Call for Proposals – TEI 2024 – Texts, Languages and Communities

An ODD Schema for a Sustainable Encoding of Catalog Objects

Sarah Bénière[1], Hugo Scheithauer[1,2], Juliette Janès[1], Laurent Romary[1,3]

[1] ALMAnaCH – Automatic Language Modelling and ANAlysis & Computational Humanities, Inria Paris
[2] Université Paris Sciences & Lettres
[3] Directorate for Scientific Information and Culture, Inria

Long Paper Proposal

April 2024

Keywords

Sales Catalogs, TEI ODD, Digital Scholarly Edition, SegmOnto Controlled Vocabulary, Document Layout Segmentation

Abstract

Sales catalogs are a valuable resource for art historians as they are the witnesses of the circulation of works of art. The organization of information within the catalogs is consistent and structured, which makes it interesting material for automatic processing tasks. This long paper proposal presents our reflection on the structuration of the content of sales catalogs in TEI-XML. This consideration is part of a wider reflection within the framework of the DataCatalogue research project[1] (Inria, BnF, INHA) on an automated workflow processing sales catalogs from digitization to publication.

The collections from the BnF and the INHA are composed of over 28,000 sales catalogs and keep growing with, for example, legal submissions or private acquisitions. Most catalogs are digitized and may be transcribed *via* Optical Character Recognition (OCR), leaving us with an image and plain text. However, in order to make the most of this substantial amount of data, we need to structure and make it as interoperable as possible. We are building our workflow out of a sample of 713 catalogs—mostly in French—with particular attention to chronological diversity (18th-21st centuries) and variety in the types of sales (numismatics, works of art, furniture, books, etc.).

[1] DataCatalogue is an ongoing project funded by Inria and the French Ministry of Culture, involving the ALMAnaCH research team at Inria Paris, the French National Library (BnF), and the French National Institute for Art History (INHA). After an experimental phase in 2021-2022, the project was renewed for a second phase in 2023-2024. For further information, see the DataCatalogue GitHub organization: https://github.com/DataCatalogue.

The DataCatalogue ODD[2] is based, on the one hand, on the grounding work from the TEI community on the representation of physical objects and catalogs (Château-Dutier & Corbières, 2021; Gabay *et al.*, 2021b; Nelson, 2016). In 2021, Gabay *et al*. suggested the creation of a `<catalogueEntry>` element, associated with two other new elements: `<catalogueDesc>` and `<catalogueItem>`. It is based, on the other hand, on the SegmOnto controlled vocabulary (Gabay *et al.*, 2021a) designed for Document Layout Segmentation (DSL). Our structuration pipeline involves macro-segmentation as its first step, resulting in the creation of an annotated DataCatalogue dataset,[3] which was also integrated as a subset of the LADaS (Layout Analysis Dataset with SegmOnto)[4] corpus (Clérice *et al.*, 2024) within the framework of the COLaF project.[5] Thanks to our trained object detection model, the catalog items are identified by the MainZone:Entry class (Figure 1).

Thus, we created a table of equivalence between the SegmOnto classes from the LADaS corpus—and especially the DataCatalogue subset[6]—including the new `<catalogueDesc>`, `<catalogueEntry>` and `<catalogueItem>` elements,[7] as well as the existing TEI elements from the Guidelines. A `<catalogueEntry>` corresponds to a category of objects appearing in the catalog—for example “Egyptian Antique”—and generally contains a `<head>` and at least one `<catalogueItem>`. The note for each object is described with standard TEI vocabulary, specified in the ODD to appear within the `<catalogueItem>` element (Figure 2). The `<catalogueDesc>` element is optional and contains a short paragraph giving further information which are common to all items in the entry and do not appear in its title.

Our ODD schema aims to further define the proposed TEI elements from Gabay *et al.* (2021b) and to advocate for the shared usage of the SegmOnto controlled vocabulary simultaneously with the TEI standard for document layout analysis in general, based on the DataCatalogue corpus as an example.

---

[2] Available at: https://github.com/DataCatalogue/datacat-tei/tree/main/ODD.
[3] Dataset and segmentation model available at: https://app.roboflow.com/datacatalogue/macro-segmentation/overview. Description of the classes and annotation guide available (in French) at: https://github.com/DataCatalogue/datacat-object-detection-dataset/blob/main/DataCat_AnnotationGuide.md.
[4] Available at: https://universe.roboflow.com/colaftextes/segmonto.
[5] COLaF (*Corpus et Outils pour les Langues de France*) is an Inria research project aiming at creating open corpora and tools focusing on standard French and other languages of France as well. https://colaf.huma-num.fr/.
[6] Available at: https://github.com/DataCatalogue/datacat-tei/blob/main/SegmOnto-to-TEI.md.
[7] Our work on the definition of these new elements is available at: https://github.com/DataCatalogue/datacat-tei/tree/main/catalogueElements.

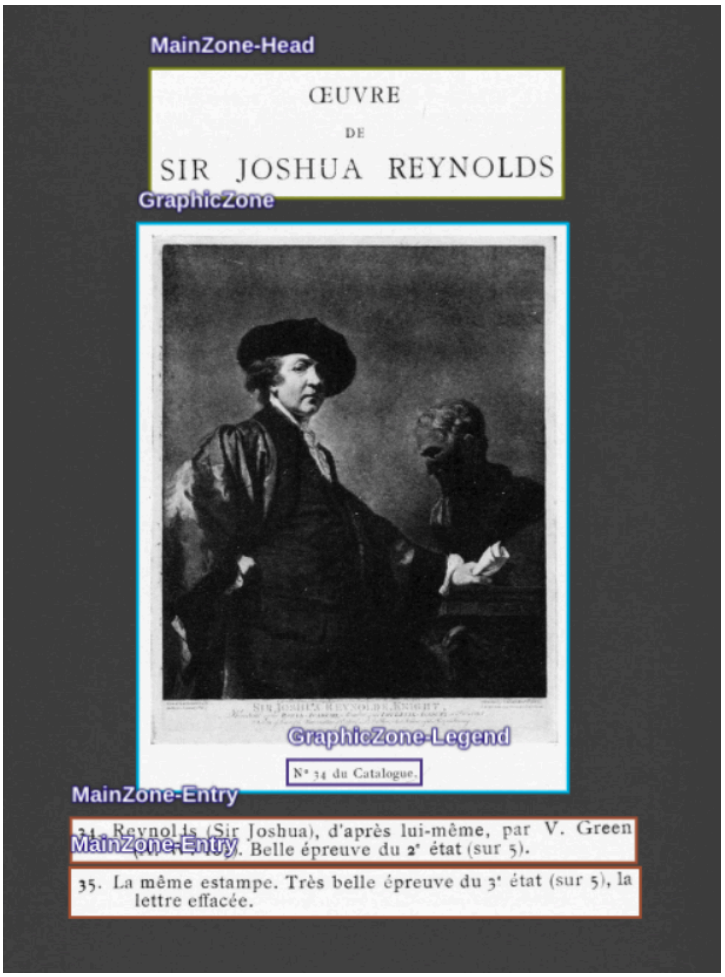

Figure 1. *Example of an annotated page, with the corresponding labels*. Bibliothèque de l'Institut National d'Histoire de l'Art, collections Jacques Doucet. Lair-Dubreuil, 1924, CV09567_19241219. https://bibliotheque-numerique.inha.fr/idurl/1/61813.

```
<catalogueEntry>
    <head>ANTIQUITÉS ÉGYPTIENNES, <lb/>GRANIT, ALBATRE ORIENTAL, SERPENTINE, <lb/>CRAIE,
        IVOIRE.</head>
    <!-- <catalogueDesc/> -->
    <!-- <catalogueItem n=""/> -->
    <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>
            <!-- <width/> -->
        </dimensions>
        <!-- <condition/> -->
    </catalogueItem>
</catalogueEntry>
```

Figure 2. *Example of content for the <catalogueItem> element*.

Bibliography

Château-Dutier, E., & Corbières, C. (2021). *A Broader <object> Content Model for Art History*. Next-Gen TEI 2021 – TEI Conference and Members' Meeting. https://hal.science/hal-03654979.

Clérice, T., Janès, J., Scheithauer, H., Bénière, S., Romary, L., & Sagot, B. (2024). *Layout Analysis Dataset with SegmOnto*. DH 2024 – Annual Conference of the Alliance of Digital Humanities, Washington, D.C. https://inria.hal.science/hal-04513725.

Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021a). *SegmOnto: Common Vocabulary and Practices for Analyzing the Layout of Manuscripts*. 1st International Workshop on Computational Philology – ICDAR 2021. https://hal.science/hal-03336528.

Gabay, S., Topalov, B., Corbières, C., Rondeau du Noyer, L., Joyeux-Prunel, B., & Romary, L. (2021b). *Automating Artl@s – Extracting Data From Exhibition Catalogues*. EADH 2021 – 2nd International Conference of the European Association for Digital Humanities. https://hal.science/hal-03331838.

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

About the Authors

Sarah Bénière is a research and development engineer in the ALMAnaCH team at Inria, Paris. She holds an M.A. in Cultural Studies and in Digital Technologies Applied to History from the École nationale des chartes. She is working on data structuration and TEI modeling of archival documents.

PhD student in the Inria ALMAnaCH team, **Hugo Scheithauer** holds a master's degree in art history and in technologies numériques appliquées à l'histoire (digital technologies applied to history) at the École nationale des chartes. He works on document layout analysis, automatic text recognition, and information extraction for historical documents.

Juliette Janès is an R&D engineer in the ALMANaCH team at Inria, Paris. She holds a master's degree in Digital Technologies applied to History from the École nationale des Chartes. Currently, she is involved in corpus production and NLP tools development for the COLaF project, which focuses on the languages of France.

Laurent Romary is a research director and the director for scientific information and culture at Inria, Paris. He has conducted multiple research projects in computer science, linguistics, and digital humanities and developed a strong interest in standardization and open access. He also identified the need to implement shared infrastructures for open science (e.g. DARIAH).