



HAL
open science

A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents

Juliette Janes, Sarah Bénéière, Benoît Sagot, Thibault Clérice

► To cite this version:

Juliette Janes, Sarah Bénéière, Benoît Sagot, Thibault Clérice. A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents. TEI 2025 - New Territories - TEI Conference and Members' Meeting 2025, Sep 2025, Krakow, Poland. <hal-05232691>

HAL Id: hal-05232691

<https://inria.hal.science/hal-05232691v1>

Submitted on 1 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

A TEI-based Layout Annotation System for a Deeper Automatic Encoding of Documents



Juliette Janès, Research & Development Engineer, ALMAnaCH (Inria)

Sarah Bènière, Research & Development Engineer, ALMAnaCH (Inria)

Benoît Sagot, Research Director, ALMAnaCH (Inria)

Thibault Clérice, Researcher, ALMAnaCH (Inria)

Context and Groundwork



Institutional Context



Corpus et Outils pour les Langues de France (COLaF)

- French project
- Led by Inria
- “Contribute to the development of free corpora and tools for French and other languages of France”

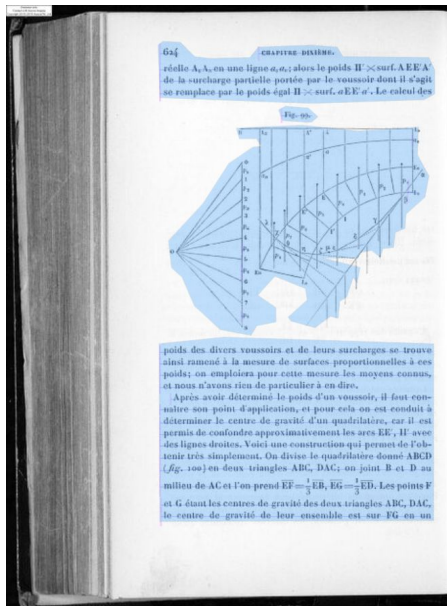
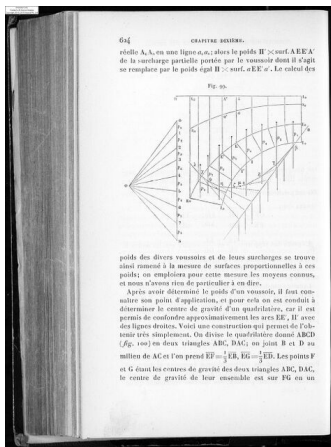
Association with mass textual corpora production projects like Persée



Advancing frontier Research In the arts and humanities (ATRIUM)

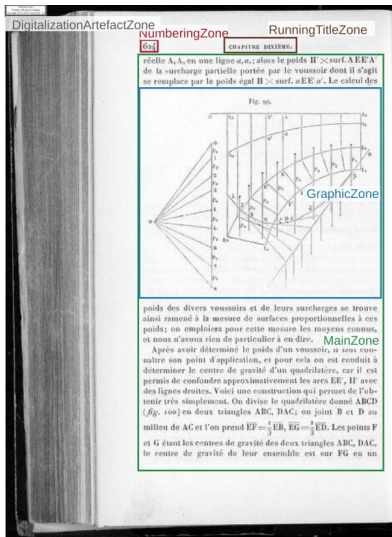
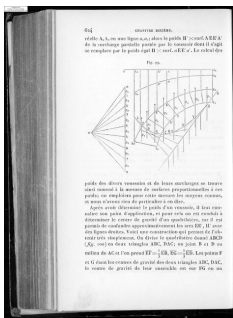
- European project
- Led by DARIAH, partnering with Inria
- “Facilitating access to digital research infrastructures and advancing frontier knowledge in the arts and humanities — across disciplines, languages and media”

A Basic ATR Pipeline



1 644
2 CHATITRE DIXIEME.
3 réelle Ae, en une ligne a, ae; alors le poids II' >XSUTIAEE'A
4 de la surcharge partielle portée par le vousoir dont il s'agit
5 se remplace par le poids égal II — surt. «EB' a». Le calcul des
6 °
7 r
8 29
9
10 F.
11 FIE. 90.
12 poids des divers vousoirs et de leurs surcharges se trouve
13 ainsi ramené à la mesure de surfaces proportionnelles à ces
14 poids; on emploiera pour ceule mesure les moyens connus.
15 et nous n'avons rien de particulier à en dire.
16 Après avoir déterminé le poids d'un vousoir, il faut con-
17 naître son point d'application, et pour cela on est conduit à
18 déterminer le centre de gravité d'un quadrilatère, car il est
19 permis de confondre approximativement les arcs EE', II' avec
20 des lignes droites. Voici une construction qui permet de l'ob-
21 tenir très simplement. On divise le quadrilatère donné ABCD
22 (fig. 10°) en deux uiangles ARC. DAC; on joint B et D au
23 milieu de AC et l'on prend EF — $\frac{1}{3}$ EB, EG=JED. Les points E
24 et G étant les centres de gravité des deux triangles ARC, DAC,
25 le centre de gravité de leur cusemble est sur FG en un

An ATR Pipeline with Document Layout Analysis (DLA)



```

<!--
  DESCRIPTION="block type MarginTextZone"/>
<!--
  ID="BT16576" LABEL="DamageZone" DESCRIPTION="block type DamageZone"/>
<!--
  ID="BT16577" LABEL="DigitalizationArtefactZone"
  DESCRIPTION="block type DigitalizationArtefactZone"/>
<!--
  ID="BT16578" LABEL="DropCapitalZone"
  DESCRIPTION="block type DropCapitalZone"/>
<!--
  ID="BT16579" LABEL="GraphicZone" DESCRIPTION="block type GraphicZone"/>
<!--
  ID="BT16581" LABEL="NumberingZone" DESCRIPTION="block type NumberingZone"/>
<!--
  ID="BT16582" LABEL="QuireMarksZone"
  DESCRIPTION="block type QuireMarksZone"/>
<!--
  ID="BT16583" LABEL="RunningTitleZone"
  DESCRIPTION="block type RunningTitleZone"/>
<!--
  ID="BT16584" LABEL="SealZone" DESCRIPTION="block type SealZone"/>
<!--
  ID="BT16585" LABEL="StampZone" DESCRIPTION="block type StampZone"/>
<!--
  ID="BT25114" LABEL="MainZone" DESCRIPTION="block type MainZone"/>
<!--
  ID="LT6928" LABEL="DefaultLine" DESCRIPTION="line type DefaultLine"/>
<!--
  ID="LT6927" LABEL="HeadingLine" DESCRIPTION="line type HeadingLine"/>
<!--
  ID="LT6928" LABEL="InterlinearLine"
  DESCRIPTION="line type InterlinearLine"/>
<!--
  ID="LT8033" LABEL="default" DESCRIPTION="line type default"/>
</Tags>
<Layout>
<Page WIDTH="470" HEIGHT="640" PHYSICAL_IMG_NUM="3" ID="eSc_duanypage">
<PrintSpace VP0S="160" VP0S="41" WIDTH="470" HEIGHT="640">
<TextBlock VP0S="160" VP0S="41" WIDTH="22" HEIGHT="11"
  ID="eSc_textblock_4a82328" TAGREFS="BT16581">
<Shape>
<Polygon POINTS="160 41 160 54 182 54 182 41"/>
<TextLine ID="eSc_line_7b8eeab2" TAGREFS="LT8033"
  BASELINE="top" VP0S="161" VP0S="38"
  WIDTH="20" HEIGHT="22">
<Shape>
<Polygon
  POINTS="179 38 179 38 172 38 168 38 165 38 161 39 161
  />
</Shape>
<String CONTENT="644" VP0S="161" VP0S="38"
  WIDTH="20" HEIGHT="22" W0="0.97327693580526"/>
</TextLine>
</TextBlock>
<TextBlock VP0S="162" VP0S="60" WIDTH="296" HEIGHT="485"
  ID="eSc_textblock_8561f03b" TAGREFS="BT25114">
<Shape>
<Polygon
  POINTS="162 60 162 169 162 545 458 545 458 60"/>
</Shape>
<TextLine ID="eSc_line_76f06046" TAGREFS="LT8033"
  BASELINE="top" VP0S="161" VP0S="54"
  WIDTH="282" HEIGHT="18">
<Shape>
<Polygon
  POINTS="442 57 439 54 435 54 431 54 428 54 424 54 421
  />
</Shape>
<String
  CONTENT="réelle Ae, en une ligne a, ae: alors le poids II x sur. AEEA</>de la surcharge
  partielle portée par le voussoir dont il s'agit se remplace par le pods égal II x
  sur. aEeA. Le calcul
  </figure>
</figures>
</body>
  
```

```

<!--
  ID="BT16576" LABEL="DamageZone" DESCRIPTION="block type DamageZone"/>
<!--
  ID="BT16577" LABEL="DigitalizationArtefactZone"
  DESCRIPTION="block type DigitalizationArtefactZone"/>
<!--
  ID="BT16578" LABEL="DropCapitalZone"
  DESCRIPTION="block type DropCapitalZone"/>
<!--
  ID="BT16579" LABEL="GraphicZone" DESCRIPTION="block type GraphicZone"/>
<!--
  ID="BT16581" LABEL="NumberingZone" DESCRIPTION="block type NumberingZone"/>
<!--
  ID="BT16582" LABEL="QuireMarksZone"
  DESCRIPTION="block type QuireMarksZone"/>
<!--
  ID="BT16583" LABEL="RunningTitleZone"
  DESCRIPTION="block type RunningTitleZone"/>
<!--
  ID="BT16584" LABEL="SealZone" DESCRIPTION="block type SealZone"/>
<!--
  ID="BT16585" LABEL="StampZone" DESCRIPTION="block type StampZone"/>
<!--
  ID="BT25114" LABEL="MainZone" DESCRIPTION="block type MainZone"/>
<!--
  ID="LT6928" LABEL="DefaultLine" DESCRIPTION="line type DefaultLine"/>
<!--
  ID="LT6927" LABEL="HeadingLine" DESCRIPTION="line type HeadingLine"/>
<!--
  ID="LT6928" LABEL="InterlinearLine"
  DESCRIPTION="line type InterlinearLine"/>
<!--
  ID="LT8033" LABEL="default" DESCRIPTION="line type default"/>
</Tags>
<Layout>
<Page WIDTH="470" HEIGHT="640" PHYSICAL_IMG_NUM="3" ID="eSc_duanypage">
<PrintSpace VP0S="160" VP0S="41" WIDTH="470" HEIGHT="640">
<TextBlock VP0S="160" VP0S="41" WIDTH="22" HEIGHT="11"
  ID="eSc_textblock_4a82328" TAGREFS="BT16581">
<Shape>
<Polygon POINTS="160 41 160 54 182 54 182 41"/>
<TextLine ID="eSc_line_7b8eeab2" TAGREFS="LT8033"
  BASELINE="top" VP0S="161" VP0S="38"
  WIDTH="20" HEIGHT="22">
<Shape>
<Polygon
  POINTS="179 38 179 38 172 38 168 38 165 38 161 39 161
  />
</Shape>
<String CONTENT="644" VP0S="161" VP0S="38"
  WIDTH="20" HEIGHT="22" W0="0.97327693580526"/>
</TextLine>
</TextBlock>
<TextBlock VP0S="162" VP0S="60" WIDTH="296" HEIGHT="485"
  ID="eSc_textblock_8561f03b" TAGREFS="BT25114">
<Shape>
<Polygon
  POINTS="162 60 162 169 162 545 458 545 458 60"/>
</Shape>
<TextLine ID="eSc_line_76f06046" TAGREFS="LT8033"
  BASELINE="top" VP0S="161" VP0S="54"
  WIDTH="282" HEIGHT="18">
<Shape>
<Polygon
  POINTS="442 57 439 54 435 54 431 54 428 54 424 54 421
  />
</Shape>
<String
  CONTENT="réelle Ae, en une ligne a, ae: alors le poids II x sur. AEEA</>de la surcharge
  partielle portée par le voussoir dont il s'agit se remplace par le pods égal II x
  sur. aEeA. Le calcul
  </figure>
</figures>
</body>
  
```

SegmOnto

- Controlled vocabulary for Layout Analysis
- Generic labels for the description of multiple types of document
- Label syntax ⇨ Type : Subtype
- Widely used in DH projects (standardisation)

Table 1. SegmOnto's zones.
Labels syntax should be `Region(:subtype)?`

Region	Suggested subtype
CustomZone	
DamageZone	corrosion; hole; mold ...
DigitazationArtefactZone	ruler; colorTarget...
DropCapitalZone	historiated; flourished...
GraphicZone	illustration; headpiece...
MainZone	column; block
MarginTextZone	upper; lower; note...
MusicZone	stave; neums
NumberingZone	page; folio; item...
QuireMarksZone	signature; catchwords...
RunningTitleZone	
SealZone	
StampZone	postal; curatorial...
TableZone	header; column
TitlePageZone	



Gabay, Simon, Ariane Pinche, Kelly Christensen, and Jean-Baptiste Camps. "SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles." *Journal of Data Mining and Digital Humanities* (2024).
<https://hal.science/hal-04343404v2>.

92 NumberingZone

RD MIAW RunningTitleZone

Lamasba was typical of all African irrigation systems in this particular respect: it only reflected its own environment. The more isolated Tripolitanian wadi settlements, like the village attested in the *Tablettes Albertini*, would have been closer to the parallel of the Jabal Amûr: a more restricted physical environment, scarcer resources, greater isolation and lesser external influence, and hence a tendency to a closed property régime.

The relationships between the large and small proprietors in the system are not, of course, specified on our inscription. But we can make some probable deductions about their social and labour relations. Most of the plots were rather small and were probably worked by peasant family units. Columella does give an estimate of the labour requirements for the process of *ablaquatatio* described above; he remarks that one man could easily cope with 50-80 trees, depending on their size¹. Given family units of production, there is no reason why plots in the median range of 500-600 K could not be worked without recourse to external sources of labour. Proprietors holding up to 4000 K could not have provided all the labour they needed even for this one operation alone, and must have acquired it by means of relations of dependence within the community itself. As stated above, some of the potential for dependence was inherent in the extreme polarization of land itself. But this polarization reflects two different trends. On the one hand, properties are fragmented and dispersed through the irrigation system as a form of insurance. Economic protection is afforded for families by owning some properties on *scalae* close to the source and others located further away. This would tend to minimize 'upstream' - 'downstream' irrigation conflict. The trend towards decentralization of holdings, however, is counteracted by the grouping of properties into familial blocks, a process of consolidation of property by intermarriage. In fact, at Lamasba a dozen families controlled nearly three-quarters (72%) of all the land in the system as recorded.

MainZone

TABLE 10
FAMILIAL GROUPINGS OF PROPERTY AT LAMASBA

No.	Family	Amount of Property	Order (1 = the largest)
1.	Aelii	1 113 K*	7
2.	Aemilii	3 231 K	2
3.	Apulei	797 K	11
4.	Caecilii	854 K	9
5.	Dentilii	640 K	12
6.	Fuficii	1 125 K	6
7.	Germani	4 015 K	1
8.	Italii	1 905 K	4
9.	Manilii	1 650 K	5
10.	Marii	902 K	8
11.	Sextilii	850 K	10
12.	Valerii	1 913 K	3
		Total = 19 005 K	

* Assuming a half interest of Aelius Victor and Valeria Fortunata in Col. IV. 17.

TableZone

MarginTextZone

¹ Cf. Columella, *R.R.*, 11.2.40, *Arbores quoque tempus est ablaquatatis circumfodere, et operire: una opera novellas circumfodit arbores octuaginta, mediores LAV, magnas quinquaginta. Ut 11.2.182, Duo iugera tres operae commode occubant, arboresque quae intererunt ablaquatant.*

LADaS 1.0

- “Layout Analysis Dataset with SegmOnto”
- Both a set of guidelines and a dataset
- Guidelines derived from SegmOnto with TEI-inspired subzones
- Backward compatibility with SegmOnto

Table 2. LADaS subtypes.
Zones in Italics are new to SegmOnto.

Zones Type	LADaS Subtypes
<i>FormZone</i>	
MainZone	Head, P, Lg, Sp, List, Entry, Date, Signature, Maths, Other
MarginTextZone	Notes, ManuscriptAddendum
<i>FigureZone</i>	Head, Figdesc
GraphicZone	Head, Figdesc, TextualContent, Part, Decoration
TableZone	Head
PageTitleZone	Index
StampZone	Sticker



Cl rice, Thibault, Juliette Jan s, Hugo Scheithauer, Sarah B ni re, Laurent Romary, and Beno t Sagot. “Layout Analysis Dataset with SegmOnto.” DH2024 - Reinvention & Responsibility, Washington, D.C., United States, 2024. <https://inria.hal.science/hal-04513725>.

Lamasba was typical of all African irrigation systems in this particular respect; it only reflected its own environment. The more isolated Tripolitanian wadi settlements, like the village of *Albertini*, would have been closer to the parallel of the Jabal Am r: a more restricted physical environment, scarcer resources, greater isolation and lesser external influence, and hence a tendency to a closed property r gime.

The relationships between the large and small proprietors in the system are not, of course, specified on our inscription. But we can make some probable deductions about their social and labour relations. Most of the plots were rather small and were probably worked by peasant family units. Columella does give an estimate of the labour requirements for the process of *abluqueatio* described above; he remarks that one man could easily cope with 50-80 trees, depending on their size¹. Given family units of production, there is no reason why plots in the median range of 500-600 K could not be worked without recourse to external sources of labour. But proprietors holding up to 4000 K could not have provided all the labour they needed even for this one operation alone, and must have acquired it by means of relations of dependence within the community. But this polarization reflects two different trends. On the one hand, properties are fragmented and dispersed through the irrigation system as a form of insurance. Economic protection is afforded for families by owning some properties on *scalae* close to the source and others located further away. This would tend to minimize ‘upstream’ - ‘downstream’ irrigation conflict. The trend towards decentralization of holdings, however, is counteracted by the grouping of properties into familial blocks, a process of consolidation of property by intermarriage. In fact, at Lamasba a dozen families controlled nearly three-quarters (72%) of all the land in the system as recorded.

TABLE 10
FAMILIAL GROUPINGS OF PROPERTY AT LAMASBA

No.	Family	Amount of Property	Order (1 = the largest)
1.	Aelii	1 113 K*	7
2.	Aemilii	3 231 K	2
3.	Apulei	707 K	11
4.	Caccilii	854 K	9
5.	Dentilii	640 K	6
6.	Fuficii	1 125 K	6
7.	Germanii	4 015 K	1
8.	Iulii	1 905 K	4
9.	Manilii	1 650 K	5
10.	Marii	902 K	8
11.	Sextilii	850 K	10
12.	Valerii	1 913 K	3
		Total = 19 005 K	

*Assuming a half interest of Aelius Victor and Valeria Fortunata in Col. IV. 17.

¹ Cf. Columella, *R. R.*, 11.2.40. *Arbores quoque tempus est abluqueatus circumfodere, et operire: una opera novellus circumfodiet arbores octuaginta, medicos LVV, magnus quinquaginta. Cui 11.2.182. Duo iugera tres operae commode occubunt, arboresque quae intererunt abluqueabunt...*

The LADaS Dataset

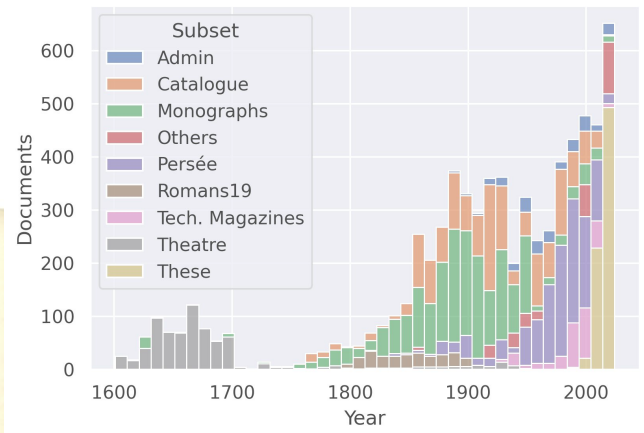
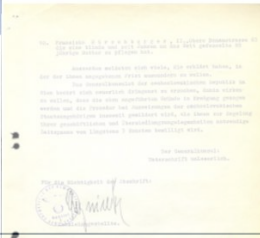
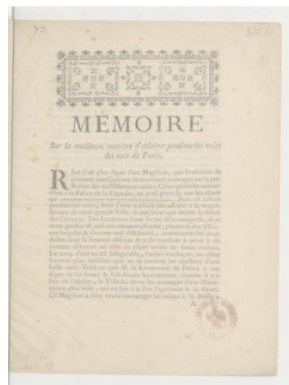


Table 3. LADaS' subsets.
F/NF stands for Fiction/Non-fiction, A/NA for Academic/Non-Academic.

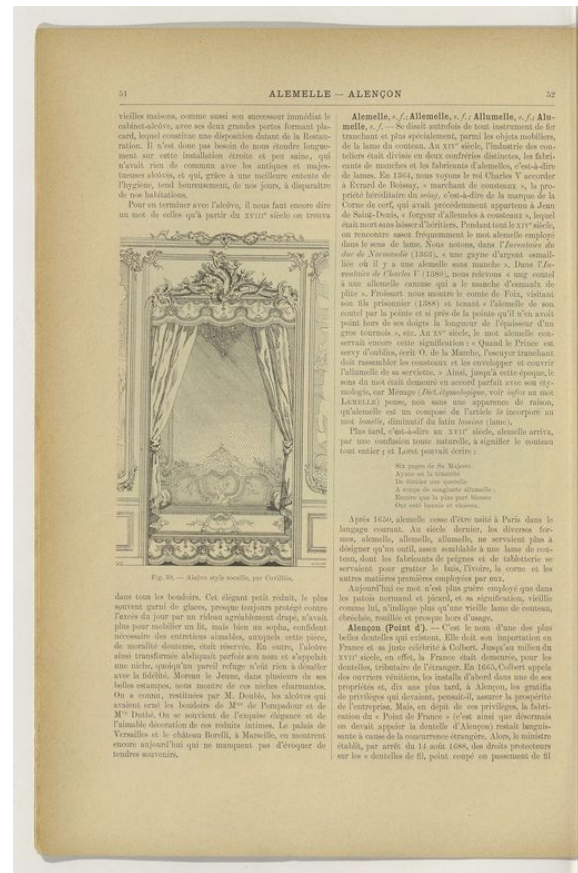
Subset	Provenance	Acquisition	Status	Fiction	Academic	Century	Pages
Admin. Rep.	Various	Harvested	Mixed	NF	NA	19-21	229
Catalogues	INHA-BNF	Donation	Digitised	NF	NA	19-20	1437
Fingers		Donation	Digitised	Mix.	NA	21	125
Magazines Tech		Harvested	Digitised	NF	NA	20-21	522
Monographies	Gallica	Harvested	Digitised	Mix.	NA	17-20	1992
Others		Production	Digitised	NF	NA	21	6
Persée		Harvested	Digitised	NF	A	20	1437
Picard	ARLP	Donation	Mixed	F	NA	21	97
Romans 19	Gallica	Harvested	Digitised	F	NA	19	240
Théâtre	Gallica	Harvested	Digitised	F	NA	17-20	1337
Thèses	Theses.fr	Harvested	Digit. Born	NF	A	21	772
Typewriter	DAHN-EHRI	Donation	Digitised	NF	NA	20	98
All							8287





From LADaS 1.0 to LADaS 2.0

- Confusion in the construction of the labels (graphical and semantic criteria)
- Zones not defined precisely enough ⇨ inconsistent annotations between the annotators
- Leading to confusion of the YOLO models in detecting and labelling zones correctly
- Result: automatic TEI encoding not valid



Building the LADaS Annotation Guidelines



LADaS 2.0 — A Two-Level Annotation System

Maria Georgescu, "Arta epocii brâncovenesti",
The art of Brâncoveanu's epoch
Editura Macarie, Târgoviste, 1996, contains 224 pages with 32 plates.

MainZone-P

Denis CĂPRĂROIU

The issuing of the volume - The Art of Brâncoveanu's epoch - by Mrs. Maria Mioara Georgescu, researcher at "Curtea Domneasca" National Art Museum - from Târgoviste and one from the didactic staff of the "Valahia" University, keen and reputed investigator of the wonderful creations belonging to the Cultural National patrimony, undoubtedly signified an editorial event of the year 1996.

This book, result of some long researches and own remarks upon the diverse aspects of the style in "Brâncoveanu" type epoch, wants to be a welcome synthesis of the "Brâncoveanu" type art, which helps to be made an assembly image of the epoch.

In the first chapter, representing the volume introduction, the author presents the stage of the researches and the purpose of the book, concisely showing how the art of "Brâncoveanu" type epoch was presented in the Romanian historiography, one from the main preoccupation being and the division of the epoch in accordance with the stylistics features. According to these features can be established three distinctive stages:

- the "Cantacuzino - Brâncoveanu" type stage, "classicizable" (the last two decades of the XVII-th century), is the crystallizing period of the "Brâncoveanu" type style, when the traditional fund and the influences received are merged in an indivisible synthesis, which leads to the creation of a style defined as "Brâncoveanu" type style;
- the "Brâncoveanu" type stage "baroque" (the end the XVII-th century and the first two decades of theXVIII-th century), stage of art maximum bloom;
- the post "Brâncoveanu" type stage, which is unfolded after 1720 until the second half of the XVIII-th century, when the legacy of the previous epoch is undertaken by the most important creators both at cult level and popular level.

The second chapter presents the hystorical frame, having four subchapters in which are shown: the situation of Wallachia in the context of the international relationships from the South-East of Europe in the

214

Level-1

- Semantic description of the page layout
- Derived from SegmOnto (relying on *Codicologia*)
- New zones for more recent types of document components (e.g. code, forms)

Level-2

- Developed by LADaS with a systematic comparison with other DLA datasets
- More detailed and graphical description of the page layout (e.g. headings, paragraphs)
- Based on the TEI Guidelines (*labels and definitions*) and Cambridge Online Dictionary (*definitions*)

Label syntax

- Structure: **Level1-Level2**
- Level-2 labels depend on Level-1 labels

Consistent Description of the Labels

- Definition
- [Level-1] MUST/MAY be specified by
- [Level-2] MUST/MAY specify
- Special case(s)
- Potential post-processing
- Suggested TEI mapping
- Example(s)

1.1.1 Main Body of Text (MainZone)

Definition: The MainZone is the main area containing the text. It excludes any paratext and can either be a single block or multiple columns. (Fig. 2, 3)

MUST be specified by: N/A

MAY be specified by: -Head, -P, -PLabelled, -PStructured, -PQuoted, -PStyled, -Item, -Lg, -Address, -Signed, -Ab, -Continued, -Maths

Special case(s): N/A

Potential post-processing: N/A

Suggested TEI mapping: N/A

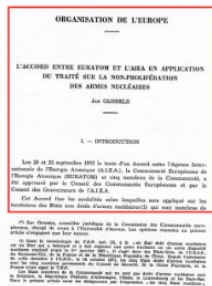


Figure 2. MainZone example

2.1.3 Paragraphs (P)

Definition: A (prose) paragraph is a short part of text, consisting of at least one sentence and beginning on a new line, which can be indented. The classic P paragraph is not numbered. (Fig. 16)

MUST specify: N/A

MAY specify: MainZone, MarginTextZone, TitlePageZone, GraphicZone

Special case(s): According to our annotation criteria, prose poetry is made up of continuous text blocks with few line breaks. It is not as visually distinct from a standard paragraph P as Lg, which consists of text segments with frequent line breaks. Therefore, prose poetry is annotated as prose—usually with the P tag, or PQuoted when indentation applies. Additional post-processing will be necessary to insert TEI <1> tags in documents identified as prose poetry.

Potential post-processing: N/A

Suggested TEI mapping: <p>

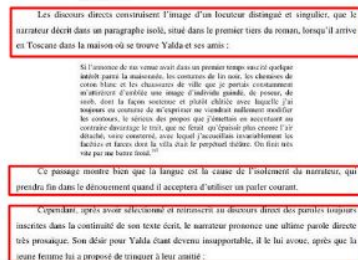


Figure 16. -P example

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring

Les discours directs construisent l'image d'un locuteur distingué et singulier, que le narrateur décrit dans un paragraphe isolé, situé dans le premier tiers du roman, lorsqu'il arrive en Toscane dans la maison où se trouve Yalda et ses amis :

Si l'annonce de ma venue avait dans un premier temps suscité quelque intérêt parmi la maisonnée, les costumes de lin noir, les chemises de coton blanc et les chaussures de ville que je portais constamment m'attiraient d'emblée une image d'individu guindé, de poseur, de snob, dont la façon soutenue et plutôt châtiée avec laquelle j'ai toujours eu coutume de m'exprimer ne viendrait nullement modifier les contours, le sérieux des propos que j'émettais en accentuant au contraire davantage le trait, que ne ferait qu'épaissir plus encore l'air détaché, voire consterné, avec lequel j'accueillis invariablement les facettes et farces dont la villa était le perpétuel théâtre. On finit très vite par me battre froid.¹⁸⁷

Ce passage montre bien que la langue est la cause de l'isolement du narrateur, qui prendra fin dans le dénouement quand il acceptera d'utiliser un parler courant.

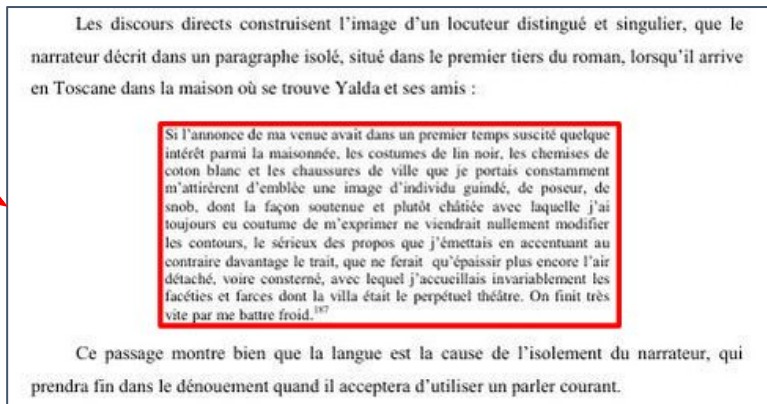
Cependant, après avoir sélectionné et retranscrit au discours direct des paroles toujours inscrites dans la continuité de son texte écrit, le narrateur prononce une ultime parole directe très prosaïque. Son désir pour Yalda étant devenu insupportable, il le lui avoue, après que la jeune femme lui a proposé de trinquer à leur amitié :

-P

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring



Les discours directs construisent l'image d'un locuteur distingué et singulier, que le narrateur décrit dans un paragraphe isolé, situé dans le premier tiers du roman, lorsqu'il arrive en Toscane dans la maison où se trouve Yalda et ses amis :

Si l'annonce de ma venue avait dans un premier temps suscité quelque intérêt parmi la maisonnée, les costumes de lin noir, les chemises de coton blanc et les chaussures de ville que je portais constamment m'attirèrent d'emblée une image d'individu guindé, de poseur, de snob, dont la façon soutenue et plutôt châtiée avec laquelle j'ai toujours eu coutume de m'exprimer ne viendrait nullement modifier les contours, le sérieux des propos que j'émettais en accentuant au contraire davantage le trait, que ne ferait qu'épaissir plus encore l'air détaché, voire consterné, avec lequel j'accueillais invariablement les facéties et farces dont la villa était le perpétuel théâtre. On finit très vite par me battre froid.¹⁹⁷

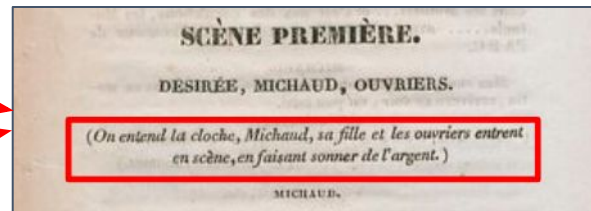
Ce passage montre bien que la langue est la cause de l'isolement du narrateur, qui prendra fin dans le dénouement quand il acceptera d'utiliser un parler courant.

-PQuoted

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring



-PStyled

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring

BOSSUET, Jacques-Bénigne, *Sermon du mauvais riche (Carême du Louvre, 1662)*, 1662, 219.

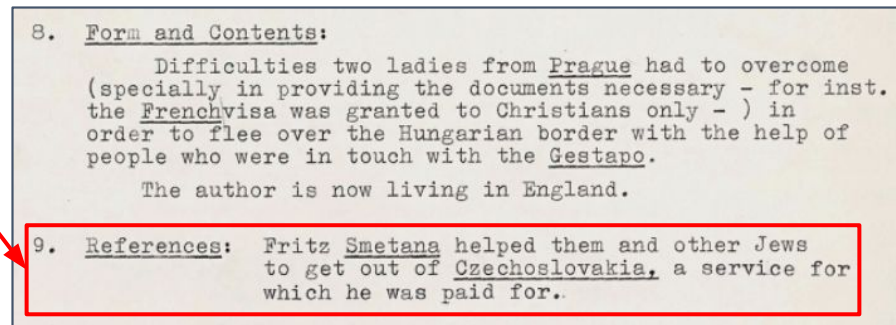
(82) *Mais disons les choses par ordre. **Premièrement**, chrétiens, c' est une fausse imagination des âmes simples et ignorantes, qui n' ont pas expérimenté la fortune, que la possession des*

-PStructured

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring



-PLabelled

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - **Typography (e.g. bold, italics, letter case)**
 - Punctuation marks
 - Entry-like internal structuring

Numbered
Starts with an uppercase letter
Ends with a full stop

Numbered
Can start with a lowercase letter
Semicolons and full stops
Consecutive elements

DRAFT AGREEMENT BETWEEN THE UNITED NATIONS AND THE INTERNATIONAL MONETARY FUND

Article I
GENERAL

1. This agreement, which is entered into by the United Nations pursuant to the provisions of Article 63 of its Charter, and by the International Monetary Fund (hereinafter called the Fund) pursuant to the provisions of article X of its Articles of Agreement, is intended to define the terms on which the United Nations and the Fund shall be brought into relationship.

2. The Fund is a specialized agency established by agreement among its member Governments and having wide international responsibilities, as defined in its Articles of Agreement, in economic and related fields within the meaning of Article 57 of the Charter of the United Nations. By reason of the nature of its international responsibilities and the terms of its Articles of Agreement, the Fund is, and is required to function as, an independent international organization.

-PLabelled

Mais Antoine Prost met en garde, et insiste sur les aberrations et les dangers qui menacent « les nouveaux visages de l'histoire. Nos contemporains invoquent à tout propos un *devoir de mémoire* qui peut passer pour un triomphe de l'histoire », trompant les historiens qui en « retirent parfois le sentiment flatter d'une plus grande utilité sociale » (Prost, 2000 : 4). Quatre raisons sont avancées pour montrer combien il est erroné d'inscrire l'historien dans cette dynamique mémorielle :

1. L'historien ne se contente pas d'accumuler des faits, des événements : il explique et raconte cette histoire et il travaille.
2. L'historien a conscience qu'aucun événement, plus qu'un autre, mérite au détriment de l'histoire d'être sauvé : faire de l'histoire, faire œuvre d'histoire, c'est organiser un passé en vue d'en rendre perceptible la cohérence.
3. L'historien ne peut se permettre d'être sous cette emprise affective terrain du devoir de mémoire. Il est un homme de science, du côté de la connaissance et du savoir. Il y a « mise à distance, rationalisation, volonté de comprendre et d'expliquer. Ce qui est toujours incompatible avec la mémoire » (Prost, 2000 : 5-6).
4. L'historien a en charge un devoir de mémoire, qui repose sur une affirmation identitaire qui « vise un événement comme fondateur par un groupe. Par là, il exclut potentiellement ceux qu'il ne concerne pas directement » (Prost, 2000 : 7).

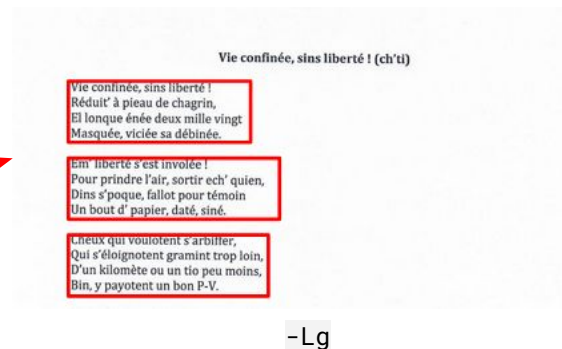
Finalement, l'histoire, loin de l'affectif et du communautarisme, « ne consiste pas à cultiver le souvenir d'un passé lourd de ressentiment ou d'identités (...) : elle est effort pour comprendre ce qui s'est passé et pourquoi cela s'est

-Item

Methodology for Constructing the LADaS Level-2

Use Case: Paragraph-like Elements from the Main Body of Text

- Choosing which zones and labels to create:
 - Analysis of the whole LADaS 1.0 corpus and labels
 - Matching LADaS Level-2 to TEI elements (e.g. <p> and -P)
- Set of differentiating criteria:
 - Position of the text on the page
 - Indentation
 - Line break
 - Typography (e.g. bold, italics, letter case)
 - Punctuation marks
 - Entry-like internal structuring



The Continued Case

- Non TEI element for interrupted textual content
- Annotation for cut text in specific contexts:
 - At the top of a page or column
 - Interrupted by another zone
 - ⇒ Impossible to determine the exact Level-2 element to which it belongs
- Post-processing:
 - Associate each text portion with its preceding zone
 - ⇒ Reconstruct the document in TEI as accurately as possible

l'époque. Le caricaturiste dépeignit avec élégance et audace les modifications du contexte urbain de Rio de Janeiro au début du XX^e siècle et le rapport de ses habitants à la modernité, les évolutions sociales et de mœurs, les innovations technologiques et la vie politique nationale.

Autre dessinateur extrêmement prolifique, également peintre, écrivain et chroniqueur, Benedito Bastos Barreto, dit Belmonte, publia inlassablement au cours de la première moitié du XX^e siècle. Il créa dans les années 1920 le personnage de Juca Pato, archétype masculin de



LADaS 2.0 Labels

Table 4. LADaS Level-1.

Classification	Labels	Short Definition
Primary Text Zones	MainZone	Main area containing text
	MarginTextZone	Any text in the margins
	TitlePageZone	Entire page with document bibliographic informations
Media	GraphicZone	Non textual content
	FigureZone	Various different textual elements (code snippets...)
	TableZone	Tables
	FormZone	Forms
	MusicZone	Sheet Music
Material Zones	DigitisationArtefactZone	Elements external to the document
	NumberingZone	Numberings
	RunningTitleZone	Running Titles
	StampZone	Stamps
	QuireMarksZone	Quire signature, catchwords...

Table 5. LADaS Level-2.

Classification	Labels
Text	Head
	HeadStructured
	P
	PLabelled
	PStructured
	PQuoted
	PStyled
	Item
	Lg
	Dateline
	Address
	Signed
	Ab
Continued	
Media and Other	Part
	Decoration
	Maths
	ManuscriptAddendum
	Field
	DropCapital

Examples of Fully Annotated Pages

1 - MainZone-Head

ORGANISATION DE L'EUROPE

L'ACCORD ENTRE EURATOM ET L'AIEA EN APPLICATION DU TRAITÉ SUR LA NON-PROLIFÉRATION DES ARMES NUCLÉAIRES

Jan GJJSSELS

I — INTRODUCTION

Les 20 et 22 septembre 1972 le texte d'un Accord entre l'Agence Internationale de l'Energie Atomique (A.I.E.A.), la Communauté Européenne de l'Energie Atomique (EURATOM) et cinq membres de la C.E.E. a été approuvé par le Conseil des Communautés Européennes et par le Conseil des Gouverneurs de l'A.I.E.A.

Cet Accord fixe les modalités selon lesquelles sera appliqué l'article 10 de l'Accord sur le statut de l'A.E.C.S. relatif aux territoires des Etats non dotés d'armes nucléaires (1) qui sont membres de

(*) Jan GJJSSELS, conseiller juridique de la Commission des Communautés Européennes, chargé de cours à l'Université d'Anvers. L. 7 - MargInTextZone-F-Labelled article n'engage que leur auteur.

(1) Dans la terminologie du T.N.P. (art. IX, § 3) « un Etat doté d'armes nucléaires est un Etat qui a fabriqué et a fait exploser une arme nucléaire ou un autre dispositif nucléaire explosif avant le 1^{er} janvier 1967 ». Il s'agit de : MargInTextZone-F-Labelled de cette dernière à l'O.N.U. le 26 octobre 1971, les cinq Etats dotés d'armes nucléaires sont les cinq membres permanents du Conseil de Sécurité, ni la Chine Populaire, ni la France n'ont signé le T.N.P.

Les Etats membres de la Communauté qui ne sont pas dotés d'armes nucléaires sont la Belgique, la République Fédérale d'Allemagne, l'Italie, le Luxembourg et les Pays-Bas. Dans le présent article nous les désignons par l'expression « les cinq Etats membres ».

NumberingZone RunningTitleZone

81 CHAPITRE 2. L'INPAINTING : ÉTAT DE L'ART ET LIMITES

GraphicZone

FIGURE 2.10 — Complétion par une méthode basée patch locale en fonction de la taille du patch. De gauche à droite et de bas en haut, nous avons des μ GraphicZone-Head 10 pixels.

cohérence de la complétion. Elles requièrent aussi plusieurs d'itérations pour atteindre une convergence et ainsi raffiner la partie masquée. MainZone-Continued

SUN *et al.* (Sun+06) proposent de séparer la propagation de la texture de la propagation de la structure. Pour cela, ils inventent d'abord l'utilisateur à indiquer les structures incomplètes en les traçant. Ensuite, chaque pixel appartenant au tracé et qui ne reçoit un label correspondant à un pixel omni du tracé; l'attribution du label est faite par une énergie à minimiser. La texture est ensuite complétée par une approche gloutonne (CPM) de la sous-section précédente.

WEXLER *et al.* (Wex07) proposent une complétion en calculant, pour un pixel du masque, une moyenne pondérée des pixels provenant de plusieurs patches qui se superposent. Ils définissent également une fonction de cohérence globale qui μ MainZone-F chaque pixel, la similarité entre les patches qui le recouvrent et le patch de la zone voisine :

$$\text{Cohérence} = \frac{1}{\text{pet}} \int_{\text{pet}} \psi(\Psi_p, \Psi_q) \text{MainZone-Maths}$$

où sim_p est une fonction de similarité coup MainZone-Continued

$$\text{sim}_p(\Psi_p, \Psi_q) = \exp\left(-\frac{\text{sim}(\Psi_p, \Psi_q)}{\sigma}\right) \text{MainZone-Maths}$$

où σ est l'écart-type du noyau gaussien et sim une fonction de similarité. L'intérêt de cette méthode multi-résolutions est d'utiliser une image construite à une résolution donnée comme initialisation pour la résolution supérieure. Cette MainZone-Continued l'algorithme PatchMatch est l'approche classique d'imaging μ MainZone-F notamment utilisée dans le procédé Content-aware Fill de Photoshop. Elle sera notamment détaillée dans la section 2.4.

ROYALLE

SCENE TROISIEME

CLEANDRE

Que ie dois bien faire pitié,
De souffrir les rigueurs d'un sort si tyrannique!
J'aime Aidor, j'aime Angelique,
Mais l'Amour cede à l'amitié,
Et l'on n'a jamais vu sous les loix d'une Belle
D'Amant si malheureux, ny d'amy si fidelle.

Ma bouche ignore mes desirs,
Et de peur de se voir trahy par imprudence
Mon cœur n'a point de confiance
Auc mes yeux, ny mes soupirs,
Mes vœux pour fa loauté sont muets, et ma flame
Non plus que son objet ne fort point de mon ame.

Key Takeaways



Conclusion and Perspectives

- Currently in the process of reannotating all the LADaS dataset
- Enhancing and automating LADaS2TEI
- Investigating post-processing possibilities (e.g. making the distinction between `<bibl>` and `<p type="structured">`)

17. *P. de Solerti, clerc. XIII^e siècle. Sceau rond de 27 mm en bronze. Un lion et un dragon à corps d'oiseau se combattant. (N^o 85.)*

18. *Prieur des Carmes d'Aix. Fin du XIII^e siècle. Sceau en navette de 47 et 28 mm en bronze. Le Christ en croix entre la Vierge et saint Jean. (N^o 86.)*

Useful Links



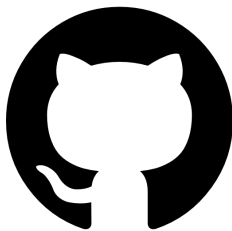
Access the LADaS 1.0
dataset



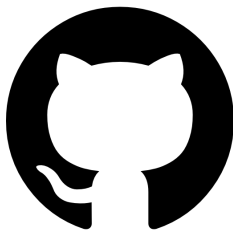
Read the LADaS
Annotation Guidelines



Read more about the
LADaS 2.0 upgrade



Check out LADaS2TEI



Check out the LADaS
object detection model



Slides and abstract of
this presentation

Thank you for your attention!



Contact: juliette.janes@inria.fr / sarah.beniere@inria.fr