



**HAL**  
open science

## Towards automatic TEI encoding via layout analysis

Juliette Janes, Ariane Pinche, Claire Jahan, Simon Gabay

### ► To cite this version:

Juliette Janes, Ariane Pinche, Claire Jahan, Simon Gabay. Towards automatic TEI encoding via layout analysis. Fantastic future 21, 3rd International Conference on Artificial Intelligence for Libraries, Archives and Museums, AI for Libraries, Archives, and Museums (ai4lam), Dec 2021, Paris, France. hal-03527287

**HAL Id: hal-03527287**

**<https://hal.science/hal-03527287v1>**

Submitted on 15 Jan 2022

**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 Attribution 4.0 International License

# Towards automatic TEI encoding via layout analysis

Juliette Janes, Ariane Pinche, and Claire Jahan

École nationale des chartes | PSL (France)

Simon Gabay

Université de Genève (Switzerland)

The forefront of research on textual documents (may they be manuscripts and prints) is slowly moving from text recognition to automatic encoding. Quickly transforming images into XML-TEI documents is therefore the next important obstacle that needs to be tackled to offer enhanced mining options to digital libraries users.

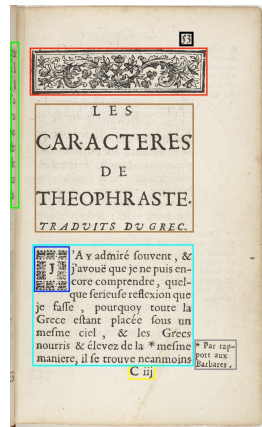


Figure 1: La Bruyère, *Les Caractères de Théophraste*, 1688, p. 53. In black the page number, in red a headpiece, in brown the text, in cyan the text, in blue the drop capital, in yellow the signature, in grey a marginal note, in green noise.

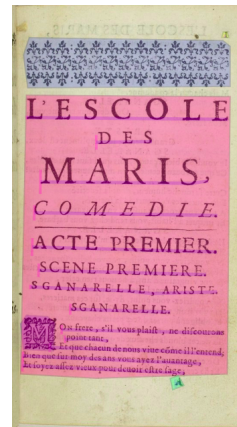


Figure 2: Example of a page with zones automatically recognised: Molière, *L'Ecole des Maris*, 1661, p. 1. In yellow the page number, in purple the headpiece, in pink the text, in blue the drop capital, in green the signature.

If extensive research has been carried on complicated and highly standardised layouts such as manuscripts (Gabay, Rondeau Du Noyer, et al. 2020) or exhibition catalogues (Gabay, Topalov, et al. 2021) with dedicated tools like GROBID (Khemakhem 2020), more simple documents still await for an easily trainable and implementable solution. Most of the pages contain simple elements such as page numbers, marginal notes, signatures... (cf. fig. 1) which all have a different status than the actual body of the text, and therefore need to be disentangled from one another. For instance, finding all the mentions of Andromache in the play of the same name is a nonsense if we do not separate running titles (in which the name of the eponym heroins appear) from the dialogues.

Recent experiments show that a standardised description of the page and state-of-the-art models for layout analysis could offer a limited, yet efficient solution (cf. fig. 2) to numerous scholars for two reasons. On the one hand powerful open-source OCR engines are now available (Kiessling 2019) via user-friendly interfaces (Kiessling et al. 2019): it will become easier to train models tailored to one’s need in a near future. On the other hand research institutions are now investing in OCR infrastructures (Gabay 2021; Romary et al. 2021), which will accelerate the production of data locally, without the direct support of GLAMs.

Using a common vocabulary to annotate zones called *SegmOnto*<sup>1</sup> (Gabay, Camps, et al. 2021) (that is still evolving), we have developed a generic workflow to analyse the layout, OCRise the text, and convert the ALTO output into minimally encoded TEI files (cf. table 1). This workflow is currently being tested on three different datasets: one of medieval manuscripts (cf. table 5), one of 17<sup>th</sup> c. literary prints (cf. table 4), and one of 19<sup>th</sup> c. catalogues (cf. table 6).

The key aspect of the workflow being zones detection, we have experimented two neural architectures for the layout analysis, and possible combinations of data to increase the efficiency. One model for each set has been trained, then two combinations: the medieval and the 17<sup>th</sup> c. data on the one hand, the 17<sup>th</sup> c. data and the 19<sup>th</sup> c. on the other hand (cf. table 2).

A second experiment has redimensioned the image taken as input in the

Segmonto zone	Corresponding TEI element
Damage	<damage>
Decoration	<figure>
DropCapital	<hi>
Figure	<figure>
Main	<p>
Margin	<note>
MusicNotation	<figure>
Numbering	<fw type="Numbering">
RunningTitle	<fw type="RunningTitle">
Seal	<figure>
Signatures	<fw type="Signatures">
Stamp	<figure>
Table	<table>
Title	<p>

Table 1: SegmOnto zones and their corresponding TEI element.

<sup>1</sup><https://segmonto.github.io>.

	Medieval data	17 <sup>th</sup> c.	19 <sup>th</sup> c.	Combination 1	Combination 2
MIoU	0.6017	0.1965	0.2092	0.4883	0.1964
FWIoU	0.7445	0.7794	0.7157	0.7537	0.7322
MAcc	0.9805	0.9846	0.9613	0.9886	0.9834
Acc	0.9805	0.9846	0.9613	0.9886	0.9834

Table 2: Architecture 1: efficiency of complex models measured with Mean Intersection-Over-Union (MIoU), frequency-weighted intersection over union (FWIoU), Mean accuracy (MAcc) and accuracy (Acc). Combination 1 gathers Medieval data and 17<sup>th</sup> c., Combination 2 gathers 17<sup>th</sup> c. and 19<sup>th</sup> c. data.

VGSL specs (1 200→1 800)<sup>2</sup>, which significantly improves the results for both main categories (cf. table 3): intersection over union (*i.e.* zone detection) and accuracy (*i.e.* zone classification).

	Medieval data	17 <sup>th</sup> c.	19 <sup>th</sup> c.	Combination 1	Combination 2
MIoU	0.6127	0.1685	0.3662	0.5386	0.3269
FWIoU	0.7754	0.8002	0.7244	0.7845	0.7821
MAcc	0.9905	0.9886	0.9651	0.9917	0.9869
Acc	0.9905	0.9886	0.9651	0.9917	0.9869

Table 3: Architecture 2.

Lower scores for medieval data and 19<sup>th</sup> c. can easily be explained by the relatively smaller size of the training set and the greater complexity of the layout. If the results obtained with combined datasets are hard to interpret without the existence of a test for layout analysis, scores do not suffer a significant drop: the combination seems to be relatively efficient.

## Acknowledgements

This paper would not have been possible without the help of Thibault Clérice and Jean-Baptiste Camps (ENC), the SegmOnto technical and semantic working groups, the Kraken and eScriptorium development teams.

## References

Gabay, S. (2021). *FOrmes Numérisées et Détection Unifiée des Écritures*. <https://www.unige.ch/lettres/humanites-numeriques/index.php/?cID=188>.

<sup>2</sup>the architecture of the second training is therefore the following 1,1800,0,3 Cr7,7,64,2,2 Gn32 Cr3,3,128,2,2 Gn32 Cr3,3,128 Gn32 Cr3,3,256 Gn32 Lbx32 Lby32 Cr1,1,32 Gn32 Lby32 Lbx32.

- Gabay, S., J.-B. Camps, et al. (Sept. 2021). “SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)”. In: *Proceedings of the 1st International Workshop on Computational Paleography, IWCP@ICDAR 2021*. 1st International Workshop on Computational Paleography IWCP. Lecture Notes in Computer Science. Lausanne (Switzerland): Springer.
- Gabay, S., L. Rondeau Du Noyer, et al. (July 2020). “Quantifying the Unknown: How many manuscripts of the marquise de Sévigné still exist?” In: *Digital Humanities DH2020*. DH2020 Book of Abstracts. Ottawa, Canada: ADHO. URL: <https://hal.archives-ouvertes.fr/hal-02898929> (visited on 11/23/2020).
- Gabay, S., B. Topalov, et al. (Sept. 2021). “Automating Artl@s - extracting data from exhibition catalogues”. In: *Digital Humanities DH2020*. EADH2021 Book of Abstracts. Krasnoyarsk, Russia: ADHO. URL: REPLACE.
- Khemakhem, M. (2020). “Standard-based Lexical Models for Automatically Structured Dictionaries”. PhD thesis. Paris: INRIA.
- Kiessling, B. (July 2019). “Kraken - an Universal Text Recognizer for the Humanities”. In: *Digital Humanities 2019 Conference Abstracts*. Digital Humanities 2019 Conference. Utrecht, The Netherlands: Alliance of Digital Humanities Organizations (ADHO). URL: <https://dev.clariah.nl/files/dh2019/boa/0673.html> (visited on 03/16/2020).
- Kiessling, B. et al. (Sept. 2019). “eScriptorium: An Open Source Platform for Historical Document Analysis”. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). Vol. 2, pp. 19–19. DOI: 10.1109/ICDARW.2019.10032.
- Romary, L. et al. (2021). *CREMMA - Consortium pour la Reconnaissance d'Écriture Manuscrite des Matériaux Anciens*. <https://www.dim-map.fr/projets-soutenus/cremma/>.

# Appendices

## Exemples

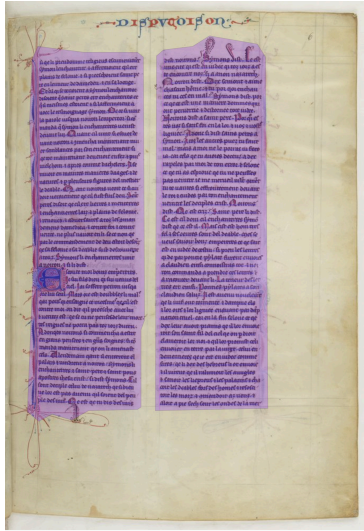


Figure 3: *Manuscrit, BnF fr. 412, f°21, 13<sup>th</sup> c.*

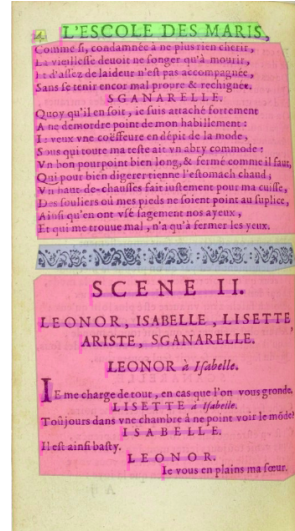


Figure 4: *Molière, L'Escole des maris, Paris, G. de Luynes, 1661*

<p>0 Hayet (Jard), Douville (S.-et-O.), 10.          4 Dechères, Paris, pl. Sorbonne, 3.          10 Ville de Paris.</p> <p><b>0.34 ORAN (Rue d') 33</b></p> <p>1-3 Hétère r. Escaudai, 3.          5-7 Hétère r. Dandaville, 50.          9 Hétère r. Lées, 24.          11 Hétère r. Lées, 29.          13 Hétère, Paris, r. Lagny, 34.          (Terme) Hétère (H<sup>er</sup>), gr. Martin, Paris, local, Barde, 46.          15 à 19 (Grosroule) Hétère, Paris, r. Oues, 15.          19 (Grosroule) Hétère, Paris, r. Lées, 8.          21 Thierry (Ver), Paris, r. Lafayette, 121.          23-25 Hétère r. Poissonnière, 46.          27 Hétère, Chagny (Cochin).          2 Hétère, Paris, r. Pajol, 33.          4 Hétère, Paris, r. Lées, 38.          6 Hétère, Paris, r. Lées, 5.          8 Lédouville, Paris, r. Solitaire, 20.          10 Thomas, Paris, r. de la Harpe, 41 (Seine).          12 Quinon, Paris, r. Oues, 12.          14 Val (M<sup>er</sup>), Paris, av. Orléans, 100.          16 Hétère, Paris, r. Bregnot, 14.          18 Noidel, Leclercq, Hétère (M<sup>er</sup>), Oues, 22.          20 Gouffon, Hétère (M<sup>er</sup>).          22 Hétère, gr. Poissonnière, Paris, r. Oues, 22.          24 Magnin (M<sup>er</sup>), Paris, r. Gardes, 32.          26 Gouffon, Paris, r. Galle, 20.          28 Courvaier (M<sup>er</sup>), Paris, r. Bancy, 20.          30 (Cant. Bouchon-Landehou) Hétère.          32 Courvaier (M<sup>er</sup>), Paris, r. Bancy, 29.          34 Sureau, Paris, r. Oues, 24.          36 Lachon, Mété, r. Roperis, 15 (S.-et-M.).          38 Hétère, Lomme par Courant (Seine).</p> <p><b>0.35 ORATOIRE (Rue de l') 3</b></p> <p>Hétère r. St-Henri, 145.          2 Hétère r. Rivoli, 128.          4 Ville de Paris.          6 Magnin, Paris, r. Oues, 6.          8 Hétère (M<sup>er</sup>), Paris, local, Capucines, 23.</p> <p><b>0.36 ORCHAMPT (Rue d') 44</b></p> <p>1 Hétère r. Baignan, 13.          2 Hétère, Paris, r. Deschamps, 3.          5-7 Hétère, Paris, r. Chantelle, 10.          8 Hétère r. Baignan, 15.          9 La Chénouy (M<sup>er</sup> 46), Paris, r. Noidel, 10.          9 Pélion, Paris, r. Bregnot, 2.</p>	<p>8-10 Bregnot, Paris, r. Anacle, 10.          12 Allée, Varennes-Saint-Hilaire (Seine).          14 André, Amiens, r. Bregnot, 82<sup>me</sup> (Seine).</p> <p><b>0.31 ORDENER (Rue) 49, 50, 51</b></p> <p>1-3 Duchonne, Paris, r. Allemagne, 120.          5-7 Hétère, Fontainebleau (S.-et-M.).          9 Hétère r. Jean Robert, 20.          11 Hétère, Paris, r. Oues, 11.          13 Hétère, Paris, r. Oues, 13.          15 à 19 Chénouy de fer du Nord.          21-25 Jean (Ver), est. Hétère r. Paris, r. Clamart, 162.          25 La Courde du Malin (H<sup>er</sup>), Paris, r. Montaigne, 5.          27 Hétère, Paris, r. Tourgny, 71.          29 Hétère, Paris, r. Oues, 29.          31-33 Hétère, Paris, r. Oues, 31.          33 Lavallette, Bussy l'Abbaye (S.-et-O.).          35 à 41 Chénouy de fer du Nord.          43 Hétère, Paris, r. Oues, 43.          45 Langelot, Paris, r. Oues, 45.          47-49 Hétère r. Maréchal, 28-32.          51 Hétère, Paris, r. Gouffon, 7.          53 Hétère r. Poissonnière, 53.          55 Hétère r. Poissonnière, 65.          57 Chénouy, Hétère St-Léger, r. Maréchal, 3 (S.-et-O.).          59 Hétère, Paris, r. Mont-Croix, 114.          61 Hétère, Paris, r. Oues, 61.          63 Hétère (Ver), Paris, r. Baigny, 42.          65-67 Hétère, Paris, r. La Harpe, 44.          69-71 Ville de Paris.          65-67 Hétère, Paris, r. Chagny, 56.          73-75 Hétère, local, Barde, 79.          77-79 Hétère (Ver), Oues, r. Compère, 8 (Seine).          77 Hétère, Paris, r. Oues, 77.          79 à 89 St. La r. Hétère Focier, Paris, r. La Harpe, 12.          91-93 Hétère, Paris, r. Oues, 91.          95 Hétère, Paris, r. Oues, 95.          103-107 Hétère, Paris, r. Lafayette, 146.          109-111 Hétère, Paris, r. Antoine Vermet, 8.          113 Hétère, Paris, r. Mont-Croix, 27.          115 Hétère r. Hémel, 14.          117-119 Hétère r. Maréchal, 112.          119-121 Hétère (M<sup>er</sup>), Paris, r. Oues, 119.          121 Hétère r. Dubouche, 56.          123 Hétère, Paris, r. Solitaire, 7.          125-127 Ville de Paris.          129 Hétère, Paris, r. Oues, 129.          131 Hétère (M<sup>er</sup> 46), Paris, r. Hétère d'Artois, 11.          133 Ville de Paris.          135-139<sup>me</sup> Hétère, Paris, r. Baigny, 14.</p>
---	---

Figure 5: *Annuaire-almanach du commerce, de l'industrie..., 1898*

<p><b>BÉNARD (AUGUSTE-SÉBASTIEN),</b>          A Paris, rue de Dunkerque, 85; élève de MM. Grosjean et Lafont.</p> <p>54. Le Tombeur.          55. Le Retour de la ville.</p> <p><b>BÉNARD (HUBERT-EGGÈNE),</b>          A Boulogne-sur-Mer, route de Calais, 3; élève de M. Claudius Gouffon.</p> <p>56. La Brise en rade de Boulogne-sur-Mer.          57. Barque de Pêcheurs du Portel, levant l'ancre par un gros temps (marine).</p> <p><b>BENNETTER (J.),</b>          Peintre de marins, à Paris, rue du Faubourg-Saint-Hippolyte, 217; élève de MM. L. Meyer et T. Guélin.</p> <p>58. Effet de Lune.          Une escadre venant en grande rade; le premier plan représente un navire marchand (brick), le second et vient de bord, pour éviter les autres, et approuve le signal de nuit.          59. Matin d'été.          Mer de Nord; au premier plan, un (Hortlogorjok), bateau norvégien de cabotage, sur la côte de la Norvège, est libéré par des pêcheurs.          40. Calme : Estre Bagholmen (Norvège).          Appartient à M. J. Deléclère, à Rouen.</p> <p><b>BENTABOLE (LOUIS),</b>          A Paris, rue Pigalle, 22.</p> <p>42. Vue de Calais.</p>
--

Figure 6: *Catalogue de l'exposition annuelle du musée de Rouen, 1860*

## Datasets

Prints' title and author	Gallica's ark	Date	No. of pages	Prints' specificities
<i>Lettres du Sieur Balzac</i> , Jean-Louis Balzac	bvt1b86262420	1624	57	None
<i>Méduse</i> , Claude Boyer	bpt6k311844g	1697	34	No Damage zone
<i>Les caractères de Théophraste</i> , Jean de la Bruyère	bvt1b86070385	1688	36	No Damage zone
<i>Histoire amoureuse des Gaules</i> , Bussy-Rabutin	bvt1b8623309s	1665	32	No Running/Title zone
<i>Le théâtre de P. Corneille</i> , Pierre Corneille	bpt6k10403751	1664	47	None
<i>Discours de la méthode</i> , René Descartes	bvt1b86069594	1637	14	None
<i>La princesse de Clèves</i> , Madame de La Fayette	bvt1b8610820b	1678	53	No Margin zone
<i>La Marianne</i> , Tristan L'Hermitte	bpt6k1511072f	1639	31	No Title zone
<i>L'Escole des femmes</i> , Molière	bvt1b8610785b	1663	37	No Damage and Margin zones
<i>George Dandin, ou le Mary confondu</i> , Molière	bvt1b8610793w	1669	66	No Margin zone
<i>Dom Garcie de Navarre</i> , Molière	+Z258398909 <sup>3</sup>	1694	20	None
<i>Statira</i> , Nicolas Pradon	bpt6k8416272	1680	38	No Rubric line
<i>Athénaïs</i> , Nicolas Pradon	bpt6k857200c	1697	34	None
<i>Les plaideurs</i> , Jean Racine	bvt1b8610811c	1669	37	No Rubric line
<i>Oeuvres de Racine. Tome Premier</i> , Jean Racine	bpt6k9905809	1676	19	No Damage zone
<i>Oeuvres de Racine. Tome Second</i> , Jean Racine	bpt6k990581p	1676	46	None

6

Table 4: 17<sup>th</sup> century prints dataset description (always with zones Main, Decoration, DropCapital, Numbering, RunningTitle, Signatures, Stamp and always with lines Default and DropCapitalLine).

Manuscript ID	Date	No. of pages	No. of columns	Running Title	Drop Capital
BnF, Arsenal 3516	13th	10	4	No	Yes
BnF, ms fr. 22549	14th	3	3	No	Yes
BnF, ms fr. 24428	13th	20	2	No	Yes
BnF, ms fr. 412	13th	49	2	Yes	Yes
BnF, ms fr. 844	13th	18	2	No	Yes
Cologne, bodmer, 168	13th	22	2	No	Yes
Vaticane, Reg. Lat., 1616	14th	41	1	No	Yes

Table 5: Medieval dataset description

Prints' title	Type	Date	No. of pages	No. of columns	Other zones
<i>Annuaire-almanach du commerce, de l'industrie...</i>	Annuaire	1898	50	2	Numbering
<i>Exposition des oeuvres de M. Courbet à l'École des Beaux Arts</i>	Exhibition	1882	19	1	Numbering
<i>Catalogue des oeuvres exposées, Société des Indépendants</i>	Exhibition	1892	5	1	Numbering
<i>Catalogue des oeuvres exposées, Société des Indépendants</i>	Exhibition	1913	7	1	Numbering
<i>Catalogue des oeuvres exposées, Société des Indépendants</i>	Exhibition	1923	5	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exhibition	1818	2	1	Numbering
<i>Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg</i>	Exhibition	1867	5	1	Numbering
<i>Catalogue de l'exposition des Beau-Arts de Nancy</i>	Exhibition	1843	5	1	Numbering
<i>Catalogue de l'exposition des Beau-Arts de Nancy</i>	Exhibition	1849	5	1	Numbering
<i>Catalogue de l'exposition des Beau-Arts de Nancy</i>	Exhibition	1892	5	1	Numbering
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1961	5	1	Running Title
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1965	4	1	Numbering Running Title
<i>Catalogue de l'exposition de la biennale de Paris</i>	Exhibition	1969	5	1	Numbering Running Title
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1853	5	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1869	7	1	Numbering
<i>Catalogue de l'exposition du Musée des Beaux Arts de Rouen</i>	Exhibition	1888	7	1	Numbering Running Title
<i>Catalogue de la Biennale de Sao Paulo</i>	Exhibition	1951	6	1	Numbering
<i>Catalogue de la Biennale de Sao Paulo</i>	Exhibition	1972	5	1	Numbering
<i>Catalogue de l'exposition de la société des amis des arts de Strasbourg</i>	Exhibition	1884	15	1	Numbering
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1884	5	1	Numbering Running Title
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1895	3	1	Numbering Running Title
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1905	3	1	Numbering Running Title
<i>Catalogue de la Biennale de Venise</i>	Exhibition	1920	5	1	Numbering Running Title Title
<i>Revue des Autographes</i>	Manuscripts	1870	5	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1871	6	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1873	4	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1877	4	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1880	2	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1881	2	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1883	5	1	Numbering
<i>Revue des Autographes</i>	Manuscripts	1885	2	1	Numbering
<i>Catalogue de ventes de manuscrits Charavay</i>	Manuscripts	1845	6	1	Numbering
<i>Catalogue de ventes de manuscrits Laverdet</i>	Manuscripts	1856	4	1	Numbering Title
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1857	6	1	Numbering Title
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1866	7	1	Numbering Title Stamp
<i>Catalogue de vente de manuscrits Bovet</i>	Manuscripts	1887	14	1	Numbering Running Title Figure
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1857	7	1	Numbering Title
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1899	6	1	Numbering
<i>Catalogue de vente de manuscrits Kra</i>	Manuscripts	1912	9	2	Numbering
<i>Catalogue de vente de manuscrits Charavay</i>	Manuscripts	1919	8	1	Numbering

Table 6: Catalogs dataset description