



HAL
open science

An ODD Schema for a Sustainable Encoding of Catalog Objects

Sarah Bénérière, Hugo Scheithauer, Juliette Janes, Laurent Romary

► To cite this version:

Sarah Bénérière, Hugo Scheithauer, Juliette Janes, Laurent Romary. An ODD Schema for a Sustainable Encoding of Catalog Objects. TEI 2024 – Texts, Languages and Communities, Universidad del Salvador, Oct 2024, Buenos Aires, Argentina. ⟨hal-04754028⟩

HAL Id: hal-04754028

<https://hal.science/hal-04754028v1>

Submitted on 25 Oct 2024

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.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License



Logo: Alix Chagué
Inspiration: Loading Artist

An ODD Schema for a Sustainable Encoding of Catalog Objects

Sarah Bélière, Research & Development Engineer, ALMAnaCH (Inria)

Hugo Scheithauer, PhD candidate, ALMAnaCH (Inria) & EPHE (Université PSL, Paris)

Juliette Janès, Research & Development Engineer, ALMAnaCH (Inria)

Laurent Romary, Research Director, DCIS (Inria)

The DataCatalogue Project



Inria

Institutional Context


Partnering institutions

- Inria
- (French) Ministry of Culture
- National Library (BnF)
- National Institute for Art History (INHA)

Project history

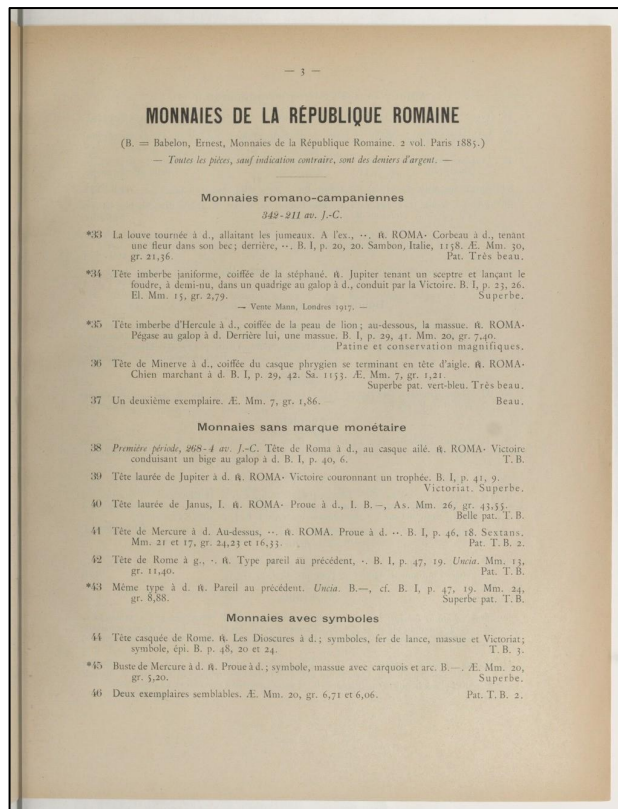
- Phase 1 (2021-2022)
- Phase 2 (2023-2024)

Scientific experts, researchers, data providers



Objective

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

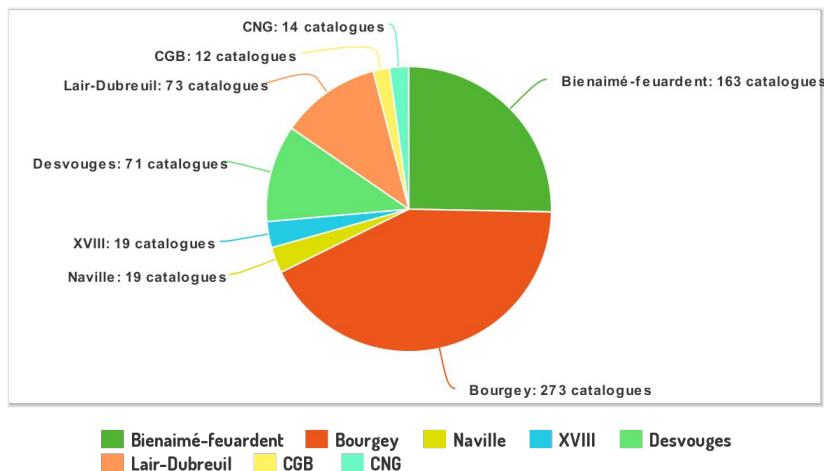


Lucien Naville, 1924

The Corpus

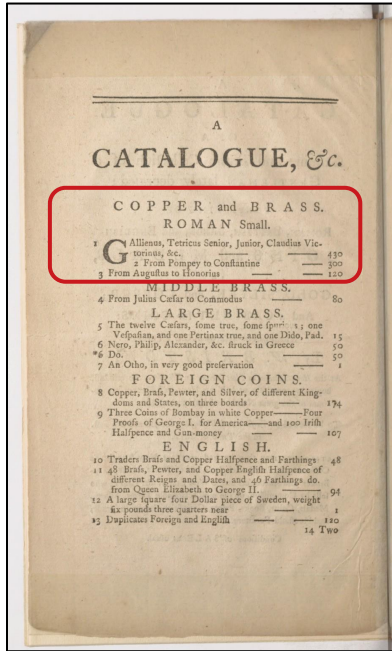


Composition

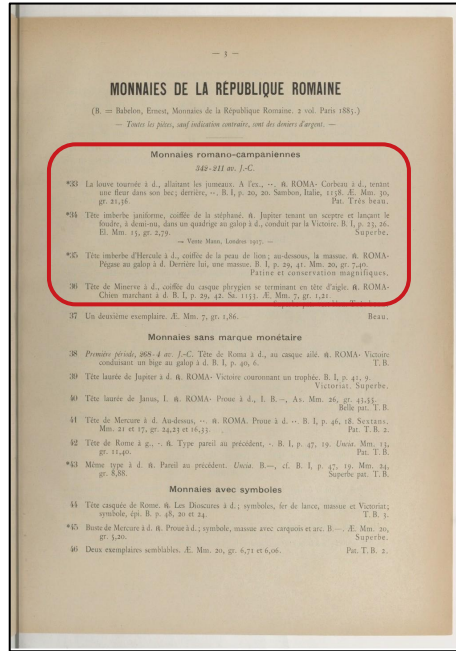


- Overall = +150k catalogs
- DataCatalogue dataset = **713 catalogs** (from BnF/INHA)
- Various representations of:
 - **Sales types** (numismatics, ancient books, antiques, fine arts, etc.)
 - **Document layouts**
 - **Time periods** (18th-21st centuries)
- Languages:
 - **French** (~95%)
 - German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

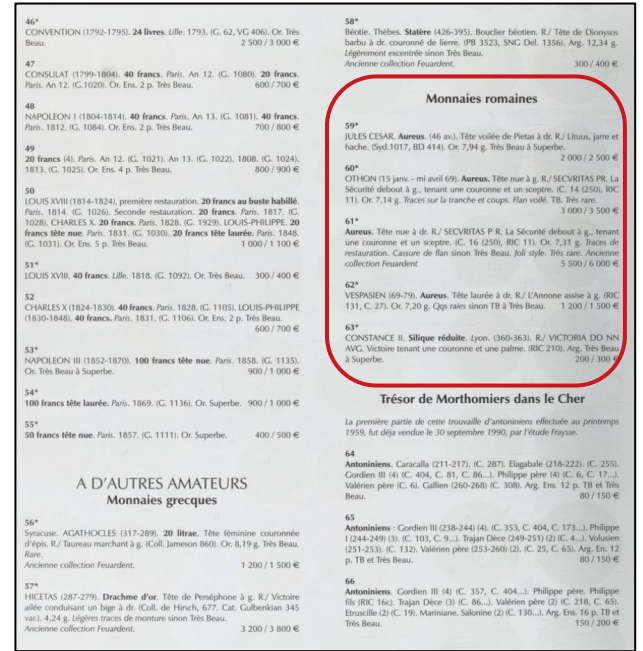
Interest – Heterogeneous Structure of Sales Catalogs



Whiston Bristow, 1762



Lucien Naville, 1924



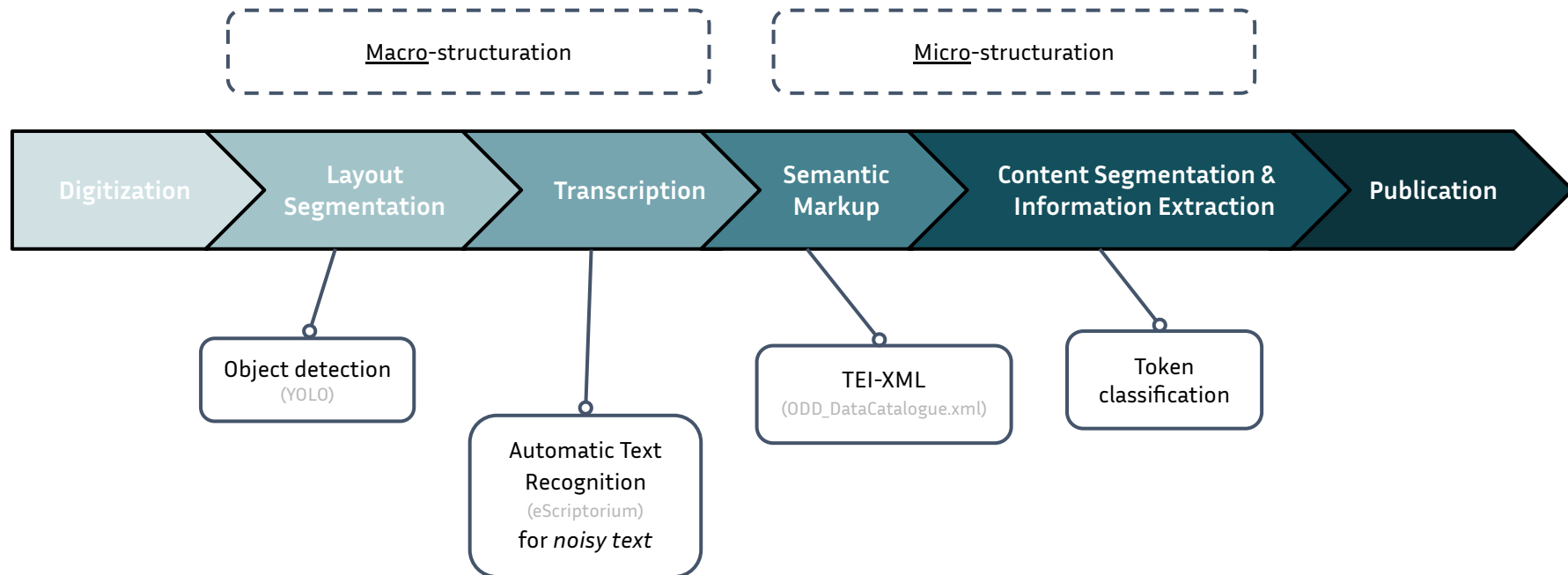
Frayse & Associés, 2011

The Workflow





The DataCatalogue Image-to-Text Workflow



Focus on Layout Segmentation





From SegmOnto to LADaS

SegmOnto



[Gabay et al. \(2023\)](#)

- Controlled vocabulary
- Layout description
- Specific syntax: **Type:Subtype**
- Documentation written in a TEI ODD

LADaS



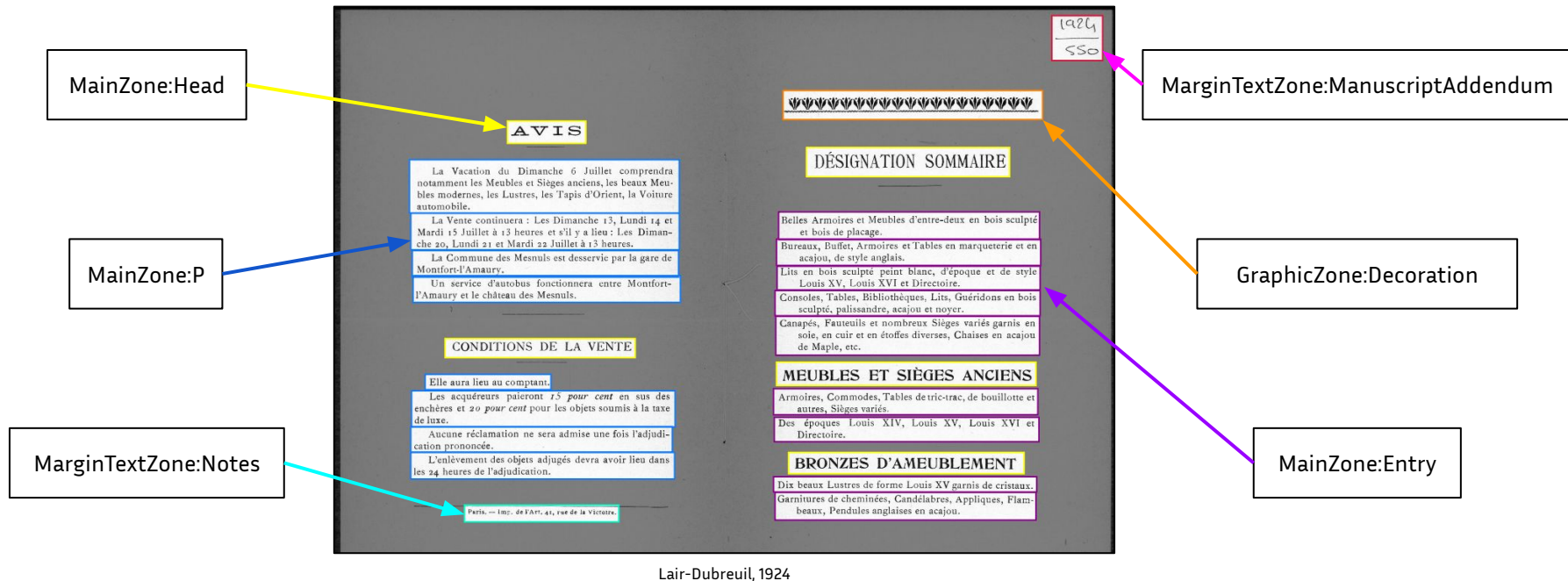
[Clérice et al. \(2024\)](#)



[Scheithauer et al. \(2024\)](#)

- “Layout Analysis Dataset with SegmOnto”
- COLaF project (Inria)
- LADaS dataset = 5071 images
- DataCatalogue subset = **1425 images**

Example of DataCatalogue LADaS Classes



From LADaS to TEI

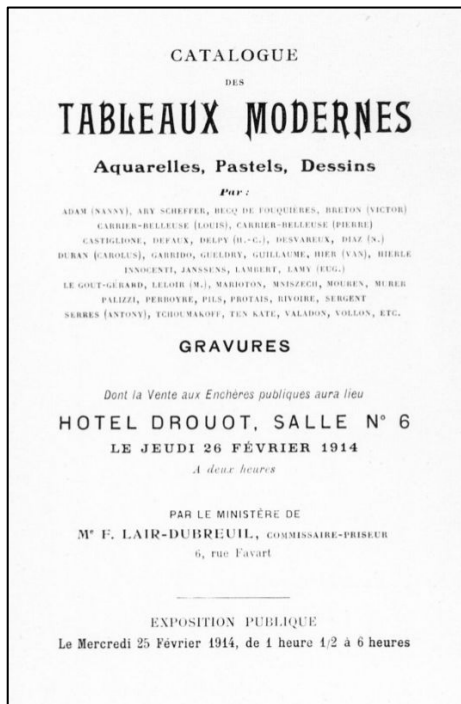
- **Table of equivalence** between LADaS and TEI-XML
- 2 “catalogue” classes:
 - MainZone:Entry
 - MainZone:P@CatalogueDesc
- 3 “catalogue” TEI elements:
 - <catalogueItem>
 - <catalogueDesc>
 - <catalogueEntry>

Class Name	TEI
GraphicZone	<figure>
GraphicZone:Decoration	<figure type="decoration">
GraphicZone:FigDesc	<figDesc>
MainZone:Entry	<catalogueItem>
MainZone:Head	<head>
MainZone:P	<p>
MainZone:P@CatalogueDesc	<catalogueDesc>
MainZone:Signature	<signed>
MarginTextZone:ManuscriptAddendum	<div type="handwritten">
MarginTextZone:Notes	<note n="{}" place="{}">
NumberingZone	<pb n="" facs="">
PageTitleZone	<front> → <titlePage>
	<catalogueEntry>

Describing Sales Catalogs in TEI



Title Page Encoded According to the P5 Guidelines



Lair-Dubreuil, 1914

```
<front>
<titlePage>
  <docTitle>
    <titlePart type="main">CATALOGUE <lb/>DES <lb/> TABLEAUX MODERNES</titlePart>
    <titlePart type="sub">Aquarelles, Pastels, Dessins</titlePart>
    <titlePart type="artists">Par : <lb/><name>ADAM (NANNY)</name>, <name>ARY SCHEFFER</name>,
      <name>BACQ DE FOUQUIERES</name>, <name>BRETON (VICTOR)</name>
      <lb/><name>CARRIER-BELLEUSE (LOUIS)</name>, <name>CARRIER-BELLEUSE (PIERRE)</name>
      <lb/><name>CASTIGLIONE</name>, <name>DEFAUX</name>, <name>DELPY (H.-C.)</name>,
      <name>DESVAREUX</name>, <name>DIAZ (N.)</name>
      <lb/><name>DURAN (CAROLUS)</name>, <name>GARRIDO</name>, <name>GUELDRY</name>,
      <name>GUILLAUME</name>, <name>HIER (VAN)</name>, <name>HIERLE</name>
      <lb/><name>INNOCENTI</name>, <name>JANSSENS</name>, <name>LAMBERT</name>, <name>LAMY
      (EUG.)</name>
      <lb/><name>LE GOUT-GÉRARD</name>, <name>LELOIR (M.)</name>, <name>MAITOTON</name>,
      <name>MNSZECH</name>, <name>MOURER</name>, <name>MURER</name>
      <lb/><name>PALIZZI</name>, <name>PERBOYRE</name>, <name>PILS</name>,
      <name>PROTAIS</name>, <name>RIVOIRE</name>, <name>SERGENT</name>
      <lb/><name>SERRIS (ANTONY)</name>, <name>TCHOUMAKOFF</name>, <name>TEN KATE</name>,
      <name>VALADON</name>, <name>VOLLON</name>, ETC.</titlePart>
    <titlePart type="sub">GRAVURES</titlePart>
    <titlePart type="sales_place">Dont la Vente aux Enchères publiques aura lieu
      <lb/><placeName>HOTEL DROUOT</placeName>, SALLE N°6 <lb/><date>LE JEUDI 26 FÉVRIER
      1914</date><lb/><time>A deux heures</time></titlePart>
  </docTitle>
  <byline>PAR LE MINISTÈRE DE <lb/><name>Me F. LAIR-DUBREUIL</name>, COMMISSAIRE-PRISEUR <lb/>
    <address>
      <addrLine>6, rue Favart</addrLine>
    </address>
  </byline>
  <metamark style="line">
    <titlePart><desc>EXPOSITION PUBLIQUE</desc>
      <lb/><date>Le Mercredi 25 Février 1914</date>, <time>de 1 heure 1/2 à 6 heures</time>
    </titlePart>
  </titlePage>
</front>
```

ExampleFile_Lair-Dubreuil_CV02553_19140226_f3.xml

TEI Modeling of Sales Catalogs

Building upon the grounding work of:



Gabay, Simon, Barbara Topalov, Caroline Corbières, Lucie Rondeau Du Noyer, Béatrice Joyeux-Prunel, and Laurent Romary. "Automating Artl@s – extracting data from exhibition catalogues." Presented at *EADH 2021 – Second International Conference of the European Association for Digital Humanities*, Krasnoyarsk, 2021. URL: <https://hal.science/hal-03331838>.

MONNAIES FRANÇAISES		
(En argent ou billon, sauf indication contraire)		
251	Hugues, fils de Robert. Portail. r/. Croix. (H. — Ciani 31). Denier. Orléans. B/TB.	700
252	Philippe I. A ₀ dans le champ. r/. Croix. (L. 46 - H. 3). Denier. Paris. 2 ^e type. B/TB à TB.	2.800
253	Louis VII. FRA-OCN. r/. Croix. (L. 139 - H. —). Denier. Paris. B/TB.	1.400
254	Mêmes types. (L. 144 - H. —). Obole. Paris. B/TB.	900
255	Croix. r/. Croix cantonnée de 2 omégas, la branche verticale terminée par un S couché. (L. 157). Denier. Senlis. TB.	1.900
256	Philippe III. Gros tournois. (L. 204 - H. 5). TB.	1.000
257	Philippe IV. Gros tournois à l'O rond. (L. 217 - H. 5). TB.	800
258	Gros tournois à l'O long. (L. 218 - H. 8). Très Beau.	1.500
259	Maille demie. (L. 221 - H. 9). Très Beau.	1.000
260	Maille tierce à l'O rond. (L. 223a var.). Superbe.	2.800

Bourgey, 1958

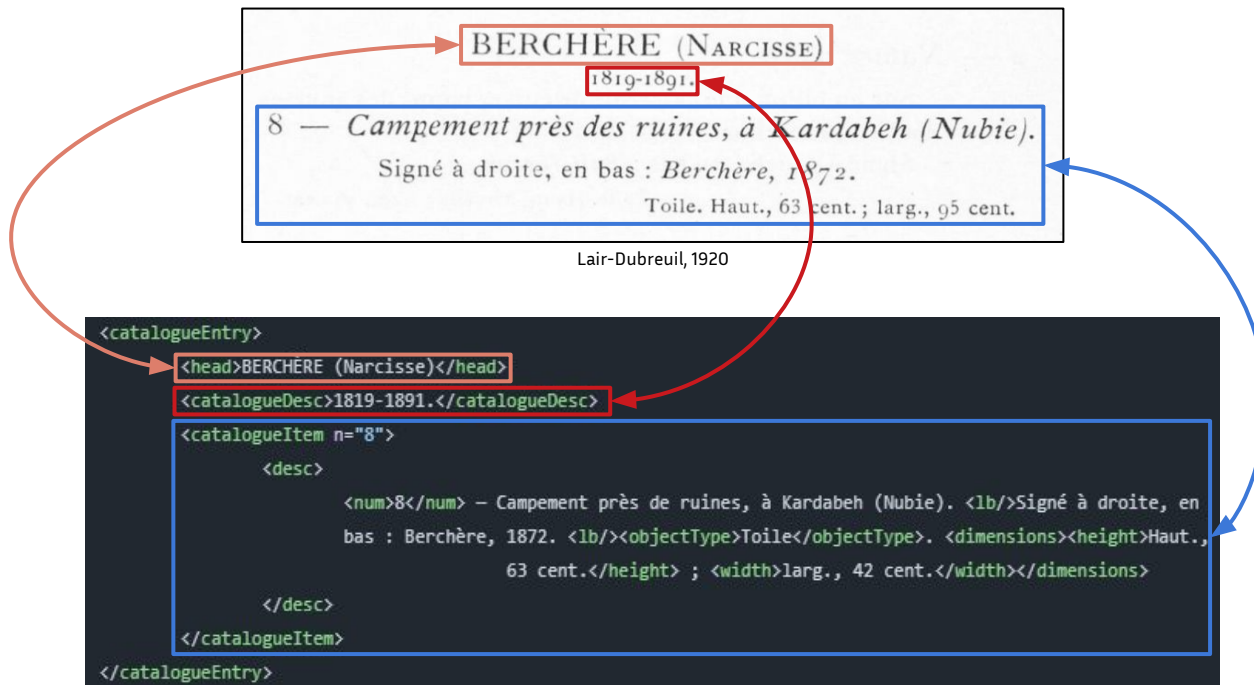


Description of Catalog-related Elements – `<catalogueEntry>`

- Closest content model ⇨ **model.divLike** (text division)
- Minimal structure:
 - **<head>** ⇨ header of the entry
 - **<catalogueDesc>** ⇨ further information, applies to all items in the entry [OPTIONAL]
 - **<catalogueItem>** ⇨ description of catalog objects (at least one)
- `<catalogueEntry>` elements can be embedded like `<div>` elements



Example of an Encoded <catalogueEntry>



ExampleFile_Lair-Dubreuil_CV05182-19200318_f6.xmm

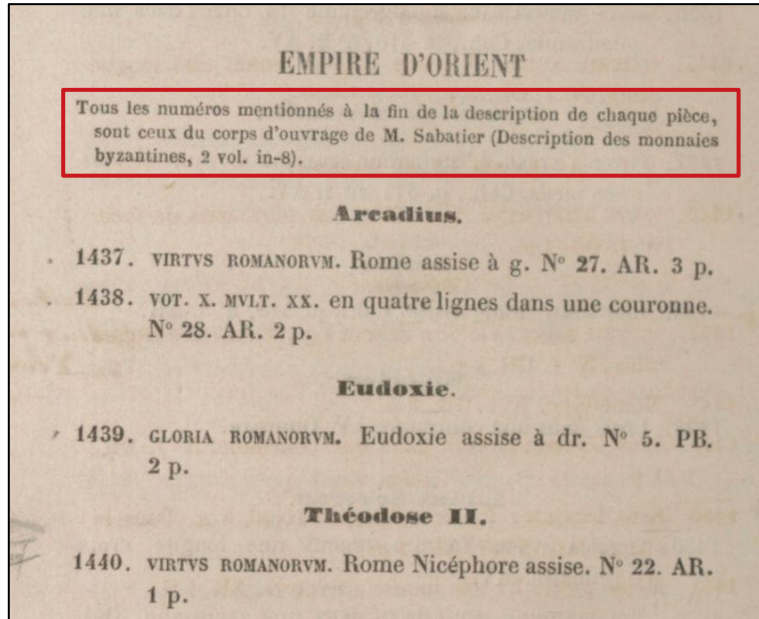
Embedding Challenge

- At which point do we switch from `<div>` to `<catalogueEntry>`?
- Do we consider “SANS NOM D’ATELIER” to be `<catalogueDesc>` or `<catalogueEntry>`?
- Does it depend on context?
- If so, can we actually automate the encoding?

```
<div>
  <head>METZ ET DOMAINES EPISCOPAUX</head>
  <div>
    <head>EVEQUES</head>
    <catalogueEntry>
      <head>SEMI-EPISCOPALES - SEMI-IMPERIALES</head>
      <catalogueEntry>
        <head>Adalbéron I et Otton I, empereur (962-964)</head>
        <catalogueDesc>SANS NOM D’ATELIER</catalogueDesc>
      </catalogueEntry>
      <catalogueEntry>
        <head>Thiéri I et Otton I (964-973)</head>
        <catalogueDesc> SANS NOM D’ATELIER</catalogueDesc>
        <catalogueItem n="411"/>
        <catalogueDesc>ATELIER DE METZ</catalogueDesc>
        <catalogueItem n="412"/>
      </catalogueEntry>
    </catalogueEntry>
  </div>
</div>
```

ExampleFile_bienaime-feuardent_12148-bpt6k9777420r_f119.xml

Description of Catalog-related Elements – <catalogueDesc>



Rollin, 1864

- Closest content model ⇨ **model.pLike** (paragraph)
- LADaS equivalent ⇨ **MainZone:P@CatalogueDesc**
- Gives **further information** about the entry
- Content of <catalogueDesc> applies to all items in the entry

Description of Catalog-related Elements – <catalogueItem>

- Closest content model ⇨ **model.descLike** (description)
- LADaS equivalent ⇨ **MainZone:Entry**
- Information is structured within <desc>:
 - **<num>** ⇨ object identifier within the catalog
 - **<objectType>** ⇨ type of object described (e.g. painting, sculpture, coin, etc.)
 - **<dimensions>** ⇨ object dimensions
 - **<condition>** ⇨ assessment of the object's physical condition
- All elements within <desc> are **optional due to the heterogeneity of the entries' presentation.**

```
<catalogueItem n="8">
  <desc>
    <num>8</num> – Campement près de ruines, à Kardabeh (Nubie). <lb/>Signé à droite, en
    bas : Berchère, 1872. <lb/><objectType>Toile</objectType>. <dimensions><height>Haut.,
    63 cent.</height> ; <width>larg., 42 cent.</width></dimensions>
  </desc>
</catalogueItem>
```

ExampleFile_Lair-Dubreuil_CV05182-19200318_f6.xmm

Conclusion & Perspectives





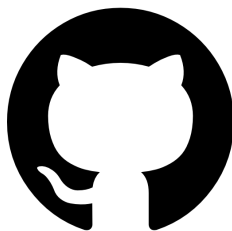
Key Takeaways

- The DataCatalogue workflow is still a work in progress:
 - Automate the transition from one step to another as much as possible
 - Resolve the embedding challenge
- Using the SegmOnto controlled vocabulary for layout segmentation ensures:
 - High compatibility with TEI
 - Standardized annotations and encoding
 - ➔ Improves DataCatalogue's commitment to make data FAIR

Useful Links



Read more about the
DataCatalogue workflow



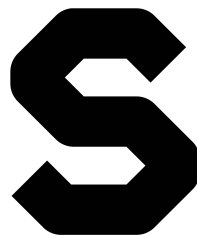
Check out the DataCatalogue
annotation guide (FR)



Test the DataCatalogue
object detection model



Access the DataCatalogue
TEI repository



Read the SegmOnto
documentation



Access the LADaS
dataset

Thank you for your attention!



Logo: Alix Chagué
Inspiration: Loading Artist