Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue

Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Javier Ortiz Suárez, Kelly Christensen, Simon Gabay, Ariane Pinche, Jean-Baptiste Camps

▶ **To cite this version:**

Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Javier Ortiz Suárez, Kelly Christensen, et al.. Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue: Restitution des travaux. DataLab de la BnF : Restitution des travaux 2022, DataLab de la BnF, Dec 2022, Paris, France. hal-03930542

HAL Id: hal-03930542
https://hal.science/hal-03930542

Submitted on 9 Jan 2023

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.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Gallic(orpor)a

Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue

B. Sagot[1] L. Romary[1] R. Bawden[1] P. J. Ortiz Suárez[1]
Kelly Christensen[1] **S. Gabay**[2] **A. Pinche**[3] J.B. Camps[3]

[1]Inria (équipe-projet ALMAnaCH)

[2]Université de Genève

[3]École nationale des chartes, Université PSL

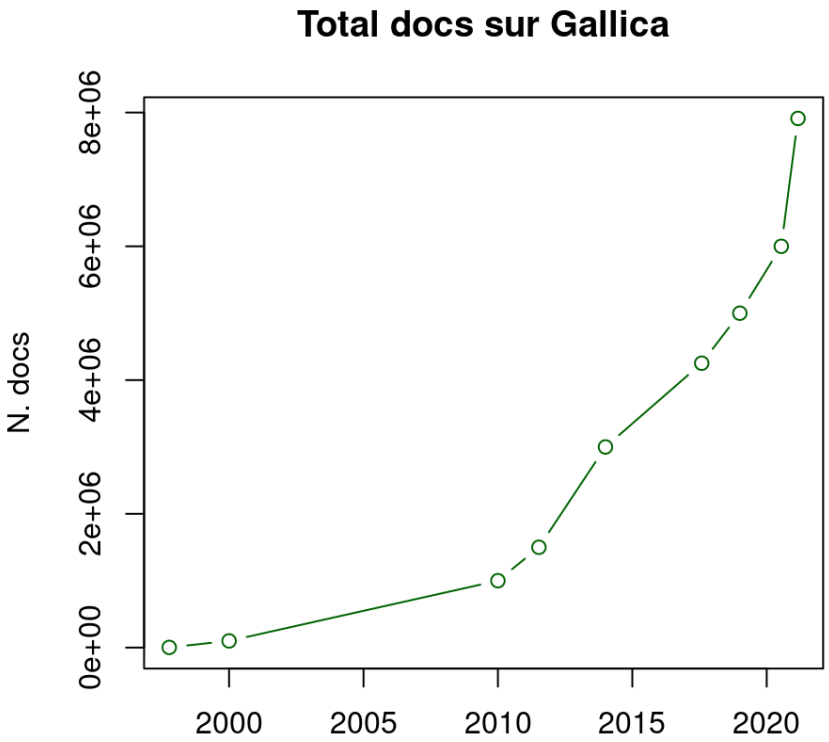
09 décembre 2022
Restitution des travaux

1. Les enjeux du projet Gallic(orpor)a
   - Les collections numérisées
   - Les objectifs du projet Gallic(orpor)a
   - Une chaîne de traitement
2. Les outils disponibles aujourd'hui
   - Projets préexistants
   - Les corpus gold Gallic(orpor)a
   - Les modèles Gallic(orpor)a
   - Les scripts Gallic(orpor)a
   - Les fichiers XML TEI générés à partir des prédictions HTR
3. Les difficultés rencontrées
   - Gérer l'hétérogénéité
   - Segmentation et premiers échecs
   - Passage à l'échelle de la production
   - Gestion du calendrier de projet
4. Conclusion

Les enjeux du projet Gallic(orpor)a
`https://github.com/Gallicorpora`

Gallica et le (Data) Deluge

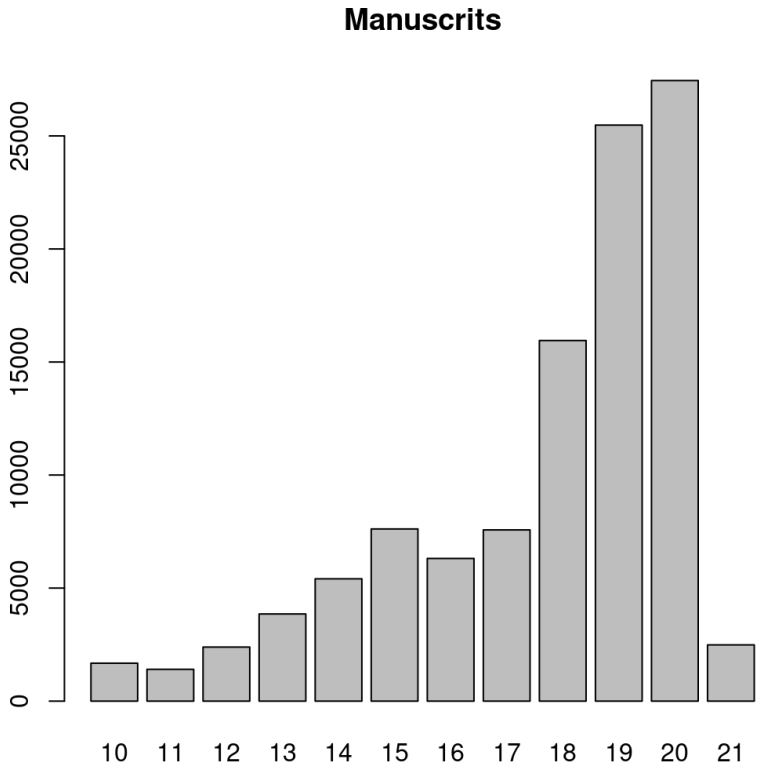
Data Deluge (Jim Gray, 2007).
1991 : 1 ; 1997 : 1 million ; 2021 : 200 millions de sites web.
Gallica : de **2500 à 8 millions de documents**



Danby, 1840. *Le Déluge*.

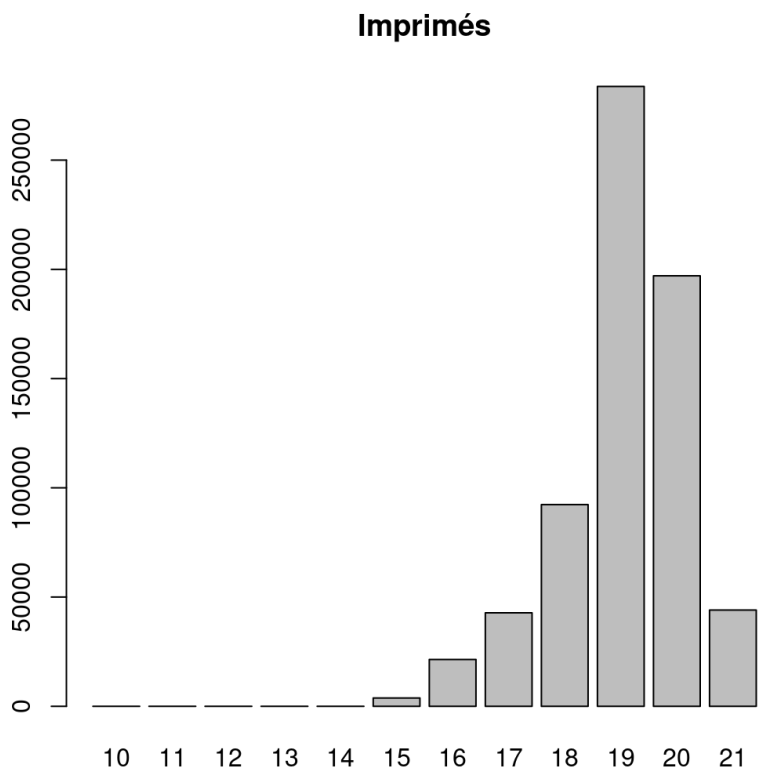
Manuscrits et imprimés sur Gallica

193 265 manuscrits



dont 52 188 avant 1800,
dont 4 avec couche texte

1 182 471 livres imprimés



dont 160 335 avant 1800,
dont 38 281 avec couche texte

Objectifs du projet Gallic(orpor)a

– Concevoir une chaîne de traitement de l'image à la **donnée structurée et enrichie**

– permettant l'**automatisation** des différentes tâches

– pour construire un corpus richement **annoté**

– qui permette **de valoriser et d'exploiter** les collections numériques de la BnF

Objectifs du projet Gallic(orpor)a

Des outils

Une chaîne de traitement unifiée, basée sur des outils *Open Source*
(Kraken, eScriptorium, Pie, CamenBERT...)

De quoi les entraîner

Un corpus de référence (*gold*, étalon-or).
(données annotées et vérifiées par l'humain)

Des résultats

Un corpus traité (*silver*)
(documents auxquels auront été appliqués les modèles)

Chaîne de traitement : Prévisionnelle

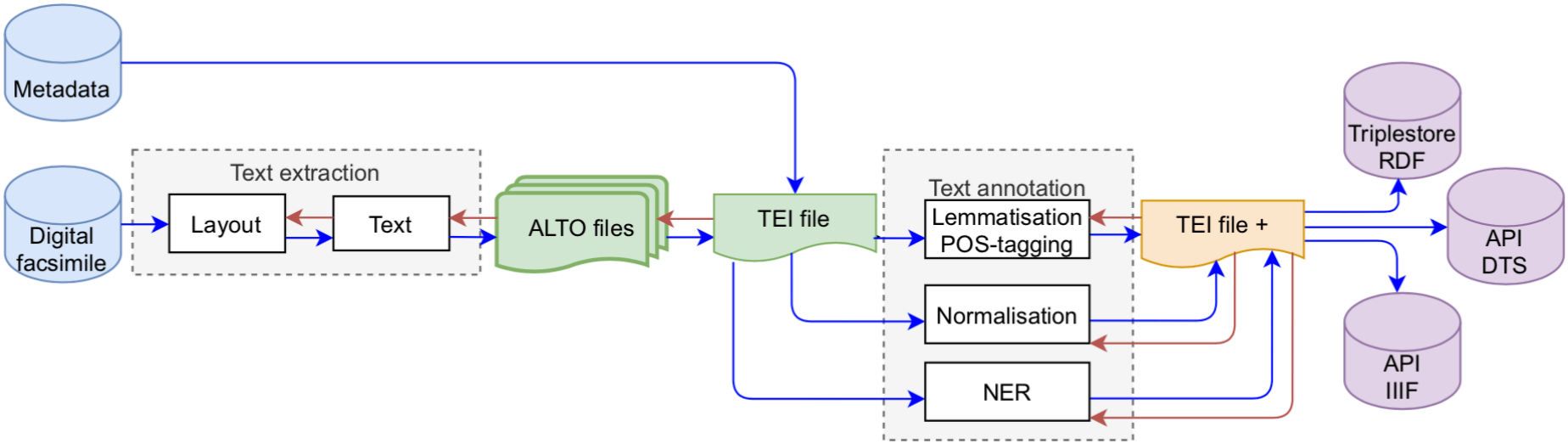


FIGURE – Chaîne de traitement du projet *Gallic(orpor)a.*

Les outils disponibles aujourd'hui

Projets préexistants en TAL

- **Deucalion** pour le traitement des documents médiévaux
Clérice (T.), Camps (J.B.) et Pinche (A.), *Deucalion, Modèle Ancien Francais (0.2.0)*, version v0.2.0, juin 2019, DOI : 10.5281/zenodo.3237455

- **E-ditiones** pour le traitement du français d'Ancien Régime
Gabay (S.), Clérice (T.), Camps (J.B.), Tanguy (J.B.) et Gille-Levenson (M.), « Standardizing linguistic data : method and tools for annotating (pre-orthographic) French », dans *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, 2020, p. 1-7

- **FREEM** pour le traitement automatique de la langue (XVI^{e}-XVIII^{e}
Gabay (S.), Ortiz Suarez (P.), Bawden (R.), Bartz (A.), Gambette (P.) et Sagot (B.), « Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime », dans *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*, Avignon, 2022, p. 154-165

Projets préexistants en HTR

- **CREMMALab** pour le traitement des documents médiévaux
  Pinche (A.), *Cremma Medieval*, juin 2022, URL : https://github.com/HTR-United/cremma-medieval (visité le 03/11/2022)

- **SegmOnto** pour la description de la mise en page
Gabay (S.), Camps (J.B.), Pinche (A.) et Jahan (C.), « SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more) », dans *Proceedings of the 1st International Workshop on Computational Paleography*, Lausann (Switzerland), 2021

- Des dépots de vérités de terrain :
  - OCR17, Jahan (C.) et Gabay (S.), *OCR17 +*, version 1.0, juil. 2021, URL : https://github.com/e-ditiones/OCR17plus
  - FONDUE-FR-PRINT-16, Gabay (S.), *Transcriptions of French 16th c. prints*, mars 2021, URL : https://github.com/FoNDUE-HTR/FONDUE-FR-PRINT-16
  - Pictocatalogs - Datasets for catalogs OCR and segmentation, Pradier (F.), *PictoCatalogs - Datasets for historical catalogs OCR and segmentation*, version 1.0, mai 2021, URL : https://github.com/PictoCatalogs/TrainingDataOCR

Les corpus gold Gallic(orpor)a pour la période médiévale

Les corpus d'entraînement pour l'HTR

– Pour les manuscrits du 15e siècle, Pinche (A.), Gabay (S.), Leroy (N.) et Christensen (K.), *Données HTR manuscrits du 15e siècle*, URL : https://github.com/Gallicorpora/HTR-MSS-15e-Siecle

– Pour les incunables du 15e siècle, Pinche (A.), Gabay (S.), Leroy (N.) et Christensen (K.), *Données HTR manuscrits du 15e siècle*, URL : https://github.com/Gallicorpora/HTR-MSS-15e-Siecle

```
<TextLine ID="eSc_line_35615467"
          TAGREFS="LT1056"
          BASELINE="1495 4225 3522 4206"
          HPOS="1477"
          VPOS="4091"
          WIDTH="2045"
          HEIGHT="180">
  <Shape><Polygon POINTS="1495 4225 1477 4137 1518 4105 1523 4105 1527 4105 1532 4105 1929 4123 1985 4100 1989 4100 1
    <String CONTENT="premierement ceste presente table en la quelle on trou"
          HPOS="1477"
          VPOS="4091"
          WIDTH="2045"
          HEIGHT="180"></String>
</TextLine>
```

Figure – Exemple de fichier XML ALTO

Les corpus gold Gallic(orpor)a pour les imprimés anciens

Les corpus d'entraînement pour l'HTR

– Pour les imprimés du 16e siècle, Gabay (S.), Pinche (A.), Vlachou-Efstathiou (m.) et Christensen (K.), *Données HTR imprimés du 16e siècle*, URL : https://github.com/Gallicorpora/HTR-imprime-16e-siecle

– Pour les imprimés du 17e siècle, Gabay (S.), Pinche (A.), Fabert (E.), Vlachou-Efstathiou (M.), Maxime (H.) et Christensen (K.), *Données imprimés du 17e siècle*, URL : https://github.com/Gallicorpora/HTR-imprime-17e-siecle

– Pour les imprimés du 18e siècle ; Gabay (S.), Pinche (A.), Fabert (E.) et Christensen (K.), *Données imprimés du 18e siècle*, URL : https://github.com/Gallicorpora/HTR-imprime-18e-siecle

Les corpus d'entraînement pour la lemmatisation

– Pour les documents du 15e au 18e siècle, Pinche (A.), Gabay (S.), Nguyen (M.) et Laroche (J.), *Gallicorpora lemmatisation dataset*, URL : https://github.com/Gallicorpora/Lemmatisation

Les corpus gold Gallic(orpor)a pour les imprimés anciens

Les données pour l'HTR ont été réalisées grâce à **eScriptorium**
Kiessling (B.), Tissot (R.), Stokes (P.) et Stökl Ben Ezra (D.), « eScriptorium : An open source platform for historical document analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, t. 2, p. 19-24

Les données de lemmatisation ont été réalisées grâce à l'interface de post-correction **Pyrrha** :
Clérice (T.), Pilla (J.) et Camps (J.B.), *hipster-philology/pyrrha : 1.0.1*, déc. 2018, DOI : 10.5281/zenodo.2325428

Les modèles Gallicorpora

– Les modèles de segmentation

- Un modèle de segmentation affiné à partir de blla.mlmodel qui est le modèle par défaut de segmentation de Kraken
  KIESSLING (B.), « Kraken - an Universal Text Recognizer for the Humanities », dans *Digital Humanities 2019 Conference Abstracts*, Utrecht, The Netherlands, 2019, URL : https://dev.clariah.nl/files/dh2019/boa/0673.html (visité le 16/03/2020)
- Un modèle entraîné à partir de YALtAi
  CLÉRICE (T.), *You Actually Look Twice At it (YALTAi) : using an object detection approach instead of region segmentation within the Kraken engine*, juil. 2022, URL : https://hal-enc.archives-ouvertes.fr/hal-03723208

Les modèles Gallicorpora

Pinche (A.) et Gabay (S.), *Segmentation and HTR Model*, URL : https://github.com/Gallicorpora/Segmentation-and-HTR-Models

- Les modèles HTR
  - Le modèle Gallicorpora+ pour les imprimés du 16e au 19e siècle (98,66%, test score).
  - Le modèle Cremma-medievalGallicorpora15, aussi appelé Cortado, pour les manuscrits et les incunables (95.54%, test score).

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

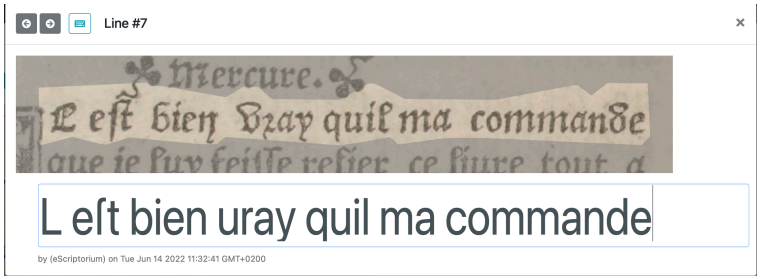


Figure – Prédiction issue du modèle Gallicorpora+

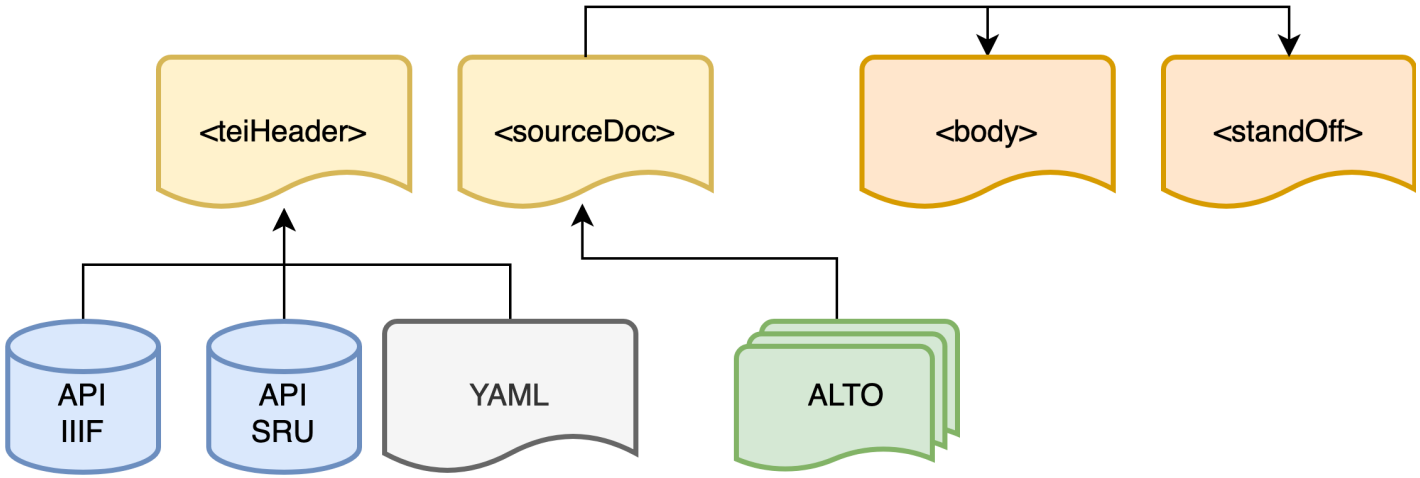
BnF, Réserve des livres rares, vélins 611, 15e s.

Figure – Prédiction issue du modèle Cortado

Les scripts Gallicorpora

– **Transformation alto → TEI**

Christensen (K.), Gabay (S.) et Pinche (A.), *AltoToTei*, URL : https://github.com/Gallicorpora/application



– **Annotator** : Injection de l'annotation linguistique (en cours de développement)

Bartz (A.), Gabay (S.) et Juliette (J.), *Annotator*, août 2022, URL : https://github.com/e-ditiones/Annotator (visité le 06/12/2022)

Les fichiers XML TEI pré-éditorialisés

Mise à disposition d'un dépôt avec :

- Un corpus *manuscrits et incunables*
- Un corpus *imprimés anciens*

Christensen (K.), Gabay (S.) et Pinche (A.), *TEI Corpus du projet Gallicorpora*, version 1.0.0, juil. 2022, DOI : 10.5281/zenodo.1234

```
/>Plaintes qu'on lift au front de ces ormes empraintes,<lb
    corresp="#f9-block_0-line_35-lineCount36"/>Nym¬</ab>
<pb corresp="#f10"/>
<ab corresp="#f10-block_0-blockCount1" type="MainZone"><lb
    corresp="#f10-block_0-line_0-lineCount1"/>I.<lb corresp="#f10-block_0-line_1-lineCount2"
    />LEMLEAV<lb corresp="#f10-block_0-line_2-lineCount3"/>Nymphes vous le ſauez, &amp;amp; vous
  qui habitez<lb corresp="#f10-block_0-line_3-lineCount4"/>Satyres dans les creux de ces
    obſcuritez.<lb corresp="#f10-block_0-line_4-lineCount5"/>Meſme le beau criſtal de ces
  viues fontaines<lb corresp="#f10-block_0-line_5-lineCount6"/>Le murmure en coulant, par
  ces herbeuſes plaiies<lb corresp="#f10-block_0-line_6-lineCount7"/>L E.<lb
    corresp="#f10-block_0-line_7-lineCount8"/>N'as to pai veu, Francin, machotter les
    brebi<lb corresp="#f10-block_0-line_8-lineCount9"/>L'herbe demibrulee, au milieu des
    herbiſẽ<lb corresp="#f10-block_0-line_9-lineCount10"/>Briſer nos chalumeaux? &amp;amp; de
  mille ruynes<lb corresp="#f10-block_0-line_10-lineCount11"/>Saccager les roufeaux de nos
  pauures caßines?<lb corresp="#f10-block_0-line_11-lineCount12"/>Au lieu d'épiz creſtez
  naiſtre ſur les fillons.<lb corresp="#f10-block_0-line_12-lineCount13"/>Des chardons
  heriſſez en pointes d'aiguillons?<lb corresp="#f10-block_0-line_13-lineCount14"/>Les porcs
  dans les ruiſſeaux? &amp;amp; troubler dans la pre<lb
    corresp="#f10-block_0-line_14-lineCount15"/>Leau que tous les bergers tenoient comme
    ſacree<lb corresp="#f10-block_0-line_15-lineCount16"/>De carmes enchantez la Lune
    enſorceler?<lb corresp="#f10-block_0-line_16-lineCount17"/>Faire tarir le lait, &amp;amp; le
  pis defenfler<lb corresp="#f10-block_0-line_17-lineCount18"/>De la vache laitiere, &amp;amp;
  de mauuaiſe œillade<lb corresp="#f10-block_0-line_18-lineCount19"/>Rendre tout le
  troupeau, &amp;amp; galeux, &amp;amp; malade:<lb corresp="#f10-block_0-line_19-lineCount20"/>Bref,
  i'eſtime celuy trois &amp;amp; trois fois heureux<lb
    corresp="#f10-block_0-line_20-lineCount21"/>Qui mourant n'a point veu vn ciel ſi
    malheureux<lb corresp="#f10-block_0-line_21-lineCount22"/>ERANCIN<lb
    corresp="#f10-block_0-line_22-lineCount23"/>Quelle greſle, quel vent, quel malheur, quel
    orag<lb corresp="#f10-block_0-line_23-lineCount24"/>Quelle eſtrange fureur, quel infame
    pillage<lb corresp="#f10-block_0-line_24-lineCount25"/>Quelle rage du Ciel, quelle nue
    d'erreur<lb corresp="#f10-block_0-line_25-lineCount26"/>Quelle mauuaiſe main, a dérobé
    l'honneur<lb corresp="#f10-block_0-line_26-lineCount27"/>Le repos, &amp;amp; la paix, la
  gloire, &amp;amp; la vaillance<lb corresp="#f10-block_0-line_27-lineCount28"/>L'heritage
  ſacré de noſtre douce France?<lb corresp="#f10-block_0-line_28-lineCount29"/>Pleurez
  villes, chaſteaux, &amp;amp; verſez larmes d'yeus<lb
    corresp="#f10-block_0-line_29-lineCount30"/>Satyres, Cheurepiez, Faunes &amp;amp;
  Demidieux</ab>
<pb corresp="#f11"/>
```

FIGURE – Exemple de fichier XML TEI généré automatiquement.

Annotation linguistique

Une série de textes OCRisée ne forme pas un corpus requêtable / exploitable. Il faut annoter les données extraites depuis le fac-similé numérique. Plusieurs tâches de TAL sont déjà opérationnelles (avec différents degrés de fiabilité) :

- Lemmatisation, POS et morphologie pour tous les états de langues
- Normalisation linguistique (=alignement avec le français contemporain) pour les périodes XVI^{e}-XVIII^{e} s.
- Reconnaissance des entités nommées (XVI^{e}-XVIII^{e} s.)
- *Entity linking* pour les lieux

Un grand merci à J. Janès, A. Bartz, Ph. Gambette, Th. Clérice pour les scripts, données et modèles.

Les difficultés rencontrées

Gérer l'hétérogénéité du corpus

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

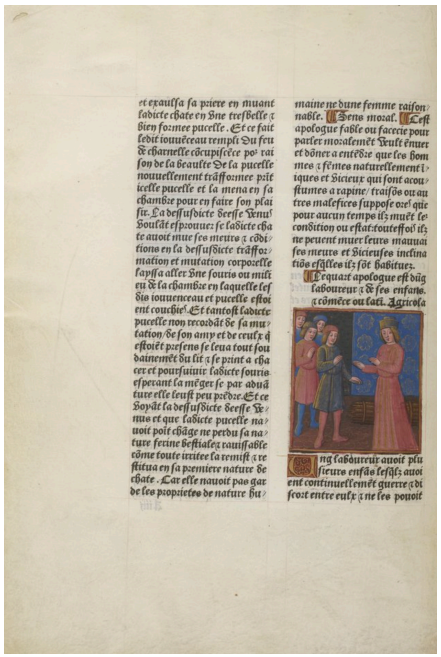


FIGURE – BnF, Réserve des livres rares, velin, 15e s.

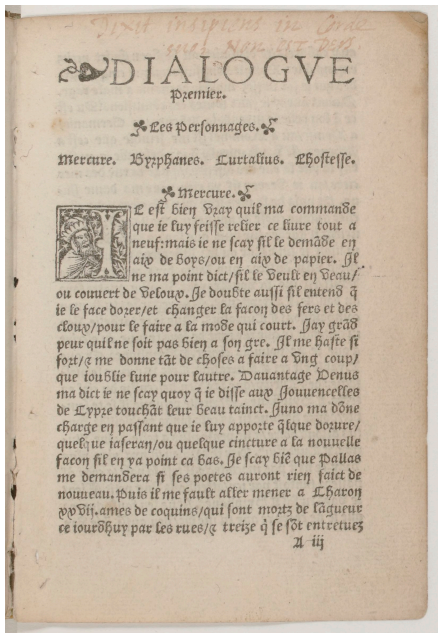


FIGURE – BnF, Réserve des livres rares, RES-Z-2442, 16e s.

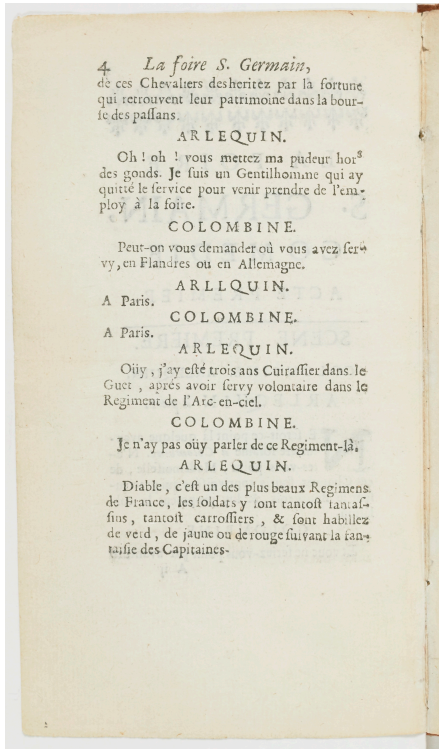


FIGURE – BnF, Arts du spectacle, réserve 8-RO-1702, 17e s.

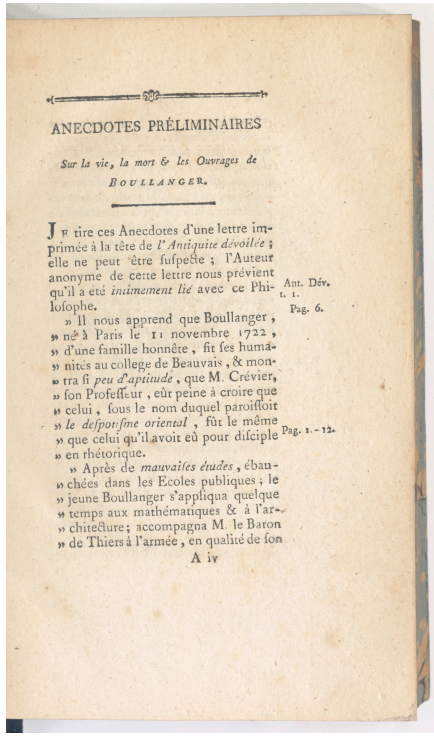


FIGURE – BnF, Droit, économie, politique, 2012-39571, 18e s.

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

Segmentation et premiers échecs

Le modèle de segmentation *Kraken* ne nous a pas permis d'obtenir des résultats satisfaisants. Problème d'hétérogénéité des mises en page ? De granularité de la description ? Ou problème méthodologique ?

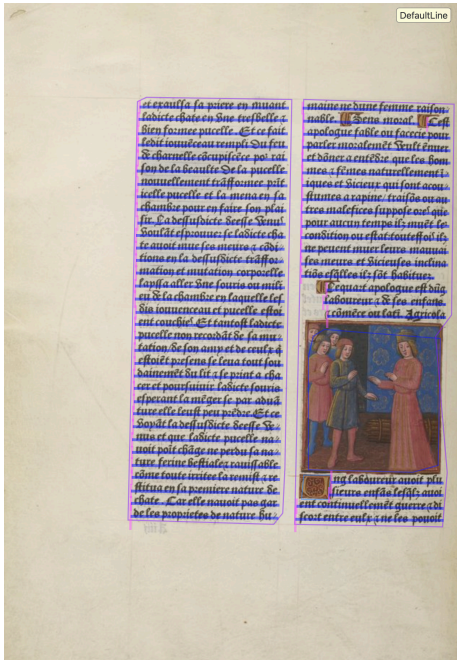


FIGURE – BnF, Réserve des livres rares, vélins 611, 15e s.

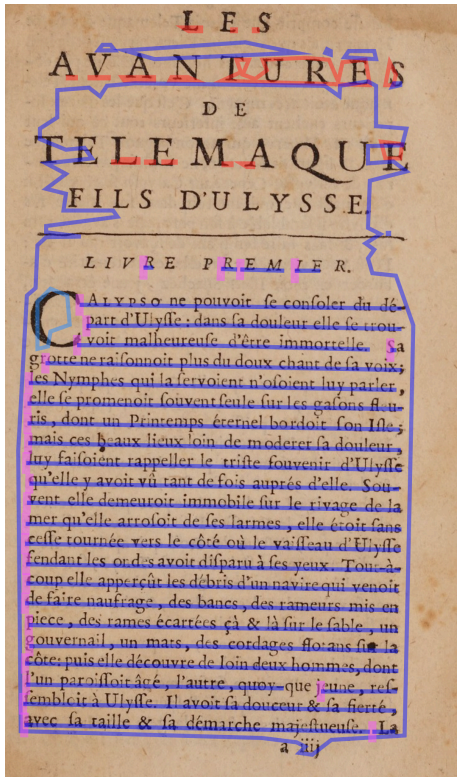


FIGURE – BnF, Réserve des livres rares, RES-Z-2442, 16e s.

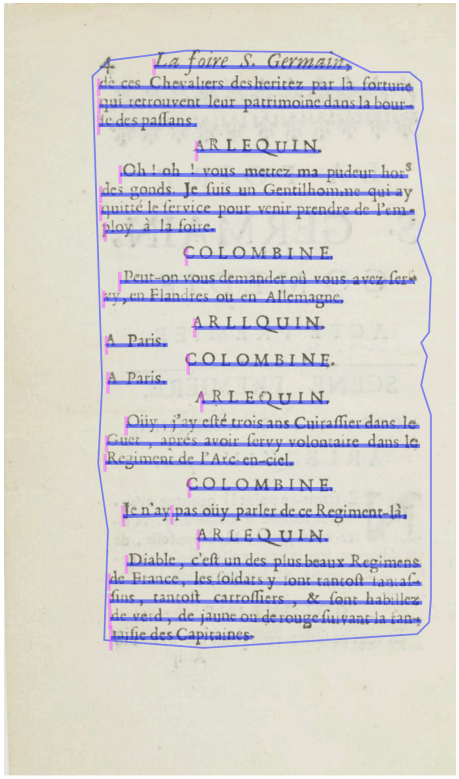


FIGURE – BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

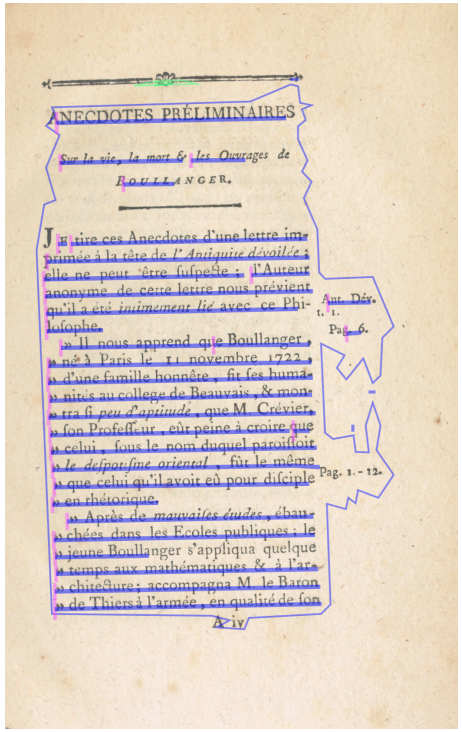


FIGURE – BnF, département Droit, économie, politique, 2012-39571, 18e s.

Une nouvelle approche : YALTAi

Une nouvelle approche pour l'analyse de la mise en page : **YALTAi**
Clérice (T.), *You Actually Look Twice At it (YALTAi) : using an object detection approach instead of region segmentation within the Kraken engine*, juil. 2022, URL : https://hal-enc.archives-ouvertes.fr/hal-03723208

- Ne fonctionne pas sur la classification des pixels des images
- Utilise de la détection d'objet en utilisant YoloV5

Nouveaux scores :

Zone	Main	Graphic	DropCapital	MarginText	Numbering	RunningTitle
Kraken	43.5	16.1	23.3	0	0	0
Yolo V5	91.7	48.4	69.2	48.3	75.8	45.6

Une nouvelle approche : YALTAi

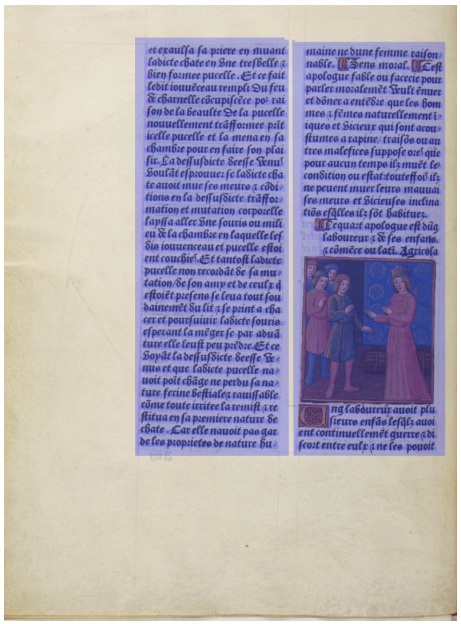


FIGURE – BnF, Réserve des livres rares, vélins 611, 15e s.

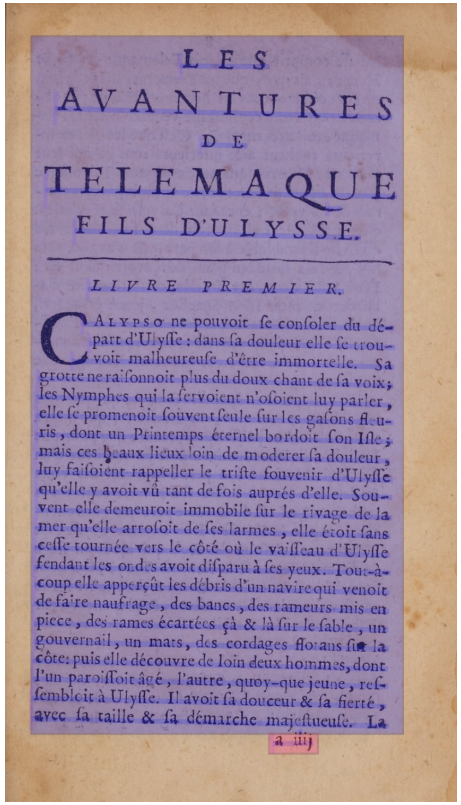


FIGURE – BnF, Réserve des livres rares, RES-Z-2442, 16e s.

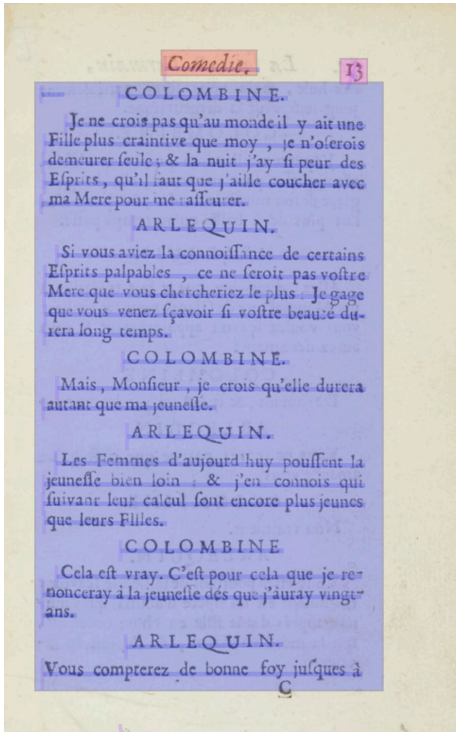


FIGURE – BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

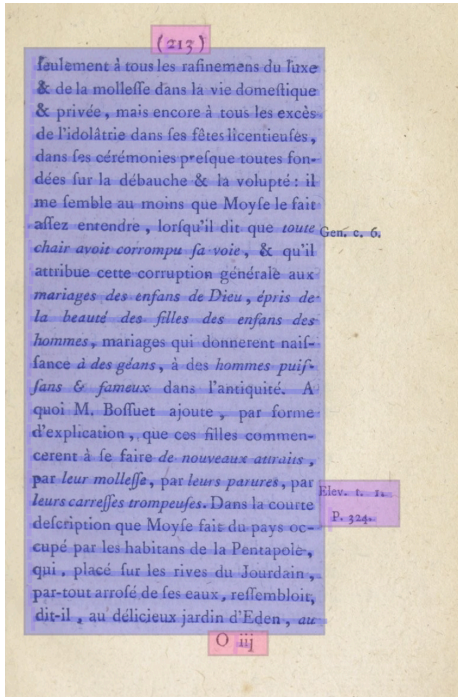


FIGURE – BnF, département Droit, économie, politique, 2012-39571, 18e s.

Passage à l'échelle

Nous avons produit un corpus silver plus réduit.

- Problèmes rencontrés :
  - Temps de génération du corpus*
  - Des prédictions Kraken imparfaites
  - Un script AltoToTei encore à l'état de prototype
  - La partie sur l'annotation est encore à intégrer dans le pipeline

*une semaine avec un serveur équipé d'une carte graphique nvidia 3090 et parallélisation des tâches TANGE (O.), « GNU Parallel - The Command-Line Power Tool », *The USENIX Magazine*, 36–1 (févr. 2011), p. 42-47, URL : http://www.gnu.org/s/parallel

Passage à l'échelle

Le passage à l'échelle soulève de nombreuses questions :

- Kraken est-il prêt pour sérialiser la production de prédiction sur des milliers de documents ?
- Le pipeline ne serait-il pas plus pertinent pour une génération à la demande ?
- La taille des fichiers XML TEI va-t-elle être acceptable ou bien faut-il proposer plusieurs sorties ?

Gestion du calendrier de projet

- Recrutement tardif des vacataires pour faire les vérités de terrain
- Le script AltoToTEI a été écrit à partir de données d'entraînement
- Le script AltoToTEI n'a pas pu être testé à l'échelle
- Mise à jour de Kraken durant l'été qui a demandé de reprendre le script AltoToTEI
- Impossibilité d'intégrer la tâche de lemmatisation avant la fin du stage

Conclusion

- Un prototype mis à disposition librement et à développer dans les années à venir...
- Un projet orienté vers la science ouverte qui donne accès à toutes les données produites
- Des données qui ont permis de créer des modèles HTR performants et réutilisés
- Une valorisation des données d'entraînement produites avec leur intégration dans le catalogue *HTR-United*

Chagué (A.), Clérice (T.) et Chiffoleau (F.), *HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages*, sept. 2021, URL : https://github.com/HTR-United/htr-united

Conclusion

- Mise à disposition de documentation :
  - Pour l'analyse de mise en page : `https://segmonto.github.io`.
  - Pour la transcription : `https://cremmalab.hypotheses.org`
  - Pour les tâches d'annotation du texte : Simon Gabay, Jean-Baptiste Camps, Thibault Clérice. *Manuel d'annotation linguistique pour le français moderne (XVI^{e} -XVIII^{e} siècles) : Version B*. 2022.
  - Pour le fonctionnement du pipeline : Kelly Christensen, *Modélisation des transcriptions ALTO avec la TEI : En complétant le pipeline du projet Gallic(orpor)a*, Paris, École nationale des chartes, 2022, `https://github.com/kat-kel/TNAH-Memoire`.

Conclusion

- Un travail qui a donné lieu à des publications et des communications :

  - Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, et al., « From FreEM to D'AlemBERT : a Large Corpus and a Language Model for Early Modern French », *Proceedings of the 13th Language Resources and Evaluation Conference, European Language Resources Association*, Jun 2022, Marseille, France. pp.3367-3374.
  - Ariane Pinche, Kelly Christensen, Simon Gabay, « Between automatic and manual encoding : Towards a generic TEI model for historical prints and manuscripts », *TEI 2022 conference : Text as data*, Sep 2022, Newcastle, United Kingdom.