SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles

Simon Gabay, Ariane Pinche, Kelly Christensen, Jean-Baptiste Camps

Journal of Data Mining and Digital Humanities

SegmOnto - A Controlled Vocabulary to Describe and Process Digital Facsimiles

Simon Gabay[1], **Ariane Pinche**[2], **Kelly Christensen**[3], **Jean-Baptiste Camps**[4]

[1]Université de Genève
[2]CNRS CIHAM UMR 5648
[3]Sciences Po Paris
[4]Ecole nationale des chartes | PSL

Corresponding author: Simon Gabay , simon.gabay@unige.ch

Abstract

Our initiative aims at designing a controlled vocabulary for the description of the layout of textual sources: *SegmOnto*. Following a codicological approach rather than a semantic one, it is designed as a generic typology, coping with a maximal number of cases rather than answering specific needs. The harmonisation of the layout description has a double objective: on the one hand it facilitates the mutualisation of annotated data and therefore the training of better models for image segmentation (a crucial preliminary step for text recognition), on the other hand it allows the development of a shared post-processing workflow and pipeline for the transformation of ALTO or PAGE files into DH standard formats, which preserves as much as possible the link between the extracted information and the digital facsimile.

I INTRODUCTION

Extracting text from a digital facsimile with Optical Character or Handwritten Text Recognition (OCR/HTR) requires not one, but (at least) two tasks: first we recognise the layout, and then the text it contains. Because a textual source is usually a composition of multiple elements, containing different kinds of data (engravings, library stamps, headings, verses... cf. fig. 1), acquiring only the raw text is not satisfactory. The forefront of research today is to produce semi-structured data as automatically as possible, rather than unstructured data. If possible, all of a document's inscribed components as well as each one's semantic status need to be captured and evaluated during the extraction process. This evaluation serves to identify textual and visual data, to distinguish the text from the paratext, and to reorder the lines of multi-columns documents (cf. fig. 2).


L'HISTOIRE
DE MAVGIS
D'AYGREMONT,
ET DE
VIVIAN SON FRERE,
EN LAQVELLE EST CONTENV COMME ledit Maugis à l'ayde d'Oriande la Fée s'amye, alla en l'Isle de Boucaut, ou il s'habilla en Diable, & comme il enchanta le Diable Roüard, & occit le Serpent qui gardoit la roche, par laquelle chose il conquist le bon cheual Bayard, & aussi conquesta le Geant Sorgalant.
A TROYES,
Chez NICOLAS OVDOT, demeurant en la rue Nostre Dame, au Chappon d'Or couronné. 1608.
Ne extra hanc Bibliothecam efferatur. Ex obedientia.

Figure 1: Title page with manuscript annotation, printer's mark, bibliographic information, library stamp...

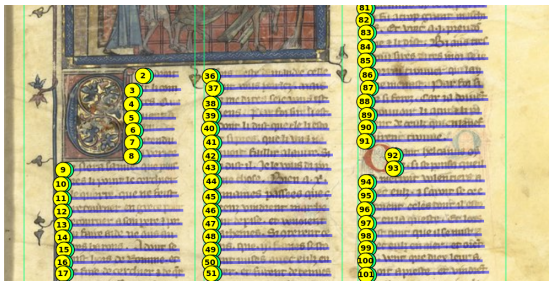


Figure 2: Ms. BnF, fr. 25550, 14th c. Lines are reordered (yellow circle) according to the layout in three columns.

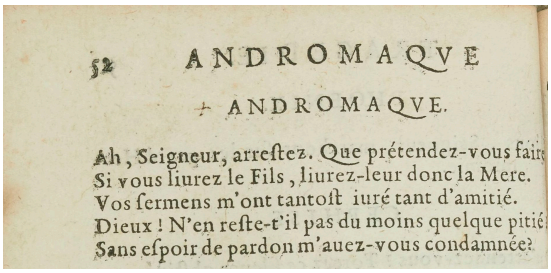


Figure 3: Racine, *Andromaque*, 1668. The first *ANDROMAQUE* is the running title, while the second is part of the play.

Semi-structured data is particularly useful for text-mining and publication purposes because it avoids introducing noise, which is detrimental to philological work on a document [Clérice, 2021]. In order to study the title character's presence in the play (such as Douguet [2015]), editorial artefacts that are not part of the text but rather of the paratext should be differenciated from the body of the text. For instance, if we look at the 1668 edition of *Andromaque* by Racine, because "Andromaque" is both the name of the main character and the title, many occurrences of this name come from the running title (cf. fig. 3), which produce noise and distort the results of the analysis by artificially over-representing the presence of Andromaque in the play, at least once every second page. Beside this computational problem, from an ecdotic point of view, a good edition would not need to display running titles, page numbers, which are sometimes wrong in the original, or present differently the name of the speakers (centered) than the verses (left-aligned). Similar conclusions can be drawn from documents with highly complex layouts, which are extremely frequent in the Middle Ages. Medieval manuscripts can include musical notations (cf. fig. 4) as well as commentaries and glosses (cf. fig. 5) that we need to differentiate from the main textual content.

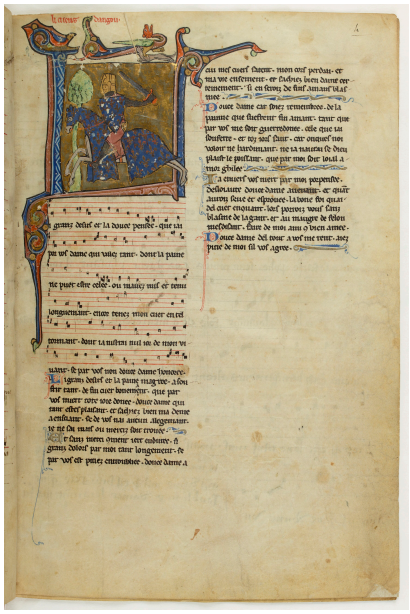


Figure 4: *Chansonnier du Roi*, Ms. BnF, fr. 844, fol. 4r

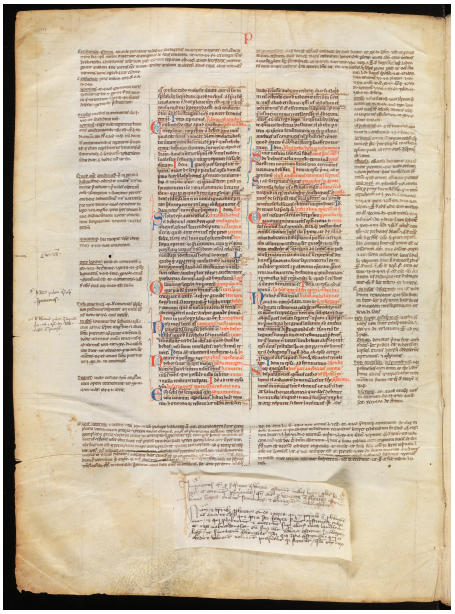


Figure 5: *Decretum Gratiani*, Sion, Archives du Chapitre, Ms. 89, fol. 3v

New methods are now available to classify different regions and zones, as well as different types of text lines on a page. First rule-based [Reul et al., 2017] and now neural-based [Kiessling, 2020,

Clérice, 2023], the efficiency of such algorithms has been increasing dramatically. Nevertheless, their application to HTR projects remains limited without a clear taxonomy for document description, particularly because every digitisation project can decide to create its own. Thus, the standardisation of practices has many interests. The following two are of special importance:

- Upstream: research teams need to share annotated documents to improve the results of HTR by increasing the amount of training data;
- Downstream: Research teams also need to share post-processing means for corpus exploration and automated document production/transformation (TEI, RDF, IIIF. . .).

Several vocabularies already exist, such as the *Vocabulaire codicologique* by Denis Muzerelle, now available online with other resources on the *Codicologia* site [Muzerelle, 1985]. Richly illustrated and offering translations of technical terms in several languages, including in a SKOS version [Geoffroy et al., 2021], the *Vocabulaire codicologique* is one of the main resources available. With more than 1500 different categories, this work is however not suitable for computational analysis, which notably requires a smaller number of classes to be efficient. On the other end of the spectrum, we find the *framework* of PageXML [Pletschacher and Antonacopoulos, 2010] which proposes an extremely simple typology of about ten zones that is simple to implement in a classification algorithm but does not meet the needs of philologists.[1] The *SegmOnto* controlled vocabulary tries to balance the pros and cons of these two solutions: a simple taxonomy which is easy to implement computationally, but adapted to philological needs.

II THE *SEGMONTO* VOCABULARY

The *SegmOnto* controlled vocabulary is based on the assumption that most textual sources can be described in the same way – whether they are (historical) prints or manuscripts (cf. fig. 6) – if we use a codicological perspective focused on material aspects. That being said, many challenges arise, especially regarding how to define a desired level of granularity. On the one hand, too many categories would unnecessarily complicate the classification task and therefore deteriorate its efficiency. On the other hand, having too few categories does not offer the proper level of description. For in-

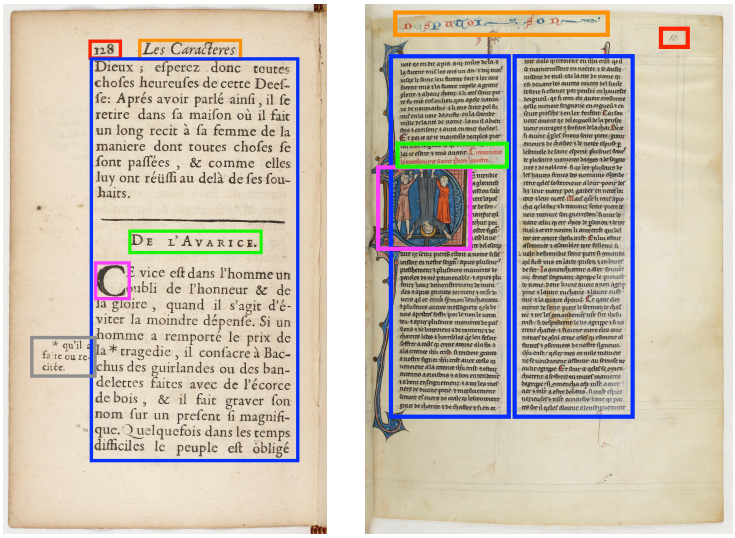


Figure 6: Similarities between the layout of a modern print (left) and a medieval manuscript (right). The running title is in orange, the pagination/foliation in red, headings/rubrics in green, drop capitals in pink, main textual zones in blue and marginal notes in grey.

[1]These are TextRegion, ImageRegion, GraphicRegion, ChartRegion, LineDrawingRegion, SeparatorRegion, TableRegion , MathsRegion, ChemRegion, MusicRegion, AdvertRegion, NoiseRegion, UnknownRegion.

stance, does it really matter if
we distinguish between a title in a modern print and a rubric in a manuscript (cf. green zones in fig. 6), both of which function as a section heading? Should we differentiate a headpiece (cf. fig. 7) from a tailpiece (cf. fig. 9)? Or group them as ornamentation that does not bear any semantic connection with the text, as opposed to illustrations such as engravings (cf. fig. 8)? Or should we group all three as decorations, which are different from drop capitals (cf. fig. 10), because the latter bear text?

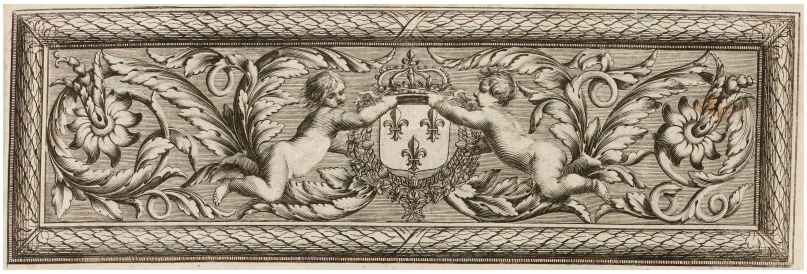


Figure 7: Headpiece

Figure 8: Engraving

Figure 9: Tailpiece

Figure 10: Drop capital

Obviously, the answer to these questions depends on each philological project, but the existence of singular needs does not inhibit the creation of common guidelines, especially if they are conceived with enough flexibility to accommodate as many situations as possible. In the broad continuum of possibilities, we have therefore tried to design a generic rather than specific controlled vocabulary, which focuses more on the material aspects (position, shape, color. . .) than on the precise semantic content of regions and text lines. Because it is impossible to encompass all the cases found in all written historical sources, the *SegmOnto* guidelines have been primarily designed for western medieval and modern documents. However, several mechanisms have however been designed to cope with problematic cases and potentially go beyond this original objective.

2.1 The *SegmOnto* syntax

The first mechanism to introduce flexibility in the annotation is the existence of a simple three-tier syntax, made of mandatory and optional parts that can be combined.

type(:subtype)?(#\d)?

The three parts are the following:

1. Types are mandatory text strings with only controlled values – the *SegmOnto* controlled vocabulary;
2. Subtypes are optional text strings, with only a suggested open list of possible values;

3. Numbers are optional integers.

Only types are defined precisely (cf. *infra*), and subtypes offer the opportunity to specify the value of the type. To annotate a GraphicZone, which is a type of region designed to annotate all the illustrations or ornamentation found in a digital facsimile, it is possible to use the following subtypes:

GraphicZone:headpiece
GraphicZone:tailpiece

Numbers do not specify the value of the type, but rather the succession (#1, #2, #3, #4) or alternation (#1, #2, #1, #2) of zones. Indeed, it is common that a manuscript or a print is organised with different columns, going from left to right in western documents (cf. fig. 2). We can document this organisation of the page in the following way:

MainZone#1
MainZone#2

It is possible to combine the label's three components in order to annotate different footnotes on a page with:

MarginTextZone:footnote#1
MarginTextZone:footnote#2

It is important to state that, at least for now, if MarginTextZone, MarginTextZone:footnote MarginTextZone:footnote#1, are closely related for a philologist, they are different for most current layout recognition systems, which would not recognise the last two categories as sub-versions of the first one. The more the annotation is precise, the more complexity it adds to the description, and therefore the more training data is required to reach a decent accuracy for the classifier.

2.2 The *SegmOnto* types

Types are divided into two different categories because a facsimile can be divided, as mentioned before, into regions (running title, page number, marginal notes. . .) and lines (heading, normal line, interlinear addition or correction. . .).

2.2.1 The SegmOnto regions

Fifteen different regions have been identified:

- MainZone is the main area containing the text, excluding any paratext, and it is either a single block or multiple columns. Thus, the MainZone is characterised by the importance of the text it bears, yet it does not exclude the presence of non-textual information (music notation, illumination. . .). When a page is divided into columns or blocks, each one is a different zone. Possible values for a subtype are:
  - MainZone:column
  - MainZone:block
- GraphicZone is a zone containing any type of graphic element, from purely ornamental information to information con-substantial to the text (*e.g.*, full-page paintings, line-fillers, marginal drawings, figures, etc.). Drop capitals are excluded from this category. Captions,

if there are any, is part of this zone, and its text line is labelled HeadingLine. If an image contains text, it is possible to label the lines as DefaultLine. Possible values for subtypes are:

- GraphicZone:illustration, cf. fig. 8.
- GraphicZone:ornamentation, cf. fig. 11, 12, fig. 7, fig. 9.
- GraphicZone:figure, cf. fig. 11 et 12.

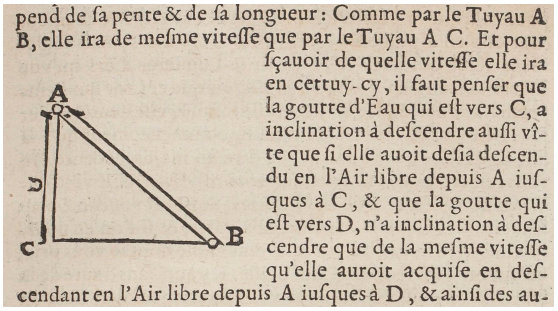


Figure 11: GraphicZone:figure

Figure 12: GraphicZone:figure

- DropCapitalZone contains any type of initial letter that occupies a space corresponding to several lines of the main text or that bears significant ornamentation. A drop capital can be a historiated, ornamented, flourished or painted initial. This zone does not encompass the whole text line to which it is attached. The letter's or letters' baseline is labelled DropCapitalLine, rather than DefaultLine, for technical reasons. Possible values for subtypes are:
  - DropCapitalZone:historiated
  - DropCapitalZone:flourished
  - DropCapitalZone:ornated
  - DropCapitalZone:voided

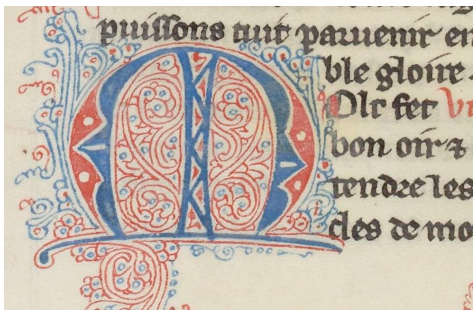


Figure 13: DropCapitalZone:flourished

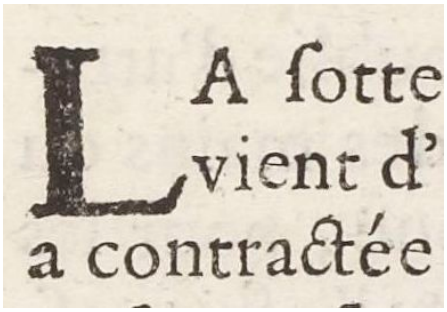


Figure 14: DropCapitalZone

- NumberingZone is a zone containing the page, the folio, or the document number, with no regard for the mark's origin (scribe, curator, etc). The zone usually is at the top of the page. Letters/numbers denoting a folio's order within a quire are annotated differently with a special zone: QuireMarksZone. Possible values for the subtypes of the NumberingZone are:
  - NumberingZone:page
  - NumberingZone:folio
  - NumberingZone:item

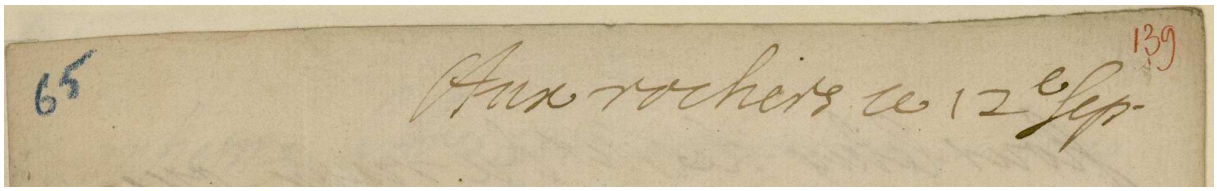


Figure 15: Double numbering: item (left) and folio (right).

- MarginTextZone characterises any text zone contained in the margins no matter its position on the page (upper, lower, inner or outer), including the space between two columns. We do not differentiate the zone's particular semantic status (gloss, addition, correction, intertextual or bibliographic reference, . . .). Possible values for subtypes are:
  - MarginTextZone:note
  - MarginTextZone:commentary
  - MarginTextZone:correction
  - MarginTextZone:variants
  - MarginTextZone:criticalApparatus

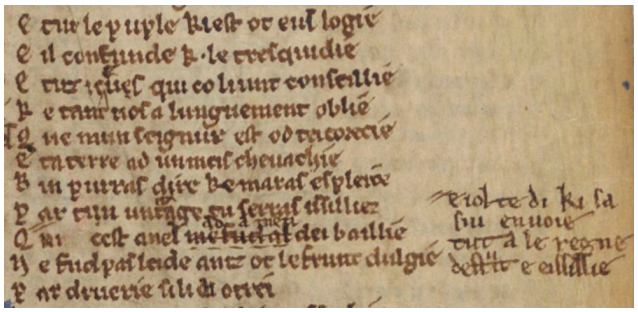


Figure 16: MainZone:addition

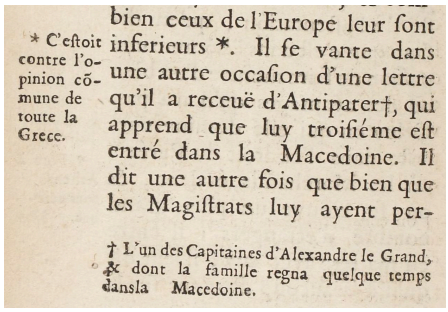


Figure 17: MainZone:commentary

- QuireMarksZone is a zone containing a quire signature (*e.g.*, a ii), catchword, or any kind of element relative to the material organisation of the source, with the exclusion of page, folio, or item numbers. The zone usually is at the bottom of the page. Possible values for subtypes are:
  - QuireMarksZone:signature
  - QuireMarksZone:catchword

Figure 18: QuireMarksZone:catchword

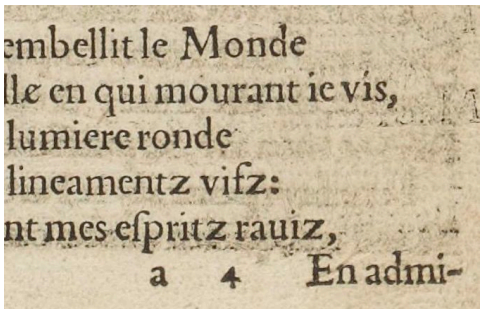


Figure 19: Signature and catchword.

- RunningTitleZone is a zone containing a running title, traditionally at the top of the page or of the double page. It can be the title (or the abbreviated title) of a document or of the current section.

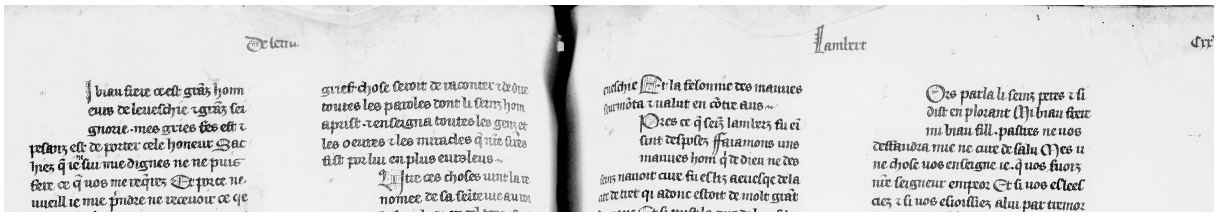


Figure 20: RunningTitle

- StampZone is a zone containing a stamp, be it a library stamp or a mark from a postal service. Possible values for subtypes are:
  - StampZone:post
  - StampZone:library

Figure 21: StampZone:post

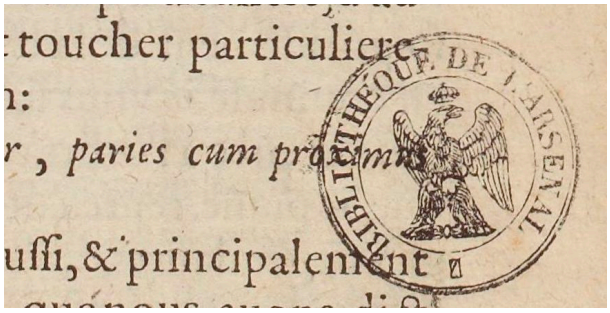


Figure 22: StampZone:library

- DamageZone characterises any area containing damage to the source, such as holes in the material (parchment, paper), blots, etc. Possible values for subtypes are:
  - DamageZone:corrosion (*corrosion*)
  - DamageZone:hole
  - DamageZone:mold
  - DamageZone:soaked
  - DamageZone:stained

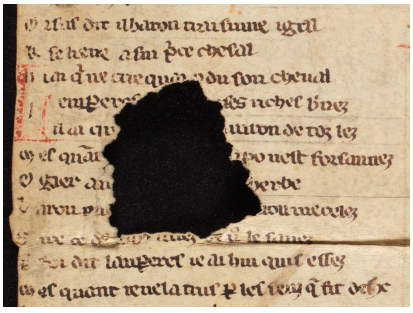


Figure 23: DamageZone:hole

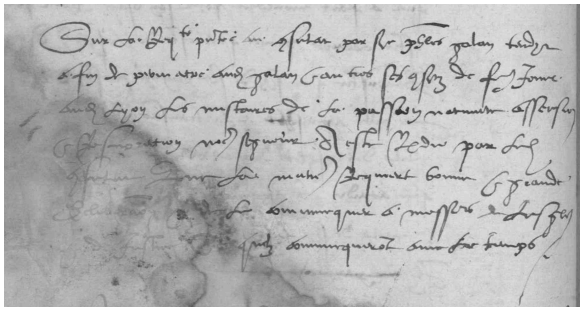


Figure 24: DamageZone:soaked

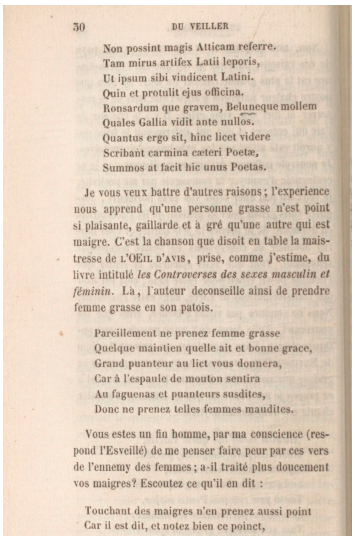
- SealZone is a zone containing a seal.
- DigitizationArtefactZone contains any type of item external to the document itself present on the image because of the digitisation process. It can be a ruler to measure the document or a color target to calibrate colours for the camera. Possible values for subtypes are:
  - DigitizationArtefactZone:ruler

– DigitizationArtefactZone:colorTarget

Figure 25: DigitizationArtefactZone

Figure 26: DigitizationArtefactZone:ruler

- CustomZone characterises any kind of zone not fitting in any of the other categories of the *SegmOnto* vocabulary, according to any convenient typology the user chooses. It can be used to identify poems, verses in a prose text, paragraphs, quotations, catalogue or dictionary entries, etc. Using subtypes is particularly recommended for this zone. Possible values for subtypes are:
  - CustomZone:verse
  - CustomZone:quotation
  - CustomZone:poem
  - CustomZone:entry
  - CustomZone:paragraph



30 DU VEILLER

Non possint magis Atticam referre,
Tam mirus artifex Latii leporis,
Ut ipsum sibi vindicent Latini.
Quin et protulit ejus officina.
Ronsardum que gravem, Beluneque mollem
Quales Gallia vidit ante nullos.
Quantus ergo sit, hinc licet videre
Scribant carmina cæteri Poetæ,
Summos at facit hic unus Poetas.

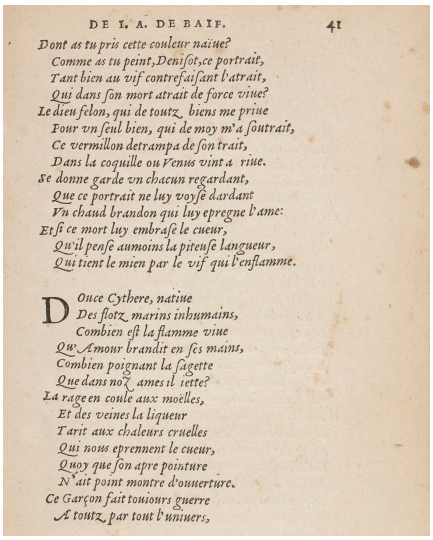
Je vous veux battre d'autres raisons ; l'experience nous apprend qu'une personne grasse n'est point si plaisante, gaillarde et à gré qu'une autre qui est maigre. C'est la chanson que disoit en table la maistresse de L'OEIL D'AVIS, prise, comme j'estime, du livre intitulé *les Controverses des sexes masculin et féminin*. Là, l'auteur deconseille ainsi de prendre femme grasse en son patois.

Pareillement ne prenez femme grasse
Quelque maintien quelle ait et bonne grace,
Grand puanteur au lict vous donnera,
Car à l'espaule de mouton sentira
Au faguenas et puanteurs susdites,
Donc ne prenez telles femmes maudites.

Vous estes un fin homme, par ma conscience (respond l'Esveillé) de me penser faire peur par ces vers de l'ennemy des femmes ; a-il traité plus doucement vos maigres? Escoutez ce qu'il en dit :

Touchant des maigres n'en prenez aussi point
Car il est dit, et notez bien ce poinct,

Figure 27: Sections in prose and in verse



DE I. A. DE BAIF. 41

Dont as tu pris cette couleur naiue?
Comme as tu peint, Deniſot, ce portrait,
Tant bien au vif contrefaiſant l'atrait,
Qui dans ſon mort atrait de force viue?
Le dieu felon, qui de toutz biens me priue
Four vn ſeul bien, qui de moy m'a ſoutrait,
Ce vermillon detrampa de ſon trait,
Dans la coquille ou Venus vint a riue.
Se donne garde vn chacun regardant,
Que ce portrait ne luy voyſe dardant
Vn chaud brandon qui luy epregne l'ame:
Et ſi ce mort luy embraſe le cueur,
Qu'il penſe aumoins la piteuſe langueur,
Qui tient le mien par le vif qui l'enflamme.

Douce Cythere, natiue
Des flotz marins inhumains,
Combien eſt la flamme viue
Qu'Amour brandit en ſes mains,
Combien poignant la ſagette
Que dans noz ames il iette?
La rage en coule aux moelles,
Et des veines la liqueur
Tarit aux chaleurs cruelles
Qui nous eprennent le cueur,
Quoy que ſon apre pointure
N'ait point montre d'ouuerture.
Ce Garçon fait touiours guerre
A toutz par tout l'vniuers,

Figure 28: Several poems

- TitlePageZone characterises the entire page, rather than a section within a page, that contains for instance headings (chapter title, act or scene number, etc.). It is distinct from other pages and is traditionally the first page of a document, especially in the case of prints. It provides bibliographic or identifying information, such as the title of the work, the production date, the names of the printer(s), publisher(s) and author(s), etc.
- TableZone is a zone containing a table of any kind. The table can be clearly drawn (with rows and columns) or not. The tables of contents are in the vast majority of cases not tables.

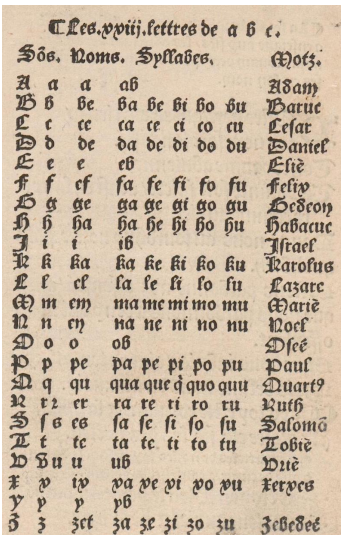


Figure 29: TableZone

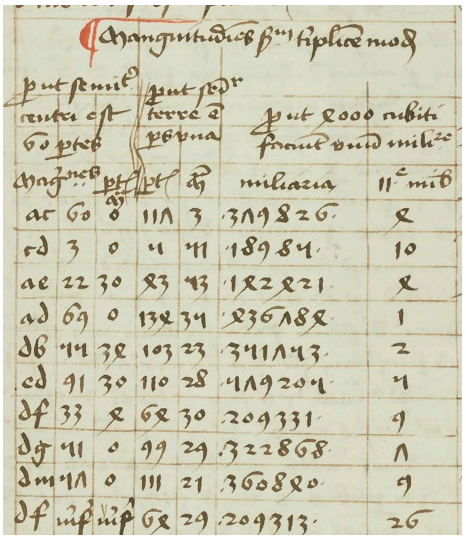


Figure 30: TableZone

- MusicZone is an area containing musical notations, such as neumes, staves, etc. It can include text. Possible values for subtypes are:
  - MusicZone:neume
  - MusicZone:stave

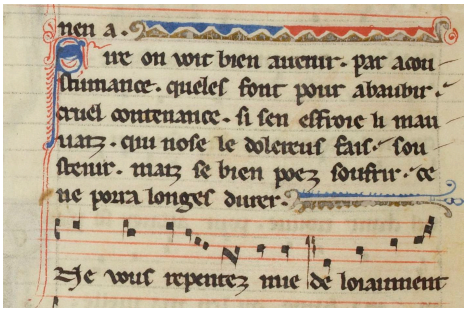


Figure 31: MusicZone:neume

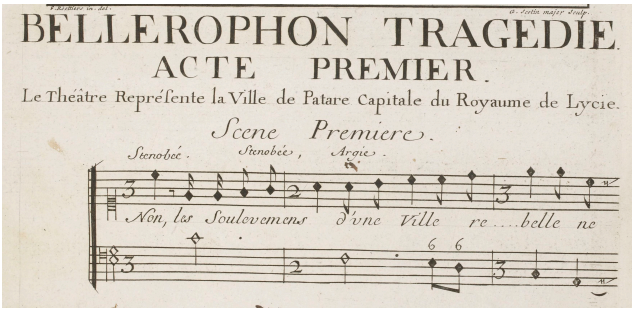


Figure 32: MusicZone:stave

2.2.2 *The SegmOnto text lines*

Six different text lines have been identified:

- DefaultLine characterises any kind of standard text line, verse or prose, regardless if it is included in the MarginZone or the MainZone, though not if it is in a DropCapitalZone. It can be used in the MusicZone for textual parts, such as the lyrics, and not the music notation;
- HeadingLine is a line containing any type of heading, which is defined as a string with a distinctive typesetting (font, size, colour, capitalisation. . .) different than that seen in the body of the text. It is not limited to the header. Rather the HeadingLine indicates the beginning of a new unit, no matter the unit's size, such as a medieval rubric, a speaker's name in a play, or the title of a poem in a collection. It is also used in the TitlePageZone. Possible values for subtypes are:
  - HeadingLine:rubric

- HeadingLine:title
- HeadingLine:incipit

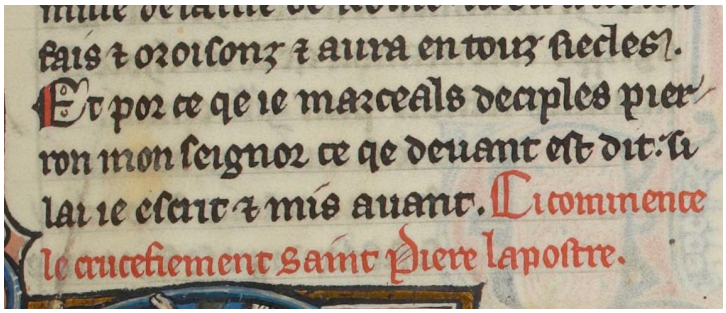
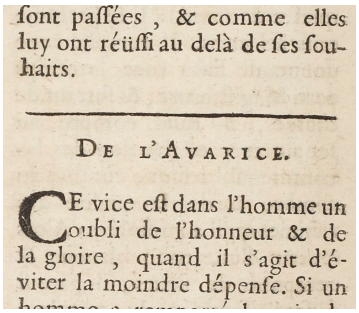


Figure 33: HeadingLine:rubric



ſont paſſées, & comme elles luy ont réüſſi au delà de ſes ſouhaits.

DE L'AVARICE.

CE vice eſt dans l'homme un oubli de l'honneur & de la gloire, quand il s'agit d'éviter la moindre dépenſe. Si un

Figure 34: HeadingLine

- CustomLine for any kind of line not fitting in any of the other categories of the *SegmOnto* vocabulary, according to any convenient typology that the user chooses. Using subtypes is particularly recommended for this zone;
- DropCapitalLine characterises a line on which rests a drop capital. It is only used in a DropCapitalZone;
- InterlinearLine is a line that is not a standard text line. Instead, it is a line that has been added between two of text lines, often in order to include a forgotten word or a gloss. Possible values for subtypes are:
  - InterlinearLine:addition
  - InterlinearLine:correction
  - InterlinearLine:gloss

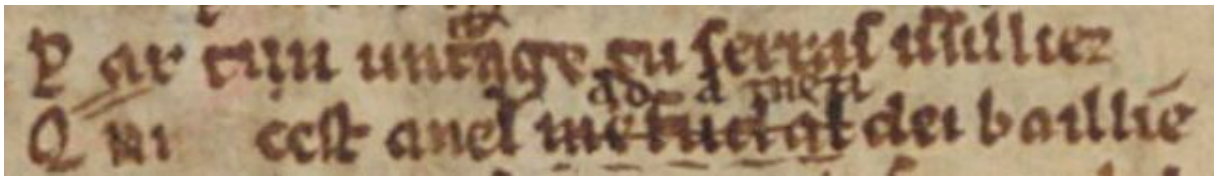


Figure 35: InterlinearLine:correction

- MusicLine characterises the central line of a musical stave.

III TOWARDS TEI

ALTO and PAGE, which are standard XML formats for OCR tasks' exports, are insufficient when it comes to text mining and digital editing (cf. ex. 1). It is therefore crucial to convert files produced by any OCR/HTR engine to a standard format such as TEI [Burnard, 2014], which offers additional elements relevant to scholars working with textual sources because of the XML elements it provides for linguistic, ecdotic and literary annotation.

```
<alto>
  <!-- Metadata -->
  <layout>
    <PrintSpace>
      <TextBlock TAGREFS="BT3246">
        <shape><Polygon POINTS="546 533 546 3049..."></shape>
        <TextLine TAGREFS="LT1125" BASELINE="573 603 2482 593...">
          <shape><Polygon POINTS="573 603 560 536..."></shape>
          <string CONTENT="predicted text of the line"/>
        </TextLine>
```

Example 1: Basic structure of an ALTO file.

Two TEI elements in particular are especially suited to accomplish two tasks that we should expect from an effective TEI publication. The first task is to store virtually all of the information that the OCR/HTR engine produced. It is undesirable to lose any information when converting data from an ALTO-XML into a TEI-XML file. The second task is to prepare the TEI publication as efficiently as possible for any future editing and annotation tasks. The <sourceDoc> accomplishes the first task because it can store all of the data procured by the HTR process and maintain a link between the predicted text and its place within the digitised image. The <body> accomplishes the second desired task because it presents the text in a structured hierarchy suitable for tools and procedures that are standard in digital publishing and literary analysis.

3.1 <sourceDoc>

Such a richly annotated and rigidly structured <sourceDoc> serves two important purposes. First, because it contains all the data that HTR models export into ALTO-XML files, the <sourceDoc> can be used to reconstruct an ALTO-XML file. Such reverse-engineered ALTO-XML files can become (re)training data. This means that researchers can take advantage of a document published according to the TEI and use its <sourceDoc> to fine tune new HTR models. Second, structured data in the <sourceDoc> can be converted into alternative formats such as json and RDF, or any additional export format. The use of a <sourceDoc> rather than a <facsimile> is recommended in order to "translate" ALTO files' topographic and textual information into a TEI schema (cf. tab. 1).

ALTO	TEI
//Page	sourceDoc/surface
//TextBlock	sourceDoc/surface/zone
//TextBlock[@TAGREFS]	sourceDoc/surface/zone[@type]
//TextBlock/Polygon[@POINTS]	sourceDoc/surface/zone/[@points]
//TextLine	sourceDoc/surface/zone/zone
//TextLine[@BASELINE]	sourceDoc/zone/zone/path[@points]
//TextLine/Shape/Polygon[@POINTS]	sourceDoc/zone/zone[@points]
//TextLine/String[text()]	sourceDoc/zone/zone/line[text()]

Table 1: Basic mapping between ALTO and TEI encodings (expressed in xPath).

3.1.1 Basic mapping

As evidenced in tab. 1, mapping data from ALTO to TEI is not straightforward and requires some basic manipulation. For instance, both an element and its attributes are not necessarily kept together in the transformation from ALTO to TEI. In example, the ALTO element <TextLine> is mapped to the TEI element <zone>. However, the attribute @BASELINE of <TextLine> is not likewise mapped to <zone>; instead, the coordinates of a baseline are mapped to the TEI element <path>, which descends from the same <zone> to which the ALTO file's <TextLine> was mapped (cf. ex. 2).

```
<TEI>
  <!-- metadata -->
  <sourceDoc>
    <surface> <!-- Page 1 -->
      <zone type="MainZone" points="546 533 546 3049...">
        <zone points="573 603 560 536...">
          <path points="573 603 2482 593..."/>
          <line>predicted text of the line</line>
        </zone>
      </zone>
    </surface>
    <surface> <!-- Page 2 -->
    </surface>
      <!-- ad lib. -->
```

Example 2: Basic structure of the sourceDoc.

Another complication is that multiple ALTO files need to become one TEI file. OCR/HTR engines produce one XML file per page of a digital facsimile, exporting into either ALTO or PAGE format. In order to not lose the relationship between these encoded pages, it is important to group all the output files' information into a single TEI document. Thus in the <sourceDoc> all the data of one ALTO-XML file is organised within one <surface> element, repeated as many times as there are ALTO-XML files.

3.1.2 Coordinates, topographic attributes and IIIF

Certain coordinates recorded in the ALTO-XML file are not always mapped directly to a similarly topographic attribute in the TEI file. For example, in the ALTO-XML file, the elements <Page>, <TextBlock>, and <TextLine> have two sets of X,Y coordinates (@HPOS/@VPOS and @WIDTH/@HEIGHT) that outline the region that the HTR model had detected . These four attributes, however, are not mapped to four similar attributes in a TEI schema. Instead, they are used to specify a region of the source image in the URI for a IIIF Image API. This URI is the value of the attribute @source for the equivalent TEI element.

The IIIF documentation provides the following syntax to create a URI:

```
{scheme}://{server}/{prefix}/{identifier}/
{region},{size},{rotation},{quality},{format}
```

Example 3: Creating a IIIF URI for a IIIF Image API

where:

- {scheme} is the transfer protocol, traditionally http or https;
- {server} is the host server, in other words the URL/web address, where the service is hosted, such as gallica.bnf.fr;
- {prefix} is the path on the host server, usually iiif for the iiif service;
- {identifier} is the name of the image file requested, traditionally an ARK;
- {region} is created with four attributes found on the <TextBlock> and <TextLine> ALTO elements (@alto:HPOS, @alto:VPOS, @alto:WIDTH, @alto:HEIGHT) separated by commas;
- {size}, {rotation}, {quality} and {format} are IIIF parameters to request the image.

In the example of a 1525 edition of the incunabulum *Menus propos Mère Sotte* by Pierre Gringore, which the BnF has digitised and given the ARK bpt6k15260973, the IIIF URI for a block of text on the tenth folio would therefore be the following:

https://gallica.bnf.fr/iiif/ark:/12148/bpt6k15260973/f10/
194,76,1368,2051/full/0/native.jpg

Example 4: IIIF URI generated from an ALTO-XML file

The IIIF URI's region parameter is provided with data from the relevant ALTO element's attributes @HPOS, @VPOS, @WIDTH, @HEIGHT. In the example above (cf. ex. 4), having first been recognised by a HTR model, that block of text would be encoded in the following ALTO format (cf. ex. 5):

```
<TextBlock  HPOS="194"
            VPOS="76"
            WIDTH="1368"
            HEIGHT="2051"
            ID="eSc_textblock_ca612caa"
            TAGREFS="BT1824">
   <Shape>
      <Polygon POINTS="209 76 194 2127 1505 2127 1562 100"/>
   </Shape>
```

Example 5: ALTO <TextBlock> of zone 1 on folio 10 of doc. bpt6k15260973

To put the IIIF URI in context, that block of text and the URI would be encoded in the following TEI format (cf. ex. 6):

```
<zone   xml:id="f10_z1"
        type="MainZone" subtype="none" n="none"
        points="209,76 194,2127 1505,2127 1562,100"
        source="https://gallica.bnf.fr/iiif/
                ark:/12148/bpt6k15260973/f10/
                194,76,1368,2051/full/0/native.jpg">
```

Example 6: TEI <zone> of zone 1 on folio 10 of doc. bpt6k15260973

In the above example of *Menus propos Mère Sotte* by Pierre Gringore, the HTR model output a document whose coordinates were encoded in pixels, which is typically the case yet nevertheless

always explicitly declared in an ALTO schema. Thus, the values of the ALTO attributes @HPOS, @VPOS, @WIDTH, @HEIGHT already and perfectly conform to the IIIF standards without any manipulation.

It is also possible to provide the IIIF URI of the entire page in a <graphic> element associated with a <surface>:

```
<surface>
   <graphic url="https://gallica.bnf.fr/iiif/ark:/12148/bpt6k15260973
                /f1/full/full/0/native.jpg"/>
   <zone   xml:id="f10_z1" ...>
```

Example 7: TEI <zone> of zone 1 on folio 10 of doc. bpt6k15260973

3.1.3 Zone and Line Types

All of an ALTO-XML file's topographic and linguistic data can be mapped to a TEI schema. Consequently, by training HTR/OCR models with a *SegmOnto* vocabulary, that codicological system will be fully available in a TEI publication, accessible to philological, linguistic, and literary analyses as well as editorial tasks. As was suggested in Ex. 6, every <zone> in a TEI publication, whether it represents a <TextBlock> or a <TextLine>, has a @type as well as a @subtype and a @n. These attributes are parsed from the decoded name with which the HTR model tagged that region of the image. In the case of Ex. 5, the model tagged the block of text as a "MainZone," to which it assigned the alphanumeric reference code "BT1824." The TEI attributes @type, @subtype and @n take parsed portions of the tag's name. In this case, there was no subtype in the tag name "MainZone." However, when a HTR model trained on a *SegmOnto* vocabulary identifies a region of text as a "MainZone:column#1", for example, that <zone> in the TEI publication would have the value "MainZone" for its attribute @type the value "column" for its attribute @subtype and the value "1" for its attribute @n.

3.2 <body>

Segmonto zone	Corresponding TEI element
CustomZone	
DamageZone	<damage>
DigitizationArtefactZone	<figure type="DigitizationArtefactZone">
DropCapitalZone	<hi type="DropCapitalZone">
GraphicZone	<figure type="GraphicZone">
MainZone	<ab>
MarginTextZone	<note type="MarginTextZone">
MusicZone	<musicNotation>
NumberingZone	<fw type="pageNumber">
QuireMarksZone	<fw type="QuireMarksZone">
RunningTitleZone	<fw type="RunningTitleZone">
SealZone	<figure type="SealZone">
StampZone	<figure type="StampZone">
TableZone	<table>
TitlePageZone	

Table 2: SegmOnto regions and their corresponding TEI element.

The transcription is stored both in the <sourceDoc> element and the <body>. The schema used for the latter is less strict than the one used for the <sourceDoc> and allows philologists to have a pre-edited text via another mapping between ALTO and TEI (cf. tab. 2).

Whereas in the <sourceDoc> every page was completely contained within an element <surface>, the pre-edited text in the <body> merely interrupts a continuous stream of text with the element <pb>, meaning page beginning. This empty element carries the attribute @corresp and points to the xml:id of the corresponding page in the <sourceDoc> (cf. ex. 8).

```
<body>
 <div>
  <pb corresp="#page5"/>
  <note corresp="#page5_zone2" type="MarginTextZone">
   <lb corresp="#page5_zone2_line1"/>79/4120
  </note>
  <pb corresp="#page6"/>
  <ab corresp="#page6_zone1">
   <hi rend="HeadingLine">
    <lb corresp="#page6_zone1_line1"/>BRADAMANTE,
    <lb corresp="#page6_zone1_line2"/>TRAGECOMDEDIE.
   </hi>
  </ab>
  <pb corresp="#page9"/>
  <fw corresp="#page9_zone1" type="RunningTitleZone">
   <lb corresp="#page9_zone1_line1"/>AV ROY.
  </fw>
  <ab corresp="#page9_zone2">
   <lb corresp="#page9_zone2_line1"/>uiuront nostre siecle, les admira¬
   <lb corresp="#page9_zone2_line2"/>bles effets de vos heroiques ver¬
   <gap reason ="sampling"/>
  </ab>
 </div>
</body>
```

Example 8: Example of a pre-editorialised <body>: Robert Garnier, *Tragédies*, Paris: Robert Estienne, 1582.

IV CONCLUSIONS AND FURTHER WORK

For several years, the expression "automated editions", defined as "presentational editions generated from both digital images of text, and their corresponding transcriptions created by artificial intelligence" [Terras et al., 2023] has become more and more frequent. Some scholars also remark (deplore?) that "editions generated via HTR [are] not considered in the scholarly editing literature" [Mühlberger and Mansutti, 2022], insisting on the role of the editor training and documenting the HTR model. This is not the place to engage in an umpteenth debate on the definition of the term "(scholarly) edition", but it seems to us that these debates maintain the confusion between the transcription, obviously useful, and the edition itself. Producing a good transcription is a complex, laborious act, which should not be devalued, but the "critical" part remains absent [Duval, 2017]. In other words, paleography is not ecdotics.

If HTR cannot automatically produce "editions" but only transcriptions, segmentation cannot produce "editions" either, but contribute to speeding up the editorial work. As the pipeline for a critical edition is necessarily "semi-automatic" [Stoekl Ben Ezra et al., 2022], without

human intervention one cannot expect more than a "pre-edition" (or a corpus). The creation of a vocabulary like *SegmOnto* is therefore a necessary condition, but not a sufficient one for the creation of proper scholarly editions.

Preliminary tests have already show the efficiency of a TEI encoding from ALTO via *SegmOnto* [Janès et al., 2021]. We are now facing large scale tests, which will help refine the use of especially subtypes, whose impact of the efficiency of the region and classification is not fully understood. Mixing data from various documents from different period or genre, to increase the efficiency, also has to be evaluated, in order to create to most efficient models.

DATA

Full documentation, with more examples for each of the types, is available and maintained online: https://segmonto.github.io. It is preferable to refer to the site rather than to this article in the event of modification of the guidelines.

ACKNOWLEDGEMENTS

We would like to thank previous interns for their precious work (Juliette Janès and Claire Jahan), the BNF Datalab for the funding, and our colleagues of the *Gallic(orpor)a* project at the INRIA Paris (Benoît Sagot, Rachel Bawden, Pedro Ortiz Suarez) and the université Gustave Eiffel (Philippe Gambette). Daniel Stökl Ben Ezra and Peter Stokes participated in the first working discussions.

References

Lou Burnard. *What is the Text Encoding Initiative? : How to add intelligent markup to digital resources*. Encyclopédie numérique. OpenEdition Press, 2014. URL https://doi.org/10.4000/books.oep.426.

Thibault Clérice. "Don't worry, it's just noise": quantifying the impact of files treated as single textual units when they are really collections. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, Virtual, India, December 2021. URL https://aclanthology.org/2021.nlp4dh-1.11.

Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. working paper, 2023. URL https://hal-enc.archives-ouvertes.fr/hal-03723208.

Marc Douguet. *La Composition dramatique : La liaison des scènes dans le théâtre français du XVIIe siècle*. PhD thesis, Université Paris 8 | Vincennes - Saint-Denis, 2015.

Frédéric Duval. Pour des éditions numériques critiques. lexemple des textes français. *Médiévales. Langues, Textes, Histoire*, 73(73):13–29, 2017. ISSN 0751-2708. URL https://doi.org/10.4000/medievales.8165.

Marc Geoffroy, Anne-Marie Eddé, Youssef Baratli, Denis Muzerelle, Philippe Bobichon, and Marie-Geneviève Guesdon. *Vocabulaire Internationale de la Codicologie SKOS - GAMS: Vokabularien und Ontologien*. GAMS Vokabularien und Ontologien. Zentrum für Informationsmodellierung - Karl-Franzens-Universität Graz, Graz, 2021. URL http://gams.uni-graz.at/archive/objects/o:voccod/methods/sdef:SKOS/get.

Juliette Janès, Ariane Pinche, Claire Jahan, and Simon Gabay. Towards automatic TEI encoding via layout analysis. In *Fantastic future 21, 3rd International Conference on Artificial Intelligence for Librairies, Archives and Museums, AI for Libraries, Archives, and Museums*, Paris, France, 2021. URL https://hal.archives-ouvertes.fr/hal-3527287.

Benjamin Kiessling. A modular region and text line layout analysis system. In *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 313–318, 2020. URL https://doi.org/10.1109/ICFHR2020.2020.00064.

Denis Muzerelle. *Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits*. Rubricae, 1. Institut de recherche et d'histoire des textes, 1985. URL http://codicologia.irht.cnrs.fr/.

Günter Mühlberger and Sara Mansutti. Could handwritten text recognition revolutionise the future of digital editions of primary sources? transkribus as a platform to facilitate the editorial workflow. In *Digital Publishing for the Humanities: New Technologies and Ideas*, Rome, Italy, 2022. URL https://doi.org/10.17617/3.PEJU8B.

Stefan Pletschacher and Apostolos Antonacopoulos. The page (page analysis and ground-truth elements) format framework. In *2010 20th International Conference on Pattern Recognition*, pages 257–260. IEEE, 2010. URL https://doi.org/10.1109/ICPR.2010.72.

Christian Reul, Uwe Springmann, and Frank Puppe. LAREX: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, pages 137–142. Association for Computing Machinery, 2017. URL https://doi.org/10.1145/3078081.3078097.

Daniel Stoekl Ben Ezra, Hayim Lapin, Bronson Brown-Devost, and Pawel Jablonski. From htr to critical edition: A semi-automatic pipeline. In *Ancient documents and automatic recognition of handwriting*, Paris, France, 2022. URL https://www.canal-u.tv/133346.

Melissa Terras, Joe Nockels, and Paul Gooding. On automating editions: The affordances of handwritten text recognition platforms for scholarly editing. *Scholarly Editing*, 2023. ISSN 2167-1257.