Automatic retro-structuration of auction sales catalogs layout and content

Hugo Scheithauer, Sarah Bénière, Laurent Romary

Call for Proposals - DH2024 - Reinvention & Responsibility

Automatic retro-structuration of auction sales catalogs layout and content

Hugo Scheithauer[1], Sarah Bénière[1], Laurent Romary[2]

[1] ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities, Inria Paris
[2] Inria, Directorate for Scientific Information and Culture

Short paper proposal

December 2023

Abstract

Sales catalogs allow us to approach historical physical materials, often missing from public collections. They constitute vast knowledge bases on the history of economics, collections, art history, iconography, and cultural heritage in general. Catalogs are highly structured printed publications that provide a factual description of every item that was sold during an auction—usually as a list of entries followed by their illustrations (fig. 1). Following their digitization and automatic transcription, catalogs are redistributed to users as pairs of images and unstructured plain text that shows limitations in extracting and processing the information it contains. Without structured data, users have access to them only at the page level and not directly at the content level. This short paper proposal aims at presenting the retro-structuration pipeline established within the framework of the DataCatalogue research project (ALMAnaCH, Inria, Paris; French National Library, or BnF; French National Institute for Art History, or INHA) to address this issue.[1] We will also advocate for a generalized use of layout segmentation, which is still missing in most digitization pipelines, as it allows to explore digitized textual documents on a broader scale with more precision and to store the semantics of the layout, which is intricately linked to the textual content.

[1] DataCatalogue is an ongoing project, funded by Inria and the French Ministry of Culture. It began with an experimental phase between 2021 and 2022, and was renewed for a second phase in 2023. See the DataCatalogue Github organization: https://github.com/DataCatalogue.

Totalling over 28,000 digitized catalogs, the digital collections from the BnF and the INHA are still growing every year, emphasizing the need for an automated pipeline to process such voluminous data. We are building our pipeline with a sample of 713 catalogs, mostly in French, representative of the overall collections in terms of chronological diversity (18th-21st century) and types of sale (e.g. numismatics, books, antiques, works of art).

Our structuration pipeline is built upon document layout analysis (DLA) and information extraction technologies. State-of-the-art DLA systems aim at detecting the layout hierarchy of a document and are currently based on two approaches: visual features with the use of pre-trained convolutional networks (Sven & Matteo, 2022); textual and visual features, with for example the transformer-based model LayoutLMv3 (Huang et al., 2022), or the conditional random fields-based (CRF) software GROBID (Lopez et al., 2021).[2] Information extraction systems leverage transformers or recurrent neural networks for named entity recognition (NER), or CRF models (Vajjala & Balasubramaniam, 2022). They allow the extraction of structured information, such as persons or locations names.

We are working with a hybrid retro-structuration pipeline, composed of two steps: the macro-segmentation relies on automatic object detection, while the micro-segmentation uses text sequence labeling models. The first step consists in accurately segmenting the layout of the catalogs and identifying three distinct sections with their own layout: the front matter (e.g. title page and general conditions of sale), the body (e.g. catalog entries and illustrations), and the back matter (e.g. table of contents and index). Macro-segmentation is only concerned with the layout of the document, and not the textual content itself. Once the layout objects are identified, the micro-segmentation step focuses on the deep semantic information levels by labeling all segments of textual information—in our case catalog entries.

To begin with, we experimented with a structuration pipeline using GROBID. However, we encountered performance issues due to the quality of the text acquired with OCR—like characters hallucinated by the transcription model when processing the page, or deconstructed words lacking proper tokenization. The quality of the transcription was therefore not consistently sufficient to create reliable textual annotated training data. As object detection models demonstrated their reliability for

[2] See the GitHub repository of GROBID: https://github.com/kermitt2/grobid.

segmenting the layout of textual documents and gained increasing popularity in digital humanities projects (Clérice, 2022; Sven & Matteo, 2022), we opted for this text-free approach and annotated the sampled dataset in order to train a dedicated macro-segmentation model for sales catalogs. The DataCatalogue macro-segmentation dataset was created by selecting two random pages from every sampled catalog, resulting in 1,426 images to annotate. It was also designed as a subset for a bigger layout annotation campaign led by the COLaF (Inria, ALMAnaCH, Multispeech) project.[3] The annotation schema is based on the SegmOnto controlled vocabulary (Gabay, Camps, et al., 2021), making the DataCatalogue subset interoperable and reusable. We then fine-tuned a pre-trained YOLOv8 model that showed satisfying performances when trained on 750 annotated images, with an mAP of 67.1%, a precision of 65.3% and a recall of 62.4% (fig. 2).[4] The model allows us to target catalog entries, and initiates the micro-segmentation step with GROBID models and state-of-the-art NER models. We are currently setting up the annotation campaign with the creation of a dedicated schema for annotating catalog entries, for example: item numbers, description, measures, conservation observations, etc. Based on the layout segmentation results, we can also detect noisy textual segments and clean them with OCR models.

A substantial part of the sales catalogs contains handwritten annotations that were not included in the original digitization campaign, although they give valuable information about the sale prices, the buyers' names, etc (fig. 3). After detecting them with our YOLOv8 model, we transcribe them using a fine-tuned HTR model (Chagué et al., 2023; Chagué & Clérice, 2022), and link them to their corresponding entries.

We also argue for a systematic use of the TEI standard when dealing with the automatic structuration of historical documents to create interoperable data. Once the segmentation is complete, transformation scripts convert inferred data into a structured TEI XML representation. Building upon previous projects suggesting new TEI elements for properly modeling the macro- and micro-structure of exhibition catalogs and their objects, we created a refined customized TEI schema for

[3] See the COLaF project website: https://colaf.huma-num.fr/projet/. The dataset will be published in a future public communication.
[4] See Ultralytic's GitHub repository for the YOLOv8 model: https://github.com/ultralytics/ultralytics. Our annotated dataset and fine-tuned YOLOv8 model is publicly available on Roboflow: https://app.roboflow.com/datacatalogue/macro-segmentation/overview.

encoding sales catalogs, which focuses on the materiality of the objects solds during auctions and captures the corresponding microstructures (fig. 4 and 5) (Dutier & Corbières, 2021; Gabay, Topalov, et al., 2021; Nelson, 2016).

Finally, we chose the TEI Publisher platform to publish a digital edition of the structured catalogs. TEI Publisher provides a fully customizable publication environment, well-known in the digital humanities community. The publication model is constantly evolving with the help of the partnering institutions, and aims at proposing an interface suitable to pave new ways of studying sales catalogs, alongside the restructured data.

MONNAIES GAULOISES ET FRANÇAISES EN OR

*121 – PROTO-HELVETES. **1/4 de statère.** Tête bouclée à dr. R/. Bige à dr. dirigé par un aurige. Inscription à l'exergue. (La Tour 39 pl. LV — Muret 9304 var. — Forrer 433 var.). B./T.B. à T.B.

122 – **Statère.** Tête d'Apollon informe. R/. Bige à dr. dirigé par un aurige. (F. 57 var.). B./T.B.

123 – ATREBATES. (Arras). **1/4 de statère.** Bateau. R/. Chêne, faucille, croissant. (LT. 8611). B./T.B.

PHILIPPE VI DE VALOIS (1328-1350)

*124 – **Écu d'or.** 2e émission. (1343-1346). Le Roi assis, tenant l'épée et appuyé sur l'écu aux fleurs de lis sans nombre, sur un trône gothique. R/. Croix fleuronnée dans une rosace quadrilobée cantonnée de 4 grands trèfles pointus. (Lafaurie 262 a — Hoffmann —). Superbe exemplaire, bien centré.

*125 – **Écu d'or.** 3e émission. (1348). Mêmes types (L. 262 b — H. —). Superbe. Flan large.

JEAN II LE BON (1350-1364)

*126 – **Mouton d'or.** (1355). Agneau pascal deb. à g., nimbé, détournant la tête vers une croix fleurdelisée à longue hampe où flotte une bannière. R/. Croix fleuronnée dans une rosace quadrilobée, cantonnée de 4 fleurs de lis. (L. 294 — H. 3). Superbe. Rare en cet état.

*127 – **Franc à cheval.** (1360). Le Roi à cheval au galop à g., armé de toutes pièces; le Roi tient la bride articulée d'une main et de l'autre brandit une épée. R/. Croix fleuronnée dans une rosace quadrilobée et cantonnée de petits trèfles pointus. (L. 297 — H. 10). Très Beau/Superbe.

— OR —

Figure 1 An illustration of a sale catalog.
French National Library, Bourgey, 1971, ark:/12148/bpt6k97781698
https://gallica.bnf.fr/ark:/12148/bpt6k97781698/f46

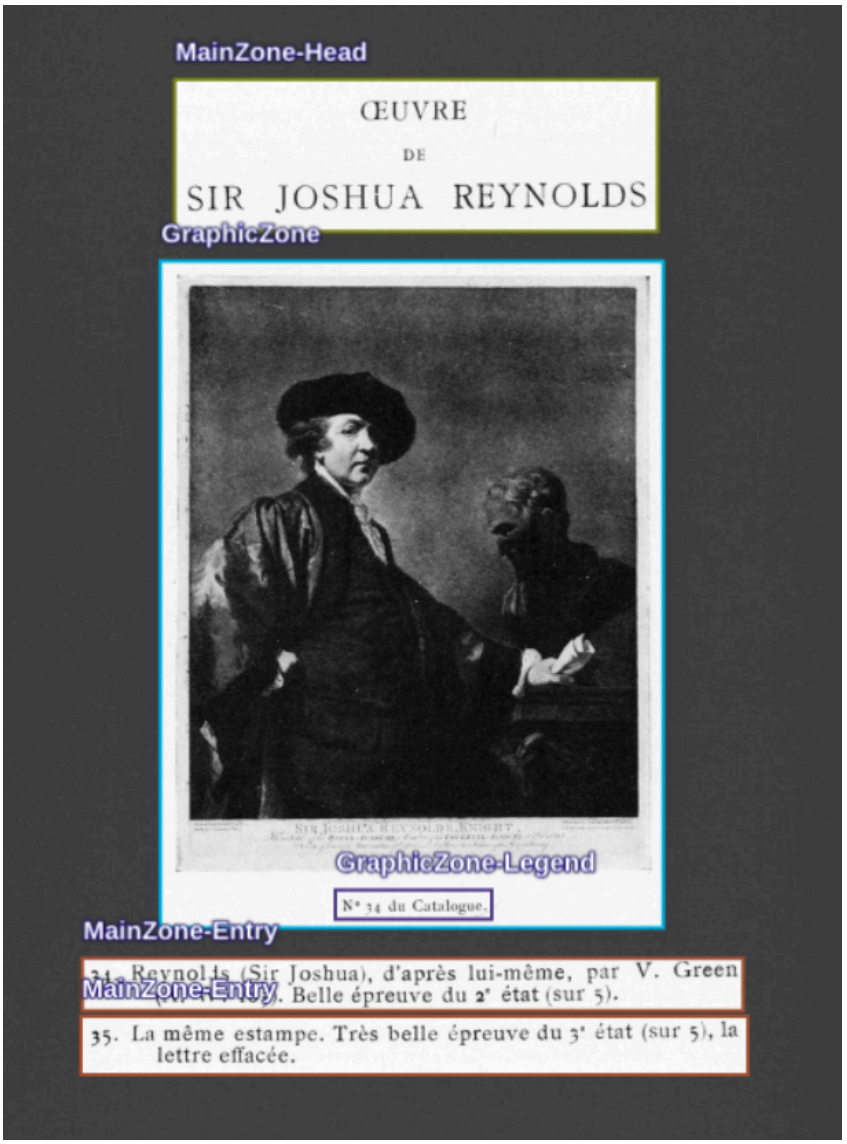
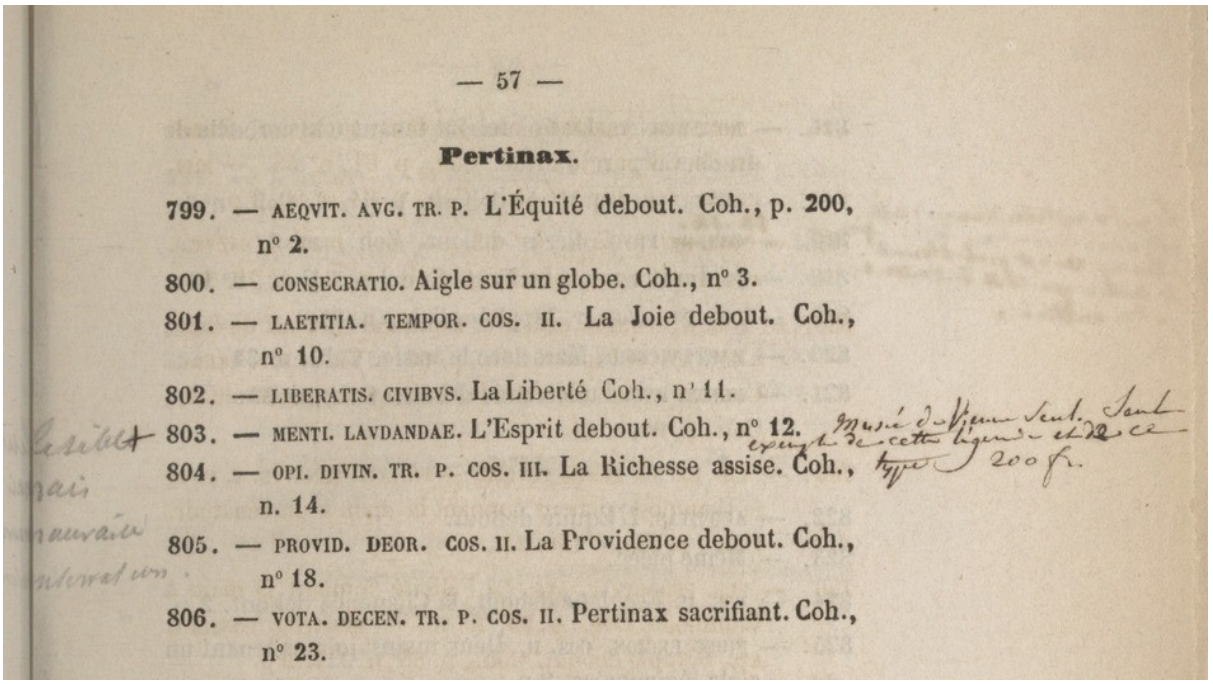


Figure 2 An annotated sale catalog macro-structure with corresponding labels.
Library of the French National Institure for Art History, Lair-Duveuil, 1924, CV09567_19241219.
https://bibliotheque-numerique.inha.fr/collection/item/61813-collection-de-m-c-estampes-du-xviiie-siecle-vente-du-19-decembre-1924



— 57 —

Pertinax.

799. — AEQVIT. AVG. TR. P. L'Équité debout. Coh., p. 200, n° 2.

800. — CONSECRATIO. Aigle sur un globe. Coh., n° 3.

801. — LAETITIA. TEMPOR. COS. II. La Joie debout. Coh., n° 10.

802. — LIBERATIS. CIVIBVS. La Liberté Coh., n° 11.

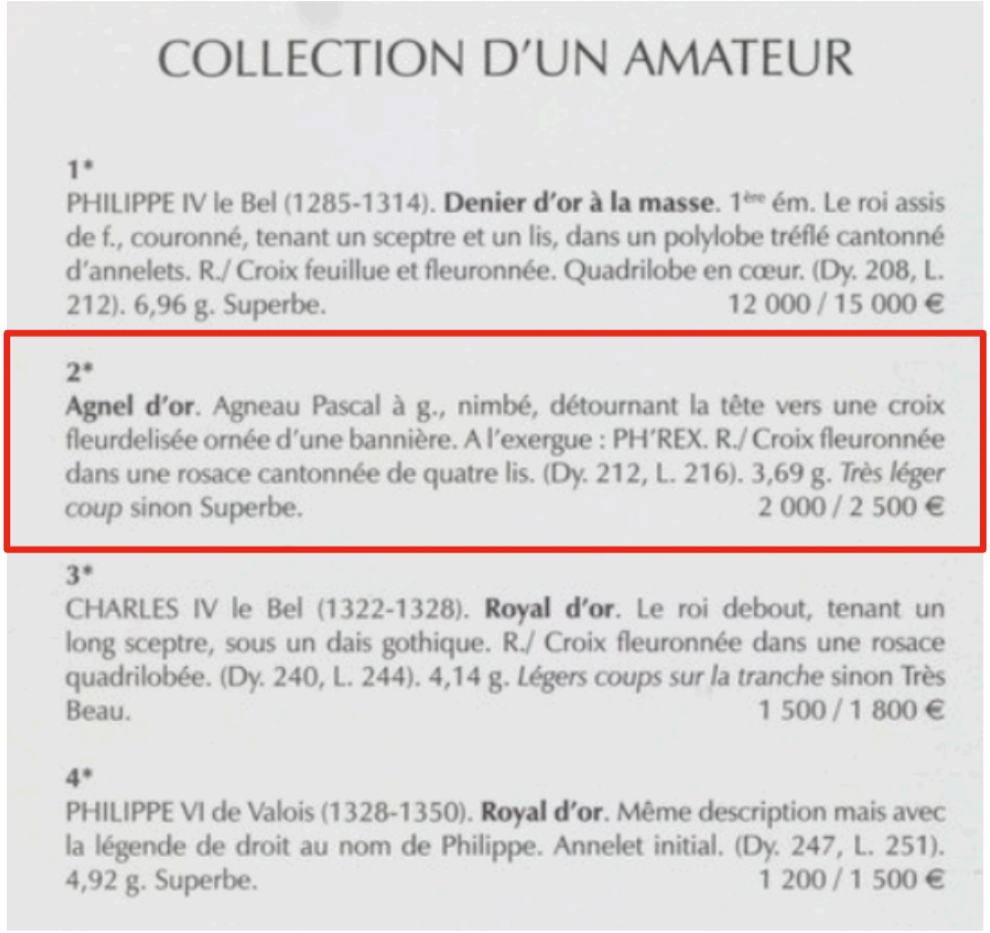
803. — MENTI. LAVDANDAE. L'Esprit debout. Coh., n° 12.

804. — OPI. DIVIN. TR. P. COS. III. La Richesse assise. Coh., n. 14.

805. — PROVID. DEOR. COS. II. La Providence debout. Coh., n° 18.

806. — VOTA. DECEN. TR. P. COS. II. Pertinax sacrifiant. Coh., n° 23.

Figure 3 Handwritten annotations on a sale catalog.
French National Library, Rollin et Feuardent, 1864, ark:/12148/bpt6k9777404n
https://gallica.bnf.fr/ark:/12148/bpt6k9777404n/f73



COLLECTION D'UN AMATEUR

1*
PHILIPPE IV le Bel (1285-1314). **Denier d'or à la masse**. 1ère ém. Le roi assis de f., couronné, tenant un sceptre et un lis, dans un polylobe tréflé cantonné d'annelets. R./ Croix feuillue et fleuronnée. 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 fleuronnée 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 fleuronnée 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 €

Figure 4 An entry in a sale catalog. Framed catalog entry in red is encoded in figure 5.
French National Library, Sabine Bourgey, 2011
https://gallica.bnf.fr/ark:/12148/bpt6k9778045h/f4

```
<!-- TEI tree -->
         <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 banière. A l'exergue : PH'REX. R./ Croix
fleuronnée 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.</condition>
                  </supportDesc>
               </objectDesc>
               <num type="currency">2000 / 2500 €</num>
            </catalogueItem>
            <!-- ... -->
         </catalogueEntry>
```

Figure 5 Customized TEI structure for representing a sales catalogue, used for the DataCatalogue project.

Bibliography

Chagué, A., & Clérice, T. (2022, June 23). *Sharing HTR datasets with standardized metadata: The HTR-United initiative*. Documents anciens et reconnaissance automatique des écritures manuscrites. https://inria.hal.science/hal-03703989

Chagué, A., Clérice, T., Norindr, J., Humeau, M., Davoury, B., Kote, E. V., Mazoue, A., Faure, M., & Doat, S. (2023, July 10). *Manu McFrench, from zero to hero: Impact of using a generic handwriting recognition model for smaller datasets*. Digital Humanities 2023: Collaboration as Opportunity. https://inria.hal.science/hal-04094241

Clérice, T. (2022). *You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine*. https://hal-enc.archives-ouvertes.fr/hal-03723208

Dutier, E. C., & Corbières, C. (2021, October 25). *A broader < object > content model for art history*. Next Gen TEI, 2021 - TEI Conference and Members' Meeting. https://hal.science/hal-03654979

Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021, September 5). *SegmOnto: Common vocabulary and practices for analysing the layout of manuscripts (and more)*. 1st International Workshop on Computational Paleography (IWCP@ICDAR 2021). https://hal.science/hal-03336528

Gabay, S., Topalov, B., Corbières, C., Noyer, L. R. D., Joyeux-Prunel, B., & Romary, L. (2021, September 21). *Automating Artl@s – extracting data from exhibition catalogues*. EADH 2021 - Second International Conference of the European Association for Digital Humanities. https://hal.science/hal-03331838

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking* (arXiv:2204.08387). arXiv. https://doi.org/10.48550/arXiv.2204.08387

Lopez, P., Du, C., Cohoon, J., Ram, K., & Howison, J. (2021). Mining Software Entities in

Scientific Literature: Document-level NER for an Extremely Imbalance and Large-scale Task. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3986–3995. https://doi.org/10.1145/3459637.3481936

Nelson, B. (2016). Curating Object-Oriented Collections Using the TEI. *Journal of the Text Encoding Initiative*, *Issue 9*, Article Issue 9. https://doi.org/10.4000/jtei.1680

Sven, N.-M., & Matteo, R. (2022). *Page Layout Analysis of Text-heavy Historical Documents: A Comparison of Textual and Visual Approaches* (arXiv:2212.13924). arXiv. https://doi.org/10.48550/arXiv.2212.13924

Vajjala, S., & Balasubramaniam, R. (2022). What do we really know about State of the Art NER? In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5983–5993). European Language Resources Association. https://aclanthology.org/2022.lrec-1.643