Chaînes d'acquisition et de pré-éditorialisation du texte

Des images à un balisage en XML TEI

Ariane Pinche [12]

[1]CNRS, [2]CIHAM - UMR 5648

25 mars 2024

Table of Contents

Qu'est-ce que le métier d'éditeur scientifique ?

- «Éditer», du latin *edere*: «mettre au jour, produire», signifie d'après le TLFi: «Assurer la reproduction, la publication et la diffusion d'une œuvre»;
- L'édition scientifique est l'héritière de méthodologies qui remontent aux premières tentatives de reconstitution d'un texte original ou du moins d'une version canonique pour des textes aux traditions complexes comme la Bible, les œuvres d'Homère, Cicéron, Virgile, Chrétien de Troyes ou encore Shakespeare.

Des nouvelles manières de voir l'édition de texte(s) ?

- Des méthodes qui diffèrent en fonction des objectifs scientifiques ?
  - Une partie des éditeurs interrogent la notion d'une version canonique du texte;
  - Dans les années 80, B. Cerquiglini défend, dans l'*Éloge de la variante*, la place du témoin manuscrit comme représentant d'un état de transmission du texte ayant sa propre valeur intrinsèque *(New Philology)*;
  - En 1991, Peter Shillingsburg définit son travail davantage comme la présentation d'un processus textuel, plutôt que comme l'établissement d'un produit final immuable;
  - Si les méthodes changent, les éditions scientifiques visent toutes l'établissement d'un texte fiable et contextualisé, quelle que soit l'alternative choisie.

Les différents niveaux d'édition numérique

- Mise à disposition d'un texte structuré;
- Mise à disposition d'un texte structuré et enrichi;
- Éditions scientifiques et/ou critiques avec plusieurs strates d'informations.

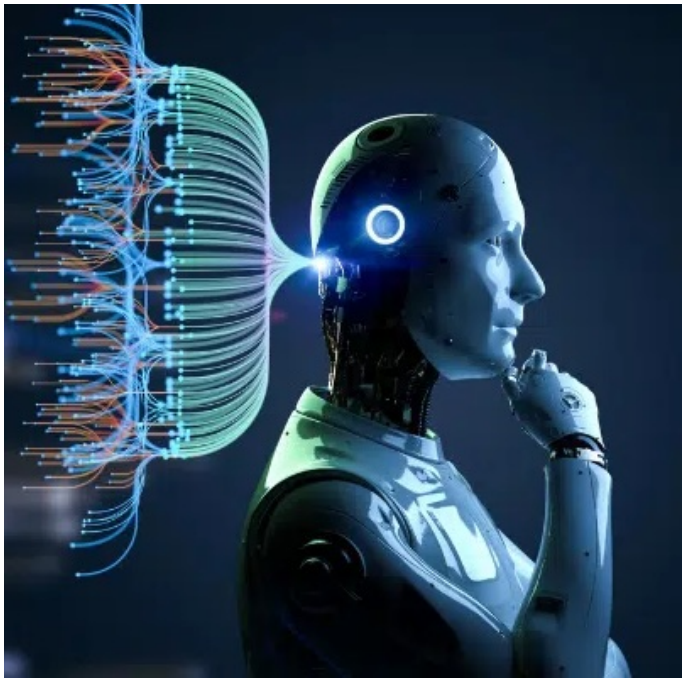
Un standard : XML TEI

Lou Burnard et Marjorie Burghart, *Qu'est-ce que la Text Encoding Initiative ?*, 2015

"La TEI a été d'abord développée, il y a plus de trente ans, comme un projet de recherche dans le champ alors émergent du «Humanities computing». L'idée originelle était de proposer un ensemble de recommandations sur la façon dont les chercheurs devraient créer des ressources textuelles «lisibles par ordinateur», qui soient adaptées aux besoins de la recherche – dans la mesure où un consensus existait sur le sujet –, mais qui soient également extensibles, puisque ces besoins changent et évoluent."

Table of Contents

L'IA, ça vous dit quelque chose ?



- Nous sommes aujourd'hui face à un bond technologique qui vient modifier les pratiques de l'édition numérique. Mais qu'est-ce que l'**intelligence artificielle** ?
- L'Intelligence artificielle (IA) désigne le développement de machines et de systèmes informatiques capables d'exécuter des tâches qui nécessitent normalement l'intelligence humaine.

L'IA, ça vous dit quelque chose ?

- Quel domaine d'application en SHS ?
  - **Vision par Ordinateur :** Analyse et interprétation d'images et de vidéos pour l'acquisition automatique de texte depuis des images.
  - **Traitement du Langage Naturel (NLP) :** Compréhension et génération de langage naturel par les machines.

Qu'est-ce que ça change pour les éditeurs numériques ?

- Des tailles de corpus plus importantes:
  - Édition d'œuvres longues;
  - Édition d'œuvres complètes;
  - Édition d'œuvres sérielles;
- Automatisation du processus d'acquisition et d'enrichissement du texte

Qu'est-que l'ATR (Automatic Text Recognition) ?

Figure: Prédiction ATR

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

Les étapes de l'ATR

- Numérisation des sources
- Chargement des images
- Segmentation des zones de l'image
- Segmentation des lignes contenant du texte
- Prédiction du texte qui se trouve sur l'image
- Export des données (txt, alto, page)

Segmentation des zones

- Segmentation des zones de l'image

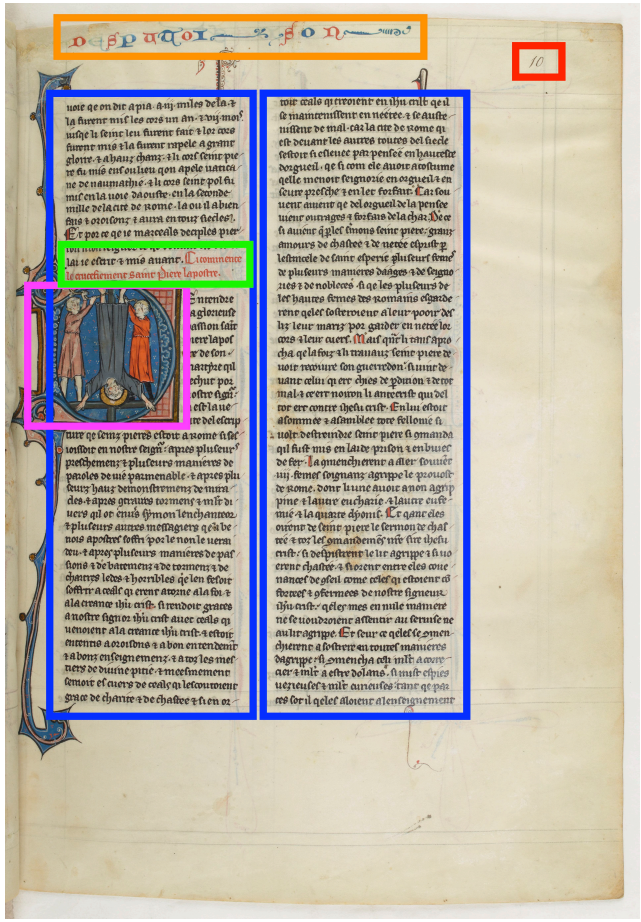
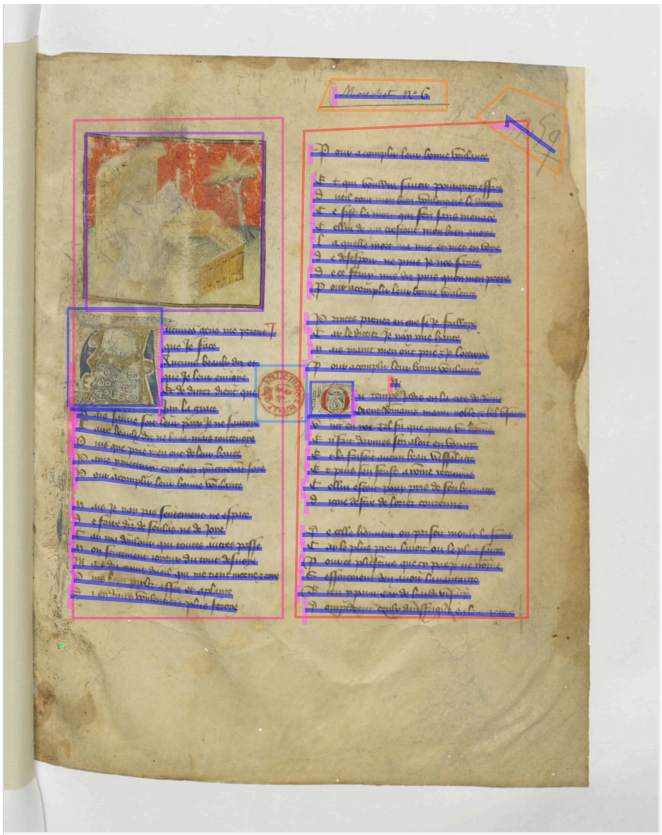


Figure: Bnf, fr. 412, fol.10r

Prédiction du texte

- Prédiction du texte qui se trouve sur l'image



```
1  ucunes gens me prient ꝑ.
2  que le face
3  Aucuns beaulx diz et
4  que le leur enuoye
5  Et de ditter dient que
6  A
7  Mais sauue soit leur paix le ne sauroye
8  lay la grace
9  Faire beaulx diz ne bons. mais touteuoye
10 Puis que prie men ont de leur bonte
11 Paine y mettray combien qũgnouĩt soye
12 Pour acomplir leur bonne uoulante
13 Mais le n ay pas sentement ne espace
14 De faire diz. de soulas ne de loye
15 Car ma doulour qui toutes autres passe
16 Mon sentement ioyeux du tout desuoye
17 Mais du grant dueil qui me tient morne ꝯcoye
18 Puis bien parler assez et aplante
19 Si en diray uoulentiers plus feroye
20 6259
21 Pour acomplir leur bonne uoulante
22 Et qui uouldra sauoir pourquoy efface
23 Dueil tout mon bien. uoulentiers le diroye
24 Ce fist la mort qui feri sans menace
25 Cellui de qui trestout mon bien auoye
26 Laquelle mort ma mis et met en uoye
27 De desespoir ne puis le nos sante
28 De ce feray mes diz puis quon men proye
29 Pour acomplir leur bonne uoulente
30 Princes prenez en gre se le failloye
31 Car le dittier le nay mie hante
32 Mais maint men ont prie ꝯ le lottroye
33 Pour acomplir leur bonne uoulante
34 u temps ladis en le cite de Rõme
35 .ii.
36 O
37 ung en yot. Tel fu que quant un hõme
38 Orent Rõmains maint noble ꝯ bel usaige
```

Figure: Bnf, fr. 12779, fol.9r

Les étapes de l'ATR

- Export des données (txt, alto, page)

```
<Layout>
        <Page WIDTH="4648" HEIGHT="3407" PHYSICAL_IMG_NR="8" ID="eSc_dummypage_">
                <PrintSpace HPOS="0" VPOS="0" WIDTH="4648" HEIGHT="3407">

                        <TextBlock HPOS="693" VPOS="321" WIDTH="1701" HEIGHT="2451"
                                ID="eSc_textblock_08b9f915" TAGREFS="BT3852">
                                <Shape>
                                        <Polygon
                                          POINTS="693 413 693 2772 2394 2772 2254 321"/>
                                </Shape>


                                <TextLine ID="eSc_line_d939596f" TAGREFS="LT1299"
                                        BASELINE="746 476 2143 428" HPOS="743" VPOS="352"
                                        WIDTH="1400" HEIGHT="156">
                                        <Shape>
                                          <Polygon
                                          POINTS="2078 388 2050 388 2021 386 1993 383 1964 383 1936 380 1908 377 1876 374 1848 374 1820 371 1811
                                          />
                                        </Shape>
                                        <String
                                          CONTENT="fors de la uille. Tant fut lassault merueilleux et"
                                          HPOS="743" VPOS="352" WIDTH="1400" HEIGHT="156"/>
                                </TextLine>
```

Figure: Exemple de fichier Alto

Table of Contents

Principales étapes d'une chaîne éditoriale théorique

Depuis l'acquisition automatique du texte, on peut mettre en place des chaînes de traitement de l'information.

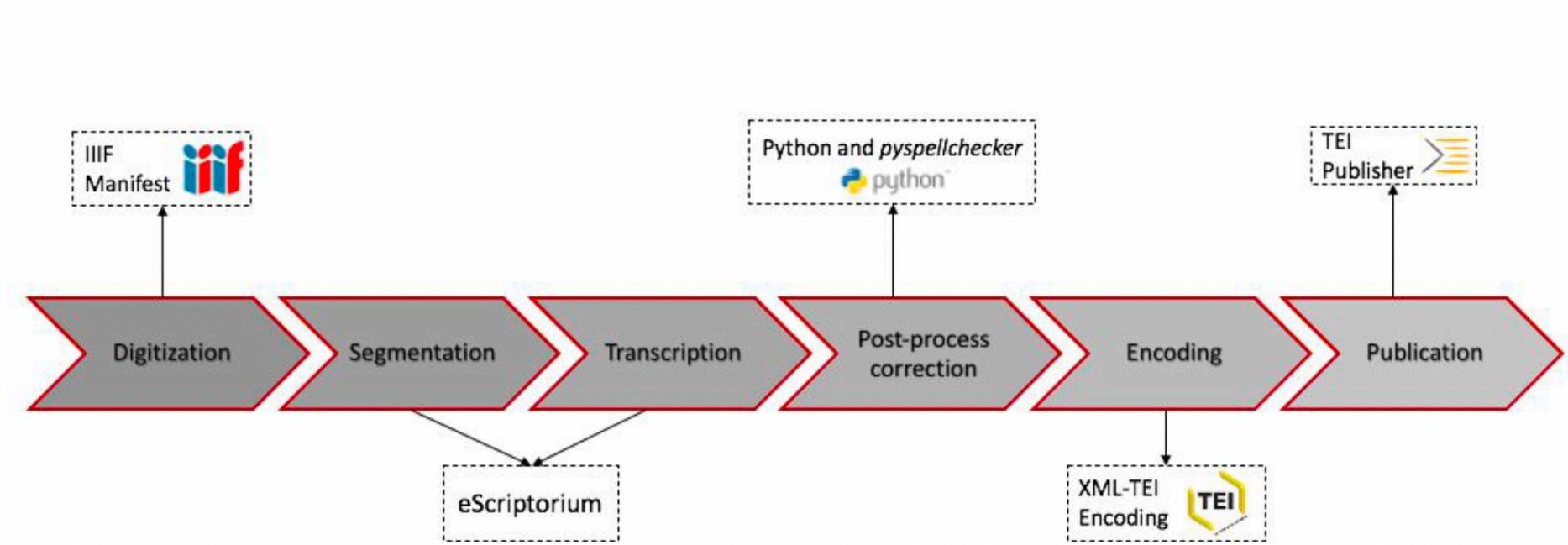


Figure: Exemple de chaîne éditoriale, Alix Chagué and Floriane Chiffoleau. "An accessible and transparent pipeline for publishing historical egodocuments". In: *WPIP21 - What's Past is Prologue: The NewsEye International Conference*. Virtual, Austria, Mar. 2021. URL: https://hal.science/hal-03173038 (visited on 12/20/2023)

Étapes détaillées

- Acquisition automatique du corpus
- Automatisation de la structuration du texte à partir de la segmentation
- Enrichissement du texte
  - Annotations linguistiques
  - Annotation des entités nommées
- Alignement des différentes versions d'un texte
- Publication du corpus en ligne

Table of Contents

Automatisation des chaînes éditoriales : Projet Gallicorpora

- Né d'une collaboration entre INRIA, l'École nationale des chartes et l'Université de Genève
- Mené entre 2021 et 2022
- Financé par le DataLab de la BnF
- **Objectifs**
  - Valoriser les collections numérisées sur Gallica
  - Produire automatiquement à partir des numérisations un corpus textuel avec des œuvres écrites entre le 15^{e} et le 18^{e} siècle
  - Assurer la compatibilité des données sur un large éventail de cas de figure
  - Réussir à trouver l'équilibre entre flexibilité et structuration stricte des données

Toutes les données et les scripts écrits durant le projet sont disponibles au lien suivant:

`https://github.com/Gallicorpora`

Projet Gallicorpora

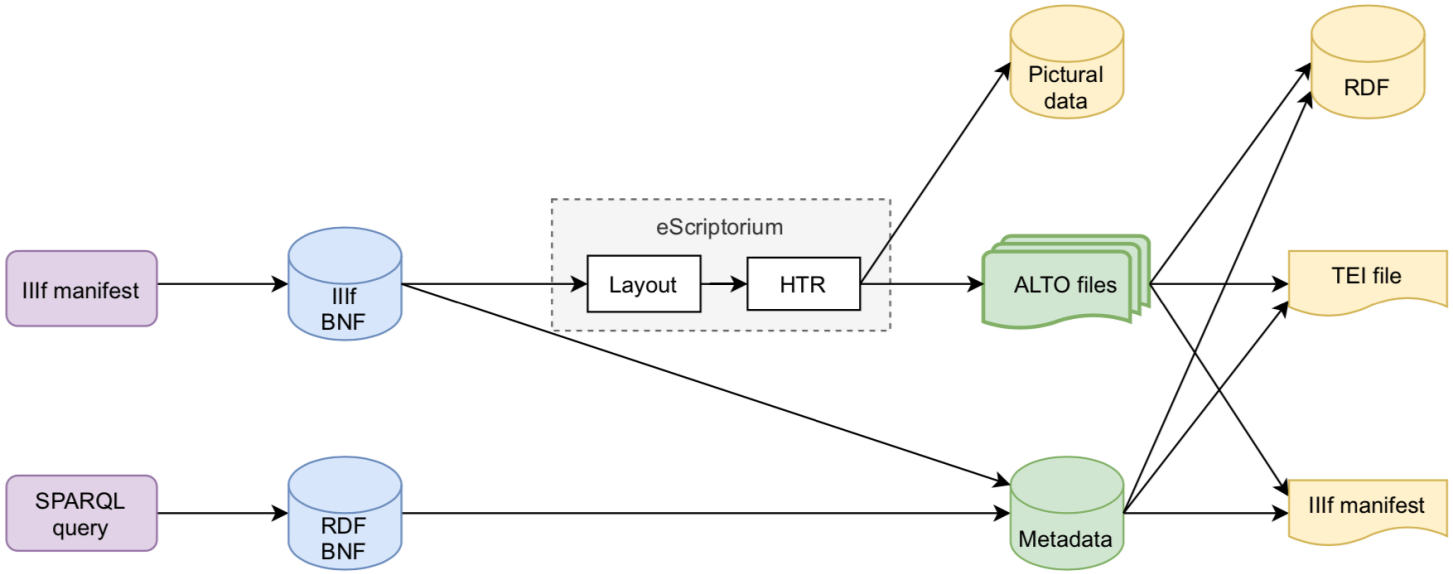
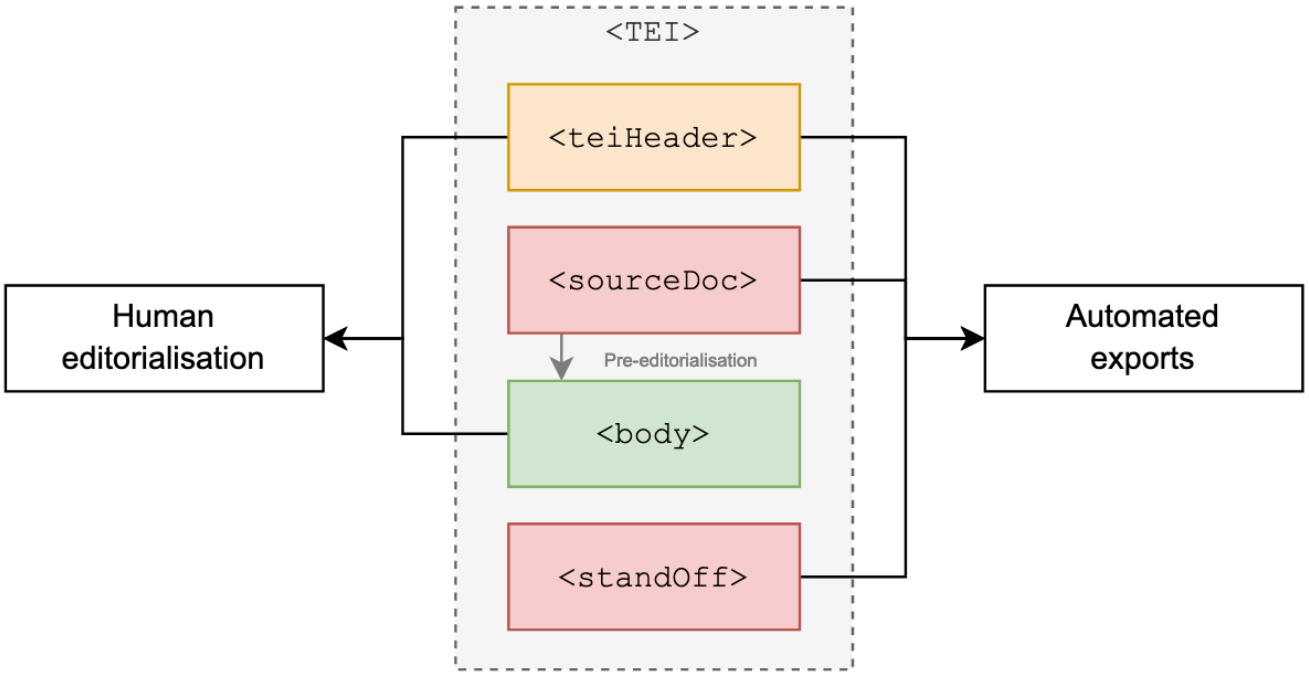


Figure: Protocole de création des données du projet Gallic(orpor)a

Modéliser en TEI les documents

- Objectif : produire un schéma générique capable de structurer une grande variété de documents
- Les fichiers TEI ont été générés par un script Pyhton écrit par K. Christensen
- Les scripts et la documentation sont disponibles au lien suivant : https://github.com/kat-kel/alto2tei



<teiHeader>

Le teiheader a été modélisé, puis rempli automatiquement à partir des métadonnées de la BnF.

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title> <!-- ... --> </title>
            <author> <!-- ... --> </author>
            <respStmt> <!-- team --> </respStmt>
        </titleStmt>
        <extent><!-- count of pages --></extent>
        <publicationStmt> <!-- project -->
        </publicationStmt>
        <sourceDesc> <!-- source--> </sourceDesc>
    </fileDesc>
    <profileDesc> <!-- language --> </profileDesc>
    <encodingDesc> <!-- models, HTR engine and
Ontology -->
```

Acquisition des métadonnées

- APIs (IIIF API and SRU API)
  - **Title** - Le Romant comique [1re partie]
  - **Responsibility** - Paul Scarron
  - **Publication** - 1655, Leide, J. Sambix (ed.)
  - **Description of the object** - français, texte imprimé
  - **Conservation** - Bibliothèque nationale de France (8-Y2-55998)
- Config file (YAML)
  - **Responsibility** : Kelly Christensen
  - **Publication** : BnF DataLab
  - **Object Description** : 20 pages

teiHeader Example

Example of teiHeader automatically generated

```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
```

From XML Alto to the <sourceDoc>

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>...
```

Mettre en place des pratiques communes pour automatiser les tâche

Le projet *SegmOnto* nous a permis de mettre en place une syntaxe simple pour décrire la mise en page des documents en nous appuyant sur la segmentation.

Zones	*Lines*
DropCapitalZone	DefaultLine
GraphicZone	DropCapitalLine
MainZone	HeadingLine
MarginTextZone	InterlinearLine
MusicZone	MusicLine
NumberingZone	
QuireMarksZone	
RunningTitleZone	
TableZone	
TitlePageZone	

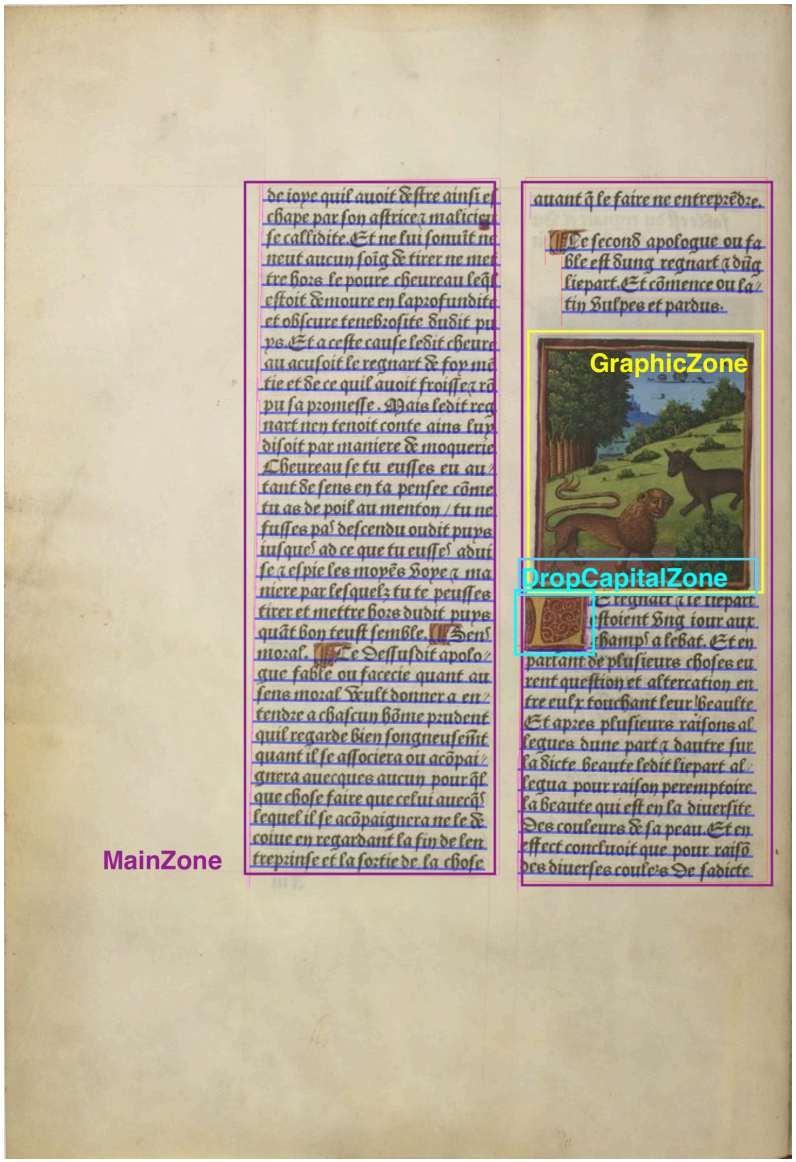


Figure: BnF, Réserve des livres rares, vélins, 611, 15e s.

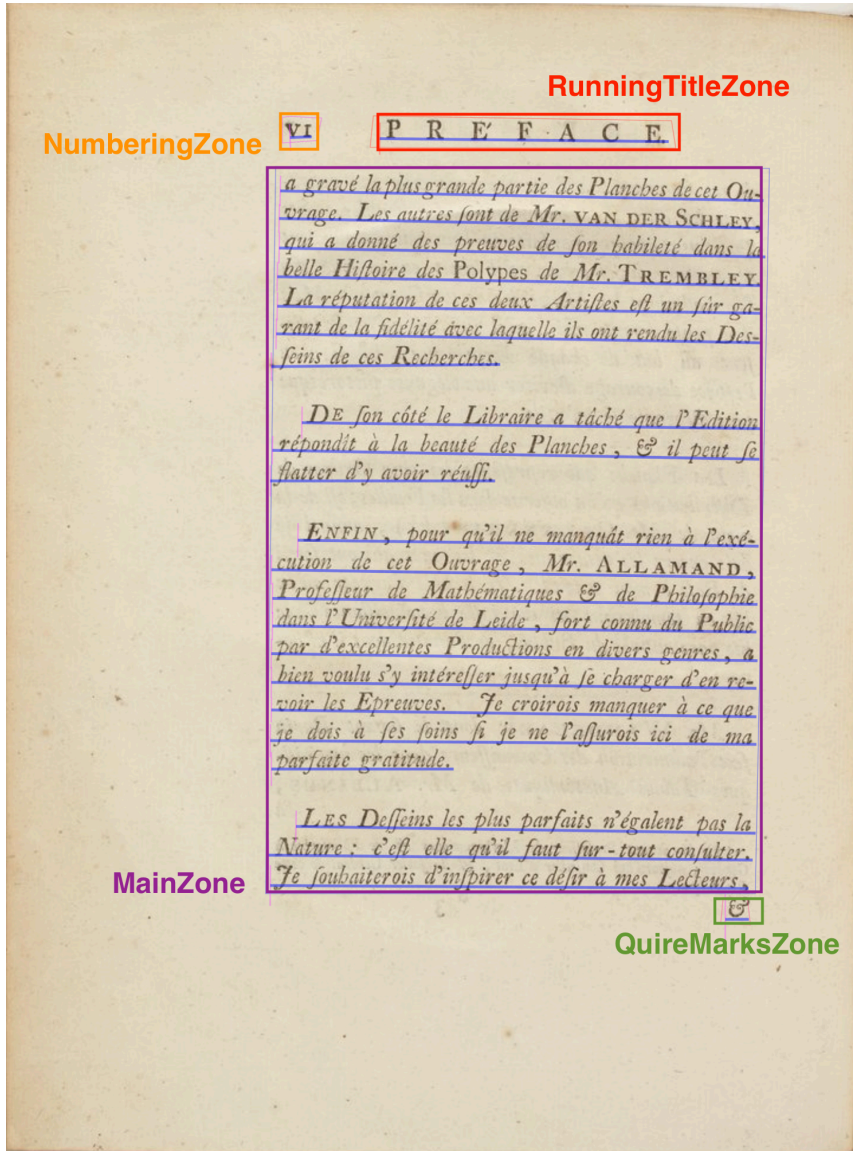


Figure: BnF, Arsenal, 4-S-1534, 18e s.

Pré-éditorialisation du texte

`<body>` et pré-éditorialisation du texte

- Créé à partir de l'analyse de la mise en page et des prédictions ATR
- Le lien entre l'image et le texte est préservé
- Livre un texte prêt à être repris, personnalisé, édité
- Limites : les données sont bruitées en raison du protocole d'acquisition automatique du texte.

Pré-éditorialisation du texte

La description normalisée de la segmentation permet une structuration de base de l'élément `<body>`

SegmOnto	TEI
NumberingZone	<fw type="NumberingZone">
QuireMarksZone	<fw type="QuireMarksZone">
RunningTitleZone	<fw type="RunningTitleZone">
MarginTextZone	<note type="MarginTextZone">
MainZone	<ab type="MainZone">
DefaultLine	<lb>
HeadingLine	<hi rend="HeadingLine">
DropCapitalLine	<hi rend="DropCapitalLine">

Correspondance entre <sourceDoc> et <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>
    ...
```

Example of body encoding

```
<pb corresp="#f23"/>
<fw corresp="#f23-eSc_textblock_a44af781-blockCount1" type="NumberingZone"><lb
    corresp="#f23-eSc_textblock_a44af781-eSc_line_543c5953-lineCount1"/>10</fw>
<ab corresp="#f23-eSc_textblock_cb6c5a03-blockCount2" type="MainZone"><hi rend="HeadingLine"
      ><lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_66797451-lineCount2"
      />BRADAMANTE,<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_637e1bb9-lineCount3"
      />TRAGECOMEDIE.<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_fb9ce0c7-lineCount4"
    />ACTE I. SCENE I.<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_d87052a6-lineCount5"
    />Charlemagne.</hi><hi rend="DropCapitalLine"><lb
      corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_e55d1a40-lineCount6"/>L</hi><lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_5598aaae-lineCount7"/>Es ſceptres des
  grands Rois vien¬<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_1e7bb869-lineCount8"
  />nent du Dieu ſupréme,<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_26556dbe-lineCount9"/>C'eſt luy qui ceint
  nos chefs d’vn<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_eda21fb8-lineCount10"
  />royal diadême,<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_d59a4e1f-lineCount11"
  />Qui nous fait quand il veut re¬<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_db058f24-lineCount12"/>gner ſur
    l’Vniuers,<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_6d2c8709-lineCount13"/>Et
  quand il veut fait cheoir no¬<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_d5bf7dce-lineCount14"/>ſtre empire à
    l'enuers.<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_a94ed2a3-lineCount15"/>Tout
  depend de ſa main, tout de ſa main procede,<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_2e0da9de-lineCount16"/>Nous n’auons rien
  de nous, c'eſt luy qui tout poſſede,<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_ea8ddb95-lineCount17"/>Monarque vniuerſel,
  &amp;amp; ſes commandemens<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_b7482d3c-lineCount18"/>Font les ſpheres
  mouuoir &amp;amp; tous les elemens.<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_587530f5-lineCount19"/>Il a mis ſur mon
  chef la Françoiſe couronne,<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_8fbc0e66-lineCount20"/>Il a fait que ma
  voix toute la terre eſtonne.<lb
```

Table of Contents

Conclusion

Dans un écosystème nativement numérique :

- L'acquisition du texte et sa structuration, ainsi que la génération des métadonnées peuvent être, pour partie, automatisées;
- Constituer des chaînes éditoriales complètes est encore relativement expérimental aujourd'hui. Il faudra à l'avenir intégrer :
  - Des annotations linguistiques
  - Des annotations sur les entités nommées
- Pour optimiser l'usage de ces corpus.