Between automatic and manual encoding

Towards a generic TEI model for historical prints and manuscripts

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

[1]École nationale des chartes — PSL [2]Inria

[3]Université de Genève

15 September 2022

The Gallic(orpor)a project

Gallic(orpor)a project

- **Who and When**
  - Born from a collaboration between the INRIA, the École nationale des chartes and Geneva University
  - Started at the end of 2021
  - Funded by the BnF dataLab
- **Objectives**
  - Valorisation of digitized collection from Gallica
  - Production of automatically generated corpus from the 15^{th} to the 18^{th} century
  - Ensure data compatibility with a wide range of use cases: between strict structured data and flexibility

Data and scripts, created during the project, are available online:

`https://github.com/Gallicorpora`

Extracting information from digital facsimiles

Designing common practices for documents annotation

Design of common practices to enable pipeline automation and data consistency

1. Harmonizing layout description
   - Using the same codicological vocabulary for description;
   - Managing the differences between different layout;
   - Results: models for page segmentation
2. Harmonizing Transcription
   - Using the same transcription rules and the same unicode characters;
   - Managing the differences between manuscript vs printed characters, gothic vs antiqua;
   - Results : two models, one for Manuscripts and incunabula : Cortado and another for incunabula and prints : **Gallicorpora+**

Models are available here :
https://github.com/Gallicorpora/Segmentation-and-HTR-Models

Designing common practices: Segmentation

The *SegmOnto* project consists of a generalist controlled vocabulary and a basic syntax to describe the layout of historical sources.

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

Example: `Zone:subZone` e.g. `MarginTextZone:footnote`

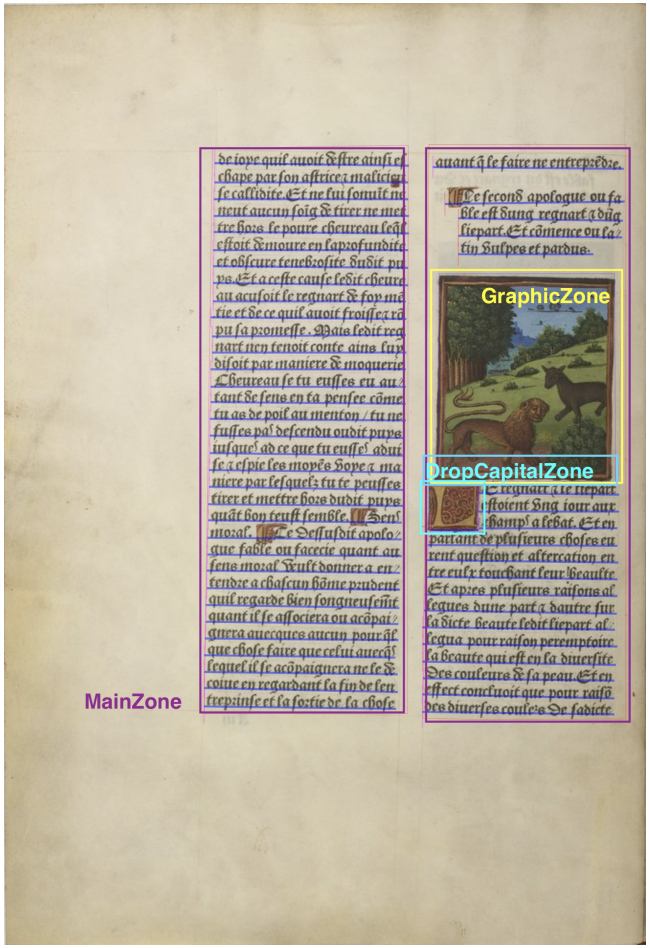


Figure: BnF, Réserve des livres rares, vélins, 611, 15e s.

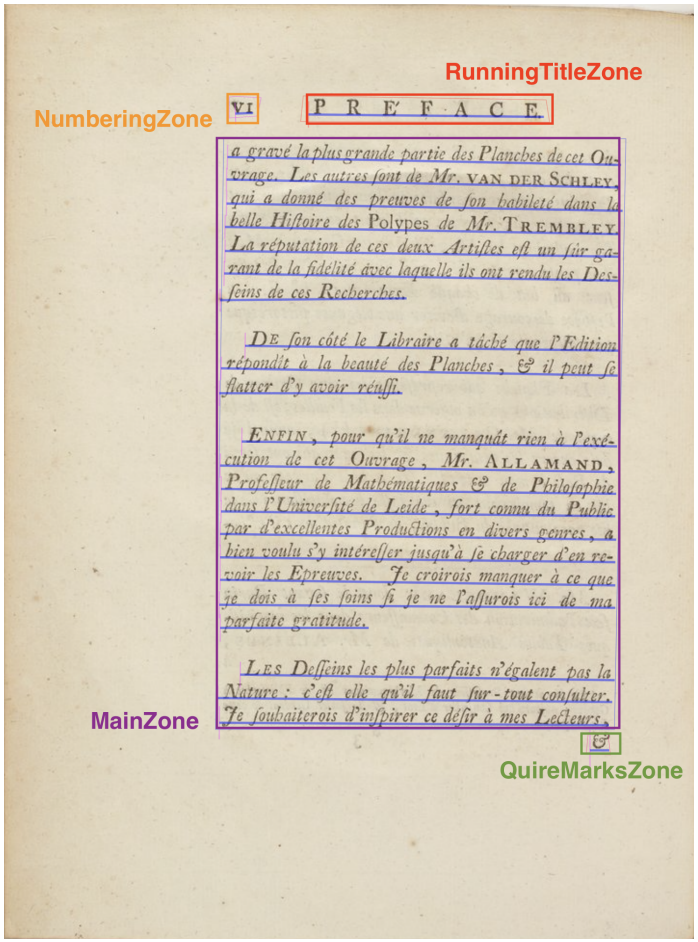


Figure: BnF, Arsenal, 4-S-1534, 18e s.

Segmentation and first failures

Training a segmentation model with *Kraken* was not successful. Is the problem due to the heterogeneity of the layouts? The granularity of the description? Or a methodological problem?

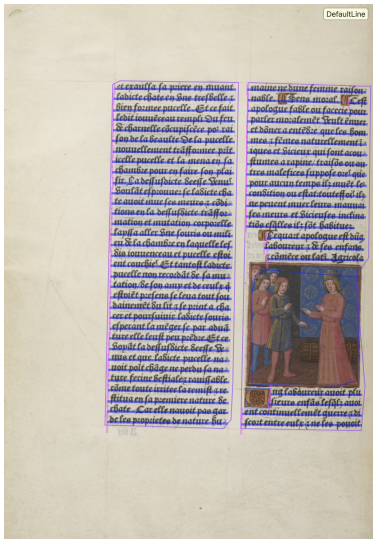


Figure: BnF, Réserve des livres rares, vélins 611, 15e s.

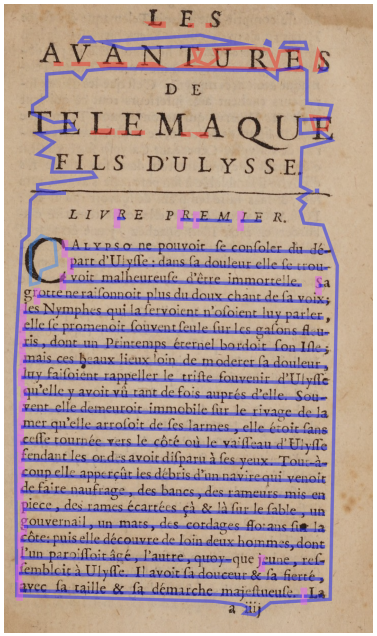


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16e s.

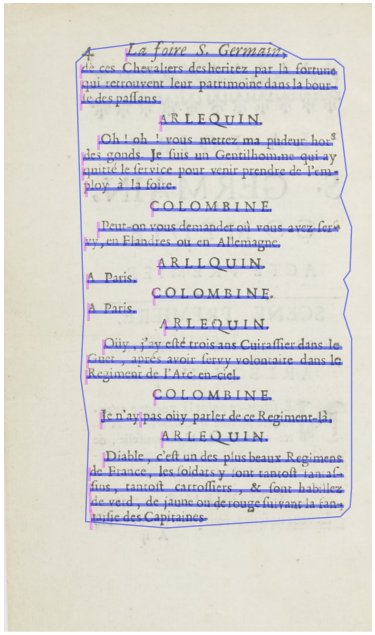


Figure: BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

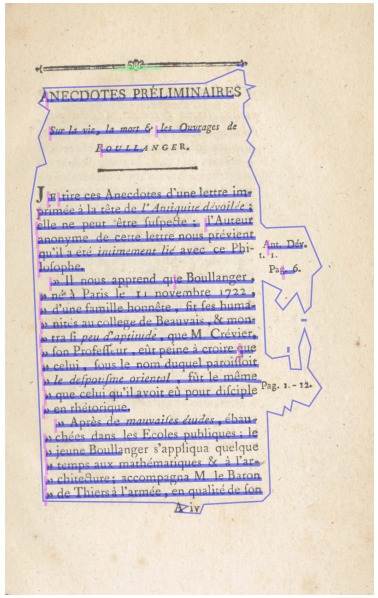


Figure: BnF, département Droit, économie, politique, 2012-39571, 18e s.

YALTAi Solution

Since segmentation is crucial to our pipeline, we needed a solution:

- A new approach : **YALTAi**, see T. Clerice, *You Actually Look Twice At it (YALTAi): Using an Object Detection Approach Instead of Region Segmentation Within the Kraken engine*
  - Different from Kraken pixel categorization
  - Using object detection, YOLOv5 solution
- New scores (average precision in percent) :

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

YALTAi Solution

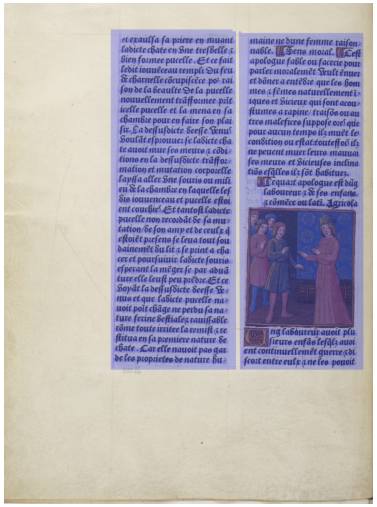


Figure: BnF, Réserve des livres rares, vélins 611, 15e s.

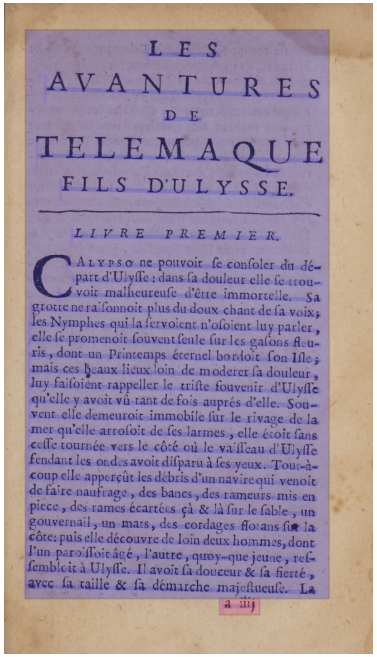


Figure: BnF, Réserve des livres rares, RES-Z-2442, 16e s.

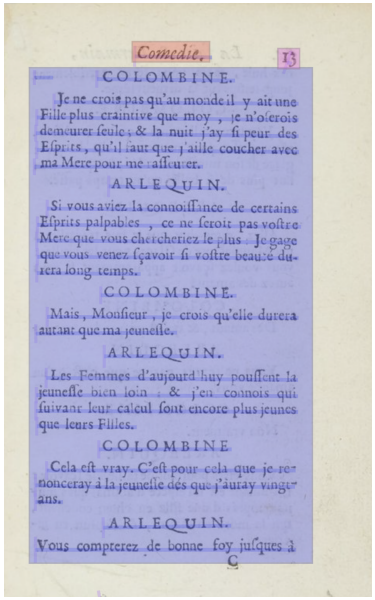


Figure: BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

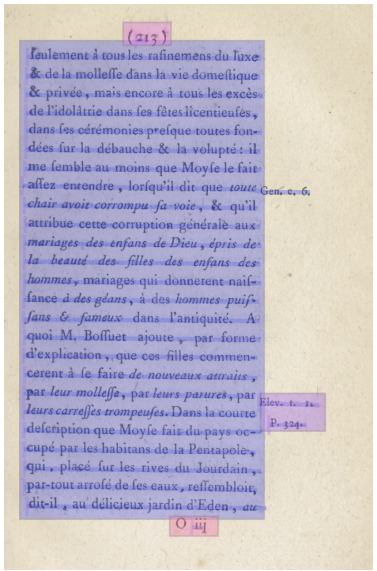
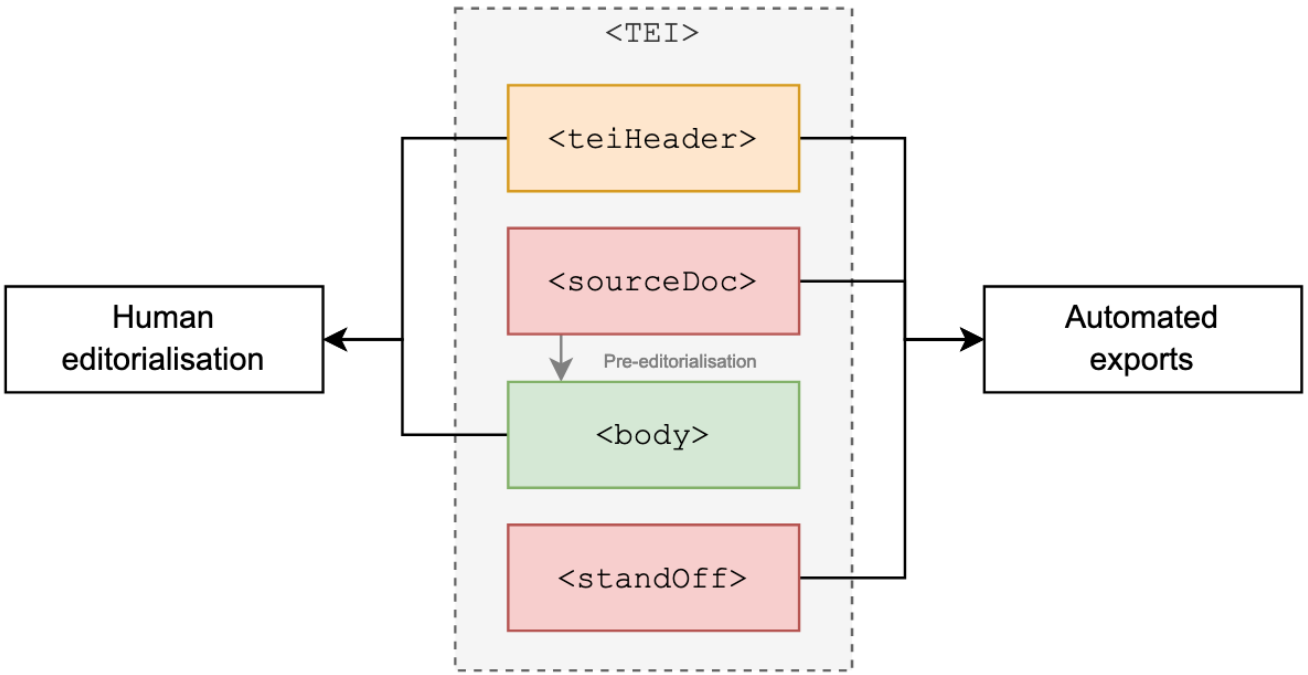


Figure: BnF, département Droit, économie, politique, 2012-39571, 18e s.

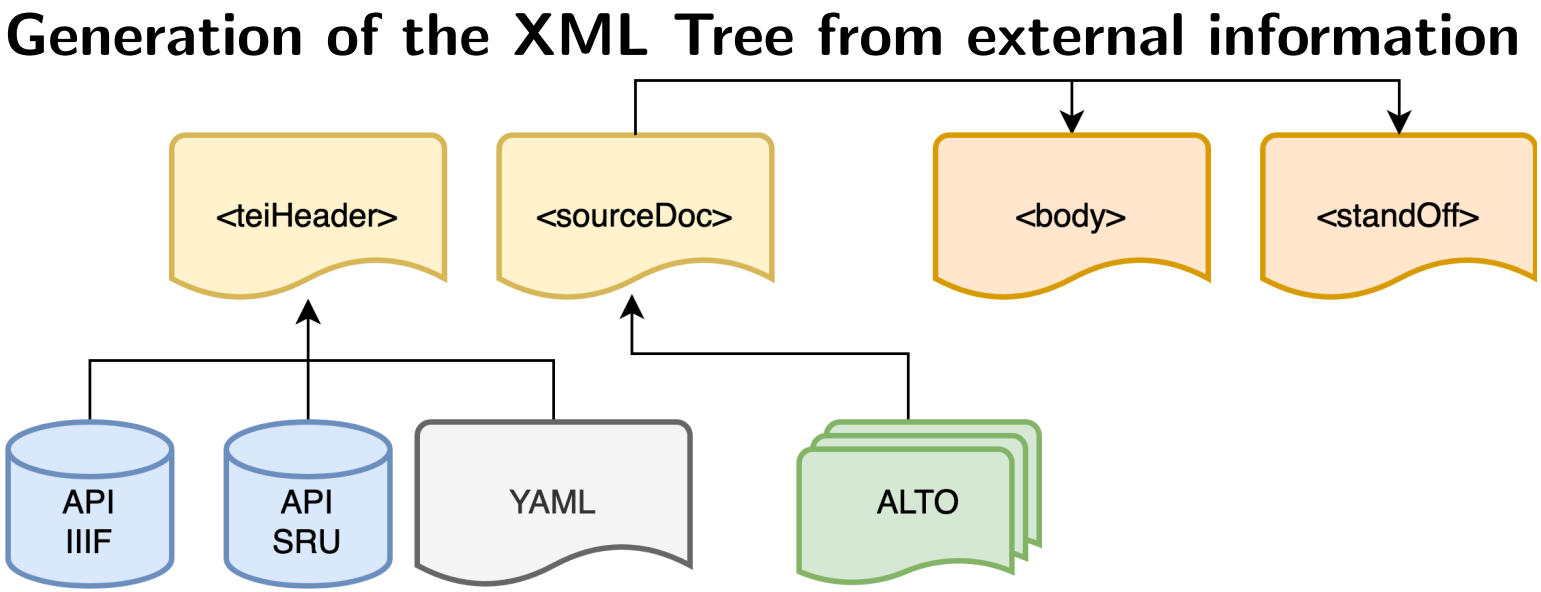
TEI Document Modelling

Main structure of the TEI model

- Aim : to produce a generic model able to receive a huge variety of documents
- TEI file generated by a python script written by K. Christensen
- Script and documentation available here : https://github.com/kat-kel/alto2tei



Automatic generation of the XML Tree



<teiHeader>

<teiHeader>

```
<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 -->
    </encodingDesc>
</teiHeader>
```

Metadata aquisition

- 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="0000000027236874X"/>
          </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
```

<sourceDoc>

From XML Alto to the <sourceDoc>

ALTO → <sourceDoc>

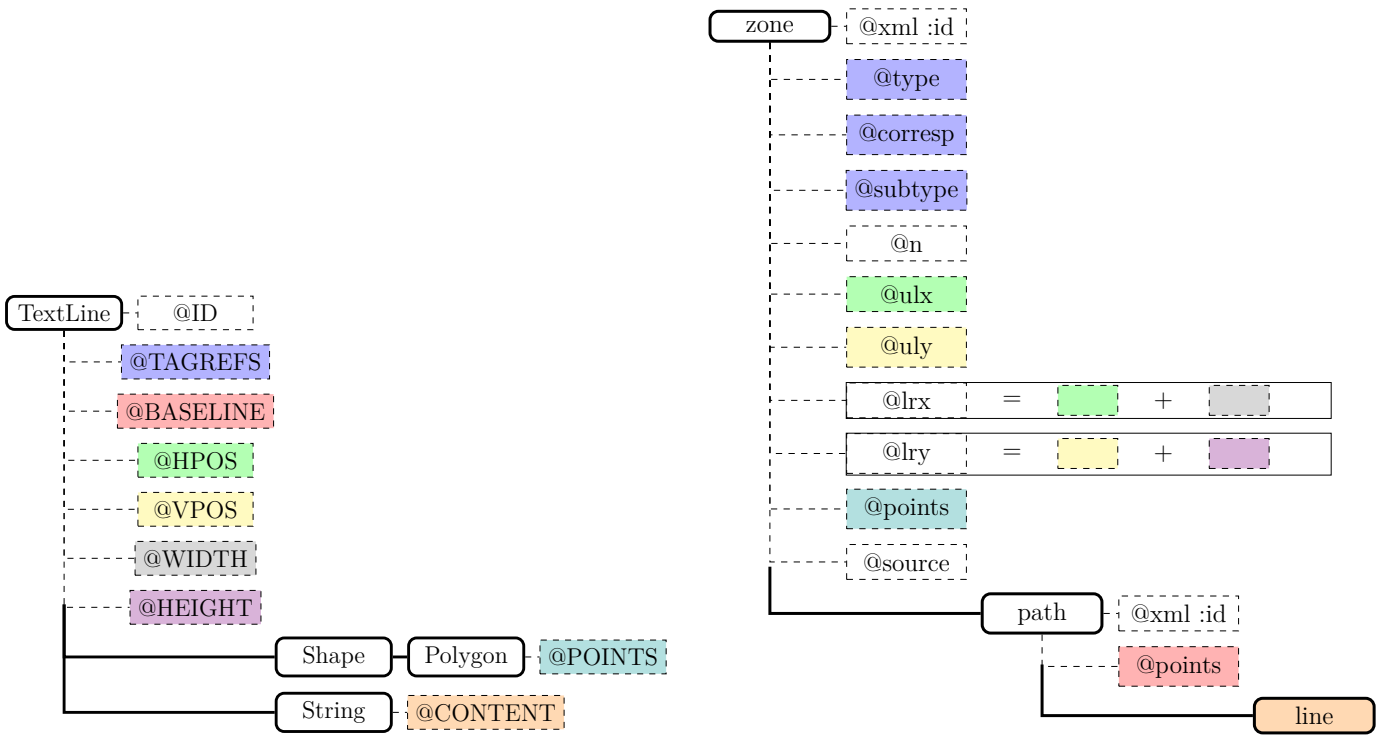
ALTO

```
<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> ...
```

From ALTO <TextLine> to TEI <zone>



<sourceDoc>

```
<sourceDoc>
  <surface><!--Page-->
    <zone type="SegmOntoZone"><!--Text Block-->
      <zone type="SegmOntoLine"><!-- TextLine
        (ex. "Text here.") -->
        <zone type="String"><!-- Segment
          (ex. "Text") -->
          <zone type="Glyph">
            <!-- Character  (ex. "T") -->
            <c>T</c>
          </zone>
          <zone type="Space"/>
        <line>Texte here.</line>
      </zone>
    </zone>
  </surface>
</sourceDoc>
```

<body>

<body> and pre-editorialized text

- Created from <sourceDoc> content
- Uses SegmOnto and layout analysis to organise the information
- The link between the image and the text is preserved
- Gives a first draft of an edition
- Made to be customized according to the problematic of each project
- Limits : As the text is produced automatically, there are always errors.

From XML ALTO to the pre-editorialisation of the <body>

Segmentation standardized description allows a basic structuring of the <body> element.

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">

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>
    ...
```

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
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_88955c54-lineCount21"/>Et que l'Aigle
  Romain perche en mes eſtendars,<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_dd0cbcd2-lineCount22"/>Guide des eſcadrons
  de mes vaillans ſoudars.<lb
    corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_9e02c03d-lineCount23"/>L'Itale m'obeit, la
  ſuperbe Alemagne,<lb corresp="#f23-eSc_textblock_cb6c5a03-eSc_line_242d0072-lineCount24"
  />Et les Rois reculez de l'ondeuſe Bretagne.</ab>
```

Conclusion

Conclusion

- The experiment is conclusive: we can automatically produce TEI XML files from the automatic acquisition of text and metadata.
- Further developments
  - To add linguistic information in our pipeline and optimise its usefulness by linking the body and the `<standOff>` element.
  - To provide a simple interface for viewing and annotating through teiPublisher