SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, Claire Jahan

SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)

Simon Gabay[1][0000−0001−9094−4475], Jean-Baptiste Camps[2][0000−0003−0385−7037], Ariane Pinche[2][0000−0002−7843−5050], and Claire Jahan[2][0000−0001−7268−9706]

[1] Université de Genève, Rue des Battoirs 7, 1205 Genève, Switzerland
{prenom.nom}@unige.ch
[2] École nationale des Chartes, 65 Rue de Richelieu, 75002 Paris, France
{prenom.nom}@chartes.psl.eu

Keywords: layout analysis · controlled vocabulary · manuscripts · old prints.

Document and layout description is a traditional task for philologists but also a need for many computational applications in document analysis, ranging from content categorisation to text recognition, and has, for this reason, been the subject of much research in computer vision [1]. For tools relying on machine learning, the existence and availability of data sets, and their interoperability are important issues.

The definition of ontologies or controlled vocabularies for the description of manuscript layout has attracted some attention. Codicological dictionaries exist [6], part of which have already been integrated into SKOS [3]. On the other hand, digitisation standards have developed their own taxonomy, such as the PAGE XML Format [7]. In between these two approaches, initiatives like the TEI [8] offer elements commonly used by editors. With the apparition of efficient layout analysers [4] and user-friendly interfaces to use them [5], the need for efficient models is rising, and so is the need for large amount of data, based on the aggregation of heterogeneous documents. For this, researchers to need to **a** agree on a common limited vocabulary, based on existing standards; **b** share common practices to ease the interoperability of their ground truth.

The SegmOnto project [2] gathers scholars from different backgrounds who have decided to tackle both issues. It mainly addresses the case of manuscripts (fig. 1 and 2), but also old prints (fig. 3 and 4). Our work is characterised by two key choices:

1. focus on common material features rather than semantic descriptions (*e.g.* marginal text, rather than gloss, commentary, note, etc.).
2. use of two levels of description: zones (main text, notes, figure, damage, seal…) and lines (default, musical, interlinear, rubric, drop capital).

It has to be noted that, as shown in the examples *infra*, that the detection of zones and lines can rely on the position on the page, but also often on the variation of hands as well as

script cursive vs block letters (fig. 5), square letters vs rashi script (fig. 7), roman vs italic (fig. 8).
module (fig. 5, 6, 7, 8)
ink blue and red (fig. 5), just red (fig. 6) or a different type of black (fig. 5).

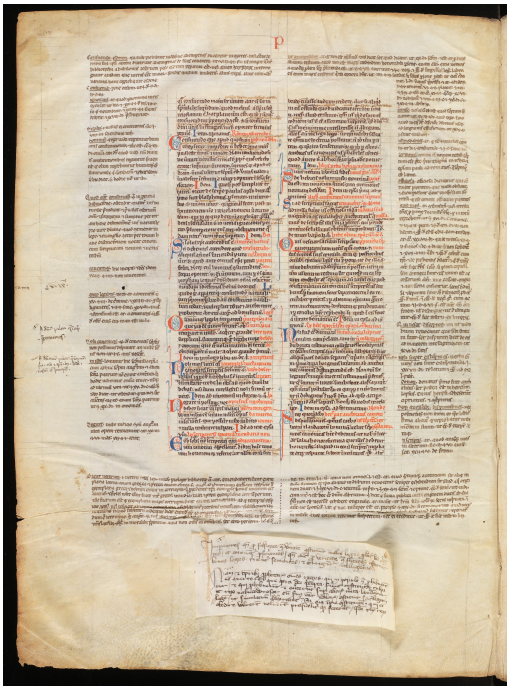

Fig. 1. *Decretum Gratiani*, Sion/Sitten, Archives du Chapitre/Kapitelsarchiv, Ms. 89, f° 3v

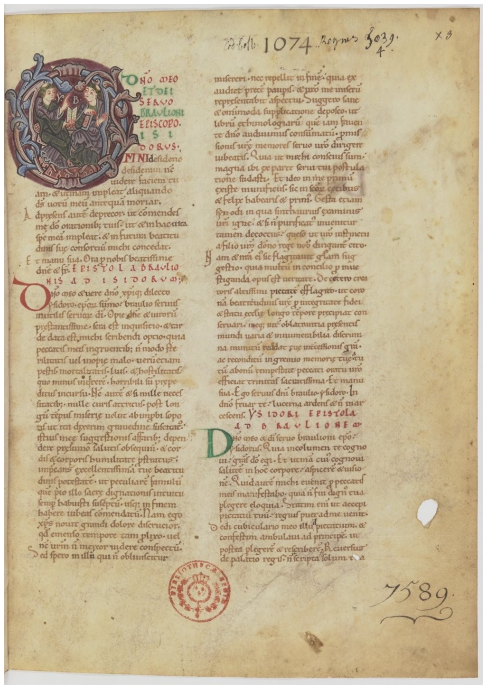

Fig. 2. Isidorus Hispalensis, *Etymologiarum*, Paris, BNF, lat. 7589, f°3r

Fig. 3. *Babylonian Talmud*, Seder Zera'im, Venice: Daniel Bomberg, [1543-44]

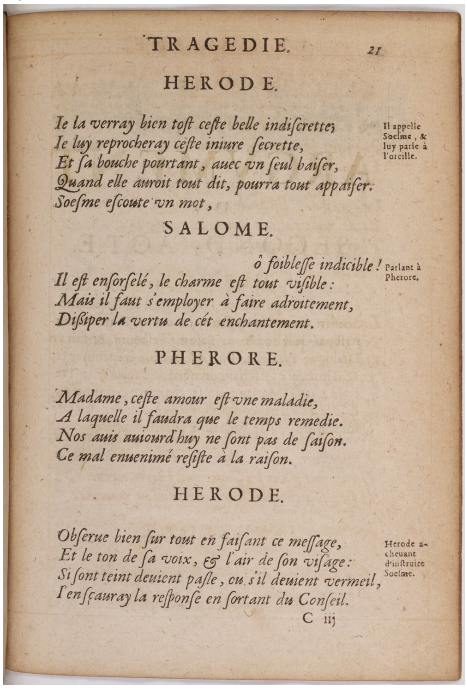

Fig. 4. Tristan L'Hermite, *La Mariane*, Augustin Courbé: Paris, 1639.

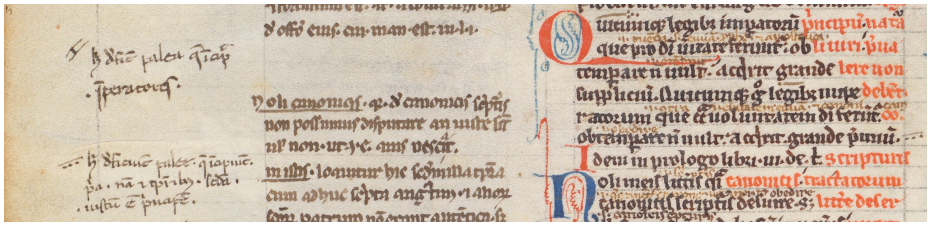

Fig. 5. *Decretum Gratiani*, Sion/Sitten, Archives du Chapitre/Kapitelsarchiv, Ms. 89, f° 3v

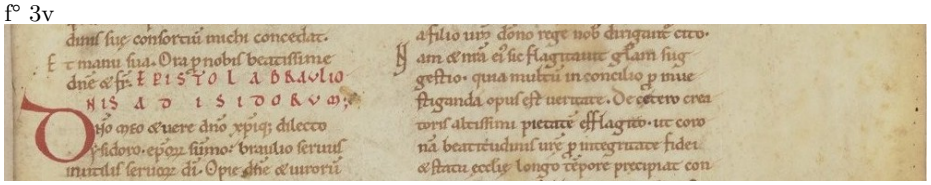

Fig. 6. Isidorus Hispalensis, *Etymologiarum*, Paris, BNF, lat. 7589, f°3r

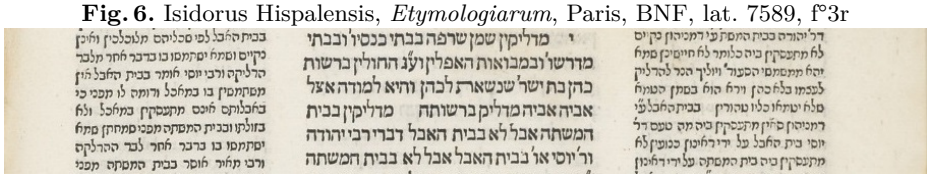

Fig. 7. *Babylonian Talmud*, Seder Zera'im, Venice: Daniel Bomberg, [1543-44]

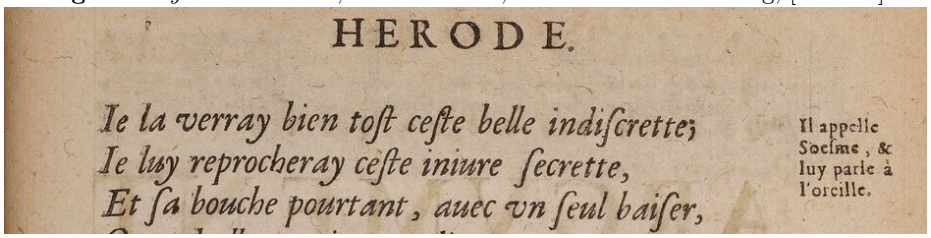

Fig. 8. Tristan L'Hermite, *La Mariane*, Augustin Courbé: Paris, 1639.

References

1. Binmakhashen, G.M., Mahmoud, S.A.: Document Layout Analysis: A Comprehensive Survey. ACM Computing Surveys **52**(6), 109:1–109:36 (Oct 2019). https://doi.org/10.1145/3355610, `https://doi.org/10.1145/3355610`
2. Camps, J.B., Gabay, S., Chagué, A., Chiffoleau, F., Stoekl, D., Salvatti, B., Albouy, S.: SegmOnto (2021), `https://github.com/SegmOnto`
3. Geoffroy, M., Eddé, A.M., Baratli, Y., Muzerelle, D., Bobichon, P., Guesdon, M.G.: Vocabulaire Internationale de la Codicologie - SKOS Vocabulaire Internationale de la Codicologie - SKOS Vocabulaire Internationale de la Codicologie - SKOS - GAMS:

Vokabularien und Ontologien. GAMS Vokabularien und Ontologien, Zentrum für Informationsmodellierung - Karl-Franzens-Universität Graz, Graz (2021), http://gams.uni-graz.at/archive/objects/o:voccod/methods/sdef:SKOS/get
4. Kiessling, B.: A modular region and text line layout analysis system. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 313–318. IEEE (2020). https://doi.org/10.1109/ICFHR2020.2020.00064
5. Kiessling, B., Tissot, R., Stokes, P., Ezra, D.S.B.: eScriptorium: An open source platform for historical document analysis. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–24. IEEE (2019)
6. Muzerelle, D.: Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits. Rubricae, 1, Institut de recherche et d'histoire des textes (1985), http://codicologia.irht.cnrs.fr/
7. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: 2010 20th International Conference on Pattern Recognition. pp. 257–260. IEEE (2010). https://doi.org/10.1109/ICPR.2010.72
8. TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2020), https://tei-c.org/Guidelines/