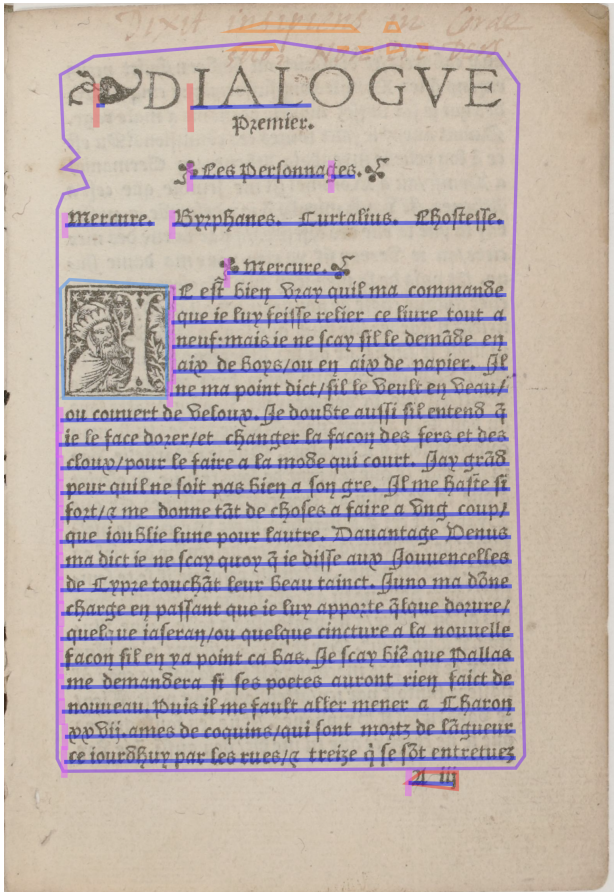

DIALOGVE
Premier.
Les Personnages.
Mercure. Byrphanes. Curtalius. Hostesse.
Mercure.
Il est bien vray quil ma commande que ie luy feisse relier ce liure tout a neuf: mais ie ne scay sil le demãde en aiz de bois/ou en aiz de papier. Il ne ma point dict/sil le veult en veau/ ou couuert de velouy. Ie doubte aussi sil entend q̃ ie le face dozer/et changer la facon des fers et des clouz/pour le faire a la mode qui court. Jay grãd peur quil ne soit pas bien a son gre. Il me haste si fort/q̃ me donne tãt de choses a faire a vng coup/ que ioublie lune pour lautre. Dauantage Venus ma dict ie ne scay quoy q̃ ie disse aux Jouuencelles de Cypre touchãt leur beau tainct. Juno ma dõne charge en passant que ie luy appoꝛte q̃que doꝛure/ quelque iaseran/ou quelque cincture a la nouuelle facon sil en va point ca bas. Ie scay bien que Pallas me demandera si ses poetes auront rien faict de nouueau. Puis il me fault aller mener a Charon xxviij. ames de coquins/qui sont morts de lagueur ce iourdhuy par les rues/q̃ treize q̃ se sõt entretuez
A iij

Figure: BnF, département Réserve des livres rares, RES-Z-2442, 16^{e} 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+

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

De l'extraction du texte à sa publication

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

L'objectif de la transformation ALTO → TEI

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>`

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

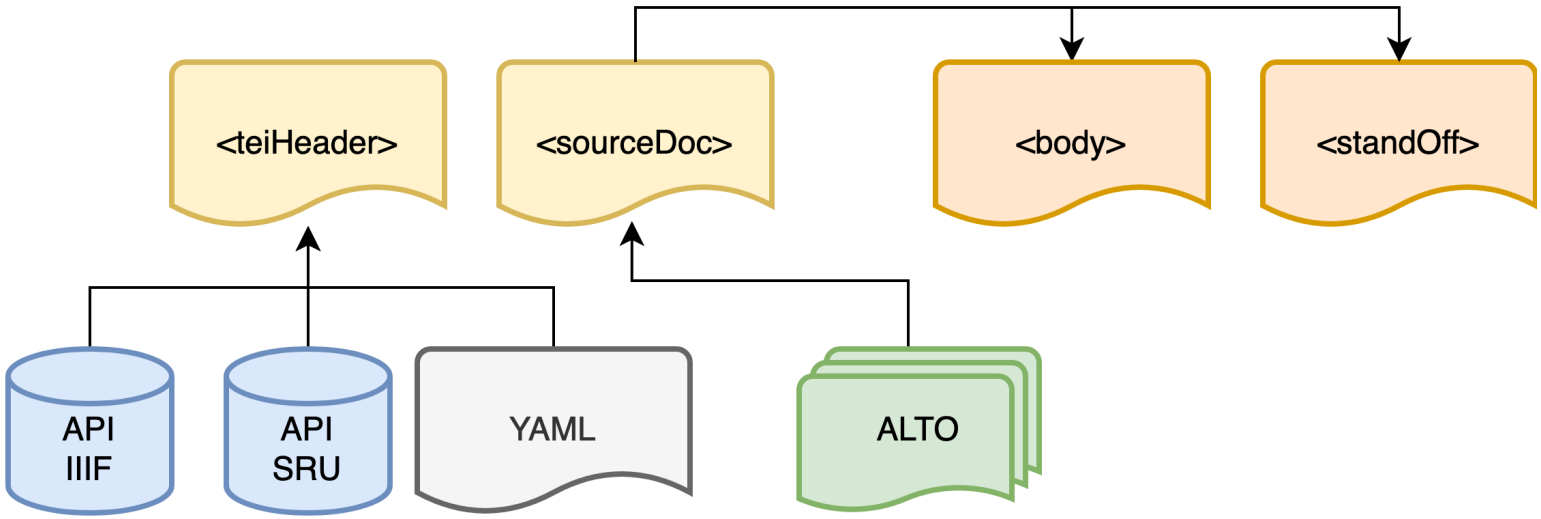
4. Texte annoté

Mapping

Conclusion

Construction automatisée de l'arbre TEI

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



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

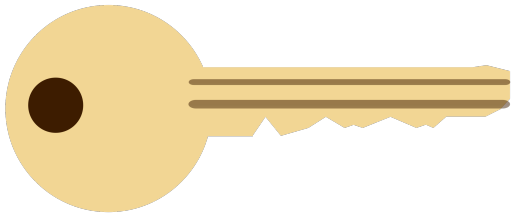
Mapping

Conclusion

Sources des métadonnées

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)

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 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

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 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

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

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

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
  - responabilité
  - publication
  - description de l'objet

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

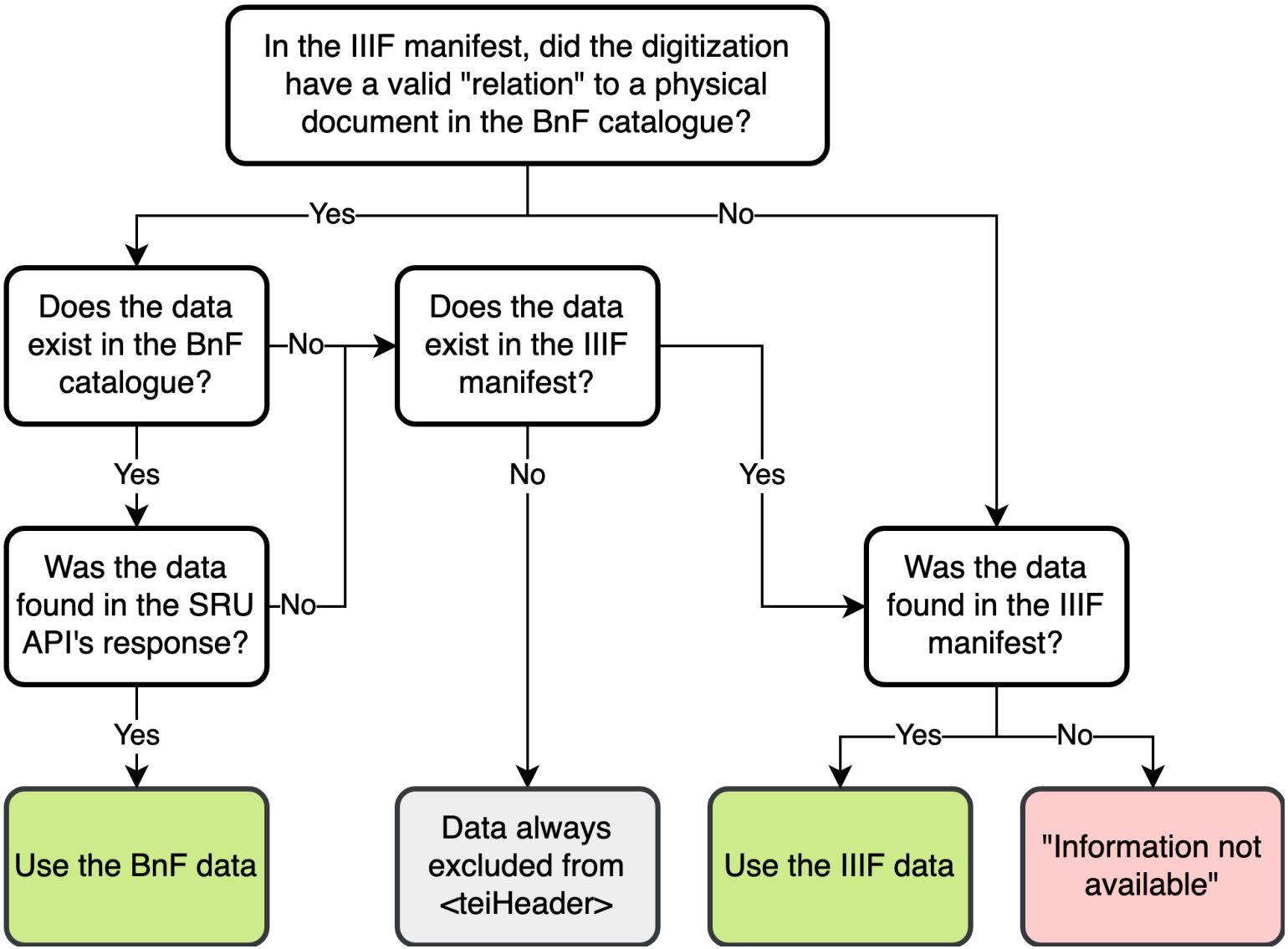
Mapping

4. Texte annoté

Mapping

Conclusion

Stratégie d'atténuation des risques



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

Résultat du mapping des métadonnées

```xml
<?xml version='1.0' encoding='UTF-8'?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="ark_12148_bpt6k6424218b">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Le Romant comique [1re partie], par Mr Scarron</title>
        <author xml:id="Sc1">
          <persName>
            <forename>Paul</forename>
            <surname>Scarron</surname>
            <ptr type="isni" target="0000000120990126"/>
          </persName>
        </author>
        <respStmt>
          <resp>Transformation from ALTO4 to TEI by</resp>
          <persName>
            <forename>Kelly</forename>
            <surname>Christensen</surname>
            <ptr type="orcid" target="000000027236874X"/>
          </persName>
          <persName>
            <forename>Simon</forename>
            <surname>Gabay</surname>
            <ptr type="orcid" target="0000000190944475"/>
          </persName>
          <persName>
            <forename>Ariane</forename>
            <surname>Pinche</surname>
            <ptr type="orcid" target="0000000278435050"/>
          </persName>
        </respStmt>
      </titleStmt>
      <extent>
        <measure unit="images" n="20"/>
      </extent>
      <publicationStmt>
        <publisher>Gallic(orpor)a</publisher>
        <authority>BnF DATALab</authority>
        <availability status="restricted" n="cc-by">
          <licence target="https://creativecommons.org/licenses/by/4.0/"/>
        </availability>
        <date when="2022-06-10"/>
      </publicationStmt>
      <sourceDesc>
        <bibl>
      </publicationStmt>
      <sourceDesc>
        <bibl>
          <ptr target="http://catalogue.bnf.fr/ark:/12148/cb31308524b"/>
          <author ref="#Sc1">
            <persName>
              <forename>Paul</forename>
              <surname>Scarron</surname>
              <ptr type="isni" target="0000000120990126"/>
            </persName>
          </author>
          <title>Le Romant comique [1re partie], par Mr Scarron</title>
          <pubPlace key="NL">Leiden</pubPlace>
          <publisher>J. Sambix</publisher>
          <date when="1655" cert="high" resp="BNF">1655</date>
        </bibl>
        <msDesc>
          <msIdentifier>
            <country key="FR"/>
            <settlement>Information not available.</settlement>
            <repository>Bibliothèque nationale de France</repository>
            <idno>8-Y2-55998</idno>
            <altIdentifier>
              <idno type="ark">bpt6k6424218b</idno>
            </altIdentifier>
          </msIdentifier>
          <physDesc>
            <objectDesc>
              <p>Texte imprimé</p>
            </objectDesc>
          </physDesc>
        </msDesc>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <langUsage>
        <language ident="fre">français</language>
      </langUsage>
    </profileDesc>
  </teiHeader>
  <sourceDoc>
    <surfaceGrp>
      <surface xml:id="f15" n="0" ulx="0" uly="0" lrx="1189" lry="2146">
        <graphic url="https://gallica.bnf.fr/iiif/ark:/12148/bpt6k6424218b/f15/ful
        <zone xml:id="f15_z1" type="MainZone" subtype="none" n="none" points="59,1
```


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

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

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>...
```


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

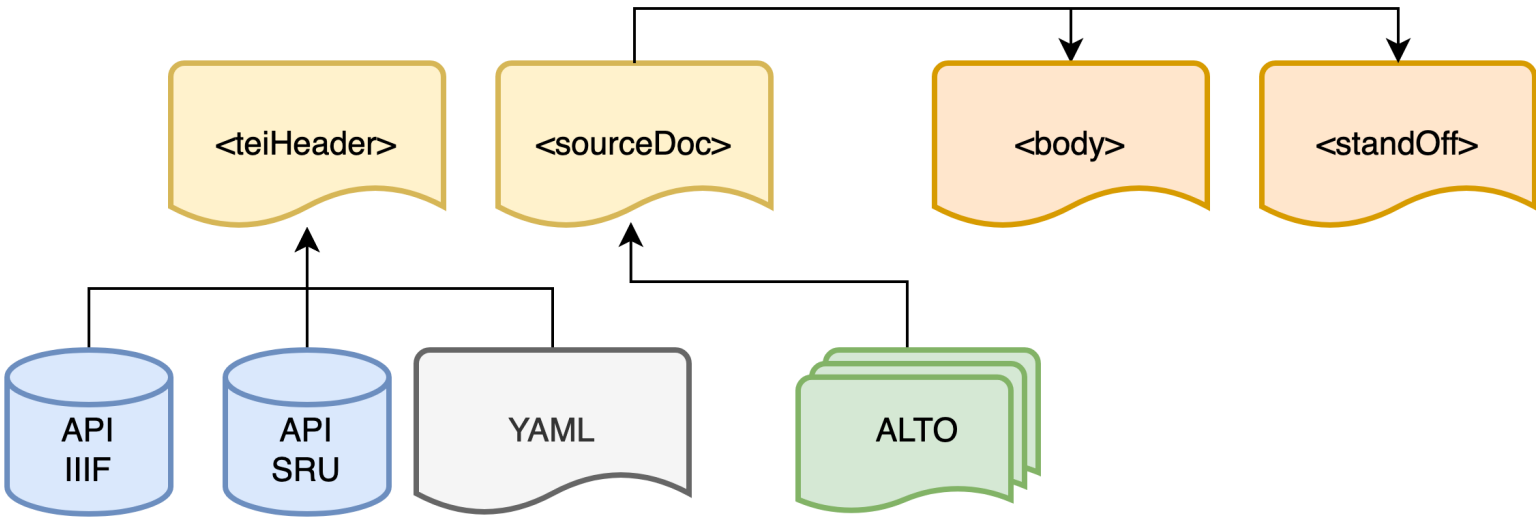
Mapping

4. Texte annoté

Mapping

Conclusion

3. Texte pré-éditorialisé (<body>)



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 <sourceDoc> → <body>

```
<sourceDoc>
<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>
```

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


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

4. Texte annoté

- Lemmatisation
- Normalisation
- Reconnaissance des entités nommées

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

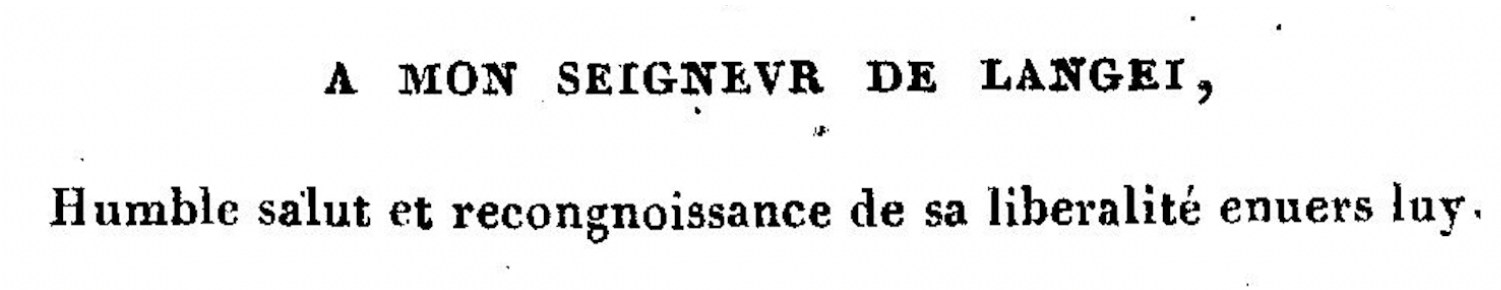
Mapping

4. Texte annoté

Mapping

Conclusion

Mapping <sourceDoc> → annotation



A MON SEIGNEVR DE LANGEI,

Humble salut et recongnoissance de sa liberalité enuers luy.

segment

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

Annotation du texte (plusieurs modèles TAL)

token	Humble	salut	et	recognoissance
lemma	Humble	salut	et	recognoissance
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

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

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