Automating Artl@s – extracting data from exhibition catalogues

Simon Gabay, Barbara Topalov, Caroline Corbières, Lucie Rondeau Du Noyer, Béatrice Joyeux-Prunel, Laurent Romary

Automating ARTL@S - extracting data from exhibition catalogues

Simon Gabay[1], Barbara Topalov[1], Caroline Corbières[2], Lucie Rondeau du Noyer[1], Béatrice Joyeux-Prunel[1], and Laurent Romary[3]

[1]université de Genève , prenom.nom@unige.ch
[2]Université Paris-Saclay (France) ,
prenom.nom@@universite-paris-saclay.fr
[3]INRIA- ALMANACH (France) , prenom.nom@inria.fr

Abstract

Catalogues, which have been published for centuries, are an extremely precious resource for scholars. Using the Artl@s database as an example, where exhibition catalogues are transformed into a georeferenced database, we question the possibility of an (almost) automatic transformation of pdfs into semantically annotated data. To do so, we present and analyse the graphic organisation of exhibition catalogues, before exploring a possible modeling into TEI (involving possible enhancement of the guidelines).

Databases for art history usually focus on images, and are primarily made by or for museums to display collections (or inventories).[1] Other databases do not focus on the work of art itself, but on a corpus about the work of art – such as databases used for provenance research[2] or by collectors monitoring the market[3]. The BASART database of the ARTL@S project[4] belongs to the second category as it aims at recording exhibitions all over the world since the invention of exhibition catalogues in the 17th century (Paris, Salon de l'Académie, 1673)[5]. BASART is essential for art history, not only because it makes available one of the most widely used sources in the discipline, at a global scale and over several centuries, but also because it makes it possible to globalise the horizons

[1] http://mandragore.bnf.fr/html/accueil.html
[2] For instance http://www.getty.edu/research/tools/provenance/search.html or https://ventesdantiques.inha.fr
[3] https://fr.artprice.com/
[4] https://artlas.huma-num.fr
[5] According to Guiffrey 1869

of research and to move on to unusual quantitative, spatial, and transnational analyses, without specific knowledge in digital technologies (Joyeux-Prunel and Marcel 2015).

If data has long been entered manually, by simply copying the content that in the exhibition catalogues describes the works exhibited, we are now designing a new workflow, based on GROBID DICTIONARIES technology (Khemakhem 2020), to semi-automate the task, that is to say to retrieve the content from pdf digitisation of the catalogues, describe it semantically, in order to transform the XML-TEI output of this new workflow into the structured BASART postGIS database. As our database is structured to deal with catalogues from very different periods of time, and from very diverse locations and languages, we intend to take advantage of the tasks involved in designing a semi-automated retrieving chain – in particular, the description necessary to train recognition algorithms –, to produce research on the history of the form of exhibition catalogues over time and space, and according to the language. The project's ambition is also to enable broader developments that will be useful for other projects dealing with comparable resources and corpora – such as auction catalogues, collection directories, or *catalogues raisonnés*.

1 Exhibition catalogues

In order to semantically describe catalogues, we need to analyse their form. This form has a history, which meets the history of the inventory, whose description is a challenge for anyone who wishes to model their structure, but also try to generalise this modeling to any type of similar documents. The *inventorium* (inventory) can be considered as an early form and the origin of the catalogue. This list of a set of objects forming a collection was originally built by recording collection items in the very order in which they were stored or displayed (Barbier, Dubois, and Sordet 2015). That is why in France, as early as the Middle Ages, inventories were primarily used to associate a given object to the place where it was preserved. The inventory is first and foremost a practical tool: it is meant to find and identify an object. Inventories also have a scientific/epistemic interest insofar as the classification of objects echoes the classifications of the natural sciences, especially in the case of curiosities cabinets (Findlen 1996). The practice of classifying allows comparison: "Unlike the inventory, the catalogue organises a set of data and arranges them by object, so as to make them comparable with each other." (Recht 1996, p. 23).

With the advent of printing, and the ability to publish inventories for a wider audience, the form of the inventory became semi-industrialised for a few particular types: the *catalogues raisonnés* of an artist's work, the inventory of a private collection or a museum, the dealer's catalogue, the auction catalogue, the exhibition catalogue. It is with this type of printed data that we believe we can develop a semi-automated pipeline for content retrieval. This process begins with the example of exhibition catalogues.

Generally the content of an exhibition catalogue is quite regular : informa-

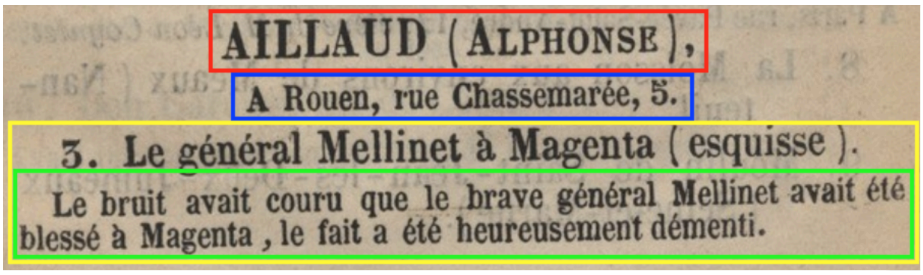

AILLAUD (ALPHONSE),
A Rouen, rue Chassemarée, 5.
3. Le général Mellinet à Magenta (esquisse).
Le bruit avait couru que le brave général Mellinet avait été blessé à Magenta, le fait a été heureusement démenti.

Figure 1: *Catalogue de l'exposition annuelle du musée de Rouen*, 1860, p. 19.

tion about the event exhibition, its title, date, supposed location, its organisers; a list of the exhibitors – first and last name, sometimes completed by biographical detail, a year and place of birth and an address. Most exhibition catalogues complete these lists of artists with a list, for each exhibitor, of the works exhibited (Joyeux-Prunel and Marcel 2015). Whereas this content is not always represented in the same way in the book-catalogue, it is actually organised according to a limited number of possibilities.

2 A Stable Layout, Over Time and Place

In the documentation we have gathered (c. 1000 catalogues since the 18th c., mainly from Western countries), the most recurrent format displays the following information: name of the artist (here, highlighted in red), information about the artist (blue) and a list of the works exhibited (yellow) with potentially additional information (green).

2.1 Type 1

Salons, yearly municipal exhibitions and academy exhibitions which were held regularly since the 18th century, usually published catalogues with a very stable layout over time:

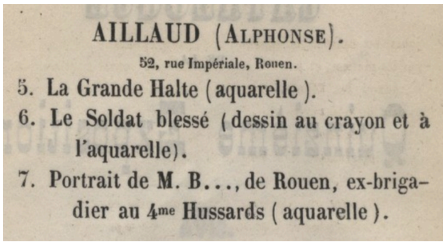

AILLAUD (ALPHONSE).
52, rue Impériale, Rouen.
5. La Grande Halte (aquarelle).
6. Le Soldat blessé (dessin au crayon et à l'aquarelle).
7. Portrait de M. B..., de Rouen, ex-brigadier au 4me Hussards (aquarelle).

Figure 2: *Catalogue de l'exposition... de Rouen*, 1853, p. 4.

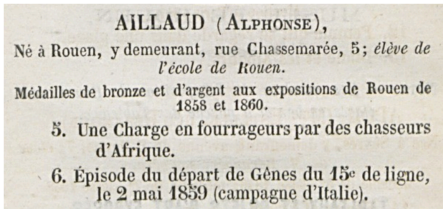
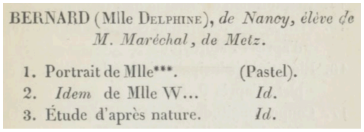

AILLAUD (ALPHONSE),
Né à Rouen, y demeurant, rue Chassemarée, 5; *élève de l'école de Rouen.*
Médailles de bronze et d'argent aux expositions de Rouen de 1858 et 1860.
5. Une Charge en fourrageurs par des chasseurs d'Afrique.
6. Épisode du départ de Gênes du 15e de ligne, le 2 mai 1859 (campagne d'Italie).

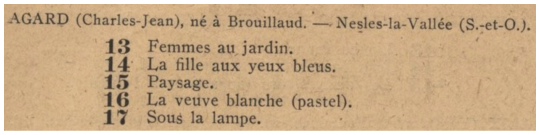
Figure 3: *Catalogue de l'exposition... de Rouen*, 1862, p. 26.

This kind of layout can be found throughout the 19th c., whatever the continent, at least until the 1940s in most cases. Data is structured the same way in most of the exhibition catalogues that have been printed – no matter the time or the place – over this period of time. It is therefore easy to build a single model for several catalogues. If we look at exhibition catalogues published in Nancy in the 1840's, Paris in the 1920's, Venice in the 1910's or São Paulo in the 1950's, we can observe that they follow a similar pattern, which should facilitate the automation of data extraction.


BERNARD (Mlle Delphine), *de Nancy, élève de M. Maréchal, de Metz.*

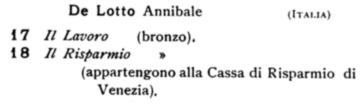
1. Portrait de Mlle***. (Pastel).
2. *Idem* de Mlle W... *Id.*
3. Étude d'après nature. *Id.*

Figure 4: *Catalogue... exposés à Nancy*, 1843, p. 3.


AGARD (Charles-Jean), né à Brouillaud. — Nesles-la-Vallée (S.-et-O.).

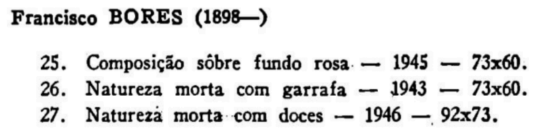
13 Femmes au jardin.
14 La fille aux yeux bleus.
15 Paysage.
16 La veuve blanche (pastel).
17 Sous la lampe.

Figure 5: *Société des artistes indépendants*, 1921, p. 17.


De Lotto Annibale (Italia)

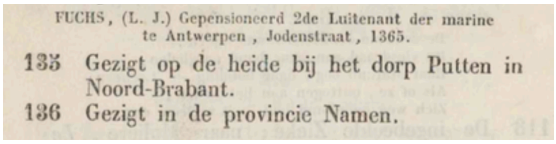
17 *Il Lavoro* (bronzo).
18 *Il Risparmio* »
(appartengono alla Cassa di Risparmio di Venezia).

Figure 6: *Esposizione Internazionale... di Venezia*, 1910.


Francisco BORES (1898—)

25. Composição sôbre fundo rosa — 1945 — 73x60.
26. Natureza morta com garrafa — 1943 — 73x60.
27. Natureza morta com doces — 1946 — 92x73.

Figure 7: *Bienal de São Paulo*, 1951, p. 54.

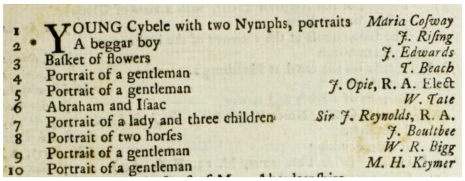

FUCHS, (L. J.) Gepensioneerd 2de Luitenant der marine te Antwerpen, Jodenstraat, 1365.

135 Gezigt op de heide bij het dorp Putten in Noord-Brabant.
136 Gezigt in de provincie Namen.

Figure 8: *Tentoonstelling van schilder- en andere werken...*, 1852, p.12.

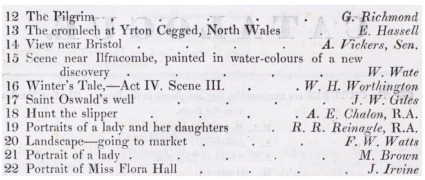
2.2 Type 2

A secondary structure coexists with the latter one – a structure that first presents the number and the title of the work, followed by the name of its author. The order is relatively opposite to that described above. This second type of layout, following the exhibition order, has been used for the *Exhibition of the Royal Academy* in London since 1780, but it can be found in other countries (e.g. Canada). It has been extremely stable over time too.



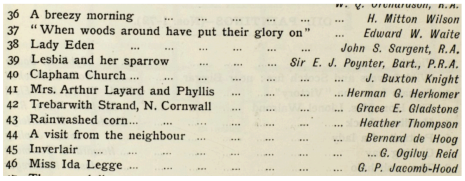
1 YOUNG Cybele with two Nymphs, portraits — *Maria Cofway*
2 * A beggar boy — *J. Rising*
3 Basket of flowers — *J. Edwards*
4 Portrait of a gentleman — *T. Beach*
5 Portrait of a gentleman — *J. Opie*, R. A. Elect
6 Abraham and Isaac — *W. Tate*
7 Portrait of a lady and three children — *Sir J. Reynolds*, R. A.
8 Portrait of two horses — *J. Boultbee*
9 Portrait of a gentleman — *W. R. Bigg*
10 Portrait of a gentleman — *M. H. Keymer*

Figure 9: *The Exhibition of the Royal Academy*, 1785, p. 1.



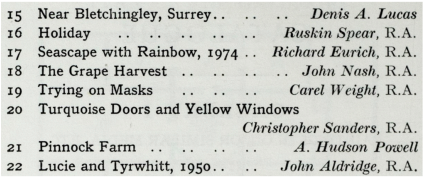
12 The Pilgrim — *G. Richmond*
13 The cromlech at Yrton Cegged, North Wales — *E. Hassell*
14 View near Bristol — *A. Vickers, Sen.*
15 Scene near Ilfracombe, painted in water-colours of a new discovery — *W. Wate*
16 Winter's Tale,—Act IV. Scene III. — *W. H. Worthington*
17 Saint Oswald's well — *J. W. Giles*
18 Hunt the slipper — *A. E. Chalon*, R.A.
19 Portraits of a lady and her daughters — *R. R. Reinagle*, R.A.
20 Landscape—going to market — *F. W. Watts*
21 Portrait of a lady — *M. Brown*
22 Portrait of Miss Flora Hall — *J. Irvine*

Figure 10: *The Exhibition of the Royal Academy*, 1831, p. 6.



36 A breezy morning — *H. Mitton Wilson*
37 "When woods around have put their glory on" — *Edward W. Waite*
38 Lady Eden — *John S. Sargent, R.A.*
39 Lesbia and her sparrow — *Sir E. J. Poynter, Bart., P.R.A.*
40 Clapham Church — *J. Buxton Knight*
41 Mrs. Arthur Layard and Phyllis — *Herman G. Herkomer*
42 Trebarwith Strand, N. Cornwall — *Grace E. Gladstone*
43 Rainwashed corn — *Heather Thompson*
44 A visit from the neighbour — *Bernard de Hoog*
45 Inverlair — *G. Ogilvy Reid*
46 Miss Ida Legge — *G. P. Jacomb-Hood*

Figure 11: *The Exhibition of the Royal Academy*, 1907, p. 8.



15 Near Bletchingley, Surrey — *Denis A. Lucas*
16 Holiday — *Ruskin Spear*, R.A.
17 Seascape with Rainbow, 1974 — *Richard Eurich*, R.A.
18 The Grape Harvest — *John Nash*, R.A.
19 Trying on Masks — *Carel Weight*, R.A.
20 Turquoise Doors and Yellow Windows — *Christopher Sanders*, R.A.
21 Pinnock Farm — *A. Hudson Powell*
22 Lucie and Tyrwhitt, 1950 — *John Aldridge*, R.A.

Figure 12: *The Exhibition of the Royal Academy*, 1975, p. 10.

We assume for the moment that these two types may have corresponded to different areas of cultural influence - on the one hand the influence of the Parisian Salon catalogue model, on the other that of a more Anglo-Saxon model. It is likely that another factor explaining these differences is simply that catalogues by work order assume a more commercial character, where it is the work for sale that is presented, and not the artist. This hypothesis is supported by the presence of Type 2 in many gallery catalogues in the early 20$^{\text{th}}$ c.

3 Beyond exhibition catalogues

According to the two types presented supra, entries can be grouped under super-entries, which potentially transmit general properties to entries that they include such as the location of the work in the exhibition (gallery 1, red wing…), or more importantly artistic forms (paintings, sculpture, architecture…).

```
CATALOGUE
__CATEGORY 1
  |__ARTIST 1
     |__Work 1.1
     |__Work 1.2
  |__ARTIST 2
     |__Work 2.1
     |__Work 2.2
__CATEGORY 2
  |__ARTIST 3
     |__Work 3.1
     |__Work 3.2
```

Figure 13: Type 1

```
CATALOGUE
__Work 1
__Work 2
__Work 3
__Work 4
__Work 5
__Work 6
__Work 7
__Work 8
__Work 9
__Work 10
__Work 11
```

Figure 14: Type 2

These two types can be found in similar documents, the structure of which is very similar to exhibition catalogues: bibliographies (Lindemann, Khemakhem, and Romary 2018), dictionaries (Khemakhem, Herold, and Romary 2018), auction catalogues (Gabay, Rondeau Du Noyer, and Khemakhem 2020), and phone directories (Khemakhem, Brando, et al. 2018).

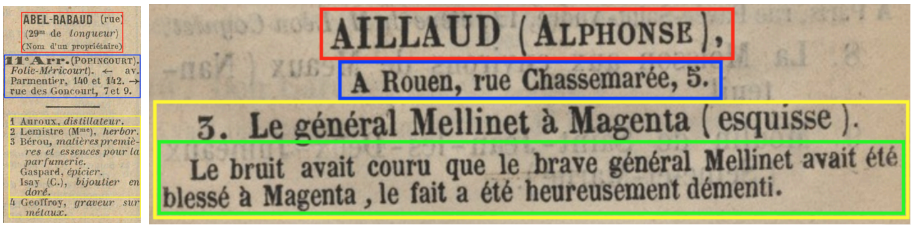


Figure 15: *Annuaire-Almanach de l'industrie*, 1901, p. 2883 compared to the *Catalogue de l'exposition annuelle du musée de Rouen*, 1860, p. 19.

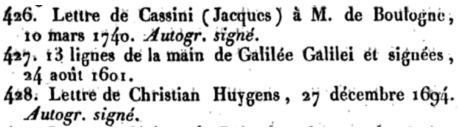

426. Lettre de Cassini (Jacques) à M. de Boulogne, 10 mars 1740. *Autogr. signé.*
427. 13 lignes de la main de Galilée Galilei et signées, 24 août 1601.
428. Lettre de Christian Huygens, 27 décembre 1694. *Autogr. signé.*

Figure 16: *Catalogue de feu M. de Bruyères Chalabre*,1833, p. 102

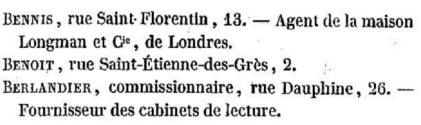

BENNIS, rue Saint-Florentin, 13. — Agent de la maison Longman et Ce, de Londres.
BENOIT, rue Saint-Étienne-des-Grès, 2.
BERLANDIER, commissionnaire, rue Dauphine, 26. — Fournisseur des cabinets de lecture.

Figure 17: *Annuaire des imprimeurs et des libraires*, 1841, p. 135

Such data has been recently identified as "encyclopedic-like" (Khemakhem, Romary, et al. 2018), *i.e.* entry-based data for which a dedicated retrieval tool has been developed: GROBID DICTIONARIES (Khemakhem 2020). Thanks to the latter, PDF files are automatically structured and annotated in XML-TEI files, which is used as a pivot format to ease encoding refinements via regexes.

The current TEI guidelines[6] propose elements allowing research to encode various kinds of lists (`<list>`, `<listPerson>`, `<listPlaces>`, `<listObject>`…) and their content (`<item>`, `<person>`, `<place>`, `<object>`…). However, some scholars have noted that encoding a catalogue as a "mere list" was not sufficient if some of its entries feature analytical or descriptive content (Nelson 2016). Therefore, following the work on dictionaries with TEI-lex 0 (Romary and Tasovac 2018) on the `<entry>` element, we would like to propose to the TEI Consortium the creation of a `<catalogueEntry>` element. The behaviour of this TEI markup would be similar to `<entry>`, transforming `<form>` and `<sense>` elements into `<catalogueDesc>` and `<catalogueItem>`. A possible encoding would therefore be the one presented *infra*.

The stakes of this type of classification are both to allow a clearer description of the content, and the possibility of spreadsheet-type retrievals, where the

[6] `https://tei-c.org/guidelines`

dependencies of a sub-entry on its entry would be maintained (in particular in the repetition of the entry for each line of the sub-entry in the case of an export in csv). This type of processing is ideal for a pivot to databases, more easily handled by users interested in quantitative and cartographic visualisation as is the case for the Artl@s project.

```
<catalogueEntry>
   <catalogueDesc>
       <person>
           <persName>
               <surname>Aillaud</surname>
               <pc>(</pc>
               <forename>Alphonse</forename>
               <pc>)</pc>
           </persName>
           <pc>,</pc>
           <birth>Né à Rouen</birth>
           <pc>,</pc>
           <residence>y demeurant rue Chasse-Marée, 5</residence>
           <pc>;</pc>
           <education>élève de l'école de Rouen</education>
           <pc>.</pc>
           <floruit>Médaille de bronze et d'argent aux expositions de ROUen de 1858 et 1860</floruit>
           <pc>.</pc>
       </person>
   </catalogueDesc>
   <catalogueItem>
       <num>5.</num>
       <title>Une charge de fourrageurs par des chasseurs d'Afrique</title>
       <pc>.</pc>
   </catalogueItem>
   <catalogueItem>
       <num>6.</num>
       <title>Épisode du départ de Gènes du 15e de ligne le 2 mai 1859 (campagne d'Italie)</title>
       <pc>.</pc>
   </catalogueItem>
</catalogueEntry>
```

Figure 18: TEI encoding of a catalogue entry.

Data and code

The ODD, the specification and an example can be found at the following address: `https://github.com/katabase/Catalogues`.

References

Barbier, F., T. Dubois, and Y. Sordet (2015). *De l'argile au nuage : une archéologie des catalogues IIe millénaire av. J.-C.-XXIe siècle [exposition, Paris, Bibliothèque Mazarine, 13 mars-13 mai 2015, Genève, Bibliothèque de Genève, 18 septembre-21 novembre 2015]*. 1 vols. Paris-Genève: Bibliothèque Mazarine-Éditions des Cendres Bibliothèque de Genève. 429 pp.

Findlen, P. (1996). *Possessing nature : museums, collecting, and scientific culture in early modern Italy*. Studies on the history of society and culture 20. Country: US ill. 23 cm. Bibliogr. p. 409-432. Index. Berkeley Los Angeles London: University of California press. 449 pp. ISBN: 978-0-520-20508-6.

Gabay, S., L. Rondeau Du Noyer, and M. Khemakhem (2020). "Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes". In: *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. AIUCD2020. Quaderni di Umanistica Digitale. Milan, Italy: Associazione per l'Informatica Umanistica e la Cultura Digitale, pp. 113–118.

Guiffrey, J. (.-1. (1869). *Collection des livrets des anciennes expositions depuis 1673 jusqu'en 1800*. 42 vols. Publisher: Liepmannssohn et Dufour (Paris). Paris: Liepmannssohn et Dufour. URL: `https://catalogue.bnf.fr/ark:/12148/cb33308085f` (visited on 06/01/2021).

Joyeux-Prunel, B. and O. Marcel (2015). "Exhibition Catalogues in the Globalization of Art. A Source for Social and Spatial Art History". In: *Artl@s Bulletin* 4.2, pp. 80–104. ISSN: 2264-2668. URL: `https://docs.lib.purdue.edu/artlas/vol4/iss2/8`.

Khemakhem, M. (2020). "Standard-based Lexical Models for Automatically Structured Dictionaries". PhD thesis. Paris: INRIA.

Khemakhem, M., C. Brando, et al. (Sept. 2018). "Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories". In: *JADH2018 "Leveraging Open Data"*. JADH2018. Tokyo, Japan. URL: `https://hal.archives-ouvertes.fr/hal-01814189` (visited on 09/11/2019).

Khemakhem, M., A. Herold, and L. Romary (May 2018). "Enhancing Usability for Automatically Structuring Digitised Dictionaries". In: *GLOBALEX workshop at LREC 2018*. Miyazaki, Japan. URL: `https://hal.archives-ouvertes.fr/hal-01708137`.

Khemakhem, M., L. Romary, et al. (Sept. 9, 2018). "Automatically Encoding Encyclopedic-like Resources in TEI". In: TEI2018. Tokyo, Japan. URL: `https://hal.inria.fr/hal-01819505` (visited on 09/11/2019).

Lindemann, D., M. Khemakhem, and L. Romary (Dec. 2018). "Retrodigitizing and Automatically Structuring a Large Bibliography Collection". In: *European Association for Digital Humanities (EADH) Conference*. Galway, Ireland: EADH. URL: `https://hal.archives-ouvertes.fr/hal-01941534`.

Nelson, B. (Sept. 24, 2016). "Curating Object-Oriented Collections Using the TEI". In: *Journal of the Text Encoding Initiative* (Issue 9). Number: Issue 9 Publisher: Text Encoding Initiative Consortium. ISSN: 2162-5603. DOI: `10.`

4000/jtei.1680. URL: http://journals.openedition.org/jtei/1680 (visited on 06/01/2021).

Recht, R. (1996). "La Mise en ordre : note sur l'histoire du catalogue". In: *Les Cahiers du Musée National d'Art Moderne* 56/57, pp. 20–35.

Romary, L. and T. Tasovac (2018). "TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources". In: *Book of AbstractsThe 18th Annual TEI Conference and Members' Meeting*. TEI 2018. Conference Name: TEI 2018: TEI as a Global Language (TEI2018) Publisher: Zenodo. Tokyo, Japan. DOI: 10.5281/zenodo.2613594. URL: https://zenodo.org/record/2613594#.YGBGHEgzZPM (visited on 03/28/2021).