Towards automatic TEI encoding via layout analysis

Juliette Janes, Ariane Pinche, Claire Jahan, Simon Gabay

Towards automatic TEI encoding via layout analysis

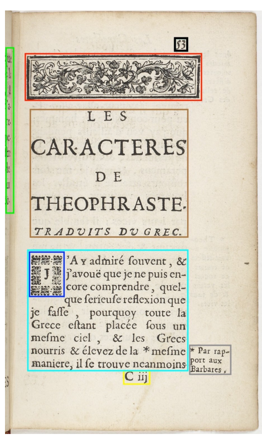
Juliette Janes, Ariane Pinche, and Claire Jahan

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.



53

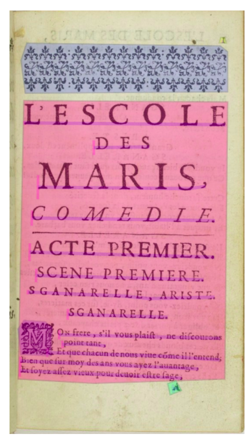
LES
CARACTERES
DE
THEOPHRASTE.
TRADUITS DU GREC.

J'Ay admiré souvent, & j'avouë que je ne puis encore comprendre, quelque serieuse reflexion que je fasse, pourquoy toute la Grece estant placée sous un mesme ciel, & les Grecs nourris & élevez de la * mesme maniere, il se trouve neanmoins

* Par rapport aux Barbares.

C iij

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



L'ESCOLE
DES
MARIS,
COMEDIE.
ACTE PREMIER.
SCENE PREMIERE.
SGANARELLE, ARISTE.
SGANARELLE.

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*[1] (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$^{\text{th}}$ c. literary prints (cf. table 4), and one of 19$^{\text{th}}$ c. catalogues (cf. table 6).

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.

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$^{\text{th}}$ c. data on the one hand, the 17$^{\text{th}}$ c. data and the 19$^{\text{th}}$ c. on the other hand (cf. table 2).

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

[1] `https://segmonto.github.io`.

	Medieval data	17^{th} c.	19^{th} 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^{th} c., Combination 2 gathers 17^{th} c. and 19^{th} c. data.

VGSL specs (1 200→1 800)[2], 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^{th} c.	19^{th} 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^{th} c. can easily 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`.

[2]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

Examples

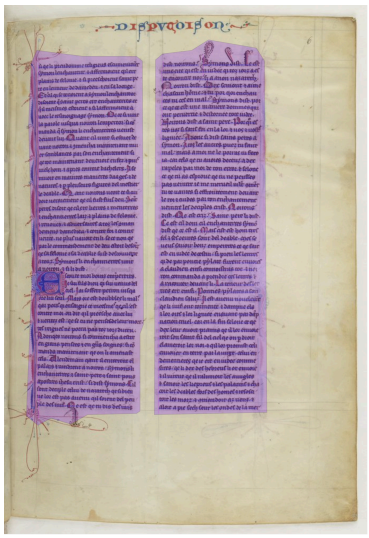


Figure 3: *Manuscrit, BnF fr. 412, f°21, 13th c.*

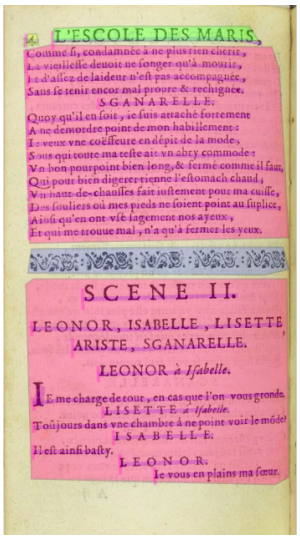


Figure 4: *Molière, L'Ecole des maris, Paris, G. de Luyne*, 1661

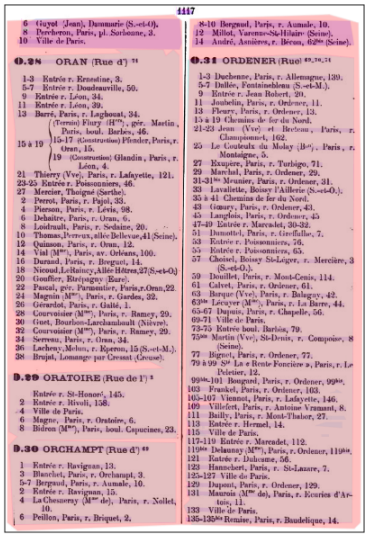


Figure 5: *Annuaire-almanach du commerce, de l'industrie...*, 1898

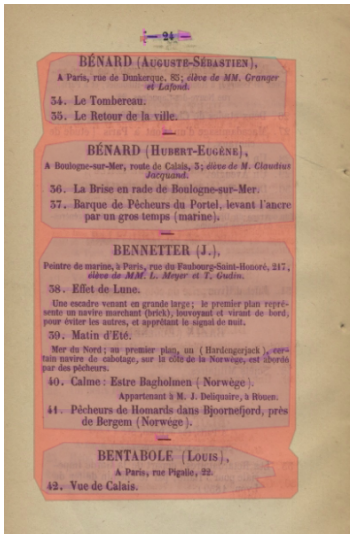


Figure 6: *Catalogue de l'exposition annuelle du musée de Rouen*, 1860

Datasets

Prints' title and author	Gallica's ark	Date	No. of pages	Prints' specificities
Lettres du Sieur Balzac, Jean-Louis Balzac	btv1b86262420	1624	57	None
Méduse, Claude Boyer	bpt6k311844g	1697	34	No Damage zone
Les caractères de Théophraste, Jean de la Bruyère	btv1b86070385	1688	36	No Damage zone
Histoire amoureuse des Gaules, Bussy-Rabutin	btv1b8623309s	1665	32	No RunningTitle zone
Le théâtre de P. Corneille, Pierre Corneille	bpt6k10403751	1664	47	None
Discours de la méthode, René Descartes	btv1b86069594	1637	14	None
La princesse de Clèves, Madame de La Fayette	btv1b8610820b	1678	53	No Margin zone
La Mariane, Tristan L'Hermite	bpt6k1511072f	1639	31	No Title zone
L'Escole des femmes, Molière	btv1b8610785b	1663	37	No Damage and Margin zones
George Dandin, ou le Mary confondu, Molière	btv1b8610793w	1669	66	No Margin zone
Dom Garcie de Navarre, Molière	+Z258398909 [3]	1694	20	None
Statira, Nicolas Pradon	bpt6k8416272	1680	38	No Rubric line
Athénaïs, Nicolas Pradon	bpt6k857200c	1697	34	None
Les plaideurs, Jean Racine	btv1b8610811c	1669	37	No Rubric line
Oeuvres de Racine. Tome Premier, Jean Racine	bpt6k9905809	1676	19	No Damage zone
Oeuvres de Racine. Tome Second, Jean Racine	bpt6k990581p	1676	46	None

Table 4: 17^{th} 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
Cologny, 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
Annuaire-almanach du commerce, de l'industrie…	Annuaire	1898	50	2	Numbering
Exposition des oeuvres de M. Courbet à l'École des Beaux Arts	Exhibition	1882	19	1	Numbering
Catalogue des oeuvres exposées, Société des Indépendants	Exhibition	1892	5	1	Numbering
Catalogue des oeuvres exposées, Société des Indépendants	Exhibition	1913	7	1	Numbering
Catalogue des oeuvres exposées, Société des Indépendants	Exhibition	1923	5	1	Numbering
Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg	Exhibition	1818	2	1	Numbering
Catalogue des peintures, sculptures et miniatures, Palais du Luxembourg	Exhibition	1867	5	1	Numbering
Catalogue de l'exposition des Beaux-Arts de Nancy	Exhibition	1843	5	1	Numbering
Catalogue de l'exposition des Beaux-Arts de Nancy	Exhibition	1849	5	1	Numbering
Catalogue de l'exposition des Beaux-Arts de Nancy	Exhibition	1892	5	1	Numbering
Catalogue de l'exposition de la biennale de Paris	Exhibition	1961	5	1	Running Title
Catalogue de l'exposition de la biennale de Paris	Exhibition	1965	4	1	Numbering Running Title
Catalogue de l'exposition de la biennale de Paris	Exhibition	1969	5	1	Numbering Running Title
Catalogue de l'exposition du Musée des Beaux Arts de Rouen	Exhibition	1853	5	1	Numbering
Catalogue de l'exposition du Musée des Beaux Arts de Rouen	Exhibition	1869	7	1	Numbering
Catalogue de l'exposition du Musée des Beaux Arts de Rouen	Exhibition	1888	7	1	Numbering Running Title
Catalogue de la Biennale de Sao Paulo	Exhibition	1951	6	1	Numbering
Catalogue de la Biennale de Sao Paulo	Exhibition	1972	5	1	Numbering
Catalogue de l'exposition de la société des amis des arts de Strasbourg	Exhibition	1884	15	1	Numbering
Catalogue de la Biennale de Venise	Exhibition	1895	5	1	Numbering Running Title
Catalogue de la Biennale de Venise	Exhibition	1905	3	1	Numbering Running Title
Catalogue de la Biennale de Venise	Exhibition	1920	5	1	Numbering Running Title Title
Revue des Autographes	Manuscripts	1870	5	1	Numbering
Revue des Autographes	Manuscripts	1871	6	1	Numbering
Revue des Autographes	Manuscripts	1873	4	1	Numbering
Revue des Autographes	Manuscripts	1877	4	1	Numbering
Revue des Autographes	Manuscripts	1880	2	1	Numbering
Revue des Autographes	Manuscripts	1881	2	1	Numbering
Revue des Autographes	Manuscripts	1883	5	1	Numbering
Revue des Autographes	Manuscripts	1885	2	1	Numbering
Catalogue de Ventes de manuscrits Charavay	Manuscripts	1845	6	1	Numbering
Catalogue de ventes de manuscrits Laverdet	Manuscripts	1856	4	1	Numbering Title
Catalogue de vente de manuscrits Charavay	Manuscripts	1857	6	1	Numbering Title
Catalogue de vente de manuscrits Charavay	Manuscripts	1866	7	1	Numbering Title Stamp
Catalogue de vente de manuscrits Bovet	Manuscripts	1887	14	1	Numbering Running Title Figure
Catalogue de vente de manuscrits Charavay	Manuscripts	1857	7	1	Numbering Title
Catalogue de vente de manuscrits Charavay	Manuscripts	1899	6	1	Numbering
Catalogue de vente de manuscrits Kra	Manuscripts	1912	9	2	Numbering
Catalogue de vente de manuscrits Charavay	Manuscripts	1919	8	1	Numbering

Table 6: Catalogs dataset description