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

Juliette Janes, Sarah Bénière, Benoît Sagot, Thibault Clérice

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inria

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

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

Sarah Bénière, 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)

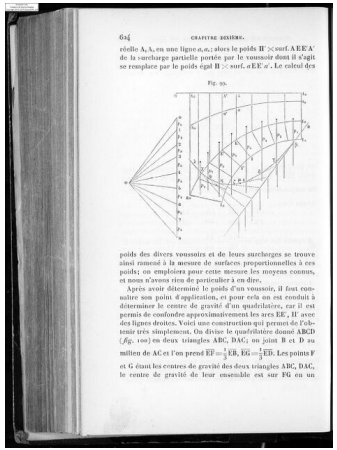
- French project
- Led by Inria
- “Contribute to the development of free corpora and tools for French and other languages of France”

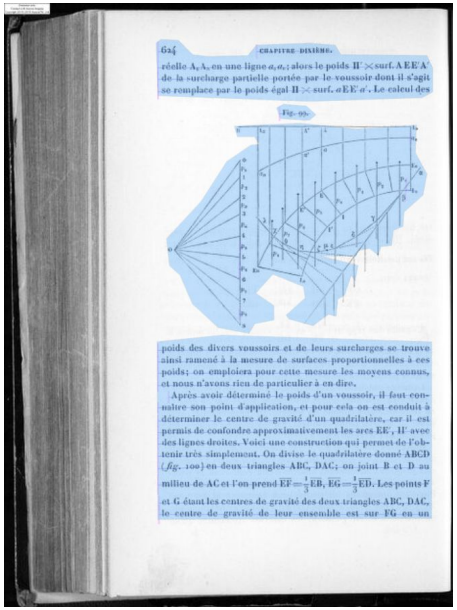
Advancing fronTier Research In the arts and hUmanities (ATRIUM)

- 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”

Association with mass textual corpora production projects like Persée

A Basic ATR Pipeline





644 CHAPITRE DIXIEME.

réelle A, A, en une ligne a, a; alors le poids Π′ × surf. A EE′A′ de la surcharge partielle portée par le voussoir dont il s'agit se remplace par le poids égal Π × surf. a EE′ a′. Le calcul des

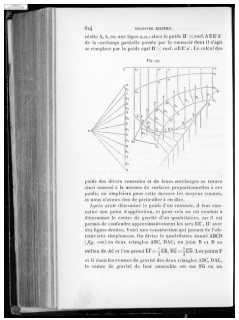
Fig. 90.

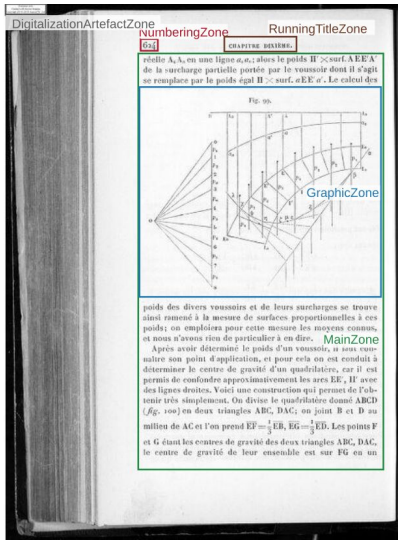
poids des divers voussoirs et de leurs surcharges se trouve ainsi ramené à la mesure de surfaces proportionnelles à ces poids; on emploiera pour cette mesure les moyens connus, et nous n'avons rien de particulier à en dire.

Après avoir déterminé le poids d'un voussoir, il faut connaître son point d'application, et pour cela on est conduit à déterminer le centre de gravité d'un quadrilatère, car il est permis de confondre approximativement les arcs EE′, Π′ avec des lignes droites. Voici une construction qui permet de l'obtenir très simplement. On divise le quadrilatère donné ABCD (*fig.* 100) en deux triangles ABC, DAC; on joint B et D au milieu de AC et l'on prend $\overline{EF} = \frac{1}{3}\overline{EB}$, $\overline{EG} = \frac{1}{3}\overline{ED}$. Les points F et G étant les centres de gravité des deux triangles ABC, DAC, le centre de gravité de leur ensemble est sur FG en un

```
1   644
2   CHATITRE DIXIEME.
3   réelle Ae, en unc ligne a, ae: alors le polds Il' XSuTI.AEE'A
4   de la surcharge partielle portée par le voussoir dont il s'agit
5   se remplace par le poids égal II — surt. «EB' «°. Le calcul des
6   °
7   r
8   29
9
10  F.
11  FiE. 90.
12  poids des divers voussoirs 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 voussoir, il faut con¬
17  naître son point d'application, et pour cela on est condult à
18  déterminer le centre de gravité d'un quadrilatère, car il est
19  permis de confondre approximativement les ares EE, Il' avec
20  des lignes droites. Voici une construction qui permet de l'ob¬
21  tenir très ximplement. On divise le quadrilatère donné ABCD
22  (ig. 10°) en deux uiangles ARC. DAC; on joint B et D au
23  milieu de AC et l'on prend EF — ¿ EB, EG=JED. Les points E
24  et G étant les centres de gravité des deux trlangles 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"/>
    <OtherTag ID="BT16576" LABEL="DamageZone" DESCRIPTION="block type DamageZone"/>
    <OtherTag ID="BT16577" LABEL="DigitizationArtefactZone"
        DESCRIPTION="block type DigitizationArtefactZone"/>
    <OtherTag ID="BT16578" LABEL="DropCapitalZone"
        DESCRIPTION="block type DropCapitalZone"/>
    <OtherTag ID="BT16579" LABEL="GraphicZone" DESCRIPTION="block type GraphicZone"/>
    <OtherTag ID="BT16581" LABEL="NumberingZone" DESCRIPTION="block type NumberingZone"/>
    <OtherTag ID="BT16582" LABEL="QuireMarksZone"
        DESCRIPTION="block type QuireMarksZone"/>
    <OtherTag ID="BT16583" LABEL="RunningTitleZone"
        DESCRIPTION="block type RunningTitleZone"/>
    <OtherTag ID="BT16584" LABEL="SealZone" DESCRIPTION="block type SealZone"/>
    <OtherTag ID="BT16585" LABEL="StampZone" DESCRIPTION="block type StampZone"/>
    <OtherTag ID="BT25114" LABEL="MainZone" DESCRIPTION="block type MainZone"/>
    <OtherTag ID="LT6926" LABEL="DefaultLine" DESCRIPTION="line type DefaultLine"/>
    <OtherTag ID="LT6927" LABEL="HeadingLine" DESCRIPTION="line type HeadingLine"/>
    <OtherTag ID="LT6928" LABEL="InterlinearLine"
        DESCRIPTION="line type InterlinearLine"/>
    <OtherTag ID="LT8033" LABEL="default" DESCRIPTION="line type default"/>
</Tags>

<Layout>
    <Page WIDTH="470" HEIGHT="640" PHYSICAL_IMG_NR="3" ID="eSc_dummypage_">
        <PrintSpace HPOS="0" VPOS="0" WIDTH="470" HEIGHT="640">
            <TextBlock HPOS="160" VPOS="41" WIDTH="222" HEIGHT="13"
                ID="eSc_textblock_4a882328" TAGREFS="BT16581">
                <Shape>
                    <Polygon POINTS="160 41 160 54 182 54 182 41"/>
                </Shape>
                <TextLine ID="eSc_line_7b0ceab2" TAGREFS="LT8033"
                    BASELINE="161 51 181 51" HPOS="161" VPOS="38"
                    WIDTH="20" HEIGHT="22">
                    <Shape>
                        <Polygon
                        POINTS="179 38 175 38 172 38 168 38 165 38 161 39 161
                        "/>
                    </Shape>
                    <String CONTENT="644" HPOS="161" VPOS="38"
                        WIDTH="20" HEIGHT="22" WC="0.973727663398580526"/>
                </TextLine>
            </TextBlock>
            <TextBlock HPOS="162" VPOS="60" WIDTH="296" HEIGHT="485"
                ID="eSc_textblock_6561f03b" TAGREFS="BT25114">
                <Shape>
                    <Polygon
                        POINTS="162 60 162 169 162 545 458 545 458 60"/>
                </Shape>
                <TextLine ID="eSc_line_76f06046" TAGREFS="LT8033"
                    BASELINE="161 68 443 67" HPOS="161" VPOS="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 A, a, en unc ligne a, ae: alors le poid
                        HPOS="161" VPOS="54" WIDTH="282" HEIGHT="18"
```

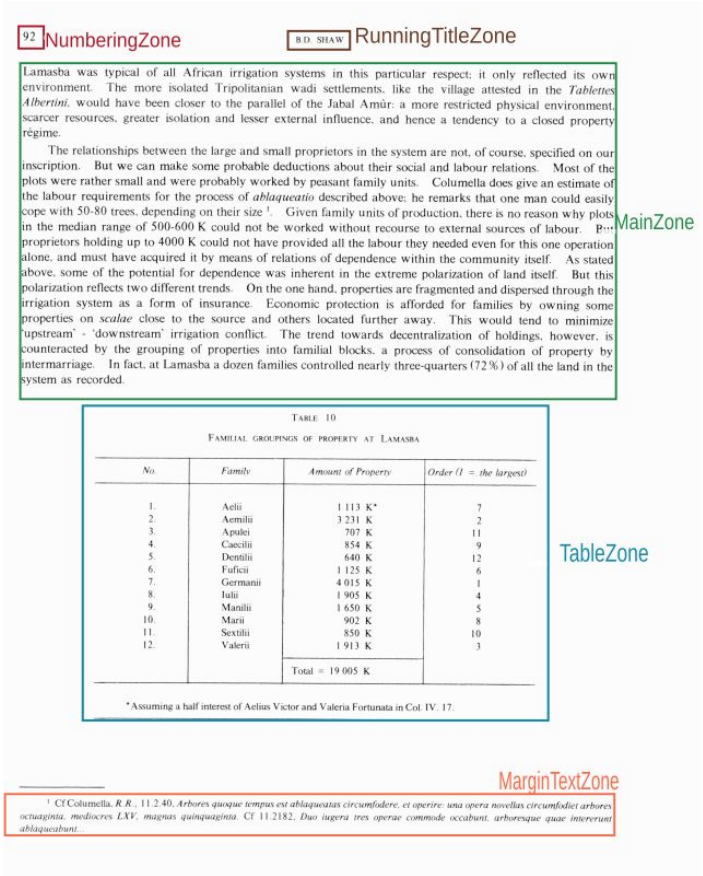
```
    <body>
        <!--...-->
        <fw type="numbering"><lb/>644</fw>
        <fw type="runningTitle"><lb/>CHAPITRE DIXIEME</fw>
        <ab type="mainZone">
            <lb/> réelle A, A en une ligne a, ae: alors le poids II x sur. AEEA<lb/>de la surchage
            partielle portée par le voussoir dont il s'agit<lb/> se remplace par le pods égal II x
            suf. aEEa. Le calcul
            <figure>
                <head><lb/>Fig. 99</head>
            </figure>
            <lb/>poids des divers voussoirs et de leurs surcharges se trouve <lb/>ainsi ramené à
            la mesure de surfaces proportionnelles à ces <lb/>poids; on emploiera pour ceule mesure
            les moyens connus. <lb/>et nous n'avons rien de particulier à en dire. <lb/>Après
            avoir déterminé le poids d'un voussoir, il faut con¬ <lb/>naître son point
            d'application, et pour cela on est condult à <lb/>déterminer le centre de gravité
            d'un quadrilatère, car il est <lb/>permis de confondre approximativement les ares EE,
            Il' avec <lb/>des lignes droites. Voici une construction qui permet de l'ob¬ <lb/>tenir
            très ximplement. On divise le quadrilatère donné ABCD <lb/>(ig. 10°) en deux uiangles
            ARC. DAC; on joint B et D au <lb/>milieu de AC et l'on prend EF — ¿ EB, EG=JED. Les
            points E <lb/>et G étant les centres de gravité des deux trlangles ARC, DAC. <lb/>le
            centre de gravité de leur cusemble est sur FG en un </ab>
        <!--...-->
```

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	



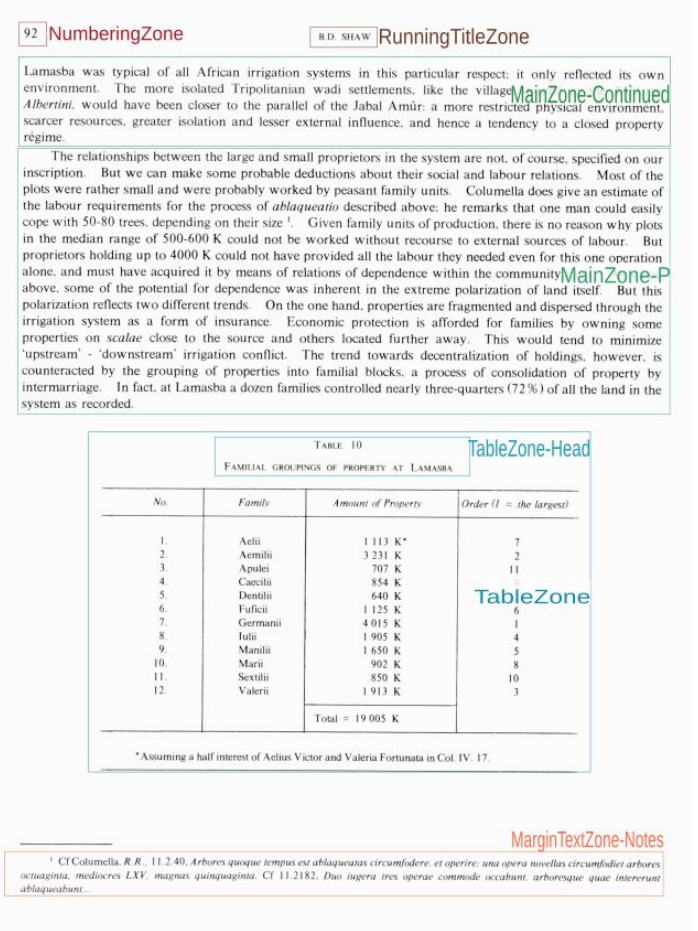
Gabay, Simon, Ariane Pinche, Kelly Christensen, and Jean-Baptiste Camps. "SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles." *Journal of Data Mining and Digital Humanities* (2024). https://hal.science/hal-04343404v2.

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

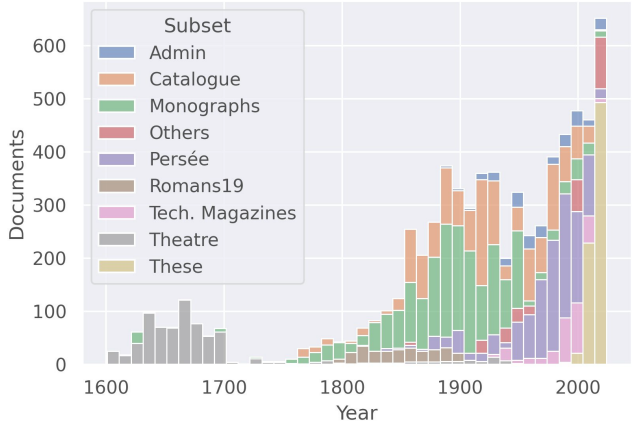
Table 2. LADaS subtypes.
Zones in Italics are new to SegmOnto.

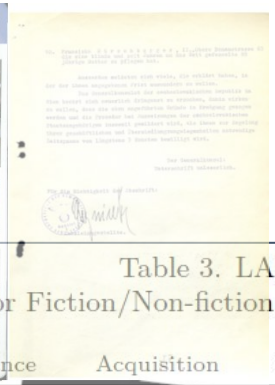
Zones Type	LADaS Subtypes
FormZone	
MainZone	Head, P, Lg, Sp, List, Entry, Date, Signature, Maths, Other
MarginTextZone	Notes, ManuscriptAddendum
FigureZone	Head, Figdesc
GraphicZone	Head, Figdesc, TextualContent, Part, Decoration
TableZone	Head
PageTitleZone	Index
StampZone	Sticker



Clérice, Thibault, Juliette Janès, Hugo Scheithauer, Sarah Bénière, Laurent Romary, and Benoît Sagot. “Layout Analysis Dataset with SegmOnto.” DH2024 - Reinvention & Responsibility, Washington, D.C., United States, 2024. https://inria.hal.science/hal-04513725.

The LADaS Dataset





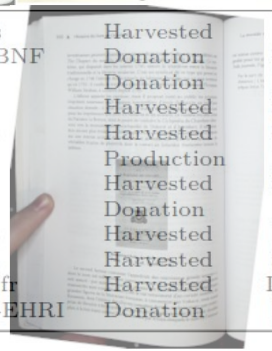
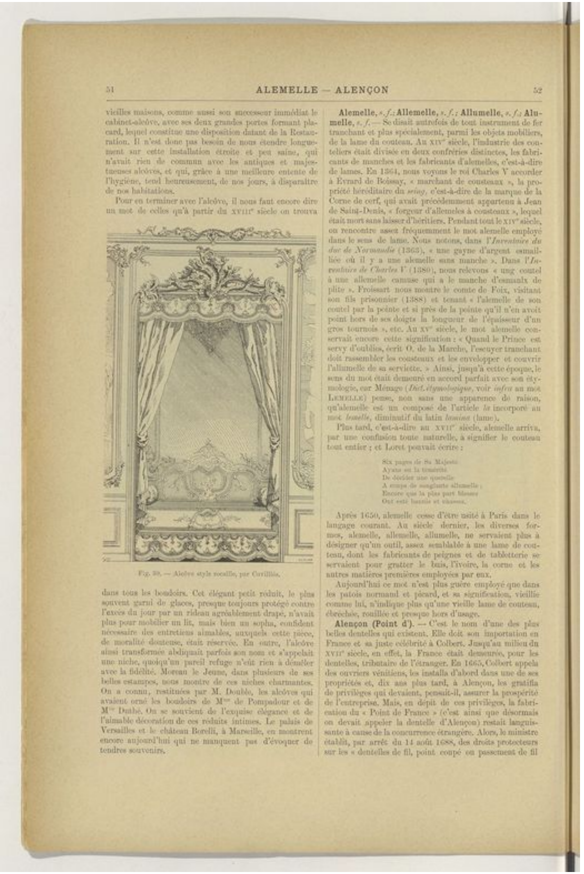


Table 3. LADaS' subsets.
F/NF stands for Fiction/Non-fiction, A/NA for Academic/Non-Academic.

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	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	Dig. Born	NF	A	21	772
Typewriter	DAHN-EHRI	Donation	Digitised	NF	NA	20	98
All							8287

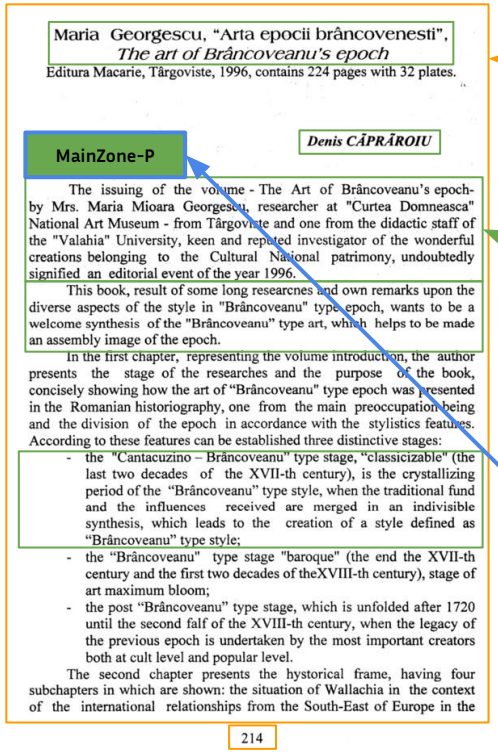
From LADaS 1.0 to LADaS 2.0

- 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


51 ALEMELLE — ALENÇON 52

Building the LADaS Annotation Guidelines

LADaS 2.0 — A Two-Level Annotation System



- **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


ORGANISATION DE L'EUROPE

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

I. — INTRODUCTION

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)

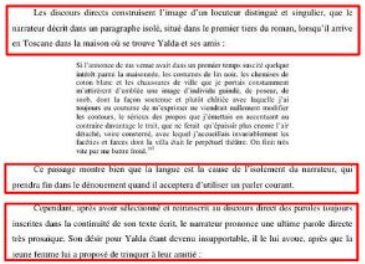
MUST specify: N/A

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 <l> tags in documents identified as prose poetry.

Potential post-processing: N/A

Suggested TEI mapping: <p>


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 :

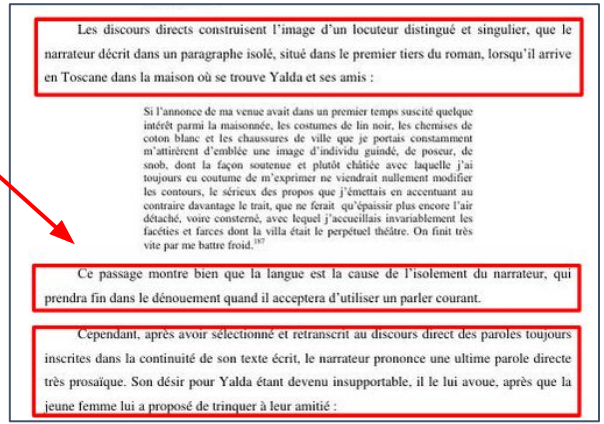
Ce passage montre bien que la langue est la cause de l'isolement du narrateur, qui prendra fin dans le dénouement quand il acceptera d'utiliser un parler courant.

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'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.[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.

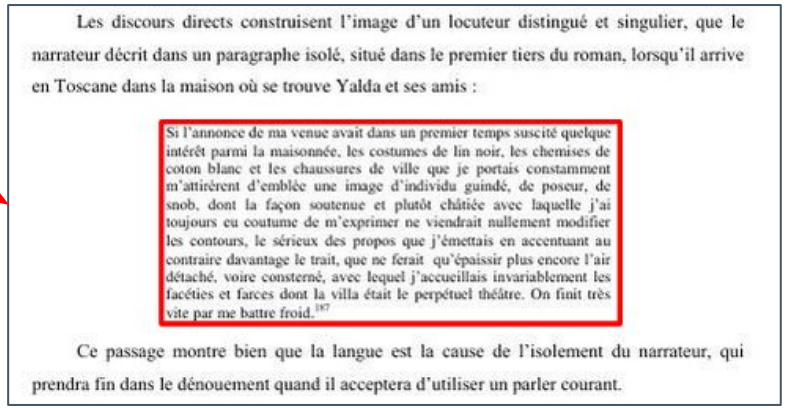
Cependant, après avoir sélectionné et retranscrit au discours direct des paroles toujours inscrites dans la continuité de son texte écrit, le narrateur prononce une ultime parole directe très prosaïque. Son désir pour Yalda étant devenu insupportable, il le lui avoue, après que la jeune femme lui a proposé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

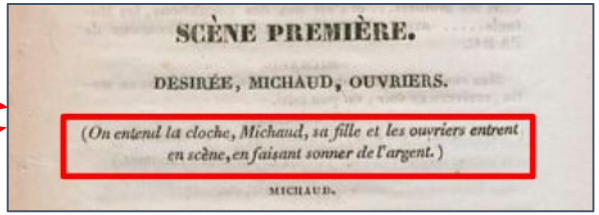
Ce passage montre bien que la langue est la cause de l'isolement du narrateur, qui prendra fin dans le dénouement quand il acceptera d'utiliser un parler courant.

`-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

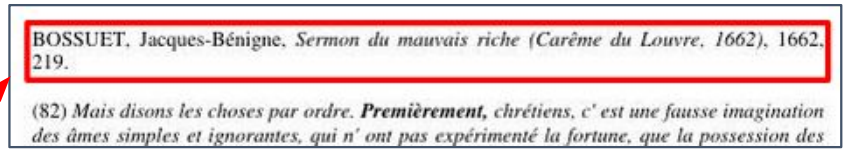


`-PStyled`

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

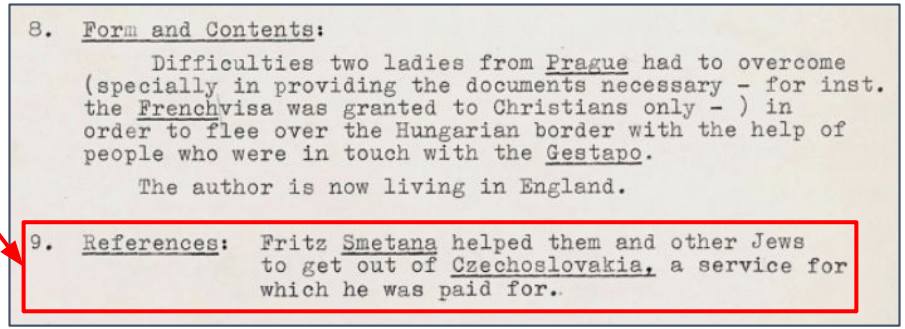


`-PStructured`

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



8. Form and Contents:

Difficulties two ladies from Prague had to overcome (specially in providing the documents necessary - for inst. the Frenchvisa was granted to Christians only -) in order to flee over the Hungarian border with the help of people who were in touch with the Gestapo.

The author is now living in England.

9. References: Fritz Smetana helped them and other Jews to get out of Czechoslovakia, a service for which he was paid for.

`-PLabelled`

Methodology for Constructing the LADaS Level-2

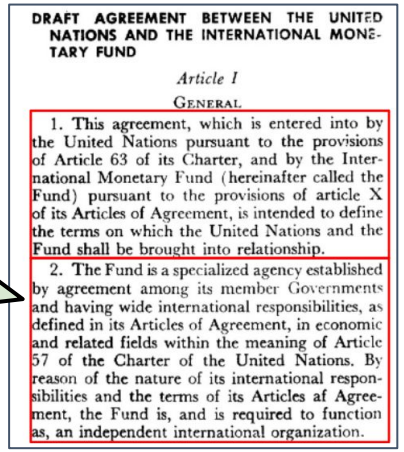
Use Case: Paragraph-like Elements from the Main Body of Text

- 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



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 af Agreement, the Fund is, and is required to function as, an independent international organization.

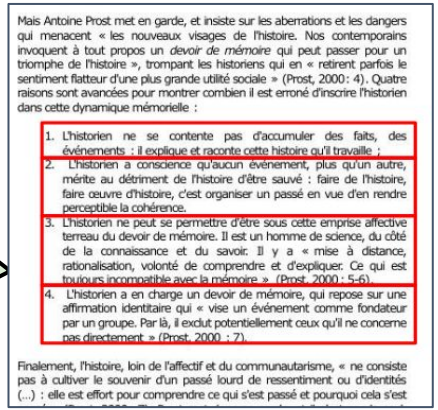
`-PLabelled`

Numbered

Can start with a lowercase letter

Semicolons and full stops

Consecutive elements



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 flatteur 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 qu'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 terreau 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).

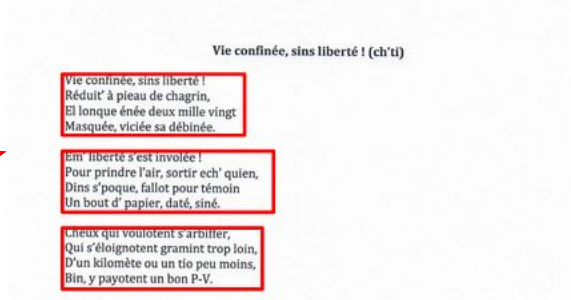
Finalement, l'histoire, loin de l'affectif et du communautarisme, « ne consiste pas à cultiver le souvenir d'un passé lourd de ressentiment ou d'identités (...) : elle est effort pour comprendre ce qui s'est passé et pourquoi cela s'est

`-Item`

Methodology for Constructing the LADaS Level-2

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

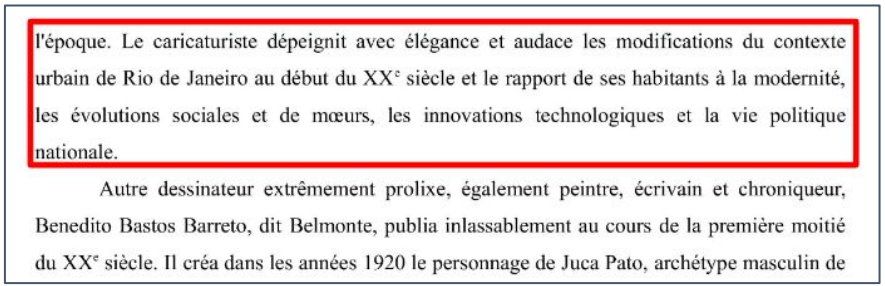
- Choosing which zones and labels to create:
  - Analysis of the whole LADaS 1.0 corpus and labels
  - Matching LADaS Level-2 to TEI elements (e.g. `<p>` and `-P`)
- Set of differentiating criteria:
  - Position of the text on the page
  - Indentation
  - Line break
  - Typography (e.g. bold, italics,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



`-Lg`

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 prolixe, 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

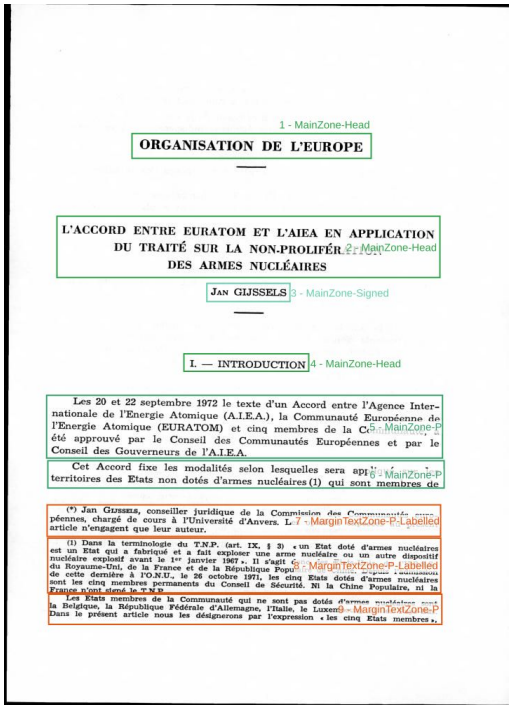
Table 4. LADaS Level-1.

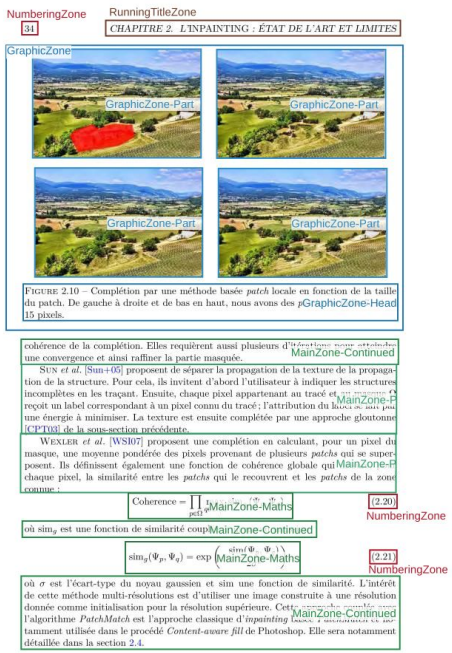
Classification	Labels	Short Definition
Primary Text Zones	MainZone	Main area containing text
	MarginTextZone	Any text in the margins
	TitlePageZone	Entire page with document bibliographic informations
Media	GraphicZone	Non textual content
	FigureZone	Various different textual elements (code snippets...)
	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...

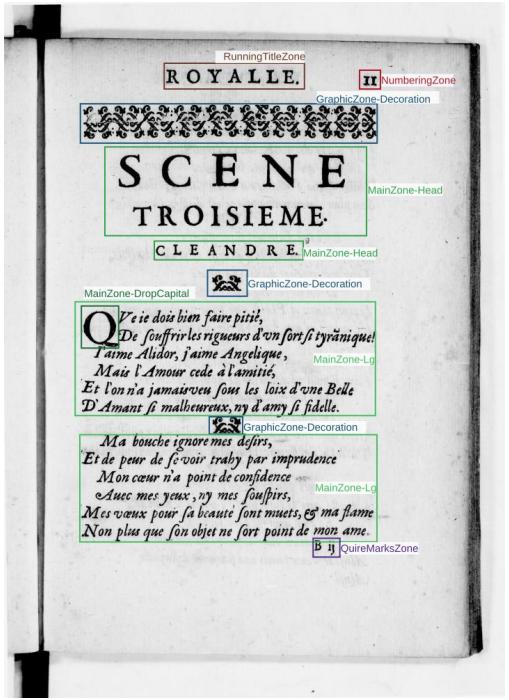
Table 5. LADaS Level-2.

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

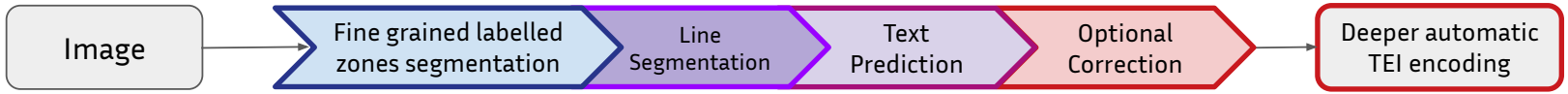


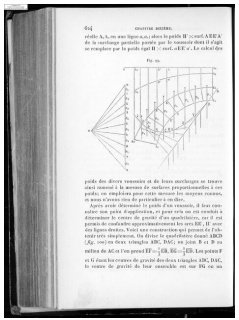


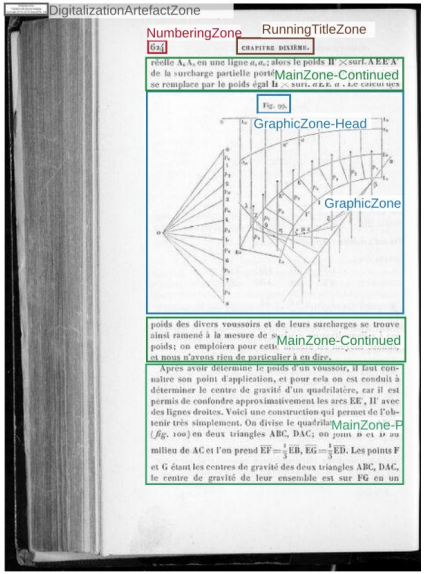


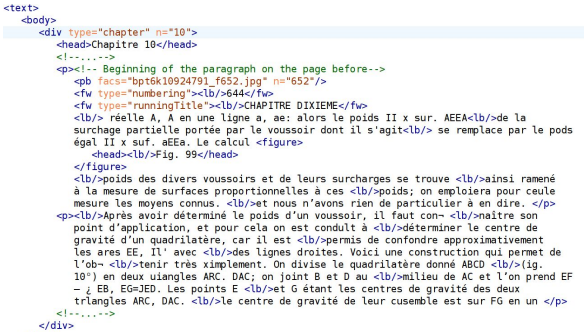
Key Takeaways

ATR Pipeline Using LADaS 2.0



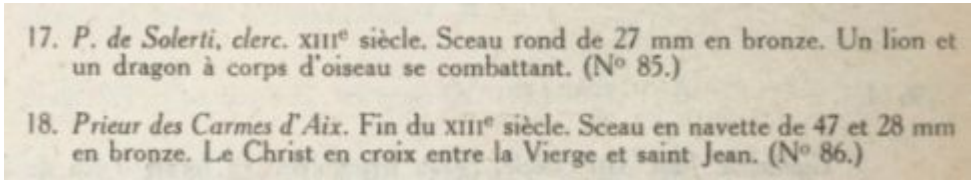






Conclusion and Perspectives

- 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° 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° 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