



Automatic Retro-Structuration of Auction Sales Catalogs Layout and Content

Hugo Scheithauer, PhD Candidate, ALMAnaCH (Inria, Paris) / École Pratique des
Hautes Études (EPHE)

Sarah Bénière, Research & Development Engineer, ALMAnaCH (Inria, Paris)



(DataCat for short)

The DataCatalogue Project





A Research & Development Partnership



{ BnF

- French **Ministry of Cultural Affairs** (*Ministère de la Culture*)
- French **National Library** (*Bibliothèque nationale de France - BnF*)
- French **National Institute for Art History** (*Institut national d'histoire de l'art - INHA*)
- French **National Institute for Research in Digital Sciences and Technology** (*Institut national de recherche en informatique et automatique - Inria*)


Scientific experts,
researchers,
data provider

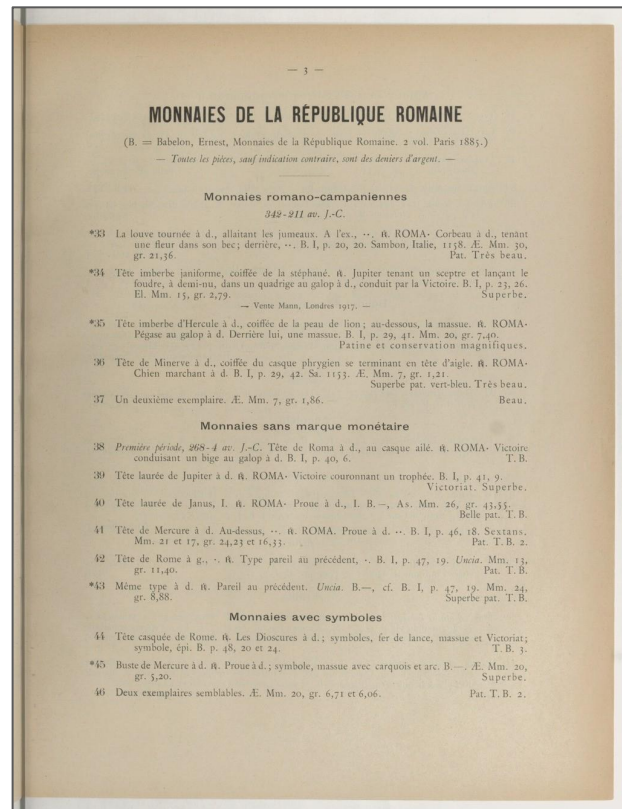
Inria

Researchers, R&D engineers,
PhD students

Objectives

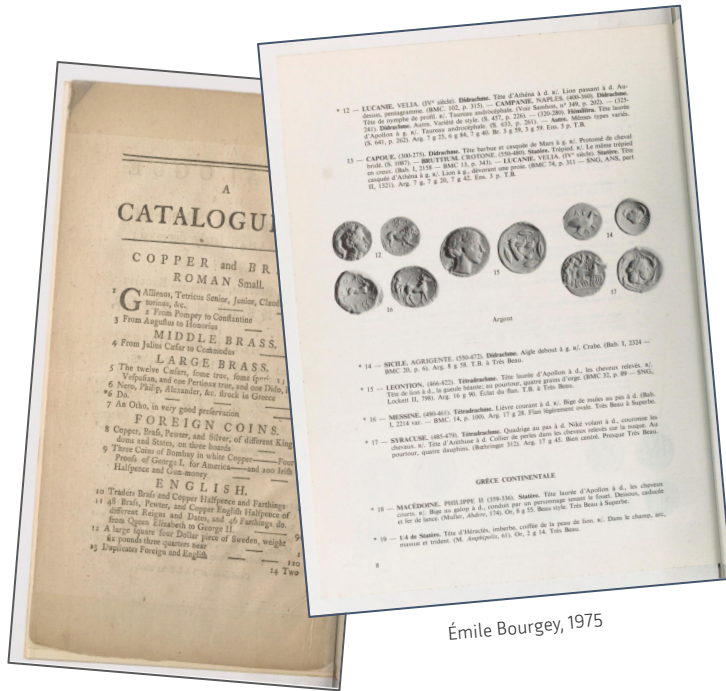
Automate the transformation of digitized sales catalogs into a structured database, using machine learning tools

- Segmentation of the layout and content
- Conversion into 
- Publication of the restructured catalogs



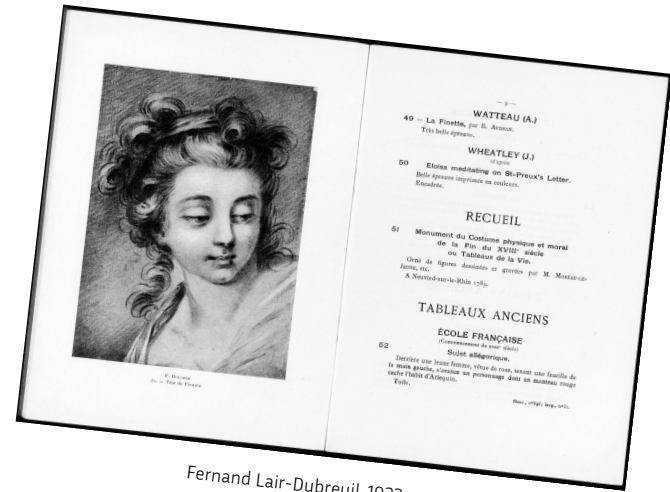
Lucien Naville, 1924

Sales Catalogs as Knowledge Bases (1/2)



Whiston Bristow, 1762

Émile Bourgey, 1975



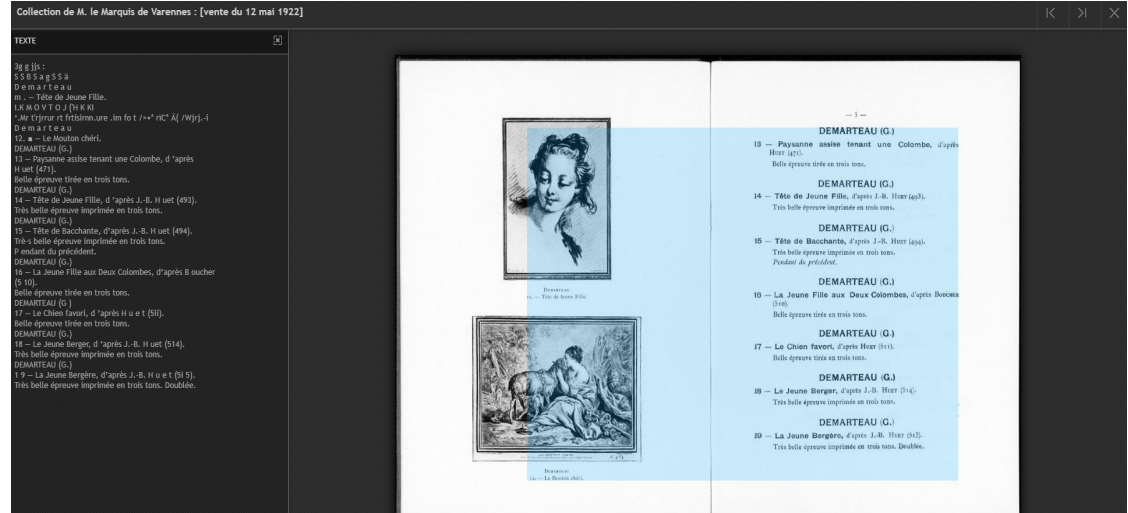
Fernand Lair-Dubreuil, 1922

- History of economics and collections
- Art history
- Iconography
- Cultural heritage in general

Sales Catalogs as Knowledge Bases (2/2)

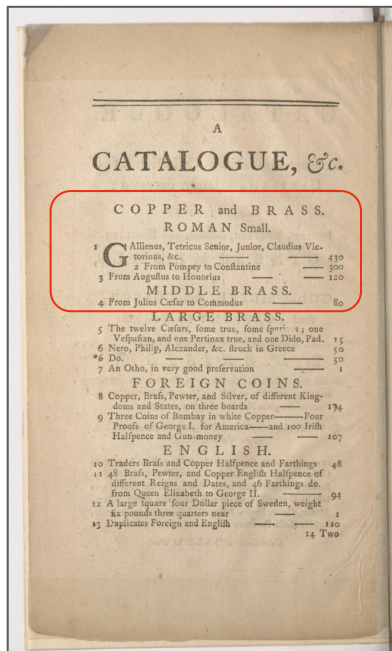
Redistributed online as pairs of images and **unstructured plain text**

➔ Makes exploration and content extraction difficult

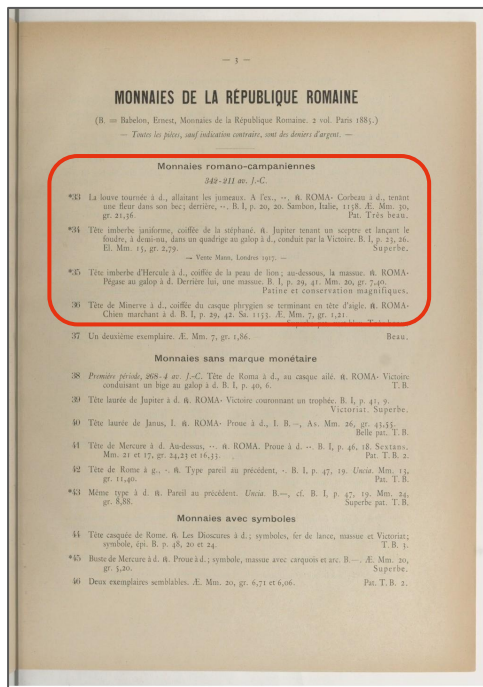


French national institute for art history's online document viewer

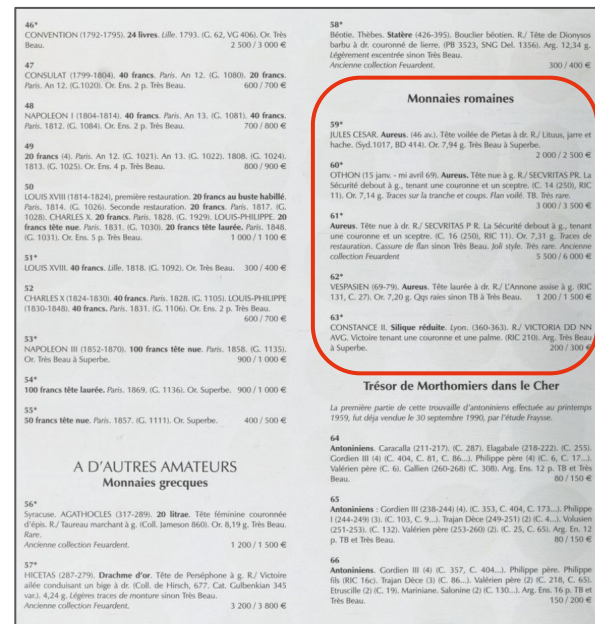
Basic Structure of a Sales Catalog



Whiston Bristow, 1762

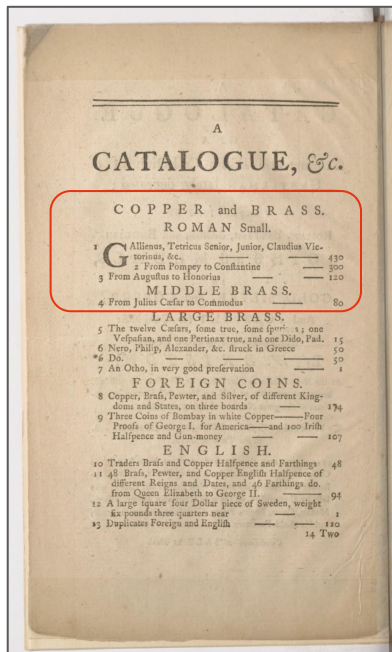


Lucien Naville, 1924

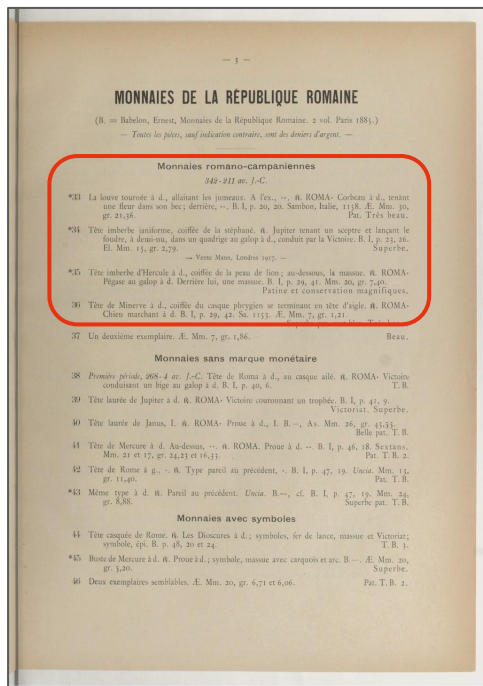


Fraysse & Associés, 2011

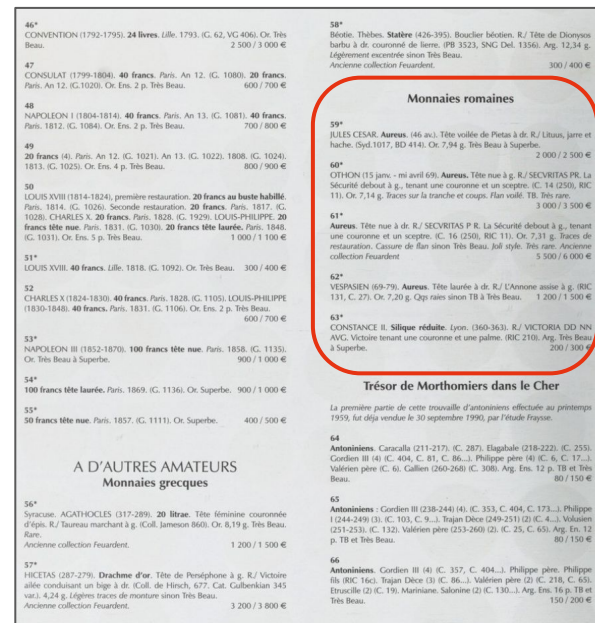
Challenges: A Heterogeneous Structure (1/2)



Whiston Bristow, 1762

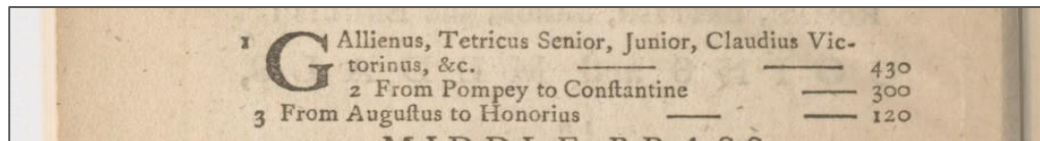


Lucien Naville, 1924



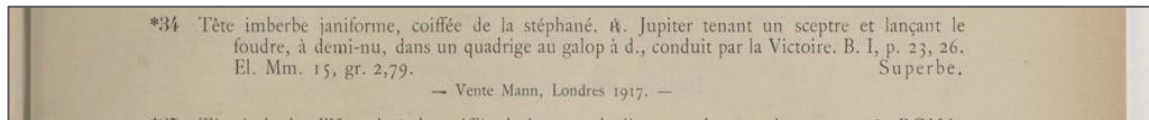
Fraysse & Associés, 2011

Challenges: A Heterogeneous Structure (2/2)



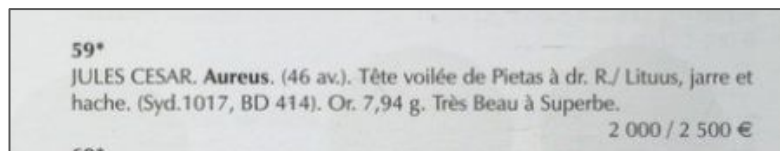
| | | |
|---|----------------------------------------------------|-----|
| 1 | G Allienus, Tetricus Senior, Junior, Claudius Vic- | 430 |
| | torinus, &c. | — |
| 2 | From Pompey to Constantine | 300 |
| 3 | From Augustus to Honorius | 120 |

Whiston Bristow, 1762.



| | | |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| *34 | Tête imberbe janiforme, coiffée de la stéphané. Æ. Jupiter tenant un sceptre et lançant le foudre, à demi-nu, dans un quadriges au galop à d., conduit par la Victoire. B. I, p. 23, 26. El. Mm. 15, gr. 2,79. Superbe. | — |
| | — Vente Mann, Londres 1917. — | — |

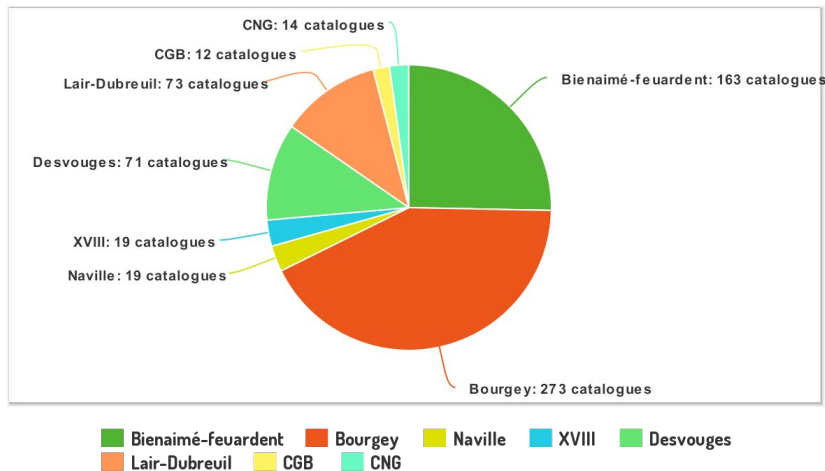
Lucien Naville, 1924.



| | | |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 59* | JULES CESAR. Aureus. (46 av.). Tête voilée de Pietas à dr. R./ Lituus, jarre et hache. (Syd.1017, BD 414). Or. 7,94 g. Très Beau à Superbe. | 2 000 / 2 500 € |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------|-----------------|

Frayse & Associés, 2011.

The DataCatalogue Corpus



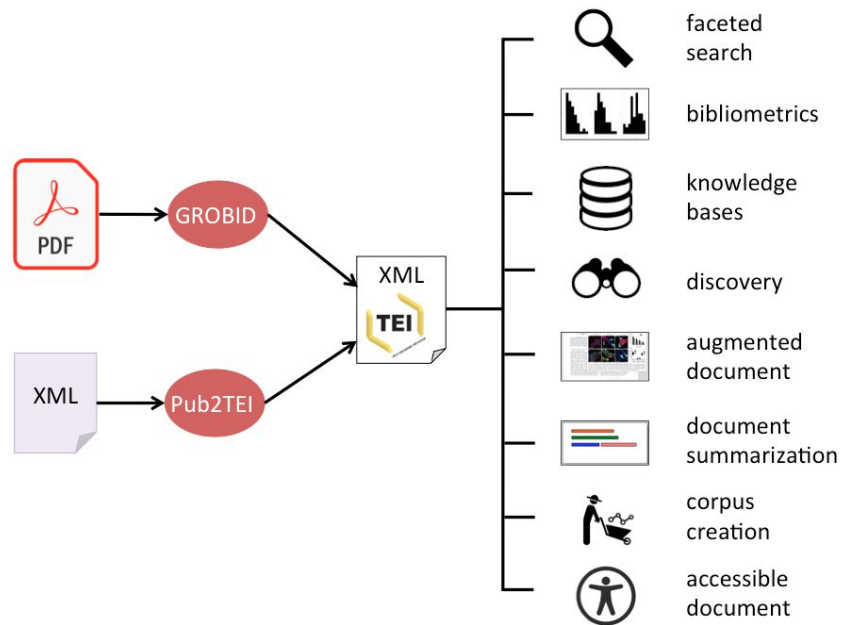
- Overall = +150k catalogs
- DataCatalogue dataset = **1426 images**
 - **713 catalogs** (from BnF/INHA)
 - **2 random pages** from each
- Various representations of:
 - **Sales types** (numismatics, ancient books, antiques, fine arts, etc.)
 - **Document layouts**
 - **Time periods** (18th-21st centuries)
- Languages:
 - **French** (~95%)
 - German and English (5%)

Phase 1 Assessment (2021-2022) : project engineering and technology bottleneck

We first tried to use **GROBID** (Generation of Bibliographic Data) to restructure the catalogs into TEI.

GROBID is an highly efficient **open source machine learning tool for extracting, parsing and restructuring PDFs into XML-TEI**. GROBID works at the document level and was designed for working with scientific publications. However, its models can be retrained on new annotated document types.

GROBID models use a combination of **textual and visual features** to restructure a document.



Source : <https://grobid.readthedocs.io/en/latest/Principles/>

Phase 1 Assessment (2021-2022): project engineering and technology bottleneck

However, GROBID models had trouble being accurate when dealing with the **noise generated by the OCR tools** used by the online libraries, such as:

- hallucinated characters
- Deconstructed tokens

```
<?xml version="1.0" encoding="UTF-8" ?>
<tei xmlns="http://www.tei-c.org/ns/1.0" space="preserve">
  <teiHeader>
    <fileDesc xml:id="0"/>
  </teiHeader>
  <text xml:lang="fr">
    <p>
      ■ ? > .
    </p>
    <p>
      » , f ,
    </p>
    <p>
      > . -y . i
    </p>
    <p>
      '
    </p>
    <p>
      .
    </p>
    <p>
      ■ ;
    </p>
    <p>
      '
    </p>
    <p>
      > i , * s : / i
    </p>
    <p>
      ■
    </p>
    <p>
      -
    </p>
    <p>
      '
    </p>
    <p>
      .
    </p>
    <p>
      ■
    </p>
    <p>
      -
    </p>
    <p>
      .
    </p>
  </text>
</tei>
```

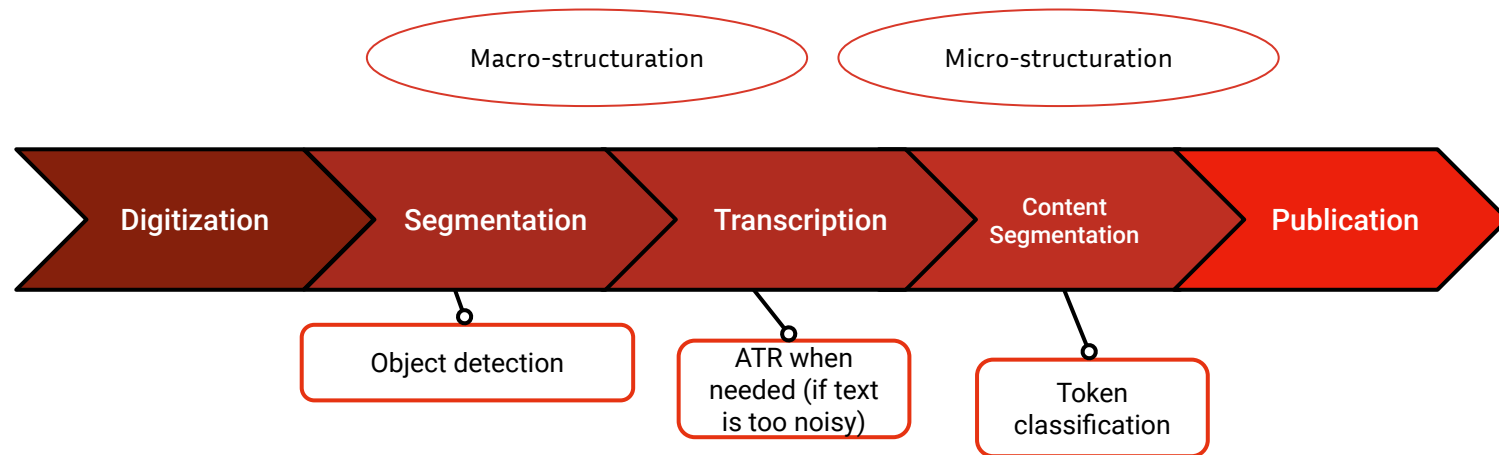
```
à g r o t e s q u e s e t f l e u r s .
<lb/>
2 -A p r e y . B o u i l l o n c o u v e r t à d e u x a n s e s à t o r e d e b r a n
<lb/>
c h a g e s e n a n c i e n n e f a i e n c e d é c o r é e e n c o u l e u r d ' o i s e a u x
<lb/>
e t c h i e n d e c h a s s e .
<lb/>
3-4 -D e l f t e t H o l l a n d e . N e u f p i e c e s : s i x p l a t s r o n d s , u n e
<lb/>
a s s i e t t e e t d e u x p e t i t e s c o u p e s e n a n c i e n n e f a i e n c e , d é c o r s
<lb/>
v a r i é s e n b l e u e t c o u l e u r .
<lb/>
«
<lb/>
5 -D e l f t ( g e n r e ) . D e u x c a c h e - p o t s à a n s e s c o q u i l l e s e n f a i e n c e
<lb/>
d é c o r é e e n c a m a i e u b l e u , f e u i l l a g e , r o c a i l l e e t p a y s a g e .
<lb/>
6-7 -D e l f t . P e t i t p o t à l a i t e t p e t i t e b o u t e i l l e à c o l à r e n f l e
<lb/>
m e n t e n a n c i e n n e f a i e n c e , d é c o r p o l y c h r o m e à f l e u r s .
<lb/>
8 -D e l f t . U n p l a t e n a n c i e n n e f a i e n c e , d é c o r e n c a m a i e u
<lb/>
b l e u , f e u i l l a g e e t a r m o i r i e a v e c l i o n .
<lb/>
9 -H i s p a n o -M a u r e s q u e . P l a t à o m b i l i c e n a n c i e n n e f a i e n c e
<lb/>
d e M a n i s s è s , d é c o r é a u c e n t r e d ' u n e r o s a c e , m a r l i a v e c
```

```
<lb/>
V E N T E
<lb/>
HOTEL DROUOT -SALLE N° 6
<lb/>
Les Lundi 20 et Mardi 21 Février 1922
<lb/>
A 2 H E U R E S
<lb/>
EXPOSITION PU BLIQUE
<lb/>
Le Dimanche 19 Février 1922 , de 2 à 6 heures
<lb/>
mm
<lb/>
^y ê S p ' Á -
<lb/>
B | ii ^ > >
<lb/>
■
<lb/>
» v r f f e ' :5pii
```

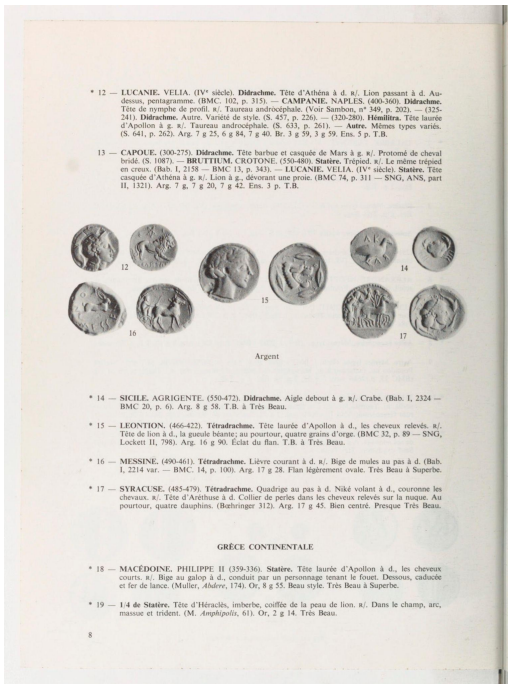
Retro-Structuration Pipeline



The DataCatalogue Workflow



From image to TEI



Émile Bourgey, 1975

Layout and content structuration

MainZone:Entry

12 — LUCANIE, VELLIA (IV^e siècle). *Didrachme*. Tête d'Albèra à d. n. Lion passant à d. Au-dessous, pentagramme. (BMC 102, p. 315). — CAMPANIE, NAPLES (400-300). *Didrachme*. Tête de romule de profil, n. Taurau androcephale. (Voir Seston, n° 349, p. 202. — (325-241). *Didrachme*. Autre. Variété de style. (S. 457, p. 226. — (230-200). *Hemilitra*. Tête laurée d'Apollon à g. n. Taurau androcephale. (S. 453, p. 261). — Autre. Mêmes types variés. (S. 441, p. 262. Arg. 7 g 25, g 84, 7 g 40. Br. 3 g 59, 3 g 59. Ems. 5 p. T.B.

MainZone:Entry

13 — CAPIQUE (300-275). *Didrachme*. Tête barbue et casquée de Mars à g. n. Protomé de cheval bridé. (S. 1087). — BRUTTIUM, CROTONE (550-480). *Statère*. Triped. n. Le même triped en croix. (Bab. I, 218. — BMC 13, p. 343). — LUCANIE, VELLIA (IV^e siècle). *Statère*. Tête casquée d'Albèra à g. n. Lion à g., dévorant une proie. (BMC 74, p. 311 — SNG, ANS, part II, 1323). Arg. 7 g, 7 g 20, 7 g 42. Ems. 3 p. T.B.

GraphicZone

GraphicZone:Head

GraphicZone:Head

GraphicZone:Head

GraphicZone:Head

GraphicZone:Head

GraphicZone:Head

GraphicZone:FigDesc

Argent

MainZone:Entry

14 — SICILE, AGRIGENTE (550-472). *Didrachme*. Aigle debout à g. n. Crabe. (Bab. I, 2324 — BMC 20, p. 6). Arg. 8 g 58. T.B. à Très Beau.

MainZone:Entry

15 — LEONTOON (466-422). *Tétradrachme*. Tête laurée d'Apollon à d., les cheveux relevés, n. Tête de lion à d., la queue blanche; au pourtour, quatre grammes d'orge. (BMC 32, p. 89 — SNG, Lockert II, 798). Arg. 16 g 90. Éclat du flan. T.B. à Très Beau.

MainZone:Entry

16 — MESSINE (490-461). *Tétradrachme*. Lièvre courant à d. n. Bège de mules au pas à d. (Bab. I, 2214 var. — BMC 14, p. 100. Arg. 17 g 28. Flan légèrement ovale. Très Beau à Supérieur.

MainZone:Entry

17 — SYRACUSE (485-479). *Tétradrachme*. Quadrige au pas à d. Nixé volant à d., cocronne les chevaux, n. Tête d'Aréthuse à d. Collier de perles dans les cheveux relevés sur la nuque. Au pourtour, quatre dauphins. (Bahringer 312). Arg. 17 g 45. Bien centré. Presque Très Beau.

MainZone:Head

GRÈCE CONTINENTALE

MainZone:Entry

18 — MACÉDOINE, PHILIPPE II (359-336). *Statère*. Tête laurée d'Apollon à d., les cheveux courts, n. Bège au galop à d., conduit par un personnage tenant le fouet. Dessous, caducée et fer de lance. (Muller, *Aldrov*, 174). Or, 8 g 55. Beau style. Très Beau à Supérieur.

MainZone:Entry

19 — 14 de *Statère*. Tête d'Héraclès, imberbe, coiffée de la peau de lion, n. Dans le champ, arc, massue et trident. (M. *Amphipolis*, 61). Or, 2 g 14. Très Beau.

NumberingZone



Macro-Structuration, or Layout Segmentation

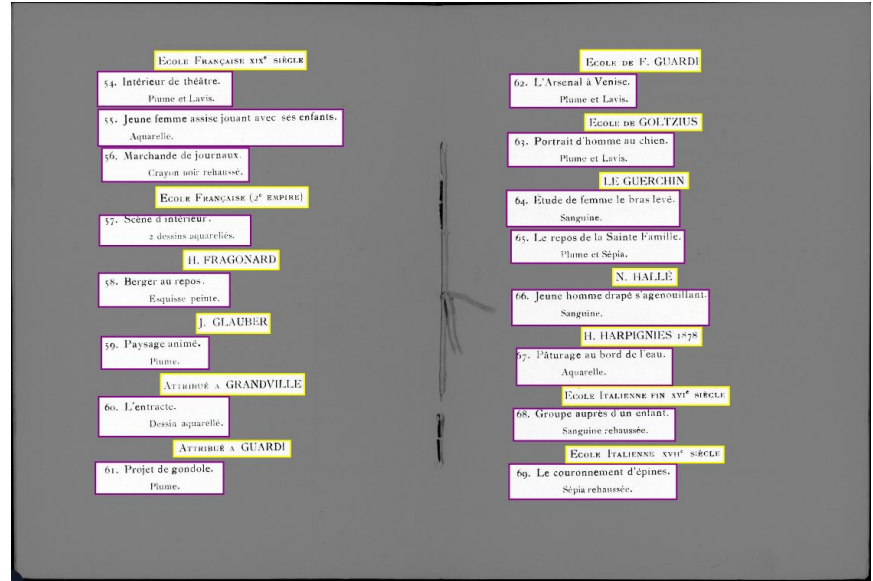
To bypass OCR noise, and based on the growing popularity and efficiency of the **YOLO model** in DH projects (Clérice, 2023 & Clérice et al., 2024), we decided to opt for an **object detection approach** for segmenting the sale catalogs' layouts.

YOLO only uses **visual features** to segment an image into zones. Each zone is also given a label by YOLO.

We annotated all images (1426) from our dataset.

```
<?xml:space="preserve">
<?xml-stylesheet type="xsl" href="xsl/transform.xsl" />
<table border="1">
| Titre | Description |
| --- | --- |
| 54. Intérieur de théâtre. | Plume et Lavis. |
| 55. Jeune femme assise jouant avec ses enfants. | Aquarelle. |
| 56. Marchande de journaux. | Crayon noir rehaussé. |
| 57. Scène d'intérieur. | 2 dessins aquarellés. |
| 58. Berger au repos. | Esquisse peinte. |
| 59. Paysage animé. | Plume. |
| 60. L'entracte. | Dessin aquarellé. |
| 61. Projet de gondole. | Plume. |
| 62. L'Arsenal à Venise. | Plume et Lavis. |
| 63. Portrait d'homme au chien. | Plume et Lavis. |
| 64. Etude de femme le bras levé. | Sanguine. |
| 65. Le repos de la Sainte Famille. | Plume et Sépia. |
| 66. Jeune homme drapé s'agenouillant. | Sanguine. |
| 67. Pâturage au bord de l'eau. | Aquarelle. |
| 68. Groupe auprès d'un enfant. | Sanguine rehaussée. |
| 69. Le couronnement d'épines. | Sépia rehaussée. |

```



Layout Annotation Using the SegmOnto Controlled Vocabulary



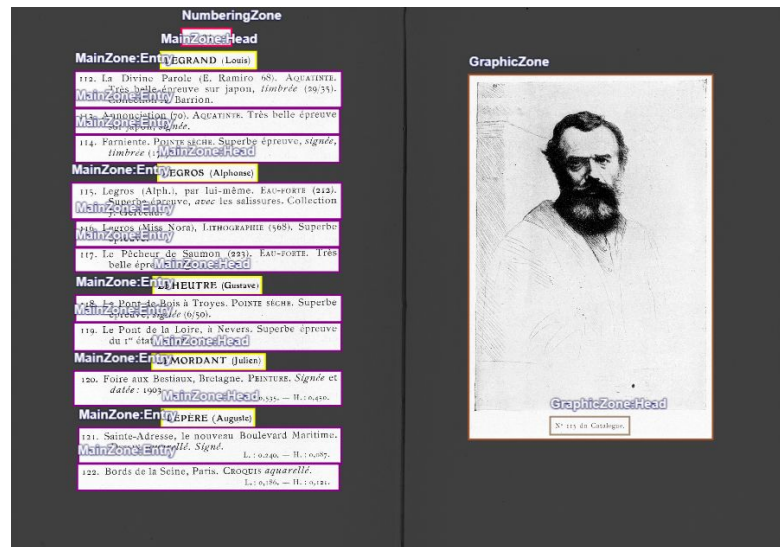
SegmOnto logo

Long paper presentation on LADaS tomorrow at 10:30 a.m. (Layout Analysis Dataset with Segmonto, by Thibault Clérice and Juliette Janes)!

→ The DataCatalogue training set was annotated following the **SegmOnto controlled vocabulary**, designed for **describing the content of books and manuscripts**.

→ It is also a subset of a wider layout analysis dataset, **LADaS (Layout Analysis Dataset with SegmOnto)**.

→ Using a controlled vocabulary ensures that our dataset follows the **FAIR principles**.



The DataCatalogue Annotation Schema: Examples

MainZone:P@CatalogueDesc and Mainzone:Entry

The screenshot displays a dark-themed interface with three entries, each highlighted in a light green box. Above each entry is a title in a pink box and a date range in a red box. The entries are as follows:

- Entry 1:** Title: **102**; Date: (1047-1050; de J.-C., 294-297.). Text: 759. MAXIMIANVS AVGVSTVS. Sa tête laurée, à droite. R. XX. MAXIMIANI AVG. SMN. en cinq lignes, dans une couronne de laurier; au-dessus, NK. (en monogramme.) (N° 704, var. inédite.) B. OR.
- Entry 2:** Title: **ALLECTVS**; Date: (1047-1050; de J.-C., 294-297.). Text: 760. IMP. C. ALLECTVS P. F. AVG. Son buste lauré, drapé et cuirassé, à droite. R. PAX AVG. La Paix debout, à gauche, tenant une branche d'olivier et un sceptre; dans le champ, à gauche, D, et à l'exergue, ML. (N° 30 var.) (Collection Du Chastel.) F.D.C. OR.
- Entry 3:** Title: **CONSTANCE CHLORE CÉSAR**; Date: (1047-1057; de J.-C., 292-304.). Text: 761. CONSTANTIVS N. C. Sa tête laurée, à droite. R. COMES AVGG. Pallas debout, à droite, tenant un sceptre et s'appuyant sur un bouclier; à l'exergue, P.T. (Inédite.) Trou rebouché. T.B. OR.

Below the third entry, there is another highlighted box with text: 762. CONSTANTIVS CAES. Sa tête laurée, à droite. R. COMITES AVGG. ET CAESS. NNNN. Les Dioscures debout, s'appuyant sur leurs hastes; celui de droite est drapé, de face, et celui de gauche drapé, de profil, leurs têtes surmontées de deux étoiles; à l'exergue, AQ.

In the bottom left corner of the interface, there is a small dark button with a white icon and the text "RESET".

DataCatalogue's SegmOnto-based annotation schema

MainZone:P@CatalogueDesc

— 102 —

759. MAXIMIANVS AVGVSTVS. Sa tête laurée, à droite.
R. XX. MAXIMIANI AVG. SMN. en cinq lignes, dans une couronne de laurier; au-dessus, NK. (en monogramme.) (N° 704, var. inédite.) B. OR.

ALLECTVS
(1047-1050; de J.-C., 294-297.)

760. IMP. C. ALLECTVS P. F. AVG. Son buste lauré, drapé et cuirassé, à droite.
R. PAX AVG. La Paix debout, à gauche, tenant une branche d'olivier et un sceptre; dans le champ, à gauche, D, et à l'exergue, ML. (N° 30 var.) (Collection Du Chastel.) F.D.C. OR.

CONSTANCE CHLORE CÉSAR
(1045-1057; de J.-C., 292-304.)

761. CONSTANTIVS N. C. Sa tête laurée, à droite.
R. COMES AVGG. Pallas debout, à droite, tenant un sceptre et s'appuyant sur un bouclier; à l'exergue, P.T. (Inédite.) Trou rebouché. T.B. OR.

762. CONSTANTIVS CAES. Sa tête laurée, à droite.
R. COMITES AVGG. ET CAESS. NNNN. Les Dioscures debout, s'appuyant sur leurs hastes; celui de droite est drapé, de face, et celui de gauche drapé, de profil, leurs têtes surmontées de deux étoiles; à l'exergue, AQ.

🏠 + RESET

MarginTextZone:ManuscriptAddendum

— 43 —

Hadrien empereur (117-138).

81 — IMP. CAESAR. TRAIAN. HADRIANVS. AVG. Buste lauré à droite avec le paludament et la cuirasse. R' P.M.TR.P.COS. III. Minerve casquée debout à gauche, mettant un grain d'encens dans la flamme d'un candélabre et tenant une haste (Cob. 1009), G.B. Belle patine. T.B. 128

Ælius César (136-138)

82 — L. AELIVS. CAESAR. Sa tête nue à droite. R' TIIB.POT. COS. II. La Piété debout à droite, tenant un grain d'encens; à ses pieds un autel allumé (Cob. 47). Or. T.B. 200

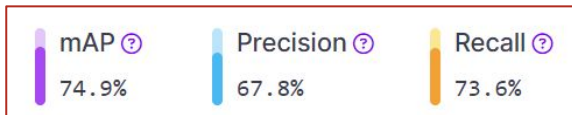
Commode empereur (180-192).

83 — M. COMMODVS. ANTONINVS. PIVS. FELIX. AVG. BRIT. Buste à droite de Commode lauré avec la cuirasse et le paludament. R' MON. AVG. (à l'exergue) P.M.TR.P.XIII.IMP.VIII.COS.V. P.P. Les trois monnaies, debout à gauche, tenant chacune une balance et une corne d'abondance (188 de J.-C.). Médailon de bronze (Cob. 376). Sans patine. B. 170

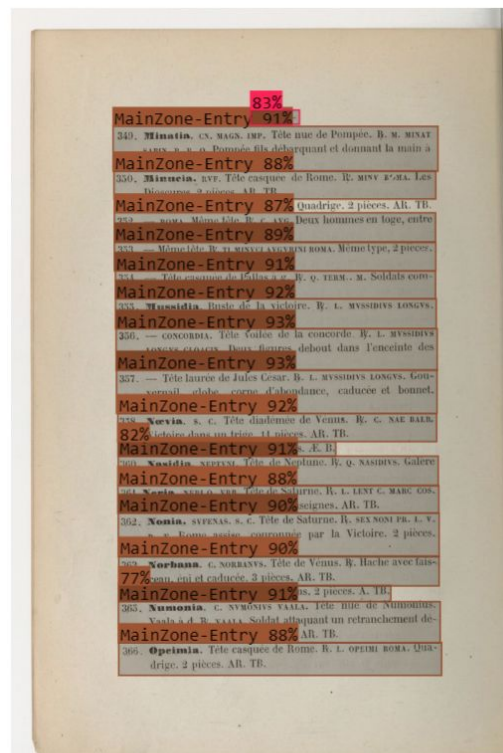
Etruscille, femme de Trajan Déce.

84 — HERENNIA. ETRVSCILLA. AVG. Son buste diadémé à droite avec le croissant. R' PVDICITIA. AVG. S.C. La Pudeur assise à gauche, se couvrant la figure de son voile et tenant un sceptre. Médailon (Cob. 48). F.D.C. 80

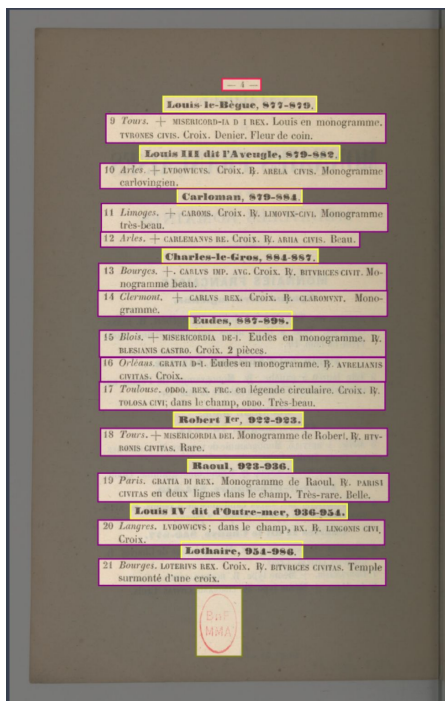
DataCatalogue's YOLO Model Performance



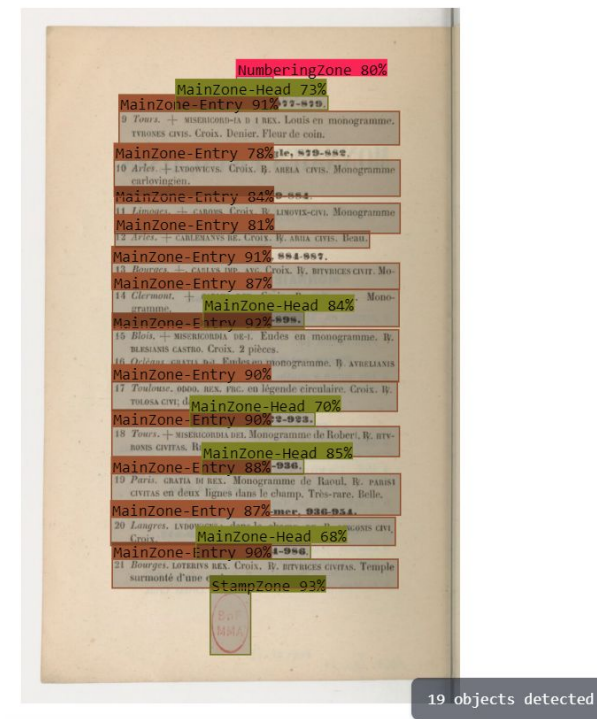
The dataset is available and the model can be tested at <https://app.roboflow.com/datacatalogue/macro-segmentation/visualize/10>.



Qualitative observations

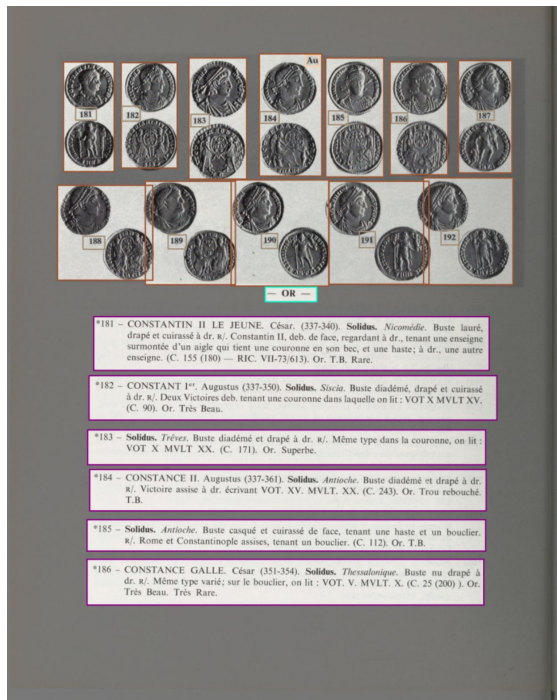


Ground truth

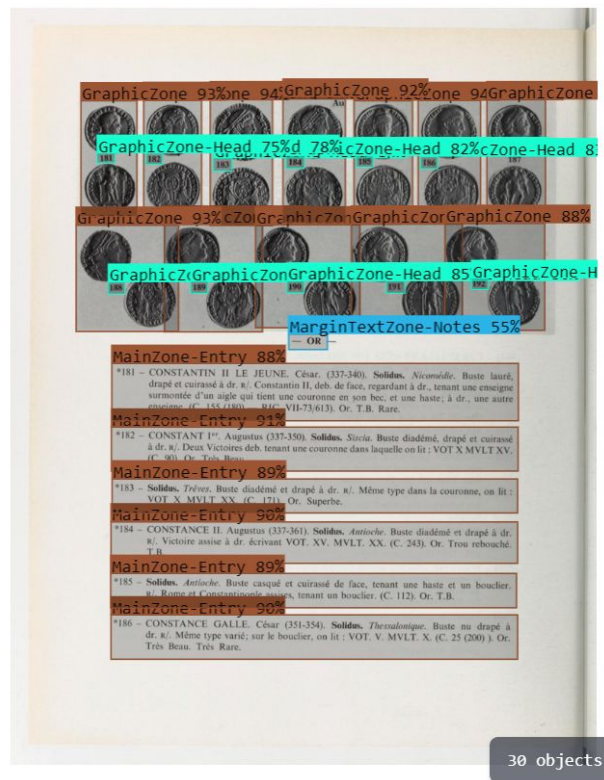


Prediction

Qualitative Observations

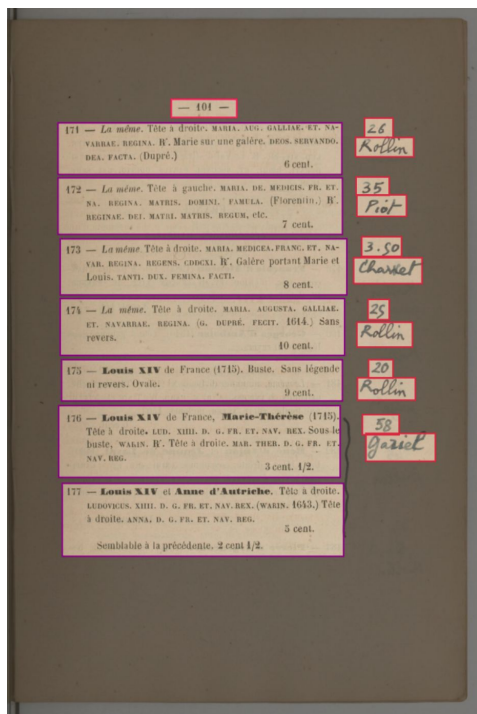


Ground truth

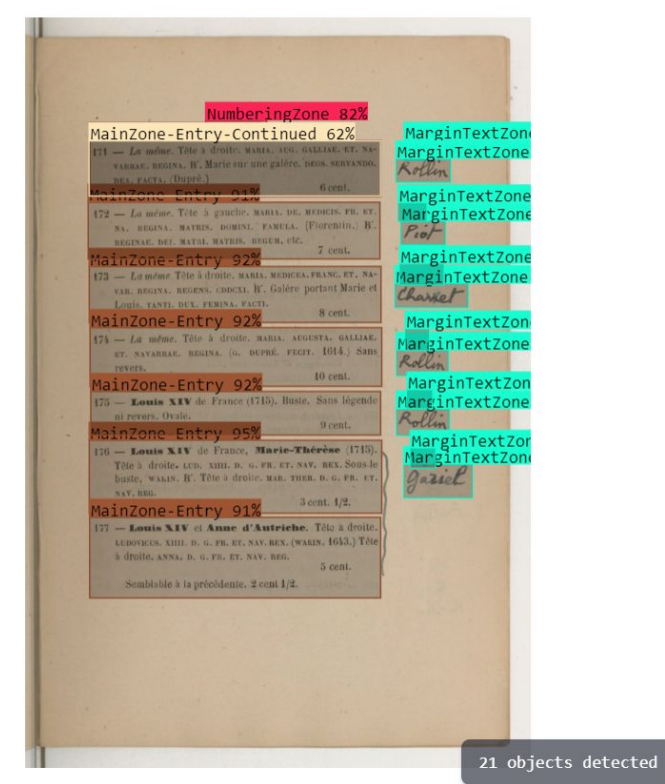


Prediction

Qualitative Observations

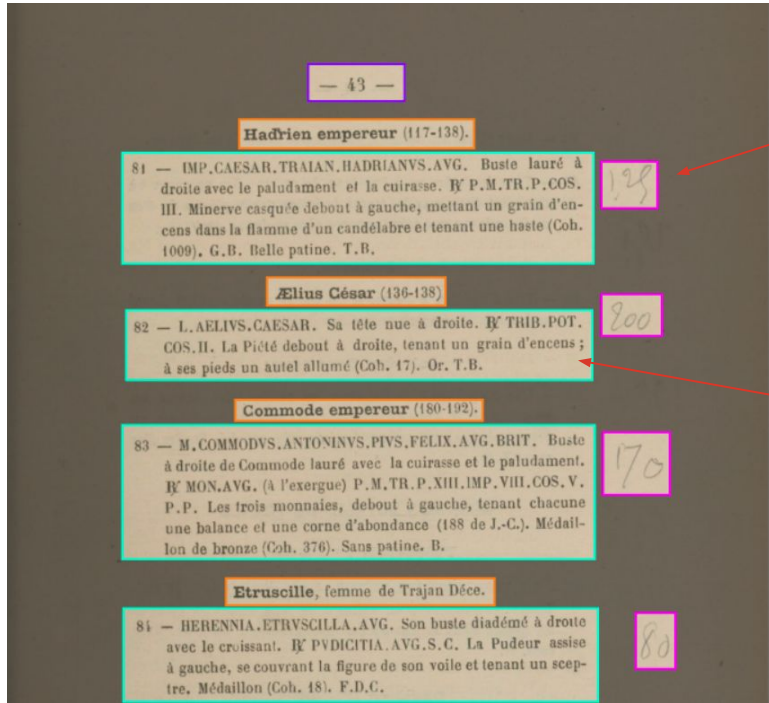


Ground truth



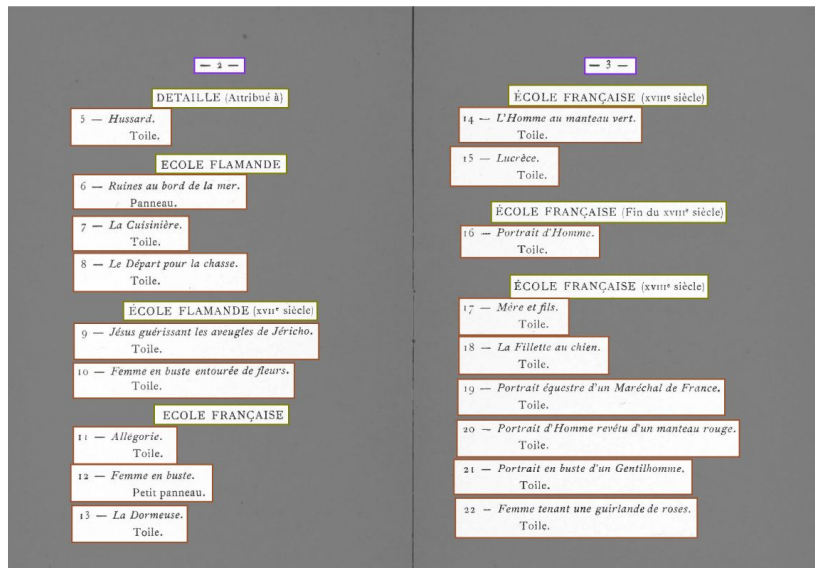
Prediction

Leveraging layout analysis for automatic text recognition



Extracting handwritten notes (previously ignored by generic OCR engines) and transcribing them with an HTR model

Semi-automatic evaluation of the quality of the OCR, and if necessary, re-ocerization

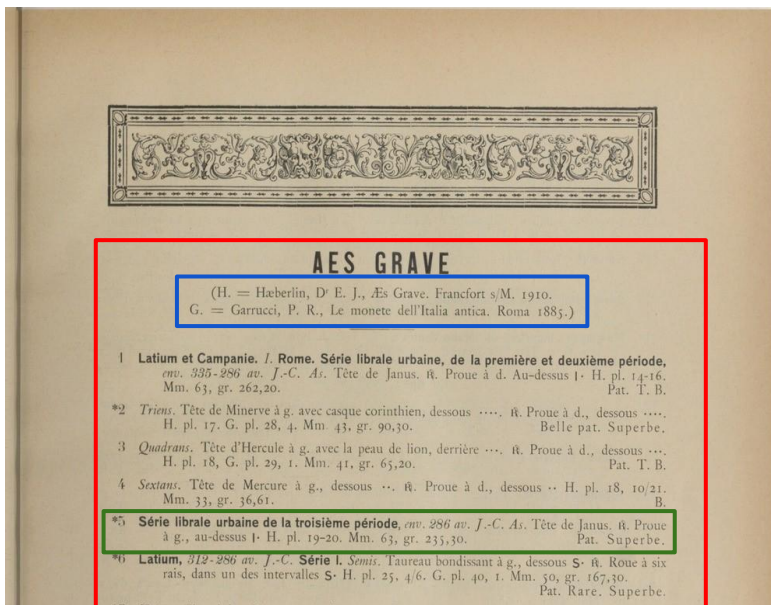


Semi-automatic
transformation of the
catalogs' segmentation
into TEI

Text zones coordinates +
transcribed text



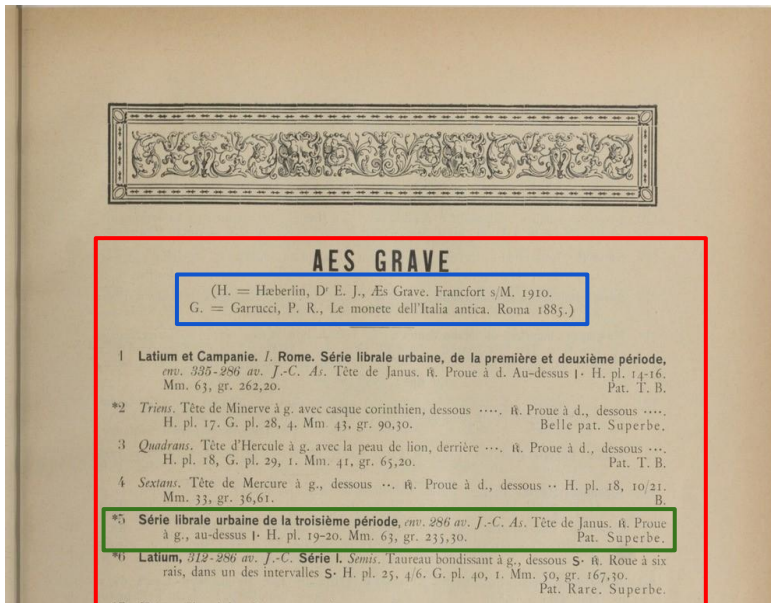
TEI Modeling of Sales Catalogs (1/2)



Built upon the grounding work of:

- Gabay, S., Topalov, B., Corbières, C., Rondeau Du Noyer, L., Joyeux-Prunel, B., & Romary, L. (2021). *Automating Artl@s – Extracting Data from Exhibition Catalogues*. <[hal-03331838](https://hal.archives-ouvertes.fr/hal-03331838)>

TEI Modeling of Sales Catalogs (2/2)



- Makes catalogs **interoperable**
- **Structured, enriched, standardized, reusable** data
- **DataCatalogue ODD** = customization of the TEI for encoding sales catalogs
- Definition of 3 new elements:
 - **<catalogueEntry>**
 - **<catalogueDesc>**
 - **<catalogueItem>**

Modeling of a Catalog Entry (1/2)

*35 Tête imberbe d'Hercule à d., coiffée de la peau de lion ; au-dessous, la massue. R. ROMA.
Pégase au galop à d. Derrière lui, une massue. B. I, p. 29, 41. Mm. 20, gr. 7,40.
Patine et conservation magnifiques.

*35

Tête imberbe d'Hercule à d., coiffée de la peau de lion ; au-dessous, la massue. R. ROMA.
Pégase au galop à d. Derrière lui, une massue.

Entry
number

Object description

B. I, p. 29, 41.

Bibliographic
information

Object dimensions

Mm. 20, gr. 7,40.

Building TEI specifications based on our
future **micro-segmentation** step (= **token
classification** task, such as **Named Entity
Recognition**)

Patine et conservation magnifiques.

Object conservation comments



Modeling of a Catalog Entry (2/2)

```
<catalogueEntry>
  <head> ANTIQUITÉS ÉGYPTIENNES, <lb/> GRANIT, ALBATRE ORIENTAL, SERPENTINE, <lb/> CRAIE,
    IVOIRE.</head>
  <!-- <catalogueDesc/> -->
  <!-- <catalogueltem n=""/> -->
  <catalogueltem n="4">
    <num>4.**.</num>
    <objectType> Albâtre oriental.</objectType>
    <desc> Deux canopes, qui présentent l'un et l'autre quatre <lb/> colonnes d'hiéroglyphes gravés
      en creux ; les couvercles <lb/> placés sur ces vases symboliques sont formés par des
      <lb/> têtes d'Isis.</desc>
    <dimensions>
      <height> Hauteur, 13 pouces.</height>
      <!-- <width/> -->
    </dimensions>
    <!-- <condition/> -->
  </catalogueltem>
</catalogueEntry>
```

Thank you !

