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**[3]

[1]Université de Genève
[2]CNRS CIHAM UMR 5648
[3]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 more physical approach rather than a strictly semantic one, it is designed as a pragmatic and generic typology, coping with most of the Western historical documents rather than answering specific needs.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 page 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. To demonstrate the capacity of *SegmOnto* to answer both these objectives, we aggregate data from multiple projects to train a layout analysis model, and we propose a prototype of a generic pipeline for converting ALTO-XMLs into XML-TEI.

I INTRODUCTION

Layout analysis is a fundamental requirement for document processing. Because a textual source is composed of multiple elements, containing different kinds of data (engravings, library stamps, headings, verses... cf. fig. 1), acquiring only the raw text is not satisfactory. To optimise our grasp on a document, all the elements it contains need to be captured and analysed during the extraction process. This analysis serves to identify textual and visual components, to distinguish the text from the paratext, to reconstruct the reading order in the case of complex layouts such as those consisting of several columns (cf. fig. 2), etc. In other words, layout analysis is a key step in document processing, in addition to transcription, to produce semi-structured data.


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 ruë Nostre Dame, au Chappon d'Or couronné. 1668.
Ne extra hanc Bibliothecam efferatur. Ex obedientiâ.

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

Semi-structured data are particularly useful because they are machine actionable: they can be

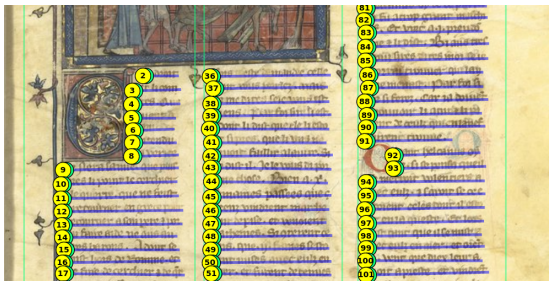


Figure 2: Ms. BnF, fr. 25550, 14th c. Lines are reordered (yellow circle) according to the layout in three columns.

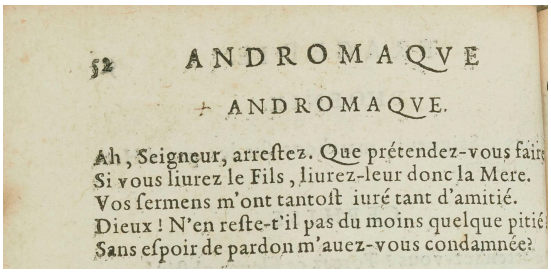


Figure 3: Racine, *Andromaque*, 1668. The first *ANDROMAQUE* is the running title, while the second is part of the play.

reorganised, filtered or converted into formats that allow their distribution or facilitate their exploration [Clérice, 2021]. For example, in order to study the character's presence in a play (such as Douguet 2022), elements that are part of the editorial paratext should be differentiated from the body of the text. If we look at the 1668 edition of *Andromaque* by Racine, because "Andromaque" is both the name of the main character and the title of the play, many occurrences of this name come from the running title (cf. fig. 3). This produces noise and distort the results of the analysis by artificially over-representing the character of Andromaque: the running title has to be filtered out. In addition to this first problem, from an ecdotic point of view, a scholarly edition does not usually include running titles or original page numbers, nor does it exactly replicate every aspect of the layout of the source (e.g. columns). For documents with (highly) complex layouts, such as medieval manuscripts, the automatic identification of musical notations (cf. fig. 4) or commentaries and glosses (cf. fig. 5) is extremely useful for speeding up editing work.

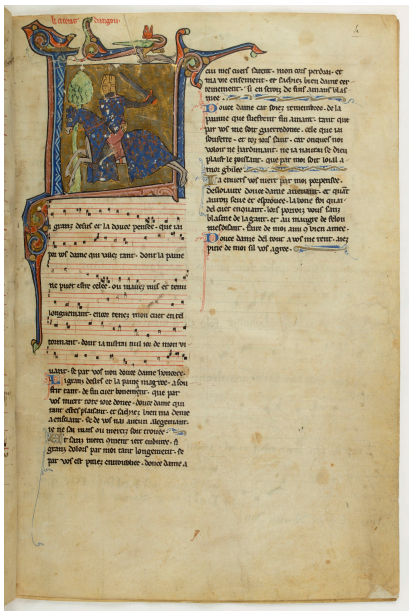


Figure 4: *Chansonnier du Roi*, Ms. BnF, fr. 844, fol. 4r.

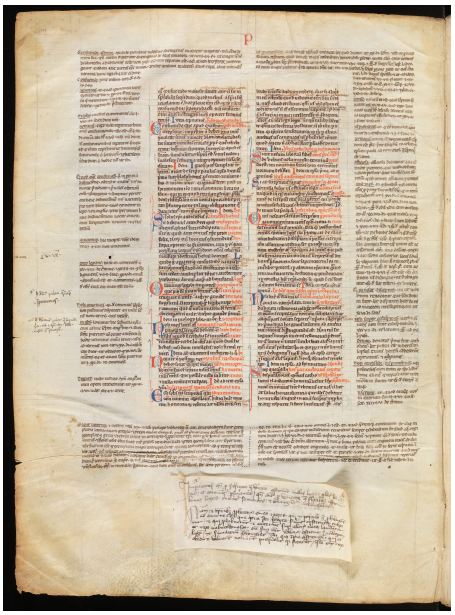


Figure 5: *Decretum Gratiani*, Sion, Archives du Chapitre, Ms. 89, fol. 3v.

New methods are now available to classify different zones, as well as different types of text line 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 OCR projects in the humanities remains limited without a clear taxonomy for

document description, primarily due to the tendency of each initiative to establish its individual approach. This lack of homogeneity impedes data sharing and the optimisation of tools reuse. Thus, the standardisation of practices has many interests, the following two being of special importance:

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

The primary aim of *SegmOnto* is to establish a controlled vocabulary for the digital humanities community to describe document layout during the segmentation phase of the OCR process. It offers a simple, concise, and predominantly layout-based vocabulary that is sufficiently generic to support a wide range of uses and document types within the humanities, from medieval manuscripts to 19th-century documents across Western Europe. Although primarily suited for books (printed or manuscript), it remains usable for other formats, such as scrolls. The limited set of classes is designed to streamline model training and annotation, yet the vocabulary remains flexible, allowing users to extend it based on project-specific needs and available training data. For example, with larger datasets or document features, more detailed annotations, such as dialogue shifts, paragraph beginnings, or continuations, can be introduced to refine the structural analysis. This vocabulary is intended to integrate seamlessly with OCR software environments, such as eScriptorium [Kiessling et al., 2019, Stokes et al., 2021].

The following sections present a state-of-the-art on layout analysis standards in both computer vision and codicology, an in-depth presentation of the *SegmOnto* vocabulary itself, the results of a first model trained on data provided by several projects, and a discussion of its potential integration with XML-TEI.

II STATE-OF-THE ART

2.1 Preceding works in computer vision

Recent advances in deep learning have significantly improved document segmentation algorithms. In 2019, when working on manuscripts in Transkribus and e-Scriptorium, it was still necessary to manually split images to separate columns and preserve the correct line order within each folio [Camps et al., 2021]. Layout analysis is now a key area of research to automatically analyse documents. Page segmentation encompasses multiple approaches, including pixel classification [Capobianco et al., 2018], as in Kraken [Kiessling, 2019]; object detection with tools like YOLO [Jiang et al., 2022] used by Prasad et al. [2020] for table detection or YALTAi [Clérice, 2023] for page segmentation; and even multi-modal methods that perform jointly the segmentation and the transcription process [Liu et al., 2019, Xu et al., 2021]). These new solutions efficiently predict zone types and therefore enable a significant sophistication in the extraction process, retaining more information than just the text contained on the page.

If new tools allow us to recognise the different zones of a page, it is still necessary to define the types of zones that we want to detect. ALTO-XML, one of the two standard formats used to store data, do not offer a consistent vocabulary grounded in text analysis standards [Stehno et al., 2003]. The other one, PAGE-XML [Pletschacher and Antonacopoulos, 2010], provides

a typology of about fifteen zone types[1] that is easy to implement in a classification algorithm. However, this typology falls short for philological needs, as it includes only a single category for all textual content zones (TextRegion).

In the absence of a real standard, research teams have designed *ad hoc* vocabularies, used primarily for evaluating or training models, usually on digital native documents like PDFs rather than on analogue (historical) sources. As a result, these vocabularies lack connections to philological standards. Various datasets have been developed to benchmark different approaches (tab. 1). In these datasets, annotations are generally divided between appearance-based analysis (sometimes also referred to as logical analysis) and semantic-based analysis. The former organises documents hierarchically by logical components, such as section headings, paragraphs or lists – based on visual attributes like location, type of writing or font, and text or image appearance [Mao et al., 2003]. In contrast, semantic-based layout analysis seeks to interpret the meaning of various regions and categorise them accordingly [Lee et al., 2019]. Often, this second approach requires the integration of textual OCR-ised information to identify elements like paragraphs or dictionary entries and their sub-components. While promising, such content-based techniques are still primarily experimental and not widely implemented in standard OCR tools.

Dataset	Size (pages)	Corpus constitution	Number of classes	Classes Classes
DocBank	500K	STEM papers	12	Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, Title
PubLayNet	360K	Medical papers	5	Text, Title, List, Table, Figure
DocLayNet	80863	financial reports, scientific articles, laws and regulations, government tenders,	11	Caption, Footnote, Formula, ListItem, Page Footer, Page Header, Picture, Section Header, Table, Text, Title
M^{6}Doc	9080	Textbook, test paper, magazine, newspaper, scientific papers, notes, book	74	(non exhaustive list) Qrcode, advertisement, algorithm, bracket, caption, catalogue, chapter title, correction, dateline, editors note, end note, figure, footer footnote, fourt-level section title, fourth-level title, index, inside, marginal note, ordered list, other question number, page number, table, table caption, table note, teasers, translator, unordered list...
PRImA	1240	Magazine, technical journals, forms, bank statements, ads	10	Text, image, line drawing, graphic, table, chart, separator, maths, noise, frame

Table 1: Non-exhaustive list of available segmentation datasets in the field of computer vision. The table is adapted from Clérice et al. [2024].

As shown in tab. 1, the labelling conventions across datasets like DocBank [Li et al., 2020], PubLayNet [Zhong et al., 2019], DocLayNet [Pfitzmann et al., 2022] and PRImA [Antonacopoulos et al., 2009] vary widely, lacking a unified approach. PubLayNet, with its 360,000 images, employs very basic labels such as text, title, list, and table. In contrast, DocBank, which contains 500,000 images, includes more specialised categories like author, caption, and equation, reflecting its focus on STEM documents. PRImA, developed for multilevel segmentation evaluation, introduces additional physical markers, such as line drawing and separator.

The M^{6}Doc dataset [Cheng et al., 2023] proposes a drastically different approach with a vast panel of possible labels, introducing 76 classes for modern documents, using concepts coming from Wang [2019] and referring to YouTube video explanations regarding magazine and newspaper

[1] the different zones are TextRegion, ImageRegion, GraphicRegion, ChartRegion, LineDrawingRegion, SeparatorRegion, TableRegion, MathsRegion, ChemRegion, MusicRegion, AdvertRegion, NoiseRegion and UnknownRegion.

layouts. Those labels go from page numbers to advertisements and marginal notes. However, this expansion brings its own challenges: complex class hierarchies can hinder annotation quality, creating ambiguities that may lead to inconsistent labelling. As reported in Cheng et al. [2023], some categories were seen as ambiguous by the annotators (e.g. table and list, where visual cues makes their differentiation challenging). This ambiguity may have introduced inconsistencies into the dataset despite the authors' review efforts. Upon examining the proposed labels more closely, several doubts emerge: for example, the difference between footer and footnote, or between fourth-level section title and fourth-level title is unclear, particularly when the segmentation operates page-by-page. These categories likely come from the earliest datasets tagged with labels derived from LaTeX files analysis. The list of labels also raises issues from a textual analysis perspective, such as treating paragraph, play and poem at the same level. A paragraph is a structural component within a text, which can be a play or a poem. Furthermore, while a poem might fit within a single page, the same is unlikely for a play, where terms like speech or replica and stanza or line group would provide more precise segmentation labels alongside paragraphs. The lack of an underlying ontological structure further complicates the annotation task. Finally, the M^{6}Doc guidelines are only available in Chinese, restricting accessibility for a broader audience.

To conclude, despite initial steps toward refining annotations, the classes provided by such projects are poorly suited for historical documents or cross-genre analysis. If the most used vocabularies are up to only five to six classes, the tentative to (dramatically) increase their number complicates model generalisation and presents annotation challenges, while lacking ties to codicological expertise or established standards in the humanities. In contrast, to describe historical documents, researchers in the humanities need a more meaningful framework. To address this gap, the *SegmOnto* controlled vocabulary aims to offer an adapted solution that operates at the level of logical structuring, assigning labels based on the physical properties represented in the source. While primarily focused on physical layout, certain zones (e.g. running title, margin note) may also carry semantic nuances.

2.2 Layout analysis vocabulary for the Humanities

Some projects focusing specifically on historical documents have developed specialised class lists to describe the various zones within their materials (cf. 2). The SCUT-CAB dataset on ancient Chinese books offers two subsets, one with 4 (physical) classes (centerfold strip, figure, page box and text) and another one with 27 (logical) classes [Cheng et al., 2022]. The *American Stories* dataset consists of historical American newspapers mainly published between 1880 and 1920 and annotated with 7 classes, like the *Japanese Historical Documents* dataset [Shen et al., 2020]. Although smaller in scale than those mentioned above (up to 500,000 pages), they all exceed 2,000 pages.

There are also projects on a smaller scale in terms of data (but not in terms of complexity), carried out by specialists in human sciences who adopt very different approaches. The *Horae* project, led by expert palaeographer Dominique Stutzmann, includes classes informed by codicological expertise, such as detailed descriptions of initials, categorised as simple, decorated, or historiated [Hazem et al., 2020]. It therefore offers a high quality dataset, but project-oriented and therefore difficult to share and reuse widely within the humanities research community. On the contrary, the *Ajax Multicommentary* project, directed by a specialist in classical philology and digital humanities (M. Romanello), uses a preliminary version of *SegmOnto*, and despite its strong focus on paratext and textual criticism (with classes such as commentary, critical apparatus, footnotes, and printed marginalia), it is therefore more reusable [Najem-Meyer and Romanello,

2022].

Dataset	Size (pages)	Corpus constitution	Number of classes	Classes Classes
SCUT CAB	4000	Chinese Ancient Books	27	bibliography, header, pagebox, book number, figure, author, title, ear note, chapter title...
Historial Japanese Dataset	2271	Japanese Historical documents (19th-20th)	7	page frame, row, title region, text region, title, subtitle, other
American Stories	2200	Historic Press	7	articles, headlines, captions, bylines, images, tables, mastheads
HORAE	500	Books of hours	13	Page, textregion, bordertext, textline, miniature, decorated border, illustrated border, initial (simple, decorated, historiated), line filer, music notation, ornamentation
AJAX Multicommentary Dataset	300	Critical edition 19th	18	Commentary, critical apparatus, footnotes, page number, text number, bibliography, handwritten marginalia, index, others, printed marginalia, table of contents, title, translation, appendix, introduction, preface, primary text, running header

Table 2: Example of datasets dealing with historical documents. The table is adapted from Clérice et al. [2024].

In the traditional field of codicology, several vocabularies already exist, such as the *Vocabulaire codicologique* by Denis Muzerelle [1985], now available online with other resources on the *Codicologia* site[2]. Richly illustrated, offering translations of technical terms in several languages, available in a SKOS version [Geoffroy et al., 2021], the *Vocabulaire codicologique* is one of the main resources available. The degree of precision of the codicological analysis is however too high for an automatic computer analysis: the different types of ruling, the presence of long lines, the difference between the different types of initials (champ, historiated, pen-flourished, etc.) are important information, but would require too large a number of classes to be efficient.

Consequently, if digital humanists aim to establish a standardised approach for describing document layout in OCR contexts, the field needs a simple, but both accurate and generic standard to optimise data sharing and reuse, as data producing is time-consuming and expensive, and training effective models relies on access to extensive datasets. Previous initiatives have either lacked sufficient focus on historical document expertise or have been overly project-specific, limiting their broader applicability.

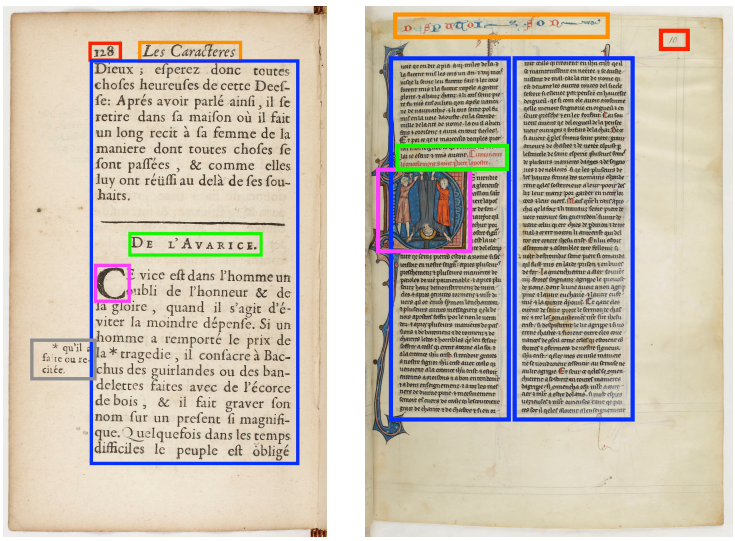


Figure 6: Similarities between the layout of a modern print (left) and a medieval manuscript (right). The running title is in orange, the pagination/foliation in red, headings/rubrics in green, drop capitals in pink, main textual zones in blue and marginal notes in grey.

In response to these observations and inspired by the *Vocabulaire codicologique*, the *SegmOnto* controlled vocabulary seeks to draw on these approaches while balancing their strengths and limitations. Its goal is to establish a simple, consistent, non-project-dependent taxonomy that

[2]https://codicologia.irht.cnrs.fr.

is computationally easy to implement yet adaptable to general philological needs. This taxonomy has been discussed since 2021 by several researchers faced with the problem of layout analysis from a practical point of view, when creating their corpora. Medieval and modern romance philology being the most represented discipline among these researchers, western manuscripts and historical prints received particular attention, but a desire to cover a wider range of cases than this initial base remained present throughout the work.

SegmOnto offers a streamlined vocabulary applicable across diverse document types, allowing for tailored adaptations if necessary. It avoids classes that require complex textual interpretation within zones, reducing ambiguities and allowing flexibility for segmentation tasks. It also proposes a solution that seeks for simplicity and independence from specific technologies. The reflections presented here are a starting point, not an end point, born from a pragmatic desire to harmonise practices, without the ambition of resolving (for now) all the problems posed by the analysis of layout.

III 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 perspective focused on material aspects. It encompasses the following fifteen zone types:

- CustomZone
- DamageZone
- DigitizationArtefactZone
- DropCapitalZone
- GraphicZone
- MainZone
- MarginTextZone
- MusicZone
- NumberingZone
- QuireMarksZone
- RunningTitleZone
- SealZone
- StampZone
- TableZone
- TitlePageZone

We also propose six types of lines: CustomLine, DefaultLine, DropCapitalLine, HeadingLine, InterlinearLine, and MusicLine.

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 instance, 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 some illustrations such as some 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?

Obviously, the answers to these questions depend on each 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 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, several mechanisms have been designed to cope with problematic cases and potentially go beyond its Western original objective.

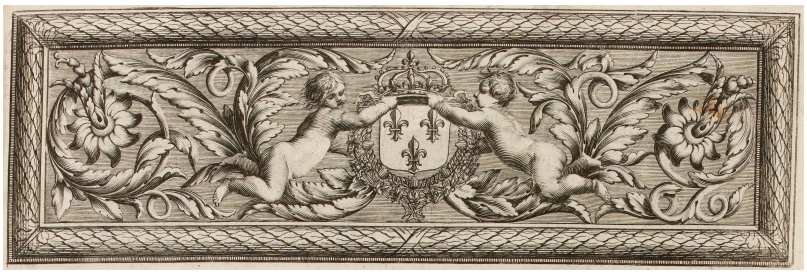


Figure 7: Headpiece

Figure 8: Engraving

Figure 9: Tailpiece

Figure 10: Drop capital

3.1 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. sec. 3.2), 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

The 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 three components of the label in order to annotate, for instance, 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 to 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 precise the annotation, the more complexity it adds to the description, and therefore more training data are required to reach a decent accuracy for the classifier.

3.2 The *SegmOnto* types

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

3.2.1 The SegmOnto regions

Our choice of zones is a combination of types taken from the possibilities offered by PageXML and *Codicologia*. From the first, we take the idea of zones covering a broad range of possibilities: images, tables, products of the digitisation process or a custom zone which can be used as a wild card. This first selection (which could eventually be extended in the future and include other zone types for mathematical or chemical formulas) is augmented of specific elements taken from the most recurrent one in codicology: running title, page or folio numbering, quire marks, etc.

A first selection is the following:

- 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 is usually 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 subtypes are:
  - NumberingZone:page
  - NumberingZone:folio
  - NumberingZone:item

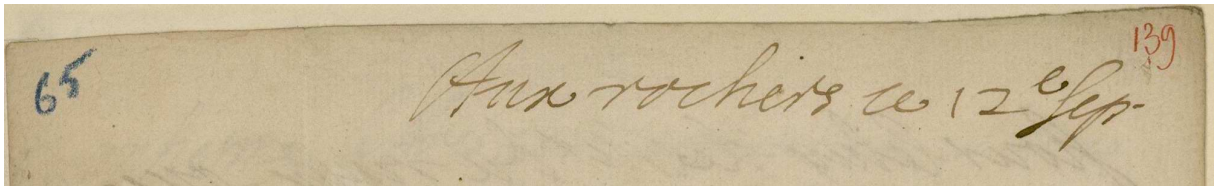


Figure 11: Double numbering: item (left) and folio (right)

- MainZone[3] is the area containing the body text, and is either a single block or multiple columns. If it is impossible to separate the body text from the paratext (e.g. commentaries or glosses), the latter should be included in the former, possibly using a commented subtype to indicate the specificity of the case (cf. fig. 13. 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, each one is a different zone. Possible values for subtypes are:

[3]The name derives from the basic segmentation options offered from older version of eScriptorium, and does not imply a functional analysis of the zone.

- MainZone:column
- MainZone:block

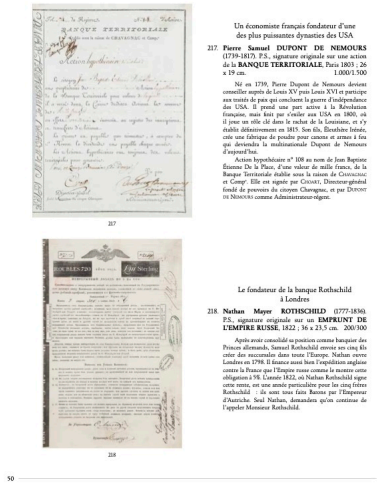


Figure 12: MainZone:block

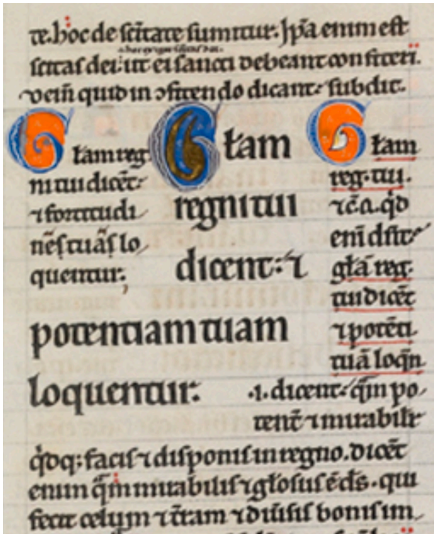


Figure 13: MainZone:commented

- 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. Possible values for subtypes are:
  - MarginTextZone:note
  - MarginTextZone:commentary
  - MarginTextZone:correction
  - MarginTextZone:addition
  - MarginTextZone:criticalApparatus

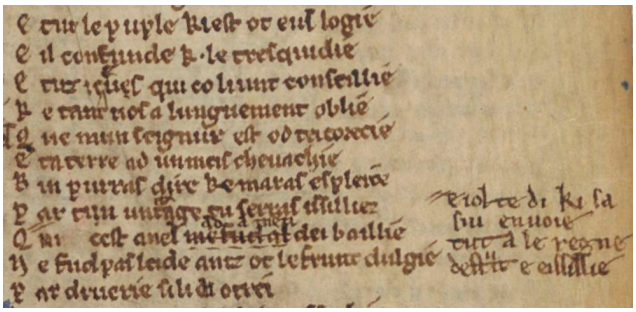


Figure 14: MarginTextZone:addition

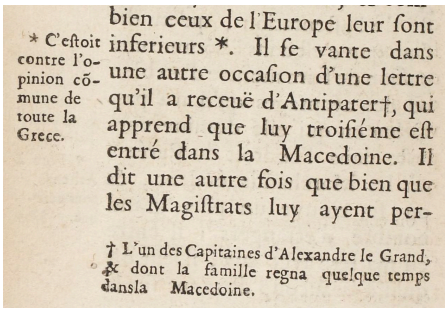


Figure 15: MarginTextZone:note

- GraphicZone is a zone containing any type of graphic element, from purely ornamental information to information consubstantial 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, are 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. 7, fig. 9.
  - GraphicZone:figure, cf. fig. 16 et 17.

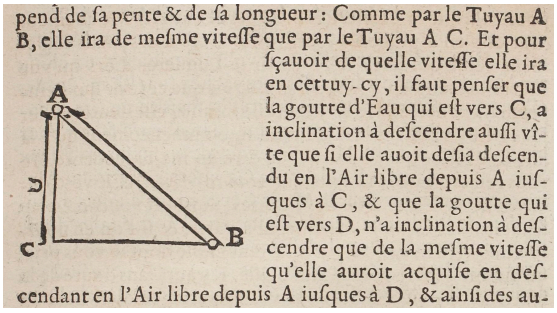


Figure 16: GraphicZone:figure

Figure 17: GraphicZone:figure

- DropCapitalZone contains any type of initial letter that occupies a space corresponding to several lines of the main text and/or that bears significant ornamentation. A drop capital can be a historiated, flourished or voided 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. Possible values for subtypes are:
  - DropCapitalZone:historiated
  - DropCapitalZone:flourished
  - DropCapitalZone:voided

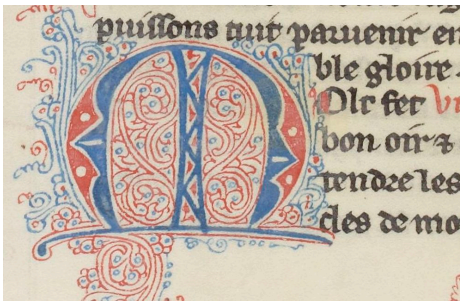


Figure 18: DropCapitalZone:flourished

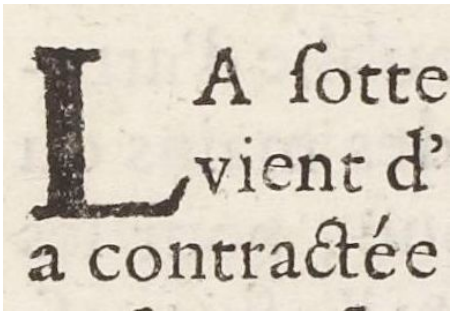


Figure 19: DropCapitalZone

- 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 the title of the current section[4].

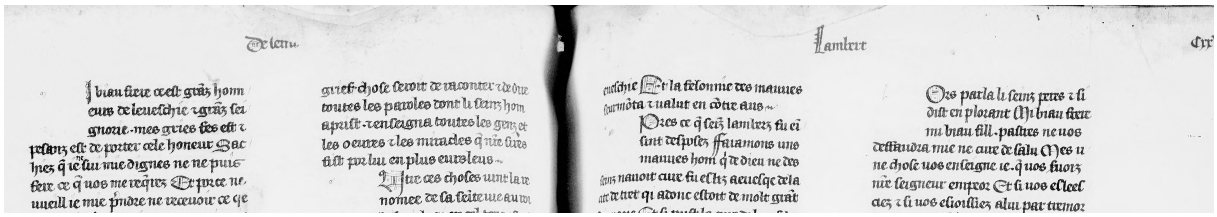


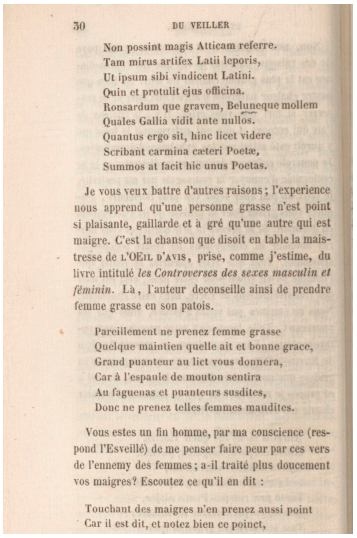
Figure 20: RunningTitle

- 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

[4]We opted for the label RunningTitleZone instead of Header, considering that the latter may also encompass elements such as NumberingZones.

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.

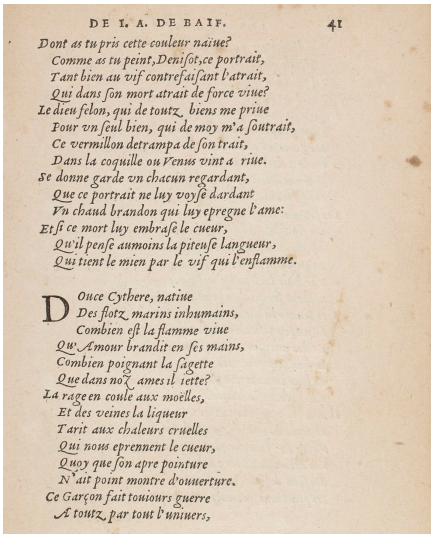
Je vous veux battre d'autres raisons ; l'experience nous apprend qu'une personne grasse n'est point si plaisante, gaillarde et à gré qu'une autre qui est maigre. C'est la chanson que disoit en table la maistresse de L'OEIL D'AVIS, prise, comme j'estime, du livre intitulé *les Controverses des sexes masculin 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 21: Sections in prose and in verse


DE I. A. DE BAIF. 41

Dont as tu pris cette couleur naïue?
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
Pour 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'uniuers,

Figure 22: Several poems

- 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 23: StampZone:post (3 times)

Figure 24: StampZone:library

- QuireMarksZone is a zone containing a quire signature, 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 is usually at the bottom of the page. Possible values for subtypes are:
  - QuireMarksZone:signature
  - QuireMarksZone:catchword

Figure 25: QuireMarksZone:catchword

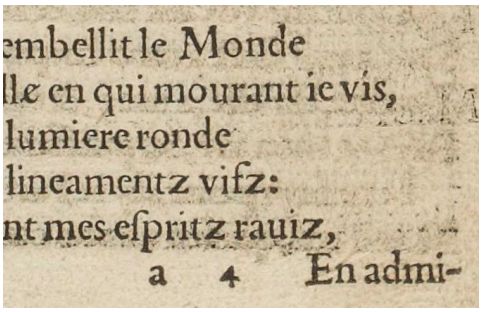


Figure 26: QuireMarksZone

- TableZone is a zone containing information structured in lines and columns, or a list organised similarly on multiple columns. 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.

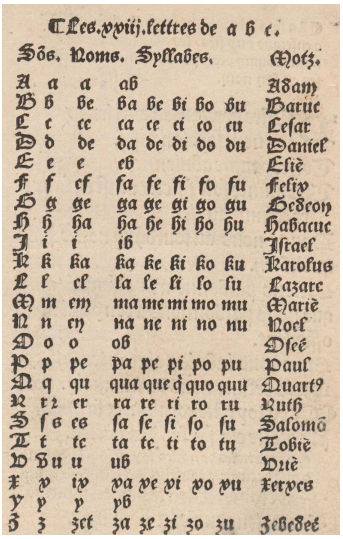


Figure 27: TableZone

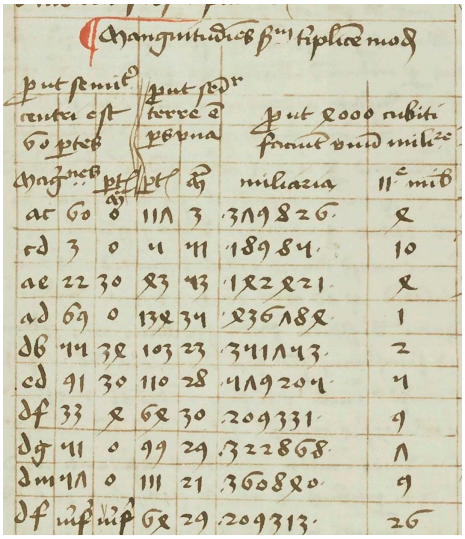


Figure 28: TableZone

- DigitisationArtefactZone 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:
  - DigitisationArtefactZone:ruler
  - DigitisationArtefactZone:colorTarget

Figure 29: DigitisationArtefactZone

Figure 30: DigitisationArtefactZone:ruler

- 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
  - DamageZone:hole
  - DamageZone:mould
  - DamageZone:soaked
  - DamageZone:stained

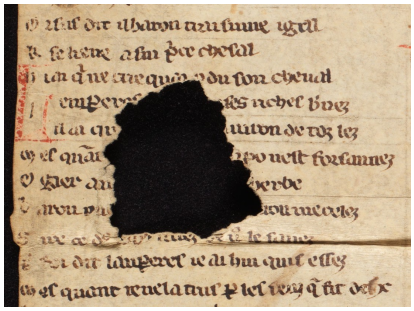


Figure 31: DamageZone:hole

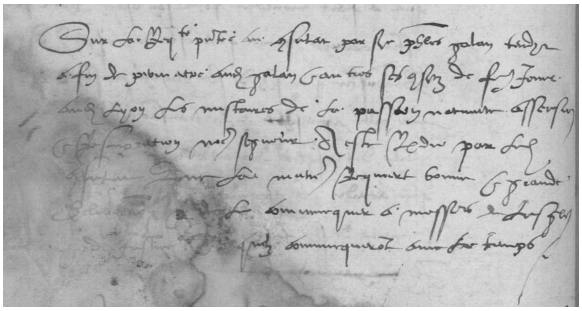


Figure 32: DamageZone:soaked

- 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.
- SealZone is a zone containing a seal.
- MusicZone is an area containing musical notations, such as neumes, staves, etc. It can include text. Possible values for subtypes are:
  - MusicZone:neumes
  - MusicZone:notes

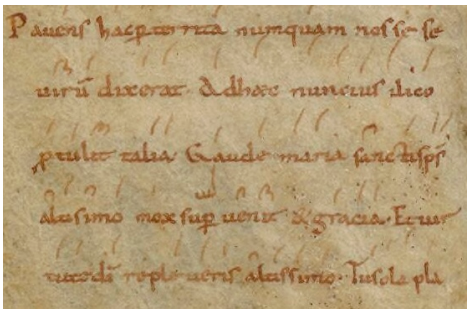


Figure 33: MusicZone:neumes

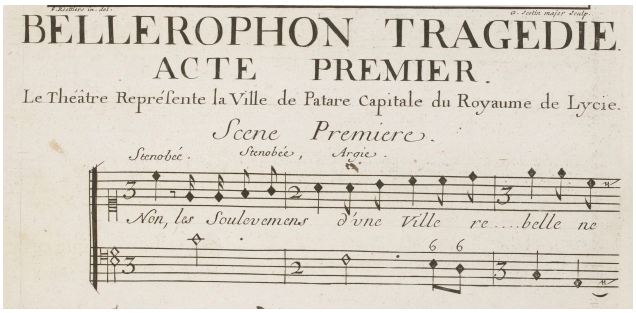


Figure 34: MusicZone:notes

3.2.2 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 MarginTextZone 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, or a portion of a line containing any type of heading, which is defined as a string with a distinctive typesetting (font, size, colour, capitalisation. . .) from the one seen in the body of the text. The HeadingLine usually 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 not limited to the header, and can be used in the TitlePageZone and the RunningTitleZone. Possible values for subtypes are:
  - HeadingLine:rubric
  - HeadingLine:title
  - HeadingLine:incipit

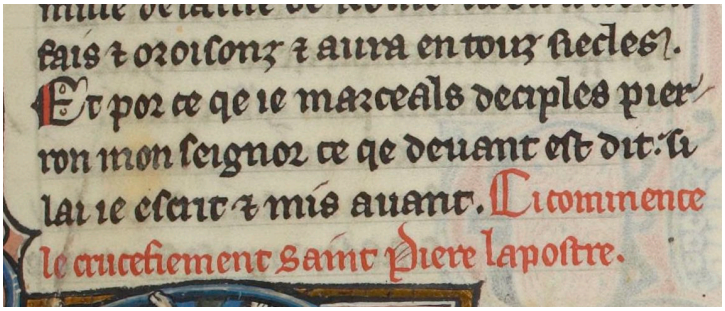


Figure 35: HeadingLine:rubric

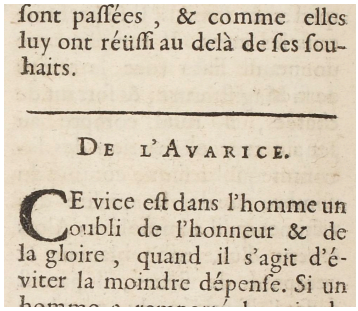


Figure 36: HeadingLine

- InterlinearLine is a line that is not a standard text line. Instead, it is a line that has been added between two text lines, often in order to include a forgotten word or a gloss. Possible values for subtypes are:
  - InterlinearLine:addition
  - InterlinearLine:correction
  - InterlinearLine:gloss

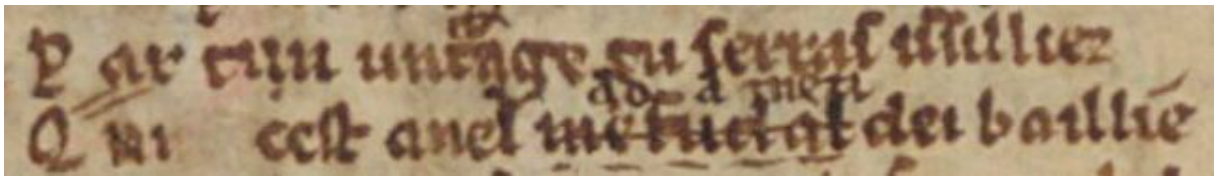


Figure 37: InterlinearLine:correction

- 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;
- MusicLine characterises the central line of a musical stave.

IV TRAINING A MODEL

In order to test the hypothesis of a performance gain through data pooling, several datasets annotated with *SegmOnto* have been produced (cf. tab. 3). In order to guarantee data quality, the datasets are controlled using continuous integration tools developed by Chagué et al. [2021], Clérice et al. [2023], which allows to report *SegmOnto* annotation errors.

Project	Type	Century	Pages	Zones	Set	Github repo
Gallic(orpor)a	Manuscript	15th c.	85	458	Train	HTR-MSS-15e-Siecle
Gallic(orpor)a	Incunable	xvth.	149	535	Train	HTR-incunable-15e-siecle
Sub-total (i)			**234**	**993**		
Gallic(orpor)a	Print	16th c.	80	233	Test	HTR-imprime-gothique-16e-siecle
SETAF	Print	16th c.	895	2 752	Train	HTR-SETAF-Pierre-de-Vingle
SETAF	Print	16th c.	404	1 365	Train	HTR-SETAF-Jean-Michel
SETAF	Print	16th c.	144	485	Train	HTR-SETAF-LesFaictzJCH
SETAF +	Print	16th c.	58	220	Train	HTR-Varia-Malingre-gothique
Sub-total (ii)			**1 581**	**5 055**		
SETAF +	Print	16th c.	202	1 062	Train	HTR-Varia-Malingre-romain
FoNDUE	Print	16th c.	223	688	Train	FONDUE-LA-PRINT-16
FoNDUE	Print	16th c.	930	2 829	Train	FONDUE-FR-PRINT-16
Gallic(orpor)a	Print	16th c.	180	591	Train	HTR-imprime-16e-siecle
Gallic(orpor)a	Print	1th c.	327	1 185	Train	HTR-imprime-17e-siecle
FoNDUE	Print	xviith.	69	246	Train	FONDUE-FR-PRINT-17
FoNDUE	Manuscript	18th c.	153	460	Train	FONDUE-FR-MSS-18
Gallic(orpor)a	Print	xviiith.	160	624	Train	HTR-imprime-18e-siecle
Sub-total (iii)			**2 244**	**7 685**		
FoNDUE	Print	19th c.	48	129	Train	FONDUE-ES-PRINT-19
FoNDUE	Print	20th c.	28	67	Train	FONDUE-IT-PRINT-20
FoNDUE	Print	20th c.	30	72	Train	FONDUE-EN-PRINT-20
FoNDUE	Print	20th c.	55	64	Train	FONDUE-FR-PRINT-20
-	Print	20th c.	47	126	Train	HN2021-OCR-Poesie-Corse
Sub-total (iv)			**208**	**458**		
FoNDUE	Print	20th-21th c.	60	197	Test	FONDUE-MLT-PRINT-TEST
Sub-total (v)			**60**	**197**		
Total			**4 327**	**14 388**		

Table 3: data details. We distinguish (i) medieval data, (ii) Renaissance data in Gothic characters (16th.), (iii) modern data in Roman script (16th-18th c.), (iv) contemporary data (19th-20th c.), (v) test data from a randomised selection of Gallica and Persée.

The release of YALTAi v. 1.0.0, which converts ALTO data to YOLO data, allows us to use the YOLO v8x model provided by Ultralytics [Jocher et al., 2023] rather than the v5x available at the beginning of the project. This development should allow us to obtain a substantial gain in the efficiency of the models produced [Ronkin and Reshetnikov, 2023]. When training with YALTAi, we use batches of 32 images, which are resized as input to 896 pixels, with a minimum of 150 epochs.

Two specialised models are trained – one with Renaissance data taken from Gothic prints ("Gothic"), another with modern prints in Latin script ("Modern") – and a global model ("Global") with all the available data (the complete experiment is available in Solfrini et al. 2024). We observe (cf. tab. 4) that both models trained on specific data perform reasonably well on documents different from that seen during training, demonstrating the effectiveness of our page modelling. Adding additional data from other centuries in training has a positive impact on the Gothic test data (cf. fig. 38), and the Global model[5] is in most cases the most efficient,

[5]The model has been published by Humeau et al. 2024.

Model	Precision	Recall	mAP50	mAP50-95
Test on gothic data only				
Gothic	0.719	0.7	0.712	0.519
Modern	0.81	**0.756**	0.777	**0.632**
Global	**0.969**	0.711	**0.789**	0.627
Test on modern data, including gothic data				
Gothic	0.72	0.497	0.462	0.327
Modern	0.738	0.657	0.673	0.52
Global	**0.872**	**0.678**	**0.774**	**0.566**
Test on all types of data				
Gothic	0.664	0.405	0.374	0.254
Modern	0.732	**0.535**	0.565	0.419
Global	**0.812**	0.526	**0.596**	**0.427**

Table 4: Results of the layout analysis models on the three test sets.

Figure 38: Comparison of ground truth and prediction with the "Global" model.

proving the effectiveness of the pooling strategy.

V TOWARDS TEI

The use of *SegmOnto* by researchers should allow for easy extraction of areas relevant to their project. Obviously, the rapid and efficient construction of corpora or editions is one of the most important practical cases, but not the only one. Some researchers want to remove the paratext to keep only the text for linguistic surveys, others want to quickly recover the signatures to study the composition of the codices, or still others want to keep only the decorations to constitute an iconographic base. The distribution and the archiving of high quality data before post-processing/cleaning within the humanities research community must also be considered.

While ALTO and PAGE are standard XML formats for exporting OCR results (cf. ex. 1), they fall short for tasks like text mining and digital editing. Therefore it seems crucial to convert files produced by any OCR engine to a more appropriate format for digitial humanists. For the past decades, the TEI has imposed itself as a standard format, as it is widely adopted and offers XML elements specifically tailored for linguistic, historical or literary purposes, making it particularly relevant for scholars working with textual sources [Burnard, 2014].

```
<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.

5.1 Modelling the TEI

5.1.1 Between transcription and edition: pre-editorialisation

Given the OHCO (Ordered Hierarchy of Content Objects, cf. Renear et al. 1996.) structure of TEI and the conception of our controlled vocabulary, we prioritise a physical representation of the text over a fully logical one. This approach makes it possible to manage the inherent diversity of textual genres (e.g. theatre, novel, accounting records, etc.): if these genres respond to (very) different logics in terms of structure, they are all contained in the same object, the codex.

More than a documentary edition [Pierazzo, 2014], the final output of our pipleine is a "ready-to-use" transcription. The result is a prestructured document, including elements like <lb> for lines, <fw> for paratextual information (running title, page numbers, etc.) or <ab> for text blocks, mimicking the structure of the information on the page. Researchers can then refine the encoding, automatically or not, using elements that are both more precise and/or more suited to their document (<p>, <header>, <persName>, <app>, etc.). This approach underpins what we define as a "pre-editorialisation" of the text [Pinche et al., 2022]. Since 2022, similar concepts have emerged, such as "proto-editions", which aim to address the "problem of mass" and provide a "reliable and consistent representation of the content of a document" [Vogeler, 2022, 2023].

From this perspective, the final file has to meet two different needs: to reverse the pipeline to recreate the data necessary for new analyses in computer vision, but also to distribute the data for further analyses – may they be distant or close. Therefore, the conversion step to TEI does not only produce a file in a new format, but a master file containing all the possible information: (i) the metadata on the document, (ii) the ALTO or PAGE data mapped into TEI, and (iii) a re-structuration of the textual content, based on the information captured during the segmentation phase. The <teiHeader> contains all the metadata, added manually or retrieved automatically from online APIs. The <sourceDoc> stores all of the data produced by the OCR process and maintain a link between the predicted text and its place within the digitised image. The <body> presents the text in a structured hierarchy suitable for exploration or publication phases, via standard tools (TXM, TEI-Publisher) or *ad hoc* scripts.

5.1.2 Digression on "automated" editions

The rapid acceleration of work and the rapid appearance of newer tools that are always more efficient should not mislead us from a philological point of view. 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 OCR [are] not considered in the scholarly editing literature" [Mühlberger and Mansutti, 2022], insisting on the role of the editor training and documenting the OCR 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].

If OCR 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.

5.2 <sourceDoc>

```
<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>.

Such a richly annotated and rigidly structured <sourceDoc> (cf. ex. 2) serves two important purposes. First, because it contains all the data that OCR models export into ALTO-XML files (cf. tab. 5), the <sourceDoc> can be used to reconstruct an ALTO-XML file. Such reverse-engineered ALTO-XML files can become (re)training data. Second, strictly structured data in the <sourceDoc> can be converted into alternative formats such as json and RDF, or any additional export format.

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

5.2.1 Basic mapping

As evidenced in tab. 5, mapping data from ALTO to TEI requires some basic manipulation. For instance, both an element and its attributes are not necessarily kept together in the transformation from ALTO to TEI (e.g. 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).

Another complication is that multiple ALTO files need to become one TEI file. OCR 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 information from the output files 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.

5.2.2 Zone and line Types

As was suggested in Ex. 2, every <zone> in a TEI publication, whether it represents a <TextBlock> or a <TextLine> in ALTO, has a @type as well as a @subtype and a @n. These attributes, deliberately generic to provide a simple encoding which can subsequently be made more complex, are parsed from the decoded *SegmOnto* type/subtype/number, with which the segmentation model tagged that zone of the image. In the case of Ex. 1, the model tagged the block of text as a MainZone to which it assigned the alphanumeric reference code 3246. In this case, there was no subtype in the tag name MainZone. However, when a layout analysis model trained on a *SegmOnto* vocabulary identifies a zone 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.

5.3 <body>

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

Table 6: *SegmOnto* zones 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-editorialised text via another mapping between ALTO and TEI (cf. tab. 6). Whereas in the <sourceDoc> every page was completely contained within an element <surface>, the pre-editorialised text in the <body> merely interrupts a continuous stream of text with the element <pb> (page beginning). This empty element carries the attribute @facs and points to the xml:id of the corresponding page in the <sourceDoc> (cf. ex. 3).

VI CONCLUSION AND FURTHER WORK

The use of the *SegmOnto* controlled vocabulary on data from several centuries, some manuscripts, others printed, did not pose a problem for its users, who did not encounter significant problems

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

Example 3: Example of a pre-editorialised <body>: Robert Garnier, *Tragédies*, Paris: Robert Estienne, 1582.

in their annotation campaigns. The ground truth they produced made it possible to train an efficient model, which in return allowed an undeniable acceleration of work in the production of better quality data. Without being perfect, the idea of a shared controlled vocabulary seems to be a success for its first conception phase [Janès et al., 2021, Solfrini et al., 2024].

However, many points still need to be improved, and we have identified three issues that need to be addressed urgently. First, significant work on subtypes, whose impact of the accuracy of the zone classification is not fully understood, must be carried out, to improve the efficiency of the taxonomy, but also to better match to the needs of researchers, particularly for the TEI conversion phase. Mixing data from various documents from different periods or genres, to increase the efficiency, also has to be evaluated, in order to create more efficient models, and cope with document types that we have not yet encountered (magazines, newspapers, etc.). Second, conversion scripts must also be reviewed to be faster, the creation of the <SourceDoc> being particularly time-consuming, which is a significant problem when the number of files to be processed is in the tens or even hundreds of thousands. Third, with our current system, conversion to TEI is done page after page, without taking into account the previous one: in order to recreate the continuity of the text (a paragraph starting on one page and ending on the next), significant thought must be carried out. Fortunately, some of these problems are currently being resolved by the LaDaS project [Clérice et al., 2024], which has taken up our work to improve the overall system, both technically and conceptually.

DATA

Our SegmOnto model, the "Capricciosa" version, is available online: https://doi.org/10.5281/zenodo.10972956.
All the scripts and a prototype corpus is available on the *Gallic(orpor)a* GitHub repositories

(https://github.com/Gallicorpora).
Online documentation, with more examples for each of the types, is available and maintained at https://segmonto.github.io. Researchers should preferably refer to the site rather than to this article in the event of modification of the guidelines.

FUNDING

This paper has been funded by the BnF DataLab grant and the FNS-Spark project No 220833.

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. Sonia Solfrini and Maxime Humeau helped us preparing the data and training the segmentatin model.

References

Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300, 2009. URL https://doi.org/10.1109/ICDAR.2009.271.

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.

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating paul meyers hagiographic hypothesis. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii49–ii71, 11 2021. URL https://doi.org/10.1093/llc/fqab033.

Samuele Capobianco, Leonardo Scommegna, and Simone Marinai. Historical handwritten document segmentation by using a weighted loss. In Luca Pancioni, Friedhelm Schwenker, and Edmondo Trentin, editors, *Artificial Neural Networks in Pattern Recognition (ANNPR 2018)*, pages 395–406, Cham, 2018. Springer International Publishing. URL https://doi.org/10.1007/978-3-319-99978-4_31.

Alix Chagué, Thibault Clérice, and Laurent Romary. HTR-United : Mutualisons la vérité de terrain! In *DH Nord 2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France, November 2021. MESHS. URL https://hal.science/hal-03398740.

Hiuyi Cheng, Cheng Jian, Sihang Wu, and Lianwen Jin. SCUT-CAB: A new benchmark dataset of ancient chinese books with complex layouts for document layout analysis. In Utkarsh Porwal, Alicia Fornés, and Faisal Shafait, editors, *Proceedings in the International Conference on Frontiers in Handwriting Recognition (ICFHR 2022)*, pages 436–451, Cham, 2022. Springer International Publishing. URL https://doi.org/10.1007/978-3-031-21648-0_30.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15138–15147, Vancouver, Canada, June 2023. URL https://doi.org/10.1109/CVPR52729.2023.01453.

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. *Journal of Data Mining and Digital Humanities*, Historical Documents and..., December 2023. URL https://dx.doi.org/10.46298/jdmdh.9806.

Thibault Clérice, Alix Chagué, and Hugo Scheithauer. Workshop HTR-United: metadata, quality control and sharing process for HTR training data. In *DH 2023 - Digital Humanities Conference: Collaboration as Opportunity*, Graz, Austria, July 2023. Alliance of Digital Humanities Organizations. URL https://inria.hal.science/hal-04094235.

Thibault Clérice, Juliette Janès, Hugo Scheithauer, Sarah Bénière, Laurent Romary, and Benoît Sagot. Layout Analysis Dataset with SegmOnto. In *DH2024 - Annual conference of the Alliance of Digital Humanities Organizations*, Washington, D.C., United States, August 2024. Alliance of Digital Humanities Organizations. URL https://inria.hal.science/hal-04513725.

Marc Douguet. *La Composition dramatique : La liaison des scènes dans le théâtre français du XVIIe siècle*. Travaux du Grand Siècle. Droz, Geneva, 2022.

Frédéric Duval. Pour des éditions numériques critiques. lexemple des textes français. *Médiévales. Langues, Textes, Histoire*, 73(73):13–29, 2017. 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.

Amir Hazem, Béatrice Daille, Louis Chevalier, Dominique Stutzmann, and Christopher Kermorvant. Hierarchical Text Segmentation for Medieval Manuscripts. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING'2020)*, pages 6240–6251, Barcelona, Spain, December 2020. International Committee on Computational Linguistics. URL https://dx.doi.org/10.18653/v1/2020.coling-main.549.

Maxime Humeau, Simon Gabay, and Ariane Pinche. Segmonto, April 2024. URL https://doi.org/10.5281/zenodo.10972956. Capricciosa version.

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, December 2021. URL https://hal.science/hal-03527287.

Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A Review of Yolo Algorithm Developments. *Procedia Computer Science*, 199:1066–1073, January 2022. URL https://doi.org/10.1016/j.procs.2022.01.135.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics. v. 8.0.0.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. In *Book of Abstracts Digital humanities 2019*, Utrecht, July 2019. Alliance of Digital Humanities Organizations. URL https://doi.org/10.34894/Z9G2EX.

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

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, Sydney, Australia, September 2019. URL https://doi.org/10.1109/ICDARW.2019.10032.

Joonho Lee, Hideaki Hayashi, Wataru Ohyama, and Seiichi Uchida. Page segmentation using a convolutional neural network with trainable co-occurrence features. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1023–1028, Sydney, Australia, September 2019. URL https://doi.org/10.1109/ICDAR.2019.00167.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain, December 2020. International Committee on Computational Linguistics. URL https://doi.org/10.18653/v1/2020.coling-main.82.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. In Anastassia Loukina, Michelle Morales, and Rohit Kumar, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/N19-2005.

Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document structure analysis algorithms: a literature survey. In Tapas Kanungo, Elisa H. Barney Smith, Jianying Hu, and Paul B. Kantor, editors, *Document Recognition and Retrieval X*, volume 5010, pages 197 – 207, Santa Clara, USA, January 2003. International Society for Optics and Photonics, SPIE. URL https://doi.org/10.1117/12.476326.

Denis Muzerelle. *Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits*. Rubricae, 1. Institut de recherche et d'histoire des textes, Paris, 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, October 2022. URL https://doi.org/10.17617/3.PEJU8B.

Sven Najem-Meyer and Matteo Romanello. Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches. In *Proceedings of the Computational Humanities Research Conference 2022*, pages 36–54, Antwerp, Belgium, December 2022. CEUR, WS. URL https://doi.org/10.48550/arXiv.2212.13924.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, volume 28, pages 3743–3751, Washington, D.C., USA, August 2022. Association for Computing Machinery. URL https://doi.org/10.1145/3534678.3539043.

Elena Pierazzo. Digital documentary editions and the others. *Scholarly editing*, 35, 2014. URL https://www.scholarlyediting.org/2014/essays/essay.pierazzo.html.

Ariane Pinche, Kelly Christensen, and Simon Gabay. Between automatic and manual encoding. In *TEI 2022 conference : Text as data*, Newcastle, United Kingdom, September 2022. URL https://hal.science/hal-03780302.

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, Istanbul, Turkey, August 2010. IEEE. URL https://doi.org/10.1109/ICPR.2010.72.

Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2439–2447, Seattle, USA, June 2020. URL https://doi.org/10.1109/CVPRW50498.2020.00294.

Allen Renear, Elli Mylonas, and David G. Durand. Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. In Nancy Ide and Susan Hockey, editors, *Research in Humanities Computing*, Oxford, 1996. Oxford University Press. URL https://hdl.handle.net/2142/9407.

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, Göttingen, Germany, June 2017. Association for Computing Machinery. URL https://doi.org/10.1145/3078081.3078097.

Mikhail Ronkin and Kirill Reshetnikov. Real-time YOLO-family comparison for blast quality estimation in the open pit conditions. In *2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pages 254–257, Yekaterinburg, Russian Federation, May 2023. URL https://doi.org/10.1109/USBEREIT58508.2023.10158813.

Zejiang Shen, Kaixuan Zhang, and Melissa Dell. A large dataset of historical japanese documents with complex layouts. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2336–2343, Seattle, USA, June 2020. URL https://doi.org/10.1109/CVPRW50498.2020.00282.

Sonia Solfrini, Simon Gabay, Maxime Humeau, Ariane Pinche, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, Geneviève Gross, and Daniela Solfaroli Camillocci. Océriser les imprimés du XVIe siècle en langue française. In *Humanistica 2024*, Meknès, Morocco, May 2024. Association francophone des humanités numériques. URL https://hal.science/hal-04555002.

Birgit Stehno, Alexander Egger, and Gregor Retti. METAe – automated encoding of digitized texts. *Literary and Linguistic Computing*, 18(1):77–88, 04 2003. URL https://doi.org/10.1093/llc/18.1.77.

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, June 2022. URL https://www.canal-u.tv/133346.

Peter Anthony Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. The eScriptorium VRE for Manuscript Cultures. *Classics@*, 18(1), 2021. URL https://ens.hal.science/hal-03991423.

Melissa Terras, Joe Nockels, and Paul Gooding. On automating editions: The affordances of handwritten text recognition platforms for scholarly editing. *Scholarly Editing*, 41, 2023. URL https://scholarlyediting.org/issues/41/on-automating-editions/.

Georg Vogeler. Edition Protoedition Reproduktion : Der digitale Wandel. *Geschichte in Wissenschaft und Unterricht*, 73, 2022. URL https://elibrary.utb.de/doi/10.5555/gwu-9%2B10-2022_02.

Georg Vogeler. Proto-editions: Historians and the "Something between digital image and digital scholarly edition". In *DH2023 - Annual conference of the Alliance of Digital Humanities Organizations*, Graz, Austria, July 2023. Alliance of Digital Humanities Organizations. URL https://zenodo.org/records/8107922.

Shao Qiang Wang. *Page design: new layout & editorial design*. Promopress, Barcelona, Spain, December 2019.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich doc-

ument understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online, August 2021. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/2021.acl-long.201.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, Sydney, Australia, September 2019. URL https://doi.org/10.1109/ICDAR.2019.00166.