



HAL
open science

SegmOnto : Vocabulaire contrôlé pour décrire les manuscrits et les imprimés

Ariane Pinche, Simon Gabay, Jean-Baptiste Camps

► To cite this version:

Ariane Pinche, Simon Gabay, Jean-Baptiste Camps. SegmOnto : Vocabulaire contrôlé pour décrire les manuscrits et les imprimés. Segmenter et annoter les images : déconstruire pour reconstruire, Nov 2022, Paris, France. hal-03930487

HAL Id: hal-03930487

<https://hal.science/hal-03930487>

Submitted on 9 Jan 2023

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 - NonCommercial - ShareAlike 4.0 International License

SegmOnto

Vocabulaire contrôlé pour décrire les manuscrits et les imprimés

Ariane Pinche¹ Simon Gabay² Jean-Baptiste Camps³

CIHAM (UMR 5648), CNRS¹

Université de Genève, Switzerland²

Centre Jean Mabillon, École nationale des chartes - PSL³

Journée Campus Richelieu, 15 novembre 2022

Outline

- 1 Introduction
- 2 Pourquoi et comment ?
- 3 Constituer un vocabulaire
- 4 Exploitation des prédictions HTR
- 5 Conclusion

Un même vocabulaire pour décrire toutes les sources



Figure – BNF, Fr. 412, f. 10^r.

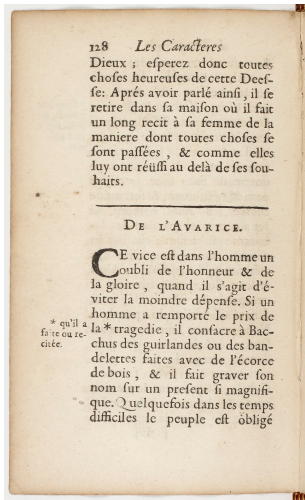


Figure – *Les Caractères*, 1688, p. 128.

Un même vocabulaire pour décrire toutes les sources

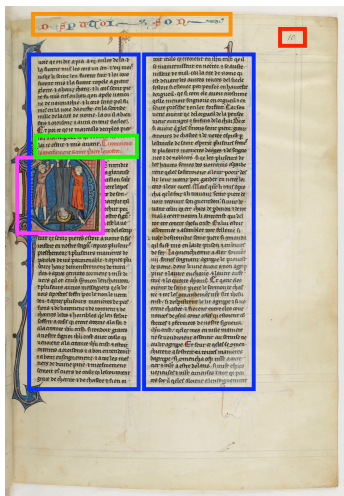


Figure – BNF, Fr. 412, f. 10^r.

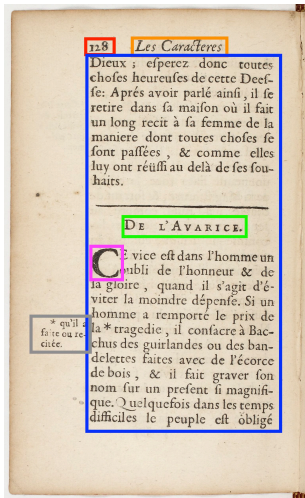


Figure – *Les Caractères*, 1688, p. 128.

Un même vocabulaire pour décrire toutes les sources



Figure – BNF, Fr. 412, f. 10^r.

- Titre courant, **numéro de page**, **corps de texte**, **capital ornée**, **rubrique**...
- Une approche diachronique (du moyen âge au XIX^e siècle)
- Une description matérielle et pas sémantique
- ▶ Standardiser pour partager

Outline

- 1 Introduction
- 2 Pourquoi et comment ?**
- 3 Constituer un vocabulaire
- 4 Exploitation des prédictions HTR
- 5 Conclusion

Pourquoi ?

Partager les données : *en amont*

Partager des données annotées pour améliorer les modèles de segmentation en mettant les données d'entraînement en commun

Partager les données : *en aval*

Partager des protocoles pour l'exploration et la production/transformation automatique des prédictions HTR

→ **Comment ? Mettre en place un vocabulaire commun pour décrire les documents**

Comment ?

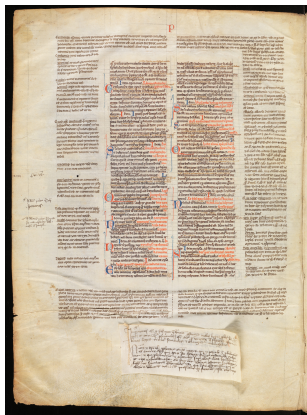


Figure – *Decretum Gratiani*, Sion, Archives du Chapitre, Ms. 89, f° 3v

La mise en page est la première source d'information pour appréhender un document. Ici, on identifie :

- deux colonnes
- plusieurs niveaux de notes
- des rubriques
- un ajout
- ...

Outline

- 1 Introduction
- 2 Pourquoi et comment ?
- 3 Constituer un vocabulaire**
- 4 Exploitation des prédictions HTR
- 5 Conclusion

Trouver des termes communs à une large communauté scientifique



- S'appuyer sur un lexique existant
- *Codicologia*, regroupe des termes utilisés par une large communauté pour décrire les documents manuscrits (+ multilingue)¹ (c. 1500 entries);
- Adaptation de l'existant : *Vocabulaire international de la codicologie - SKOS*²;

Figure – *Codicologia*

¹. <http://codicologia.irht.cnrs.fr/>.

Trouver la bonne définition

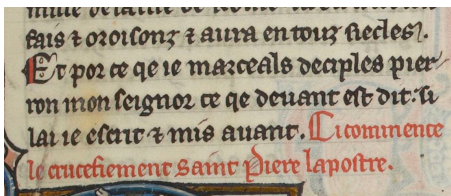


Figure – BNF, Fr. 412, f. 10^r.

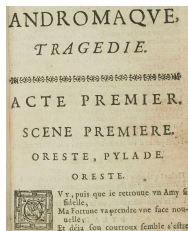


Figure – *Andromaque*, 1688, p. 1.

Comment définir la bonne catégorie pour chaque élément, même quand ils peuvent être relativement différents d'un document à un autre.

Trouver la bonne granularité



(a) Headpiece



(b) Drop capital



(c) Tailpiece



(d) Engraving

Figure – Exemple de décorations issues de *L'Ambassade de la Compagnie orientale des Provinces unies vers l'empereur de la Chine* de Jean-Baptiste Carpentier, Leiden, 1665.

Vocabulaire SegmOnto : zones et types de lignes

Zone

CustomZone
DamageZone
DigitizationArtefactZone
DropCapitalZone
GraphicZone
MainZone
MarginTextZone
MusicZone
NumberingZone
QuireMarksZone
RunningTitleZone
SealZone
StampZone
TableZone

Line

CustomLine
DefaultLine
DropCapitalLine
HeadingLine
InterlinearLine
MusicLine

Definitions

<https://segmonto.github.io>

Exemple : QuireMarksZone



Figure – BNF, Fr. 282, f. 168^v.

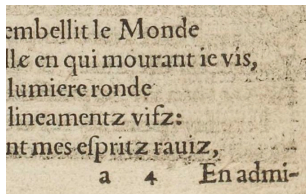


Figure – *Délie, object de plus haulte vertu*, 1544, p. 7.

Définition

QuireMarksZone est une zone qui contient une réclame (e.g., a ii), ou tout type d'élément relatif à l'assemblage des cahiers. La zone se situe le plus souvent en bas de la page.

Enjeux et problèmes

- Doit-on les distinguer de la numérotation des pages ?
- Doit-on distinguer les réclames qui utilisent un mot de celles qui utilisent des chiffres ?

Sous catégories

- QuireMarksZone:signature
- QuireMarksZone:catchword

Outline

- 1 Introduction
- 2 Pourquoi et comment ?
- 3 Constituer un vocabulaire
- 4 Exploitation des prédictions HTR**
- 5 Conclusion

ALTO vers TEI

Dans le cadre du projet, Gallic(orpor)a, créons directement des fichiers TEI à partir des fichiers ALTO de l'HTR :

```

<sourceDoc>
  <!-- Page -->
  <surface xml:id="page1" cert="gold">
    <graphic url="uri"/>
    <!-- Block of text (ex. "MainZone:column#1") -->
    <zone xml:id="page1_zone1" type="MainZone" subtype="column" n="1" points="1,2 3,4 5,6"
    ↪ source="uri">
      <!-- Mask of text line (ex. "DefaultLine") -->
      <zone xml:id="page1_zone1_line1" type="DefaultLine" subtype="none" n="1" points="1,2 3,4 5,6"
      ↪ source="uri">
        <!-- Baseline -->
        <path xml:id="page1_zone1_line1b" points="1,2 3,4 5,6"/>
        <!-- Transcription -->
        <line xml:id="page1_zone1_line1t">Some text</line>
      </zone>
    </zone>
  </surface>
</sourceDoc>

```

Exemple de <sourceDoc>.

Vers une pre-éditorialisation des corpus

Grâce à segmOnto, l'information textuelle peut être hiérarchisée dans le <body>

```

<body>
  <!-- page -->
  <pb corresp="#page1"/>
  <!-- RunningTitleZone -->
  <fw type="header">
    <lb corresp="#page1_zone1_line1"/>TITLE
  </fw>
  <!-- NumberingZone -->
  <fw type="pageNum">
    <lb corresp="#page1_zone2_line1"/>NUMBER
  </fw>
  <!-- GraphicZone -->
  <figure type="graphicZone" corresp="#page1_zone3"/>
  <!-- MainZone -->
  <ab corresp="#page1_zone4">
    <lb corresp="#page1_zone4_line1"/>A LINE
    <lb facs="#page1_zone4_line2"/>ANOTHER LINE
  </ab>
  <!-- QuireMarksZone -->
  <fw type="quireMarks">
    <lb corresp="#page1_zone5_line1"/>SIGNATURE
    <lb corresp="#page1_zone5_line2"/>CATCHWORD
  </fw>
</div>
</body>

```

Outline

- 1 Introduction
- 2 Pourquoi et comment ?
- 3 Constituer un vocabulaire
- 4 Exploitation des prédictions HTR
- 5 Conclusion**

Utilisation du vocabulaire

- Pour retrouver des éléments des documents numérisés et segmentés
 - Des éléments de "type image" : décorations
 - Des contenus textuels : titres, notes marginales, etc.
- Pour partager des données :
 - *Cremma Medieval* :
<https://github.com/HTR-United/cremma-medieval>
 - *Cremma medii aevii* :
<https://github.com/HTR-United/CREMMA-Medieval-LAT>
 - *Gallic(orpor)a* : <https://github.com/Gallicorpora>