



HAL
open science

DataCatalogue : rétro-structuration automatique des catalogues de vente

Hugo Scheithauer, Sarah Bénéière, Jean-Philippe Moreux, Laurent Romary

► To cite this version:

Hugo Scheithauer, Sarah Bénéière, Jean-Philippe Moreux, Laurent Romary. DataCatalogue : rétro-structuration automatique des catalogues de vente. Webinaire Culture-Inria, Ministère de la Culture, Nov 2023, Paris, France. hal-04360229

HAL Id: hal-04360229

<https://hal.science/hal-04360229>

Submitted on 21 Dec 2023

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

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

DataCatalogue : rétro-structuration automatique des catalogues de vente

Hugo Scheithauer¹, **Sarah Bénéière**¹, **Jean-Philippe Moreux**², **Laurent Romary**¹

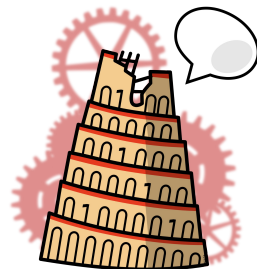
¹ ALMAnaCH - Automatic Language Modelling and ANALysis & Computational Humanities, Inria Paris

² Bibliothèque Nationale de France

Webinaire Culture-Inria - 29-11-23

{ BnF

Inria



Institut
national
d'histoire
de l'art

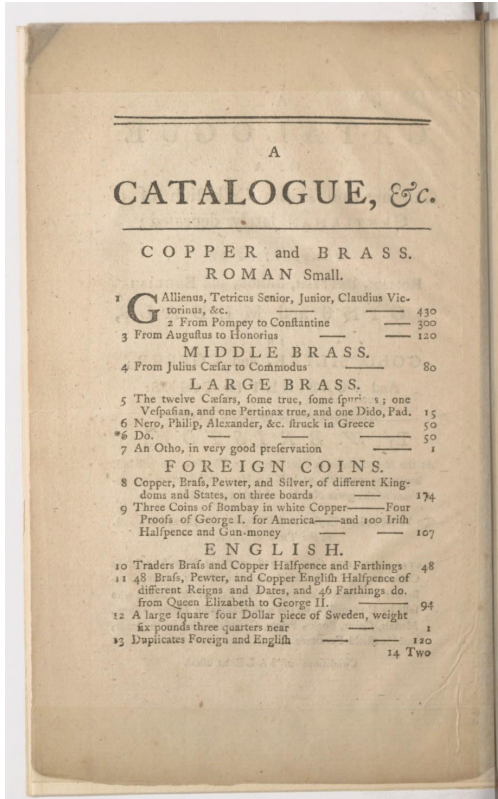


DataCatalogue : objectifs du projet

- Passer d'une **numérisation** à une **base de données textuelle et requêtable**.
- **Segmenter** les catalogues de vente et attribuer à chaque niveau d'information une **étiquette** : entrées de catalogues, numéro des entrées, description des objets, matériaux, sommes monétaires, etc.
- Produire, à partir des zones segmentées, un **encodage XML-TEI** des catalogues de vente.
- Mettre à disposition des chercheurs le corpus encodé dans une **interface de publication** permettant de **requêter** sur les zones segmentées.

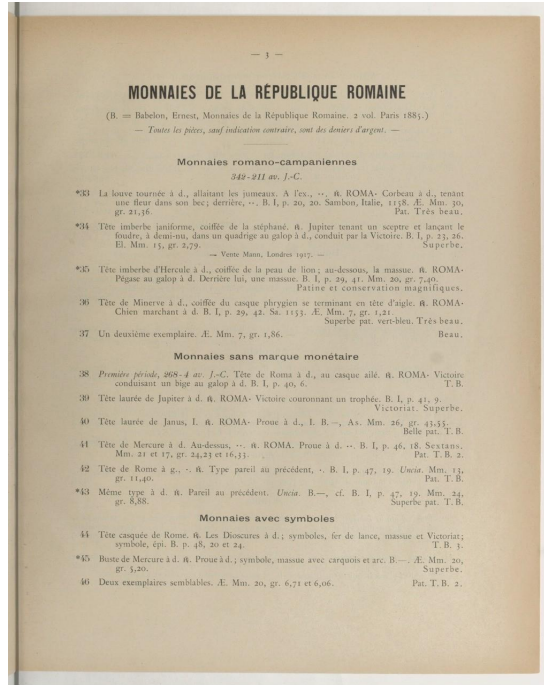


Les catalogues de vente : homogénéité de mise en page



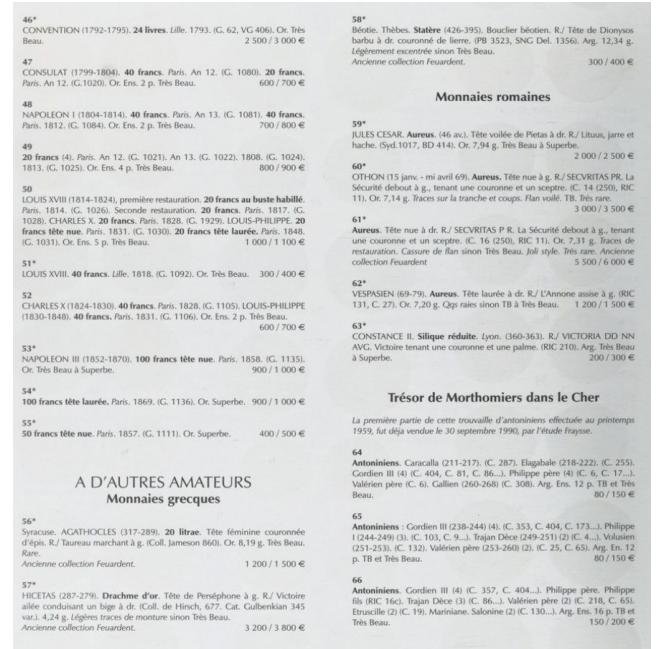
Source gallica.bnf.fr / Bibliothèque nationale de France

Whiston Bristow, 1762.



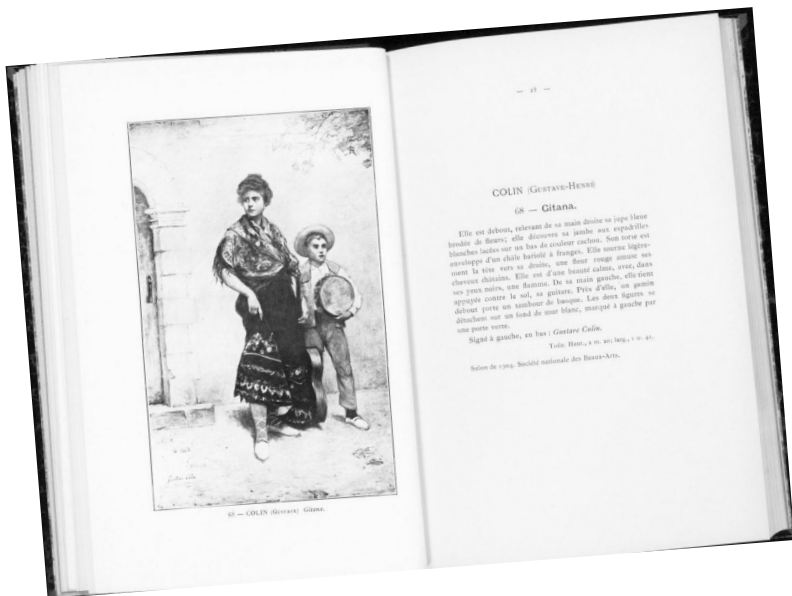
Source gallica.bnf.fr / Bibliothèque nationale de France

Lucien Naville, 1924.

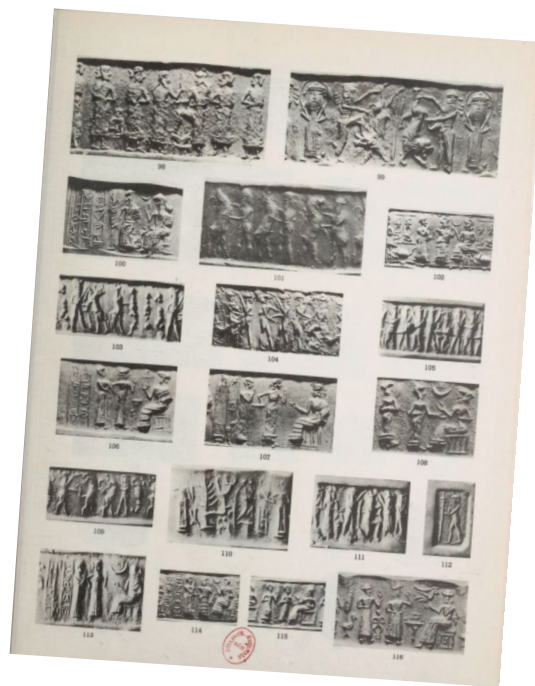


Fraysse & Associés, 2011.

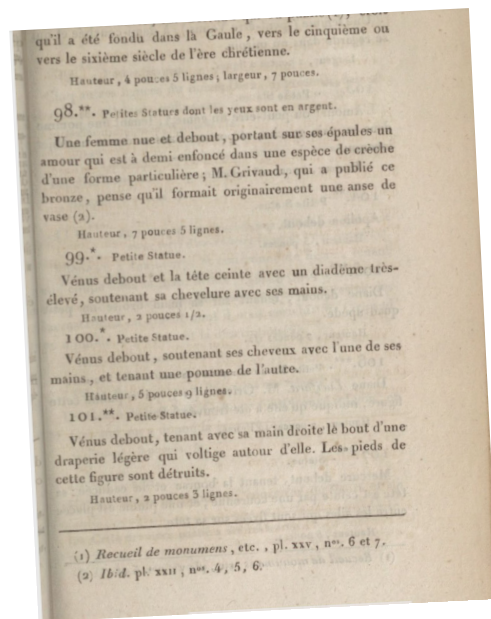
Les catalogues de vente : homogénéité de mise en page (vraiment ?)



Lair-Dubreuil, 1919 (INHA).



Bourgey, 1962 (BnF).



Dubois, 1820 (BnF).

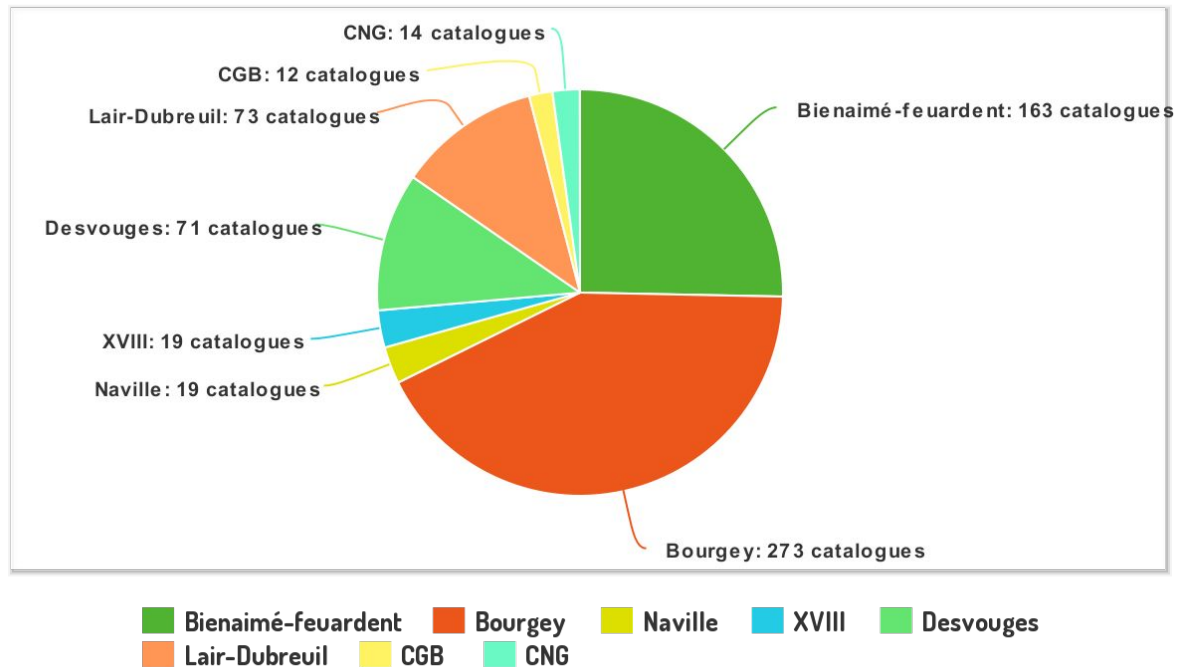
Un premier échantillonnage des catalogues de vente de la BnF et de l'INHA : 4 siècles de contenus structurés

Siècles représentés : XVIIIe,
XIXe, XXe, XXIe.

Langues : ~95% en français.

Types de vente :
numismatique, livres,
antiquités, objets d'art, objets
de luxe

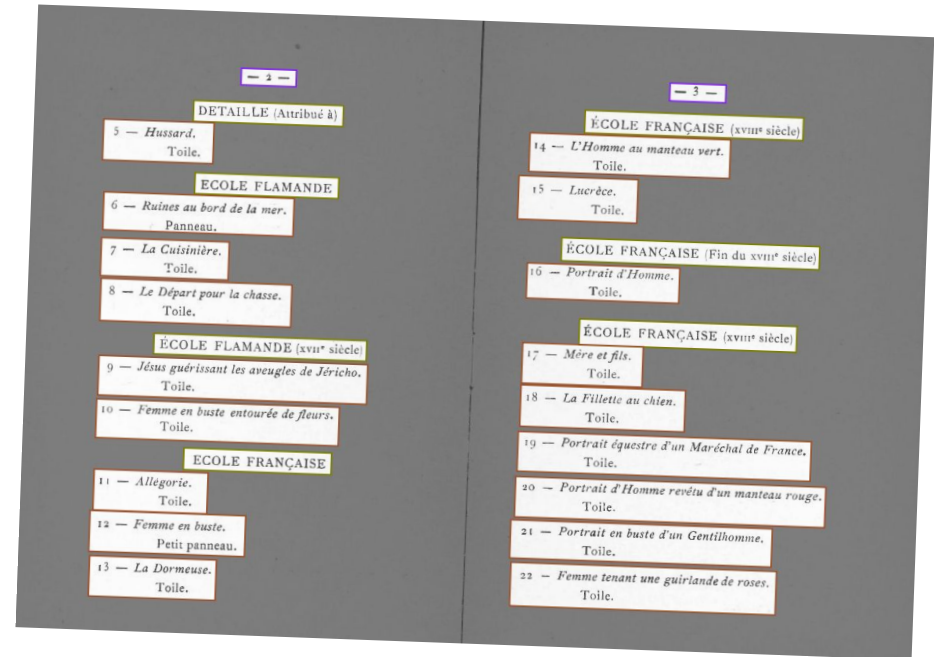
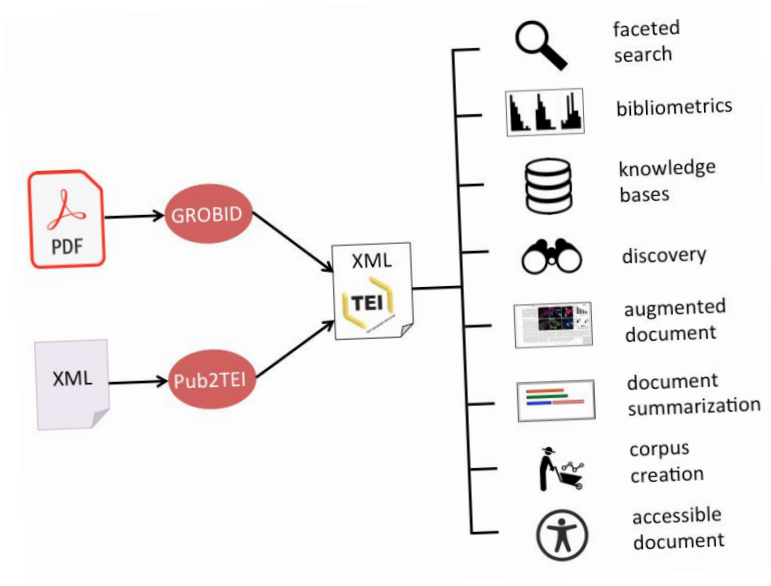
RÉPARTITION DU CORPUS



Quelles technologies pour gérer l'hétérogénéité de contenu et la diachronie des catalogues de vente ?

GROBID (GeneRation of Bibliographic Data)

Détection automatique d'objet



Quelles technologies pour gérer l'hétérogénéité de contenu et la diachronie des catalogues de vente ?

*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.

numéro d'items

*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.

description de l'objet mis en vente

informations bibliographiques

B. I, p. 29, 41.

quantités et mesures

Mm. 20, gr. 7,40.

observations de conservation

Patine et conservation magnifiques.

Quelles technologies pour gérer l'hétérogénéité de contenu et la diachronie des catalogues de vente ?

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

Lair-Dubreuil, 1926 (INHA).

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 de cygne. Sur chaque vaisseau est représenté un homme debout et drapé, placé à la proue; un rameur est à l'arrière du navire; vers l'anse est une sirène placée sur un rocher, et retournant la tête vers les vaisseaux; plus loin est un dauphin. **NIKOSΘENES ΕΡΟΙΕΙ.** *Nicosthènes a fait.*

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.

Rollin, 1836 (BnF).

1	English, Irish, Manx, American, Swedish copper and leaden casts	_____	45 0. 3. 0
2	Ditto	_____	65 0. 6. 6
3	Ancient seals and casts	_____	22 1. 11. 6
4	Four drawers, middle roman brads	_____	140 0. 14. 0
5	Ditto, large and small, from Cæsar to Constantine	_____	140 1. 14. 0

Langford, 1772 (BnF).

- 86 — **Drachme.** Mêmes types variés.-(166-88). **Drachme.** Tête radiée à dr. R/. Rose.- **CARIE.** HALICARNASSE. (1^{er} 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).

Standardiser la structuration des catalogues de vente



- Standard XML pour l'**édition de texte**, permet de rendre **lisible** et **compréhensible** un texte par un ordinateur.
- **Structuration** des catalogues en XML TEI, basé sur des **éléments existants** et standardisés, ainsi que sur des **nouveaux éléments** créés pour modéliser précisément l'objet catalogue, notamment la **notice**. Une notice structurée doit rendre compte de la matérialité de l'objet vendu.
- Une **notice** de catalogue décrit un objet matériel mis à la vente selon un ensemble de normes appartenant au monde de la **vente aux enchères**, ainsi que de la **discipline** dont provient l'objet (numismatique, archéologie, etc.).
- Le module TEI DataCatalogue se base sur les travaux de modélisation du projet **Artl@s** et du projet **Ledoux**, et cherche à **affiner** la **représentation des catalogues**.

- Simon Gabay, Barbara Topalov, Caroline Corbières, Lucie Rondeau Du Noyer, Béatrice Joyeux-Prunel, et al.. Automating Artl@s – extracting data from exhibition catalogues. EADH 2021 - *Second International Conference of the European Association for Digital Humanities*, Sep 2021, Krasnoyarsk, Russia. ([hal-03331838](#))
- Emmanuel Chateau Dutier, Caroline Corbières. A broader <object> content model for art history. Next Gen TEI, 2021 - TEI Conference and Members' Meeting, Oct 2021, Virtual, United States. ([hal-03654979](#))

D'une image à une base de données XML

COLLECTION D'UN AMATEUR

1*

PHILIPPE IV le Bel (1285-1314). **Denier d'or à la masse**. 1^{ère} ém. Le roi assis de f., couronné, tenant un sceptre et un lis, dans un polylobe tréflé cantonné d'annelets. R./ Croix feuillue et fleuronée. Quadrilobe en cœur. (Dy. 208, L. 212). 6,96 g. Superbe. 12 000 / 15 000 €

2*

Agnel d'or. Agneau Pascal à g., nimbé, détournant la tête vers une croix fleurdelisée ornée d'une bannière. A l'exergue : PH'REX. R./ Croix fleuronée dans une rosace cantonnée de quatre lis. (Dy. 212, L. 216). 3,69 g. *Très léger coup* sinon Superbe. 2 000 / 2 500 €

3*

CHARLES IV le Bel (1322-1328). **Royal d'or**. Le roi debout, tenant un long sceptre, sous un dais gothique. R./ Croix fleuronée dans une rosace quadrilobée. (Dy. 240, L. 244). 4,14 g. *Légers coups sur la tranche* sinon Très Beau. 1 500 / 1 800 €

4*

PHILIPPE VI de Valois (1328-1350). **Royal d'or**. Même description mais avec la légende de droit au nom de Philippe. Annelet initial. (Dy. 247, L. 251). 4,92 g. Superbe. 1 200 / 1 500 €

```
<catalogueEntry>
  <catalogueDesc>
    <head>Collection d'un amateur</head>
  </catalogueDesc>
  <!-- ... -->
  <catalogueItem>
    <altIdentifier>
      <idno>2</idno>
    </altIdentifier>
    <metamark>*</metamark>
    <objectDesc>
      <supportDesc>
        <support>Agnel d'or.</support>
      </supportDesc>
    </objectDesc>
    <decoDesc>
      <ab>Agneau Pascal à g., nimbé, détournant la tête vers une
        croix fleurdelisée ornée d'une bannière. A l'exergue:
        PH'REX. R./ Croix fleuronée dans une rosace cantonnée de quatre
        lis.</ab>
    </decoDesc>
    <objectDesc>
      <supportDesc>
        <support>( <measure>Dy. 212</measure>,
          <measure>L. 216</measure>).
          <measure>3,69 g.</measure></support>
        <condition>Très léger coup sinon Superbe<p></p></condition>
      </supportDesc>
    </objectDesc>
    <num type="currency">2000 / 2500 euros</num>
  </catalogueItem>
  <!-- ... -->
</catalogueEntry>
```

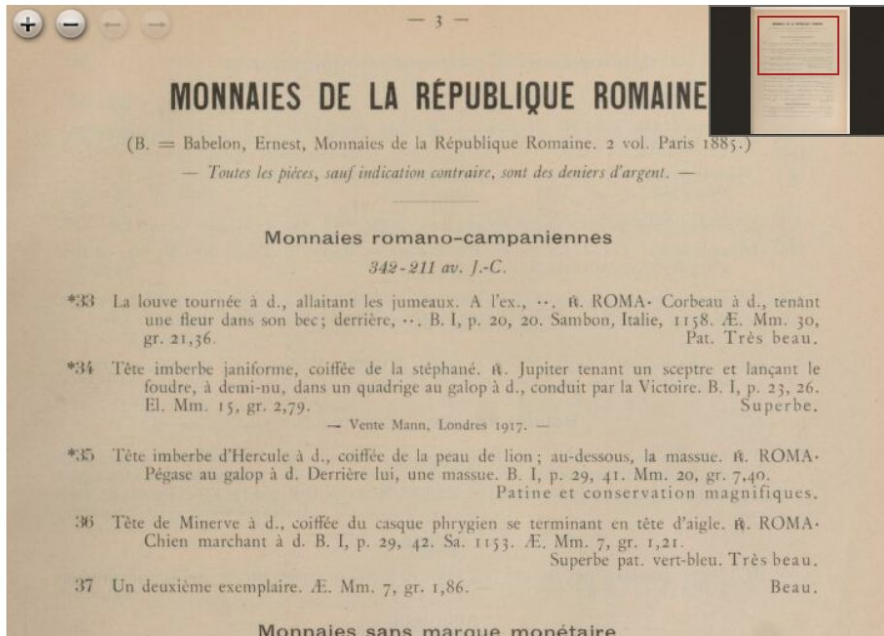
Publier les catalogues de vente structurés



- Plateforme de **publication de fichier TEI**
- *Open source*
- Permet d'**indexer** un corpus, de le **visualiser**, et de le rendre **requêtable**
- Interface entièrement **personnalisable**
- Expertise au sein de l'équipe ALMAAnaCH, Inria



AJOUTER UNE VUE



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

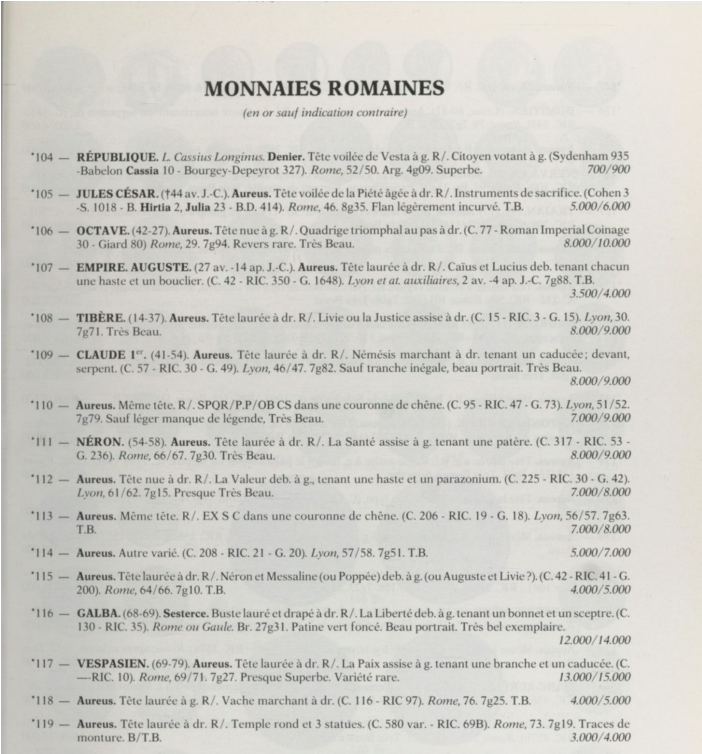
342-211 av. J.-C.

- *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. 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 quadrigé au galop à d., conduit par la Victoire. B. I, p. 23, 26. El. Mm. 15, gr. 2,79. Superbe.

Exemple de visualisation basique avec TEI Publisher d'un catalogue de vente encodé en TEI

DataCatalogue - v1 : bilan

- Modèle de données TEI pour représenter les catalogues de vente
- Annotation et entraînement de modèles GROBID pour la macro-structuration des catalogues, et la segmentation des notices
- Premières réflexions sur un modèle d'édition pour les catalogues de vente avec TEI Publisher
- Identification des défis à relever pour passer à l'échelle avec GROBID



MONNAIES ROMAINES
(en or sauf indication contraire)

*104 —	RÉPUBLIQUE. L. Cassius Longinus. Denier. Tête voilée de Vesta à g. R/. Citoyen votant à g. (Sydenham 935 -Babelon Cassia 10 - Bourgey-Depeyrot 327). <i>Rome, 52/50. Arg. 4g09. Superbe.</i>	700/900
*105 —	JULES CÉSAR. (144 av. J.-C.). Aureus. Tête voilée de la Piété âgée à dr. R/. Instruments de sacrifice (Cohen 3 -S. 1018 - B. Hirtia 2, Julia 23 - B.D. 414). <i>Rome, 46. 8g35. Flan légèrement incurvé. T.B.</i>	5.000/6.000
*106 —	OCTAVE. (42-27). Aureus. Tête nue à g. R/. Quadrigé triomphal au pas à dr. (C.77 - Roman Imperial Coinage 30 - Gard 80) <i>Rome, 29. 7g94. Revers rare. Très Beau.</i>	8.000/10.000
*107 —	EMPIRE. AUGUSTE. (27 av. -14 ap. J.-C.). Aureus. Tête laurée à dr. R/. Caius et Lucius deb. tenant chacun une haste et un bouclier. (C.42 - RIC. 350 - G. 1648). <i>Lyon et at. auxiliaires, 2 av. -4 ap. J.-C. 7g88. T.B.</i>	3.500/4.000
*108 —	TIBÈRE. (14-37). Aureus. Tête laurée à dr. R/. Livie ou la Justice assise à dr. (C.15 - RIC. 3 - G. 15). <i>Lyon, 30. 7g71. Très Beau.</i>	8.000/9.000
*109 —	CLAUDE 1^{er}. (41-54). Aureus. Tête laurée à dr. R/. Némésis marchant à dr. tenant un caducée; devant, serpent. (C.57 - RIC. 30 - G. 49). <i>Lyon, 46/47. 7g82. Sauf tranche inégale, beau portrait. Très Beau.</i>	8.000/9.000
*110 —	Aureus. Même tête. R/. SPQR/P.P/OB CS dans une couronne de chêne. (C.95 - RIC. 47 - G. 73). <i>Lyon, 51/52. 7g79. Sauf léger manque de légende, Très Beau.</i>	7.000/8.000
*111 —	NÉRON. (54-58). Aureus. Tête laurée à dr. R/. La Santé assise à g. tenant une patère. (C.317 - RIC. 53 - G. 236). <i>Rome, 66/67. 7g30. Très Beau.</i>	8.000/9.000
*112 —	Aureus. Tête nue à dr. R/. La Valeur deb. à g., tenant une haste et un parazonium. (C.225 - RIC. 30 - G. 42). <i>Lyon, 61/62. 7g15. Presque Très Beau.</i>	7.000/8.000
*113 —	Aureus. Même tête. R/. EX S C dans une couronne de chêne. (C.206 - RIC. 19 - G. 18). <i>Lyon, 56/57. 7g63. T.B.</i>	7.000/8.000
*114 —	Aureus. Autre varié. (C.208 - RIC. 21 - G. 20). <i>Lyon, 57/58. 7g51. T.B.</i>	5.000/7.000
*115 —	Aureus. Tête laurée à dr. R/. Néron et Messaline (ou Poppée) deb. à g. (ou Auguste et Livie?). (C.42 - RIC. 41 - G. 200). <i>Rome, 64/66. 7g10. T.B.</i>	4.000/5.000
*116 —	GALBA. (68-69). Sesterce. Buste lauré et drapé à dr. R/. La Liberté deb. à g. tenant un bonnet et un sceptre. (C.130 - RIC. 35). <i>Rome ou Gaule. Br. 27g31. Patine vert foncé. Beau portrait. Très bel exemplaire.</i>	12.000/14.000
*117 —	VESPASIEN. (69-79). Aureus. Tête laurée à dr. R/. La Paix assise à g. tenant une branche et un caducée. (C. — RIC. 10). <i>Rome, 69/71. 7g27. Presque Superbe. Variété rare.</i>	13.000/15.000
*118 —	Aureus. Tête laurée à g. R/. Vache marchant à dr. (C.116 - RIC 97). <i>Rome, 76. 7g25. T.B.</i>	4.000/5.000
*119 —	Aureus. Tête laurée à dr. R/. Temple rond et 3 statues. (C.580 var. - RIC. 69B). <i>Rome, 73. 7g19. Traces de monture. B/T.B.</i>	3.000/4.000

Bourgey, 1992 (BnF).

Des défis à relever pour DataCatalogue v2

```
▼<tei xml:space="preserve">
▼<teiHeader>
  <fileDesc xml:id="0"/>
  </teiHeader>
▼<text xml:lang="fr">
  , ■ ? > .
  <lb/>
  » , f ,
  <lb/>
  > . -y . i
  <lb/>
  '
  <lb/>
  .
  <lb/>
  ■;
  <lb/>
  '
  <lb/>
  > i , * s : / i
  <lb/>
  ■
  <lb/>
  -
  <lb/>
  '
  <lb/>
  .
  <lb/>
  -■
  <lb/>
  -
  <lb/>
  '
  <lb/>
```

```
<lb/>
V E N T E A PARIS
<lb/>
HOTEL DROUOT -SALLE N° 6
<lb/>
Les Lundi 2 0 et Mardi 21 F é v r i e r 1 9 2 2
<lb/>
A 2 H E U R E S
<lb/>
EXPO SITION PU BLIQU E
<lb/>
Le D im a n ch e 19 F é v r i e r 1922 , de 2 à 6 h eu res
<lb/>
mm
<lb/>
^y ê S p 'Â -
<lb/>
B | ii^ > >
<lb/>
■
<lb/>
» v r f f e ' :5pii
...

```

```
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
<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 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
...

```

Exemples “extrêmes” des scorries d’OCR rencontrés dans les PDF numérisés :
caractères non existants dans les documents originaux, mots décomposés. Ces erreurs ont des effets de cascades sur l’entraînement et l’inférence des modèles.

Des défis à relever pour DataCatalogue v2

*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

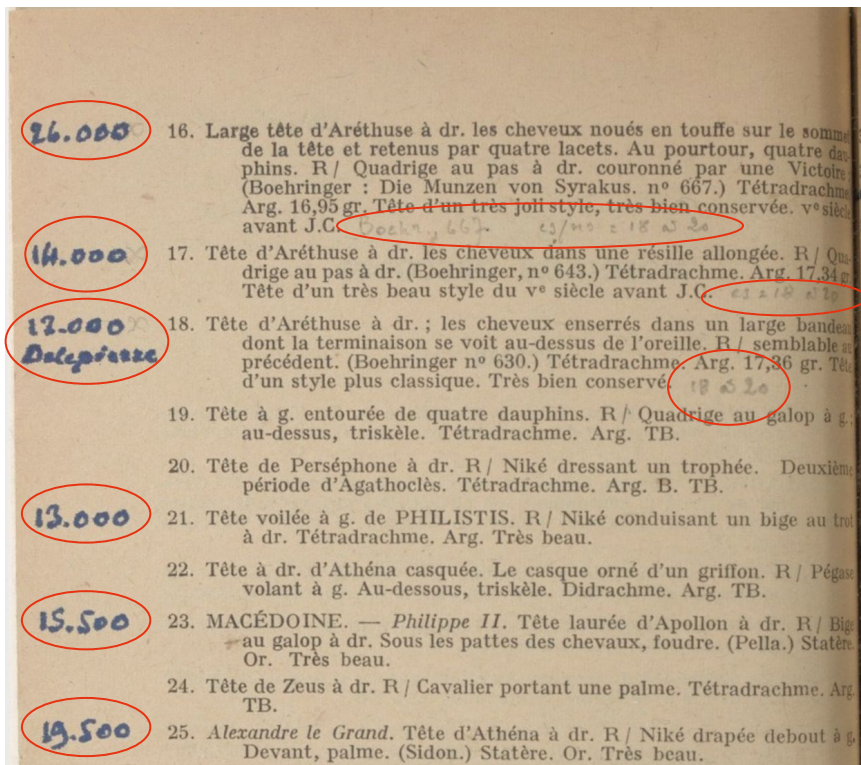
Bourgey, 1992, p. 28 (BnF).



Bourgey, 1992, p. 29 (BnF).

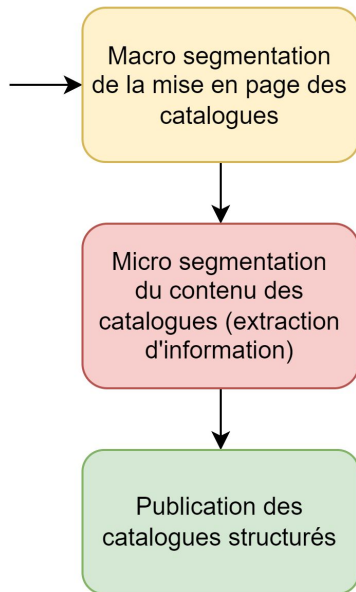
→ Connecter les notices avec leurs illustrations respectives et en rendre compte dans l'encodage TEI

Des défis à relever pour DataCatalogue v2



Segmentation, transcription et structuration des segments de texte manuscrits

DataCatalogue v2 - 2023-2024



{ BnF

Inria

Institut national d'histoire de l'art



- Recrutement de Sarah Bénéière dans l'équipe ALMA_{na}CH, Inria, en tant qu'ingénieure recherche et développement
- Création d'un projet doctoral sur l'analyse automatique de mise en page mené par Hugo Scheithauer (Paris PhD)

Collaboration et réutilisation des données

- Campagne d'annotation de la macro-structure en coordination avec le projet **COIaF** (Inria, Multispeech)
- Schéma d'annotation suivant l'ontologie **SegmOnto**
- Corpus annotés et développements mis à disposition en ligne sur GitHub et Roboflow, selon les principes de la science ouverte
- Collaboration pluri-institutionnelles entre Inria, la BnF et l'INHA. L'idée est également de créer des corpus annotés et des outils facilement réutilisables par d'autres institutions

<https://github.com/DataCatalogue>



<https://segmonto.github.io/>



<https://colaf.huma-num.fr/>



<https://htr-united.github.io/>

Breaking news: [#prixscienceouverte](#)
[#donneesrecherche](#) Catégorie Prix jeunes
chercheurs - 2023-11-29

Merci pour votre attention !

Ressources :

- GitHub DataCatalogue : <https://github.com/DataCatalogue>
- https://huggingface.co/spaces/HugoSchtr/DataCat_Yolov5
- Corpus de macro-segmentation des catalogues :
<https://app.roboflow.com/datacatalogue/macro-segmentation/overview>

Contacts :

sarah.beniere[at]inria.fr
hugo.scheithauer[at]inria.fr

