DataCatalogue : Restructurer automatiquement les catalogues de ventes

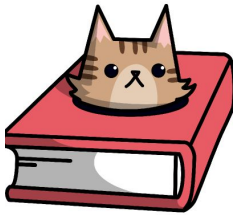
Sarah Bénière

DataCatalogue

Restructurer automatiquement les catalogues de ventes

Sarah Bénière
Ingénieure R&D (ALMAnaCH)

M2 TNAH, Panorama de projets
24 janvier 2024



Logo par Alix Chagué,
inspiré par Loading Artist

Inria

Contexte institutionnel du projet

Inria et ALMAnaCH

- *Institut national de recherche en informatique et en automatique*
- 220 **équipes-projets** réparties sur tout le territoire (10 centres de recherche)
- **Recherche et valorisation** en sciences du numérique, coordination du **Programme National de Recherche en IA** (PNRIA)

- Équipe de recherche spécialisée dans le **traitement automatique des langues** (TAL/NLP) et les **humanités numériques** (HN/DH)
- **Environ 50 membres** : chercheur·euse·s permanent·e·s, doctorant·e·s, post-doctorant·e·s, ingénieur·e·s et stagiaires

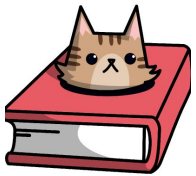
Quelques projets d'humanités numériques

LECTAUREP

- Convention Culture-Inria
- **Transcrire automatiquement** les registres de notaires conservés aux **Archives nationales** (**HTR**) pour faciliter la consultation des documents.

DataCatalogue

- Convention Culture-Inria
- Créer une **chaîne de traitement** (ou *workflow*), de la numérisation à la publication en ligne, de **catalogues de ventes** (objets d'art) conservés à la **BnF** et à l'**INHA**



EHRI

- **Horizon** (UE)
- Assister l'infrastructure dans la **publication en ligne** de ses **éditions scientifiques numériques** grâce à une **chaîne d'édition**

La Convention Culture-Inria

- Partenariat entre le **ministère de la Culture** et Inria
- Dispositif de **financement de projets de R&D en sciences du numérique appliqués à la culture**
- Objectifs :
  - Favoriser l'**interdisciplinarité**
  - Faire le **lien** entre les chercheurs en sciences du numérique et les acteurs de la culture

BnF

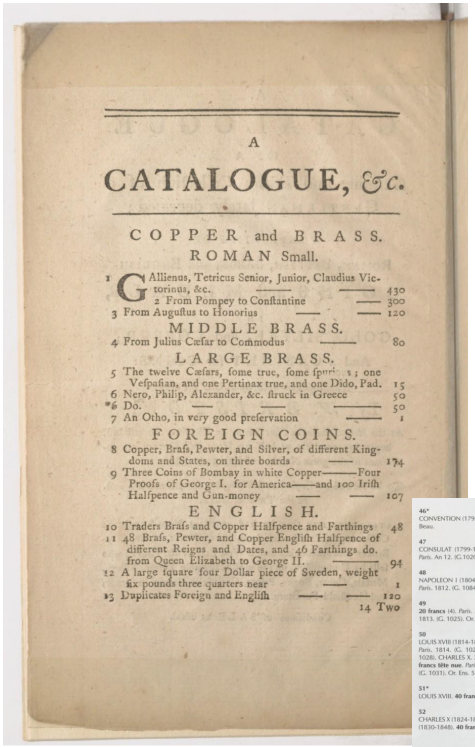
Institut national d'histoire de l'art

INHA

Le catalogue de vente

L'objet : le catalogue de vente

- Un corpus idéal pour l'**analyse automatique de mise en page**
- **Structure régulière** : liste de notices de vente régies par des niveaux de titres
- MAIS des **particularités selon les maisons de vente et les siècles**



A

CATALOGUE, &c.

COPPER and BRASS.

ROMAN Small.

1 Gallienus, Tetricus Senior, Junior, Claudius Victorinus, &c. —— —— 430
2 From Pompey to Constantine —— —— 300
3 From Augustus to Honorius —— —— 120

MIDDLE BRASS.

4 From Julius Cæsar to Commodus —— —— 80

LARGE BRASS.

5 The twelve Cæsars, some true, some spur[illegible]; one Vespasian, and one Pertinax true, and one Dido, Pad. 15
6 Nero, Philip, Alexander, &c. struck in Greece —— —— 50
*6 Do. —— —— —— —— 50
7 An Otho, in very good preservation —— —— 1

FOREIGN COINS.

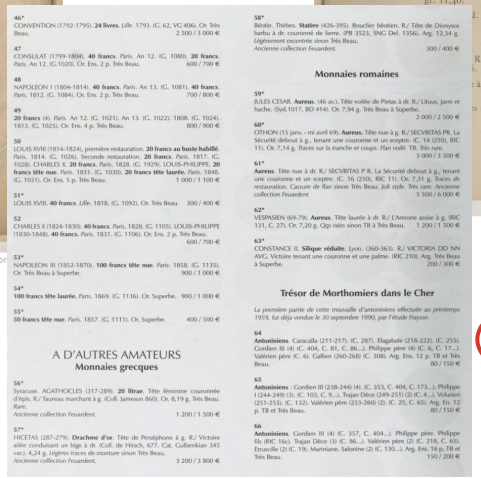
8 Copper, Brass, Pewter, and Silver, of different Kingdoms and States, on three boards —— —— 174
9 Three Coins of Bombay in white Copper—— Four Proofs of George I. for America—— and 100 Irish Halfpence and Gun-money —— —— 107

ENGLISH.

10 Traders Brass and Copper Halfpence and Farthings 48
11 48 Brass, Pewter, and Copper English Halfpence of different Reigns and Dates, and 46 Farthings do. from Queen Elizabeth to George II. —— —— 94
12 A large square four Dollar piece of Sweden, weight six pounds three quarters near —— —— 1
13 Duplicates Foreign and English —— —— 120
14 Two

Whiston Bristow, 1762

Deux colonnes



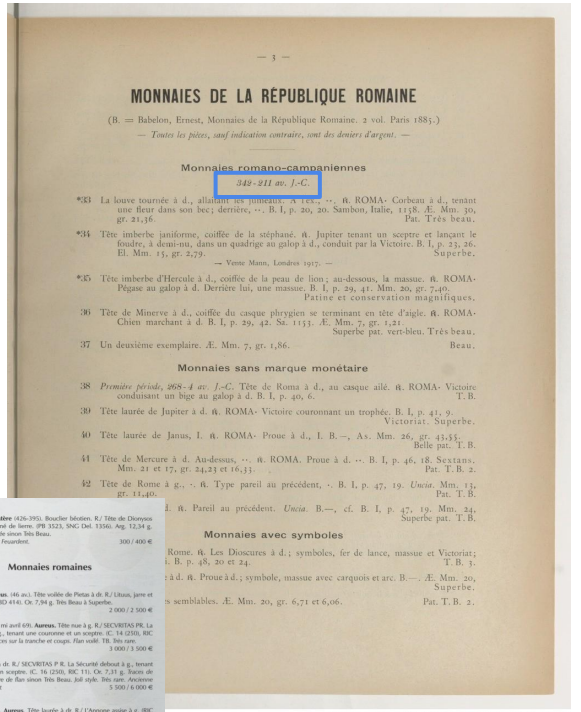
A D'AUTRES AMATEURS

Monnaies grecques

Monnaies romaines

Trésor de Morthomiers dans le Cher

Mise en page resserrée
Niveaux de titre simples



MONNAIES DE LA RÉPUBLIQUE ROMAINE

(B. = Babelon, Ernest, Monnaies de la République Romaine. 2 vol. Paris 1885.)

— *Toutes les pièces, sauf indication contraire, sont des deniers d'argent.* —

Monnaies romano-campaniennes

349-211 av. J.-C.

Monnaies sans marque monétaire

Monnaies avec symboles

Lucien Naville, 1924

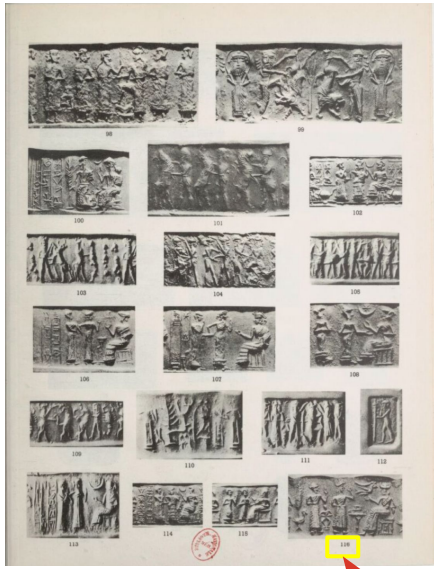
Mise en page épurée
Niveau de description

Fraysse & Associés, 2011

Une homogénéité de mise en page ?

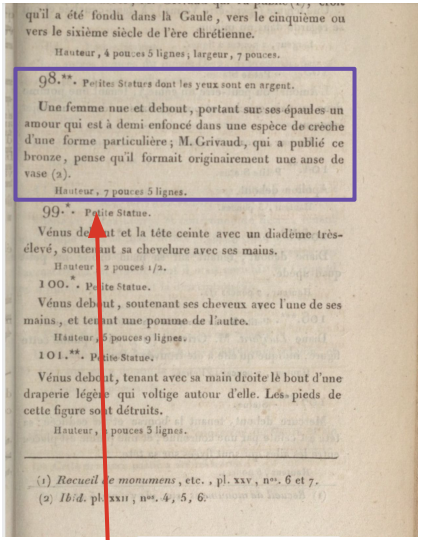
Lair-Dubreuil, 1919 (INHA)

Objet associé directement à sa notice



Bourgey, 1962 (BnF)

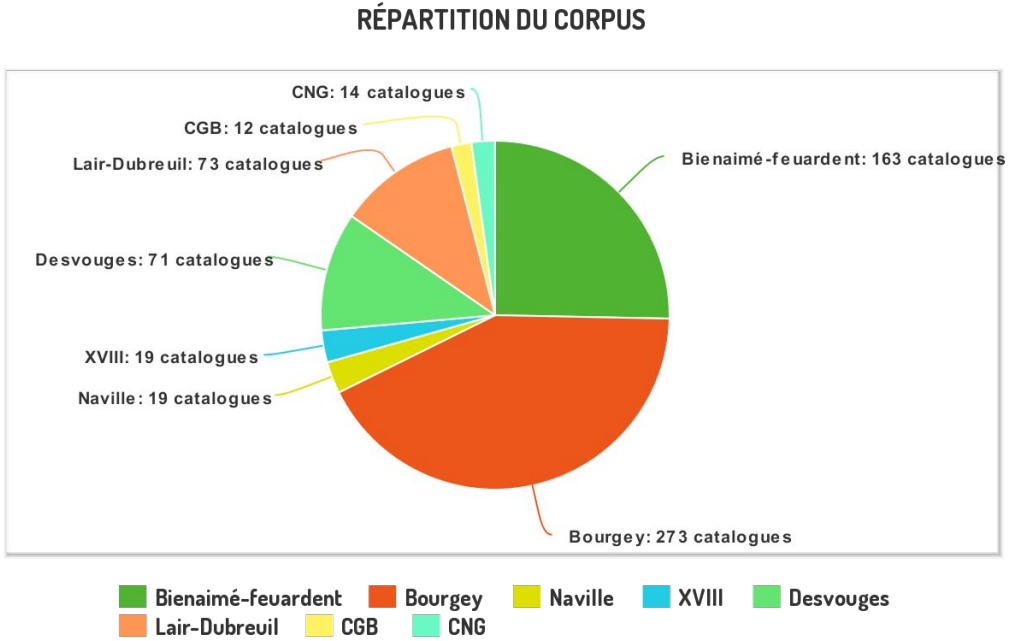
Planche d'illustrations
Reliées à leur notice par leur numéro



Dubois, 1820 (BnF)

Notices sur plusieurs lignes (voire plusieurs paragraphes)

Le corpus : un échantillon des catalogues de la BnF et l'INHA



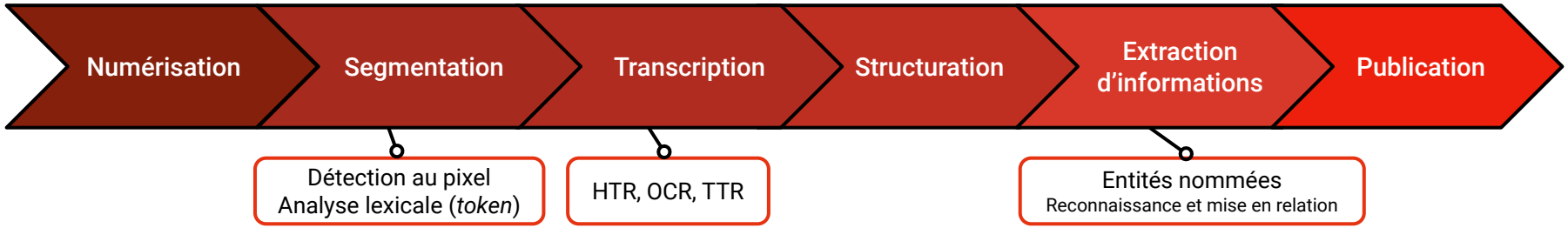
- **Échantillon de 713 catalogues** (le corpus complet compte des milliers de documents)
  - ➥ Collections de la BnF et de l'INHA + des collections privées
- **4 siècles représentés** : XVIII^{e}, XIX^{e}, XX^{e} et XXI^{e}
- **Langues** : ~95% en français
- **Types de ventes** : numismatique, livres, antiquités, objets d'art, objets de luxe

Le *workflow* DataCatalogue

Objectifs du projet

- **Dépasser la recherche plein-texte** en passant d'une numérisation à une **base de données textuelle et requêtable**
- **Segmenter** les catalogues de vente et attribuer une **étiquette** à chaque niveau d'information : *entrée de catalogue*, *numéro de notice*, *description de l'objet*, *matériaux*, *prix*, *observations de conservation*, etc.
- **Structurer l'information avec un encodage XML-TEI** à partir des zones segmentées
- **Mettre à disposition** les catalogues structurés dans une **interface de publication** permettant de **requêter** sur les zones segmentées

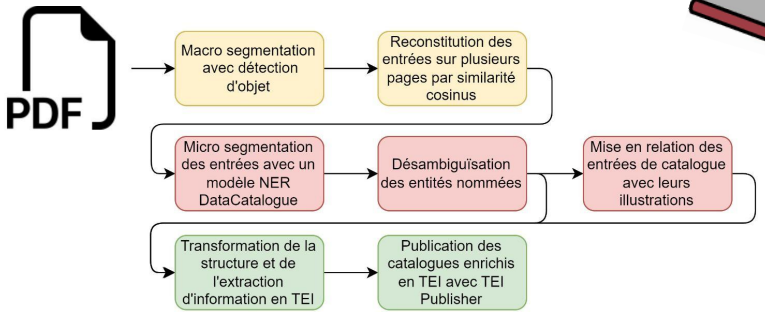
Étapes du *workflow*



- Proposer des **solutions logicielles** et des **outils** facilement appréhendables par les **partenaires institutionnels** et la **communauté HN**
- Respect des principes FAIR (*Findable, Accessible, Interoperable, Reusable*) et des valeurs de la science ouverte

Une gestion de projet agile : la feuille de route DataCat

- Feuille de route établie en tout début de Phase 2 (octobre 2023)
- Définition des **grandes étapes** puis **liste des tâches**
- *Workflow* faussement linéaire



- Campagne d'annotation pour créer un jeu de données d'entraînement
- Entraînement d'un modèle type YOLO
- Potentiellement besoin de refaire un passage d'OCR pour les segments trop bruités
- Calcul de similarité pour reconstituer les entrées étalées sur plusieurs pages

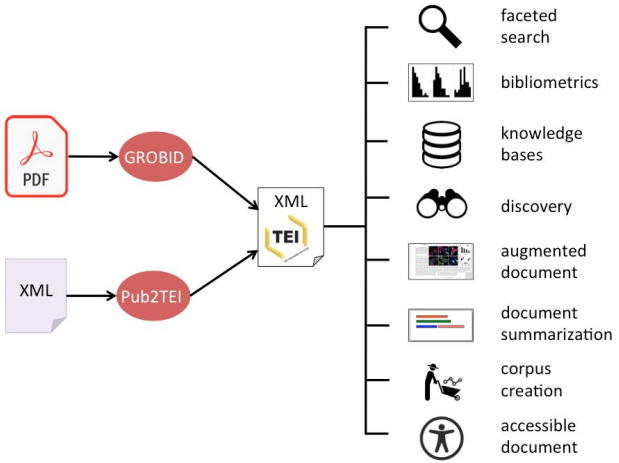
- Définition des entités nommées à segmenter : description, prix, matériaux, conditions de conservation, etc.
- Campagne d'annotation pour créer un jeu de données annotés en entités nommées
- Entraînement du modèle NER
- Désambiguïsation des entités nommées contre une base de connaissance
- Si possible, mise en place de liens entre les entrées de catalogues et leurs illustrations quand elles sont présentes

- Création d'une ODD précise rendant compte de la macro et micro structure des catalogues
- Documentation
- Transformation de l'extraction et de la classification en TEI
- Création d'un modèle de publication avec TEI Publisher
- Développement de l'application personnalisée TEI Publisher
- Si possible, implémentation d'une consultation augmentée des documents pour consulter les entrées et leurs illustrations en même temps, souvent séparées dans les volumes originaux
- Déploiement en ligne de l'application

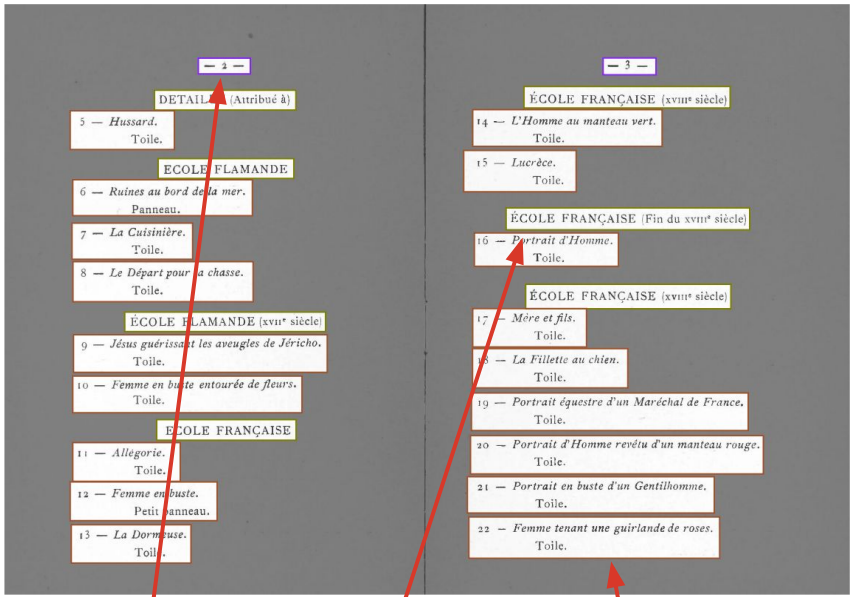
Segmentation

Deux approches de macro-segmentation : GROBID (Phase 1)

- *GeneRation Of BIbliographic Data*
- **Librairie d'apprentissage automatique** (Java)
  - ➥ Extraire, parser et structurer des documents PDF en XML-TEI
- Outil ***open source*** développé pour traiter les **publications scientifiques** (HAL, ResearchGate, etc.)
- Système de **modèles en cascade** (affinage de la segmentation)
- Permet de **créer une chaîne de traitement complète** (PDF → XML)
- **MAIS** difficile à prendre en main pour des résultats pas toujours satisfaisant dans notre cas
- **Sensible aux erreurs de transcription automatique**



Deux approches de macro-segmentation : la détection automatique d'objets (Phase 2)



Numéro de page

`NumberingZone`

Niveau de titre

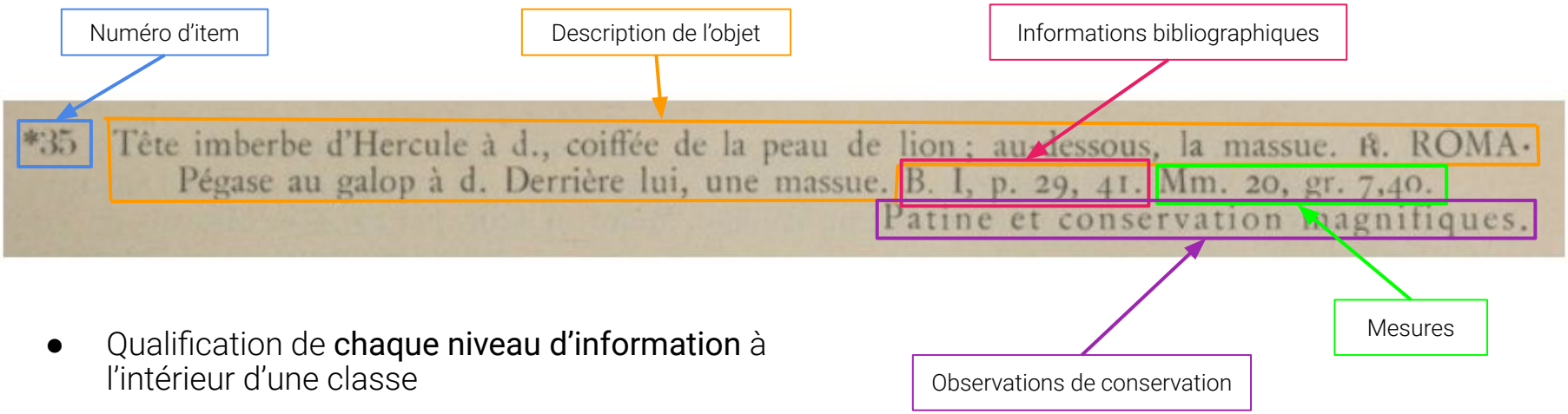
`MainZone:Head`

Notice

`MainZone:Entry`

- Méthode de **vision par ordinateur** (*computer vision*)
- Modèle YOLO (v8)
- **Apprentissage supervisé** (*supervised machine learning*) : annotations manuelles d'un échantillon aléatoire avec Roboflow
- **Détection des niveaux d'information** dans les catalogues grâce à l'image uniquement (pixel)
- Personnalisation des **classes** définies dans le vocabulaire contrôlé SegmOnto, travail collaboratif avec le projet COLaF pour le *dataset* LADaS, et création de nouvelles classes

La micro-segmentation à l'échelle de la notice



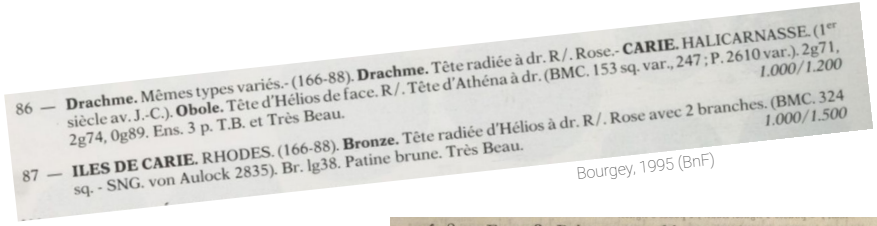
- Qualification de **chaque niveau d'information** à l'intérieur d'une classe
- Deux solutions :
  - Chaîne de traitement GROBID
  - Modèles de **reconnaissance d'entités nommées** (REN / NER)

Macro-segmentation
\+
Micro-segmentation

→ Encodage des catalogues de ventes

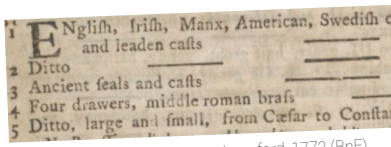
Gérer l'hétérogénéité des catalogues de ventes

- Hétérogénéité de **mise en page** et de **présentation de l'information** en fonction de l'**époque**, de la **maison de vente** et de la **discipline**
- **Absence de normes typographiques et d'écriture** (parfois au sein d'une même maison de vente)
- Nécessité de prendre en compte les **données textuelles** (pas uniquement l'image)

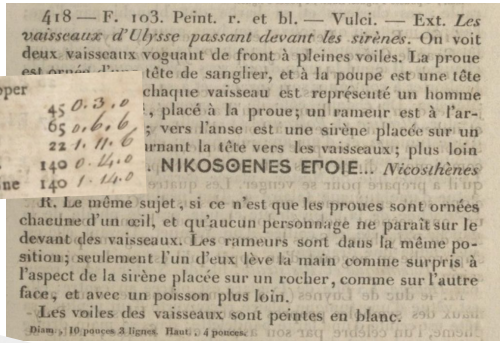

86 — **Drachme.** Mêmes types variés.- (166-88). **Drachme.** Tête radiée à dr. R/. Rose.- **CARIE.** HALICARNASSE. (1er siècle av. J.-C.). **Obole.** Tête d'Hélios de face. R/. Tête d'Athéna à dr. (BMC. 153 sq. var., 247 ; P. 2610 var.). 2g71, 2g74, 0g89. Ens. 3 p. T.B. et Très Beau. 1.000/1.200

87 — **ILES DE CARIE.** RHODES. (166-88). **Bronze.** Tête radiée d'Hélios à dr. R/. Rose avec 2 branches. (BMC. 324 sq. - SNG. von Aulock 2835). Br. lg38. Patine brune. Très Beau. 1.000/1.500

Bourgey, 1995 (BnF)


English, Irish, Manx, American, Swedish copper and leaden casts
Ditto
Ancient seals and casts
Four drawers, middle roman brass
Ditto, large and small, from Cæsar to Constantine

Langford, 1772 (BnF)

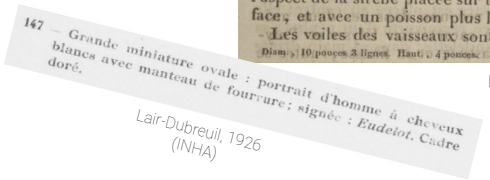

418 — F. 103. Peint. r. et bl. — Vulci. — Ext. *Les vaisseaux d'Ulysse passant devant les sirènes.* On voit deux vaisseaux voguant de front à pleines voiles. La proue est ornée d'une tête de sanglier, et à la poupe est une tête chaque vaisseau est représenté un homme placé à la proue; un rameur est à l'arvers l'anse est une sirène placée sur un rnant la tête vers les vaisseaux; plus loin NIKOSΘENES EΠOIE... *Nicosthènes*

R. Le même sujet, si ce n'est que les proues sont ornées chacune d'un œil, et qu'aucun personnage ne paraît sur le devant des vaisseaux. Les rameurs sont dans la même position; seulement l'un d'eux lève la main comme surpris à l'aspect de la sirène placée sur un rocher, comme sur l'autre face, et avec un poisson plus loin.

Les voiles des vaisseaux sont peintes en blanc.

Diam., 10 pouces 3 lignes. Haut., 4 pouces.

Rollin, 1836 (BnF)


147 — Grande miniature ovale : portrait d'homme à cheveux blancs avec manteau de fourrure ; signée : *Eudelot*. Cadre doré.

Lair-Dubreuil, 1926 (INHA)

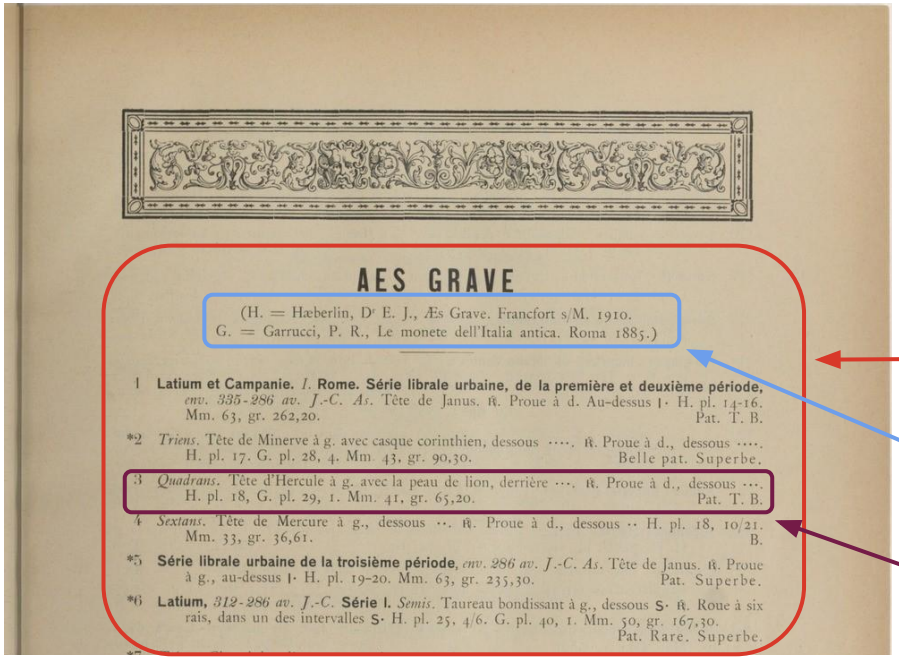
L'entraînement des modèles de segmentation (macro et micro) est basé sur l'annotation des catalogues de vente

Structuration en TEI

Standardiser la structuration des catalogues de vente

- **Standard XML interopérable** pour l'**édition de textes**
- **Modélisation** du catalogue de vente avec les éléments existants
- Introduction de **nouveaux éléments spécifiques** aux catalogues comme la structure d'une notice
- Données structurées et enrichies, standardisées, mutualisables et réutilisables

Modéliser le catalogue de vente avec un module “`catalogues`”



AES GRAVE

(H. = Hæberlin, Dr E. J., Æs Grave. Francfort s/M. 1910.
G. = Garrucci, P. R., Le monete dell'Italia antica. Roma 1885.)

1 **Latium et Campanie.** *I.* **Rome. Série librale urbaine, de la première et deuxième période,** *env. 335-286 av. J.-C. As.* Tête de Janus. ℞. Proue à d. Au-dessus I· H. pl. 14-16. Mm. 63, gr. 262,20. Pat. T. B.

*2 *Triens.* Tête de Minerve à g. avec casque corinthien, dessous ····. ℞. Proue à d., dessous ····. H. pl. 17. G. pl. 28, 4. Mm. 43, gr. 90,30. Belle pat. Superbe.

3 *Quadrans.* Tête d'Hercule à g. avec la peau de lion, derrière ···. ℞. Proue à d., dessous ···. H. pl. 18, G. pl. 29, 1. Mm. 41, gr. 65,20. Pat. T. B.

4 *Sextans.* Tête de Mercure à g., dessous ··. ℞. Proue à d., dessous ·· H. pl. 18, 10/21. Mm. 33, gr. 36,61. B.

*5 **Série librale urbaine de la troisième période,** *env. 286 av. J.-C. As.* Tête de Janus. ℞. Proue à g., au-dessus I· H. pl. 19-20. Mm. 63, gr. 235,30. Pat. Superbe.

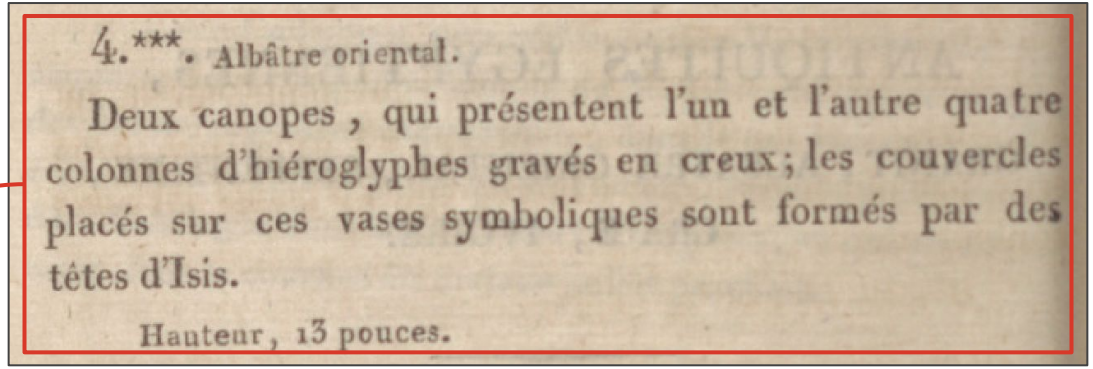
*6 **Latium,** *312-286 av. J.-C.* **Série I.** *Semis.* Taureau bondissant à g., dessous S· ℞. Roue à six rais, dans un des intervalles S· H. pl. 25, 4/6. G. pl. 40, 1. Mm. 30, gr. 167,30. Pat. Rare. Superbe.

- **Personnalisation de la TEI** (ODD) dans la mesure du possible
- Création de **3 nouveaux éléments** spécifiques décrits dans un **module** “`catalogues`” :

`catalogueEntry` : entrée de catalogue, signalée par un niveau de titre

`catalogueDesc` : informations s’appliquant à l’ensemble des notices d’une entrée

`catalogueItem` : notice de catalogue



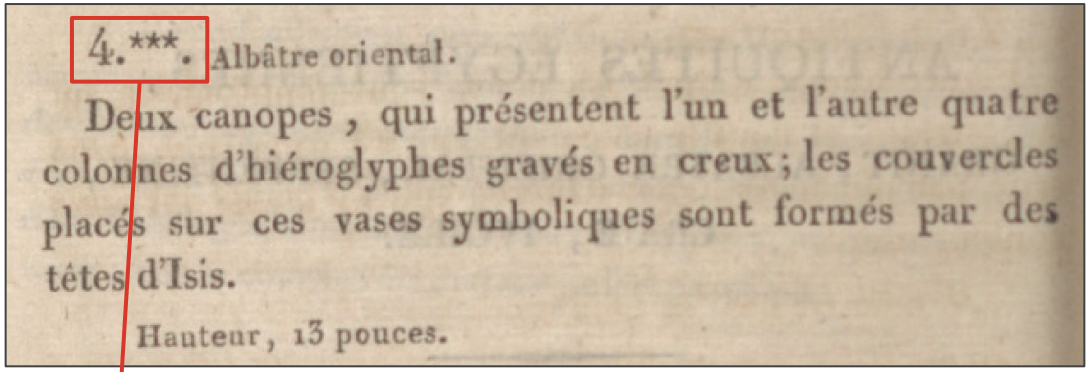
4.***. Albâtre oriental.

Deux canopes, qui présentent l'un et l'autre quatre colonnes d'hiéroglyphes gravés en creux; les couvercles placés sur ces vases symboliques sont formés par des têtes d'Isis.

Hauteur, 13 pouces.

```
<catalogueItem 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>
    </dimensions>
    <!-- <condition/> -->
</catalogueItem>
```

Représentation TEI d'une notice



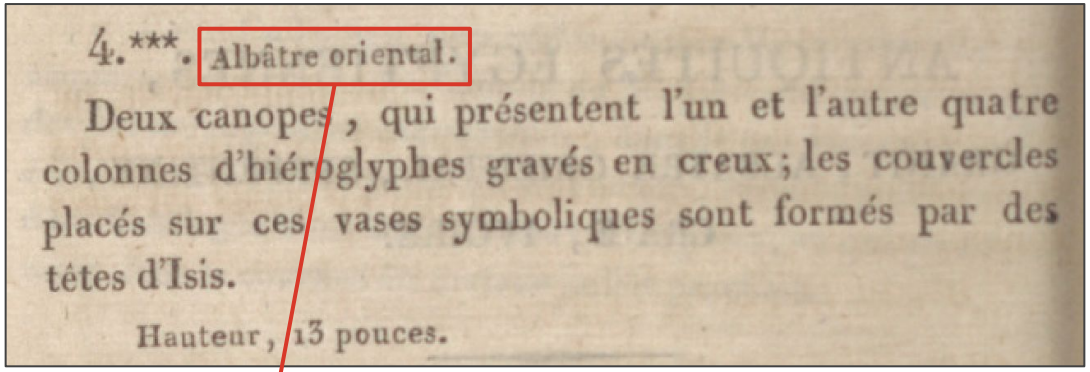
4.***. Albâtre oriental.

Deux canopes, qui présentent l'un et l'autre quatre colonnes d'hiéroglyphes gravés en creux; les couvercles placés sur ces vases symboliques sont formés par des têtes d'Isis.

Hauteur, 13 pouces.

```
<catalogueItem 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>
    </dimensions>
    <!-- <condition/> -->
</catalogueItem>
```

Représentation TEI d'une notice



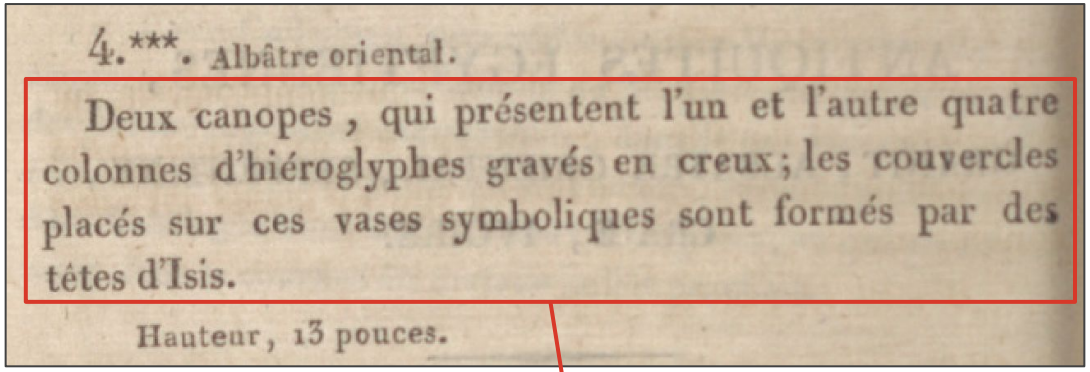
4.***. Albâtre oriental.

Deux canopes, qui présentent l'un et l'autre quatre colonnes d'hiéroglyphes gravés en creux; les couvercles placés sur ces vases symboliques sont formés par des têtes d'Isis.

Hauteur, 13 pouces.

```
<catalogueItem 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>
    </dimensions>
    <!-- <condition/> -->
</catalogueItem>
```

Représentation TEI d'une notice



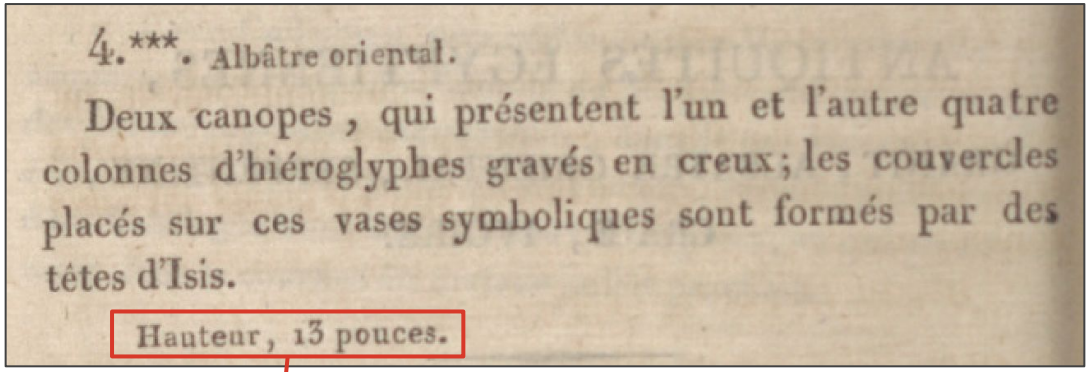
4.***. Albâtre oriental.

Deux canopes, qui présentent l'un et l'autre quatre colonnes d'hiéroglyphes gravés en creux; les couvercles placés sur ces vases symboliques sont formés par des têtes d'Isis.

Hauteur, 13 pouces.

```
<catalogueItem 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>
    </dimensions>
    <!-- <condition/> -->
</catalogueItem>
```

Représentation TEI d'une notice



4.***. Albâtre oriental.

Deux canopes, qui présentent l'un et l'autre quatre colonnes d'hiéroglyphes gravés en creux; les couvercles placés sur ces vases symboliques sont formés par des têtes d'Isis.

Hauteur, 13 pouces.

```
<catalogueItem 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>
    </dimensions>
    <!-- <condition/> -->
</catalogueItem>
```

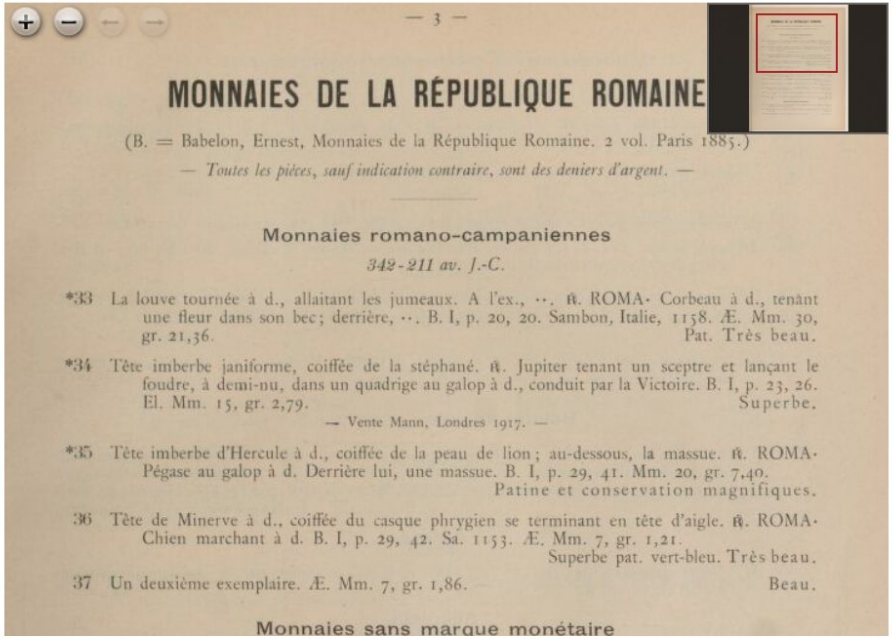
Représentation TEI d'une notice

Publication des catalogues structurés

Choix de l'outil de publication

- Plateforme ***open source*** de publication de fichiers en ligne, basé sur le système de gestion de bases de données XML **eXist-db** (fonctionne avec Docker)
- *Templates* **prédéfinis** pour être accessible à tou·te·s et éviter les répétitions de code
    - ➥ Génère des **applications prêtes à l'usage** pouvant être **personnalisées** et **améliorées**
- Particulièrement **adapté pour les fichiers TEI** et la gestion de **grands corpus**
- Accepte plusieurs types de fichiers en entrée (`.xml` ou `.docx` par exemple)
- Un **outil d'annotation intégré** permet d'encoder en TEI directement avec l'interface graphique

AJOUTER UNE VUE



— 3 —

MONNAIES DE LA RÉPUBLIQUE ROMAINE

(B. = Babelon, Ernest, Monnaies de la République Romaine. 2 vol. Paris 1885.)

— *Toutes les pièces, sauf indication contraire, sont des deniers d'argent.* —

Monnaies romano-campaniennes

342-211 av. J.-C.

*33 La louve tournée à d., allaitant les jumeaux. A l'ex., ··. ℞. ROMA· Corbeau à d., tenant une fleur dans son bec; derrière, ··. B. I, p. 20, 20. Sambon, Italie, 1158. Æ. Mm. 30, gr. 21,36. Pat. Très beau.

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

— Vente Mann, Londres 1917. —

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

36 Tête de Minerve à d., coiffée du casque phrygien se terminant en tête d'aigle. ℞. ROMA· Chien marchant à d. B. I, p. 29, 42. Sa. 1153. Æ. Mm. 7, gr. 1,21. Superbe pat. vert-bleu. Très beau.

37 Un deuxième exemplaire. Æ. Mm. 7, gr. 1,86. Beau.

Monnaies sans marque monétaire

MONNAIES DE LA REPUBLIQUE ROMAINE

(B. = Babelon, Ernest, Monnaies de la République Romaine. 2 vol. Paris 1885.) - Toutes les pièces, sauf indication contraire, sont des deniers d'argent. -

Monnaies romano-campaniennes

342-211 av. J.-C.

- *33 La louve tournée à d., allaitant les jumeaux. A l'ex., -. ROMA. Corbeau à d., tenant une fleur dans son bec; derrière, B. I, p. 20, 20. Sambon, Italie, 1158. AE. Mm. 30, gr. 21,36. Pat. Très beau.
- *34 Tête imberbe janiforme, coiffée de la stéphané. -. Jupiter tenant un sceptre et lançant le foudre, à demi-nu, dans un quadrige au galop à d., conduit par la Victoire. B.

Exemple de visualisation basique d'une page de catalogue de vente avec TEI Publisher (crédit : Hugo Scheithauer)

DiScholEd - Éditions scientifiques numériques






Trier par: Titre | Filtrer selon: Titre | Filtrer

Correspondance de Constance de Salm (1767-1845)

Ce dossier contient l'édition numérique de lettres de Constance de Salm, ainsi que l'inventaire et les index de sa correspondance.

Journal des guerres napoléoniennes

Ce dossier contient le corpus d'un journal de guerre tenu pendant les guerres napoléoniennes.

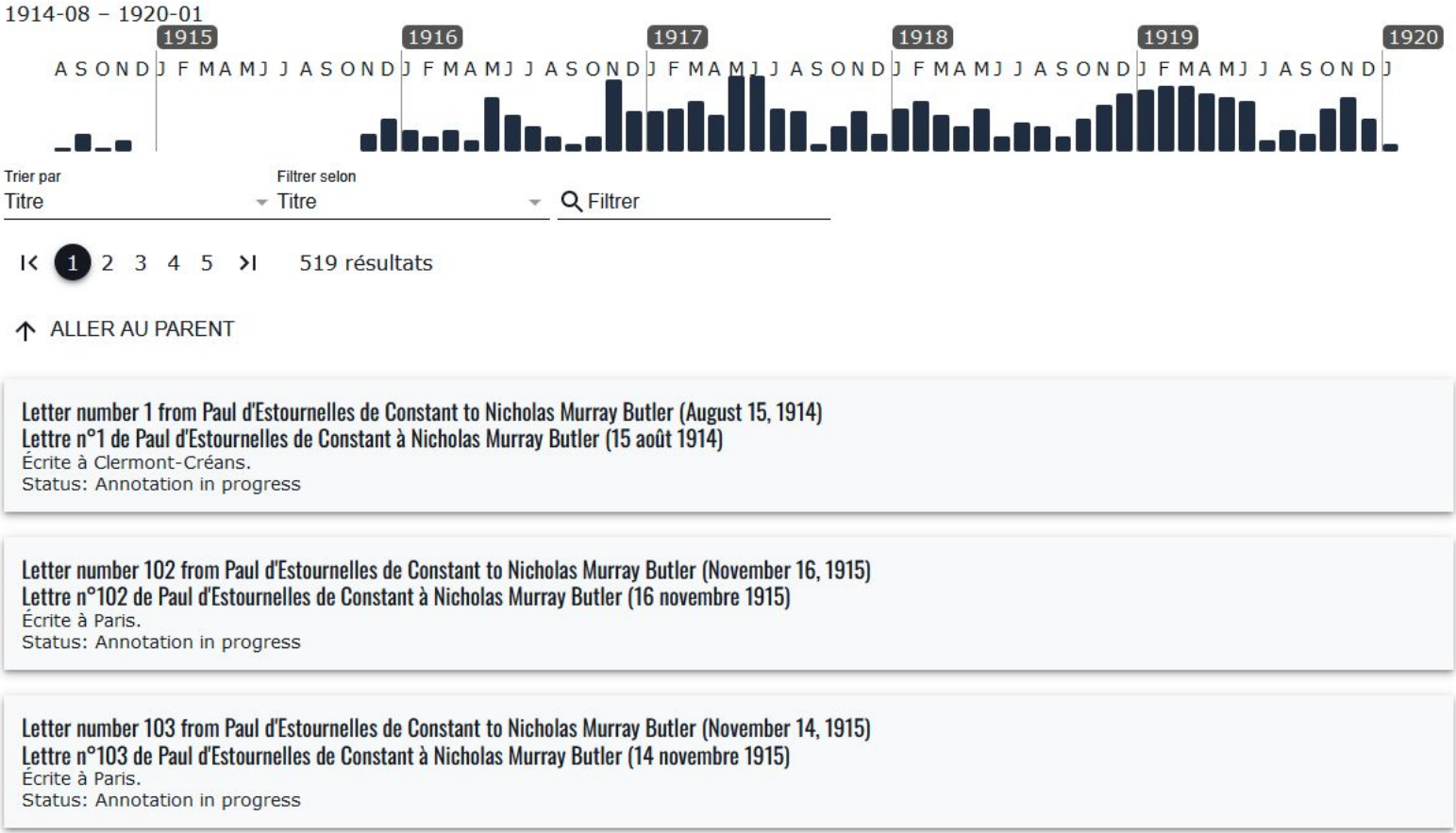
Lettres et textes: Le Berlin intellectuel des années 1800

Ce dossier contient le corpus et les index des lettres et textes des intellectuels berlinois de 1800 à 1830.

August Boeckh - Catalogue de ses manuscrits

Ce dossier contient le catalogue des manuscrits et papiers de Boeckh.

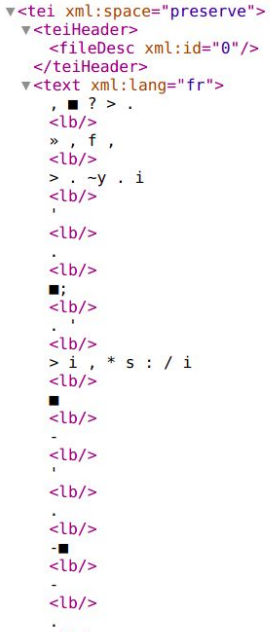
Page d'accueil de l'application DiScholEd (crédits : Floriane Chiffoleau, Manon Ovide, Samuel Scalbert)

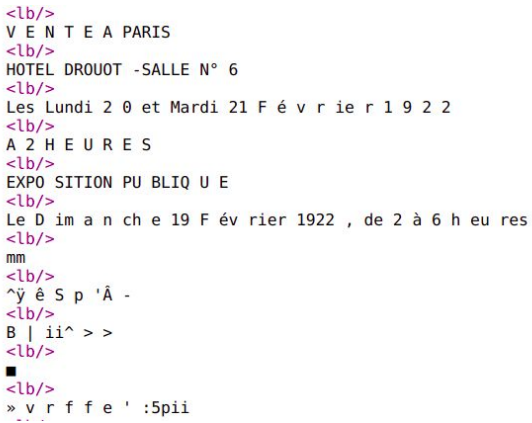


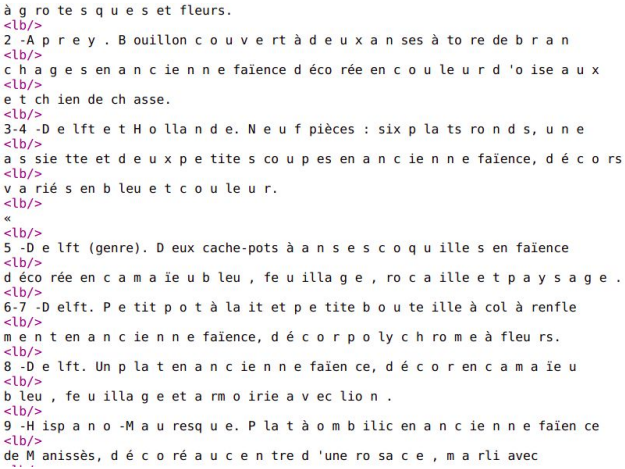
Implémentation du module **`pb-timeline`** au sein de chaque édition (crédit : Samuel Scalbert)

Défis à relever

Corriger les erreurs d'OCR







- Problèmes sur les pages avec deux colonnes, insertion de caractères non existants, décomposition des mots ➜ **répercussions sur l'entraînement** des modèles GROBID
- Entraînement d'un **modèle de détection robuste** avec des données bien annotées

Lier les notices à leurs illustrations respectives

*124 — **DOMITIEN.** (César, 69-81). **Aureus.** Tête laurée à dr. R/. La Santé nourrissant un serpent à dr. (C. 383 - RIC. 243). *Rome*, 79. 7g25. B/T.B. *3.500/4.000*

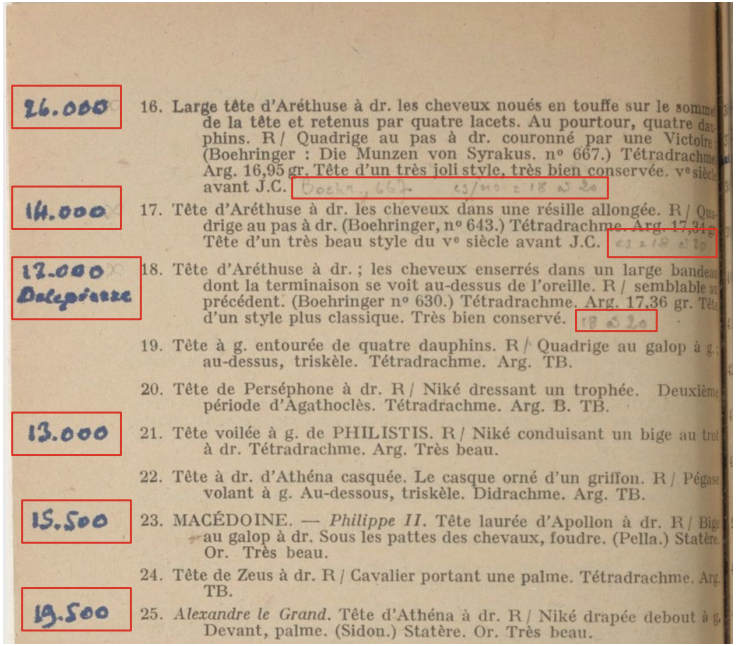
Bourgey, 1992, p.28 (BnF)

- **Lecture augmentée** des documents
- Connexion des notices et de leurs illustrations dans l'encodage TEI pour proposer une **lecture simplifiée**

Bourgey, 1992, p.29 (BnF)

Gérer les segments de texte manuscrits

- **Segmentation, transcription et structuration** des segments de texte manuscrits
- Apportent des **informations importantes** pour la recherche : prix de vente, nom d'un éventuel acquéreur, corrections apportées à la notice, etc.
- Entraînement d'un **modèle d'HTR avec eScriptorium**



Ressources en ligne

DataCatalogue

Organisation GitHub DataCatalogue : https://github.com/DataCatalogue

Présentation du jeu de données sur le site de la BnF : https://api.bnf.fr/fr/node/238

Carnet Hypotheses : https://datacat.hypotheses.org/

Corpus macro-segmenté et annoté sur Roboflow : https://app.roboflow.com/datacatalogue/macro-segmentation/overview

Guide d'annotation DataCatalogue (dérivé de SegmOnto) :
https://github.com/DataCatalogue/datacat-objet-detection-dataset/blob/main/DataCat_AnnotationGuide.md

Chaînes de traitement (1/2)

Chagué, A., Scheithauer, H., Terriel, L., Chiffoleau, F., & Tadjo-Takianpi, Y. (Juillet 2022). *Take a Sip of TEI and Relax: A Proposition for an End-to-End Workflow to Enrich and Publish Data Created with Automatic Text Recognition* [Communication]. Digital Humanities 2022, en ligne. https://inria.hal.science/hal-03739767

Chagué, A., & Romary, L. (2022). L'intelligence artificielle, une ouverture du champ des possibles. *Arabesques*, *107*, 4-5. DOI : 10.35562/arabesques.3043.

Chagué, A., Clérice, T., & Romary, L. (2022). *HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites*. https://inria.hal.science/hal-04124743

Chagué, A., & Chiffoleau, F. (Mars 2021). *An Accessible and Transparent Pipeline for Publishing Historical Egodocuments* [Communication]. What's Past is Prologue, The NewsEye International Conference, en ligne. https://hal.science/hal-03180669

Chiffoleau, F. *Pipeline for Digital Scholarly Editions*. GitHub. https://github.com/DiScholEd/pipeline-digital-scholarly-editions

Chaînes de traitement (2/2)

Frankl, M., Bryant, M., Green, J., Schellenbacher, W., & Sedlická, M. (2018). *Edition of Documents*, EHRI GA no. 654164 D.12.2 [Rapport]. European Holocaust Research Infrastructure. https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/D12%202%20Thematic%20approach%201%20Edition%20of%20documents.pdf

Moufflet, J.-F. (2022). L'intelligence artificielle au service du traitement des archives. *Arabesques*, *107*, 14-15. DOI : 10.35562/arabesques.3088.

Muñoz, T., & Viglianti, R. (2015). Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive. *Journal of the Text Encoding Initiative*, (8), 1-18. DOI : 10.4000/jtei.1270.

Rostaing, A., & Scheithauer, H. (Juin 2022). *Enrichir le patrimoine écrit archivistique grâce aux technologies numériques : ingénierie du projet LectAuRep* [Communication]. DHNord 2022, en ligne. https://hal.science/hal-03792952

Sagot, B., Romary, L., Bawden, R., Ortiz Suárez, P., Gabay, S., Pinche, A., & Camps, J.-B. *Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue*. https://github.com/Gallicorpora (Gallic(orpor)a Pipeline: https://gallicorpora.github.io/pipeline/intro/)

Scheithauer, H., Chagué, A., & Romary, L. (Novembre 2021). *From eScriptorium to TEI Publisher* [Communication]. Brace Your Digital Scholarly Edition!, Berlin. https://inria.hal.science/hal-03538115

Segmentation

Documentation officielle de GROBID : https://grobid.readthedocs.io/en/latest/

Dépôt GitHub GROBID : https://github.com/kermitt2/grobid

[Module GROBID Dictionaries] Gabay, S., Khemakhem, M., & Romary, L. (15 novembre 2018). *Les catalogues et GROBID*. HAL. https://hal.science/cel-01951107.

Documentation officielle de YOLOv8 : https://docs.ultralytics.com/fr/

Dépôt GitHub Yolov8 : https://github.com/ultralytics/ultralytics

En savoir plus sur l'apprentissage supervisé : https://www.ibm.com/fr-fr/topics/supervised-learning

Jeu de données LADaS sur GitHub : https://github.com/DEFI-COLaF/LADaS

Clérice, T. (2023). You Actually Look Twice At it (YALTAi): Using an Object Detection Approach Instead of Region Segmentation within the Kraken Engine. *Journal of Data Mining and Digital Humanities*, 1-13. DOI : 10.46298/jdmdh.9806

Structuration TEI

TEI Guidelines P5 : https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Gabay, S., Topalov, B., Corbières, C., Rondeau du Noyer, L. & Joyeux-Prunel, B. (2021). Artl@s — Extracting Data from Exhibition Catalogues. *Second International Conference of the European Association for Digital Humanities*. https://hal.science/hal-03331838.

Chateau-Dutier, E. & Corbières, C. (26 octobre 2021). Un modèle de contenu <objet> pour l'Histoire de l'Art. *Ledoux Architecture*. http://www.ledoux-architecture.fr/articles/2021/10/26/ledoux-ecrivain.html.

Publication avec TEI Publisher

Documentation TEI Publisher : https://teipublisher.com/exist/apps/tei-publisher/documentation

Chiffoleau, F. (4 décembre 2020). Publication of my Digital Edition – Working with TEI Publisher. *Digital Intellectuals*. https://digitalintellectuals.hypotheses.org/3912

Chiffoleau, F. (16 juin 2021). Publication of my Digital Edition – Developing my TEI Publisher Application. *Digital Intellectuals*. https://digitalintellectuals.hypotheses.org/4173

Chiffoleau, F. (10 décembre 2021). Publication of my Digital Edition – Online Launch of the TEI Publisher Application. *Digital Intellectuals*. https://digitalintellectuals.hypotheses.org/4399

Pierazzo, E. (2019). What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter. *International Journal of Digital Humanities*, *1*(2), 209-220. DOI : 10.1007/s42803-019-00019-3.

Scalbert, S. (29 juin 2023). Overcoming Challenges in DiScholEd's Development: A Journey of Problem-Solving and Design Enhancements. *Digital Intellectuals*. https://digitalintellectuals.hypotheses.org/4948

Scalbert, S. (2023). *Le CMS et le low-code au service des humanités numériques : l'exemple de DiScholEd, une application TEI Publisher* [Mémoire de Master 2]. École nationale des chartes. https://github.com/Samuel-Scalbert/-Memoire-TNAH/blob/main/M%C3%A9moire_TNAH_Scalbert.pdf

Exemples d’applications TEI Publisher

TEI Publisher Project Registry : https://www.e-editiones.org/map/

DiScholEd : https://dischole d.huma-num.fr/exist/apps/dischole d/index.html

Van Gogh Letters : https://teipublisher.com/exist/apps/vangogh/index.html

Surnaturel au Moyen-Âge (🇩🇪) : https://daemon.libripendis.eu/

Shakespeare’s Plays : https://teipublisher.com/exist/apps/shakespeare-pm/index.html

When the Wall Came Down : https://teipublisher.com/exist/apps/dodis-facets/index.html

TraveLab : https://teipublisher.info/exist/apps/TraveLab/Benjamin%20of%20Tudela.xml

Solomon Project : https://tei.mittelalter.uni-tuebingen.de/exist/apps/salomon/Salomon3.xml

Johann Conrad Fischer (🇩🇪) : https://johannconradfischer.com/de/travels

Merci pour votre attention !

Des questions ?

Contact: sarah.beniere[at]inria.fr

Support inspiré par Hugo Scheithauer