Segmentation de séquences de pages d'ouvrages anciens basée sur une signature structurelle des images

Maroua MEHRI *† - Pierre HÉROUX* - Petra GOMEZ-KRÄMER‡ - Rémy MULLOT‡

Financé par l'**ANR** dans le cadre du projet de recherche **DIGIDOC**

01 octobre 2015

*. Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS), Université de Rouen
†. maroua.mehri@gmail.com
‡. Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle

Problématique

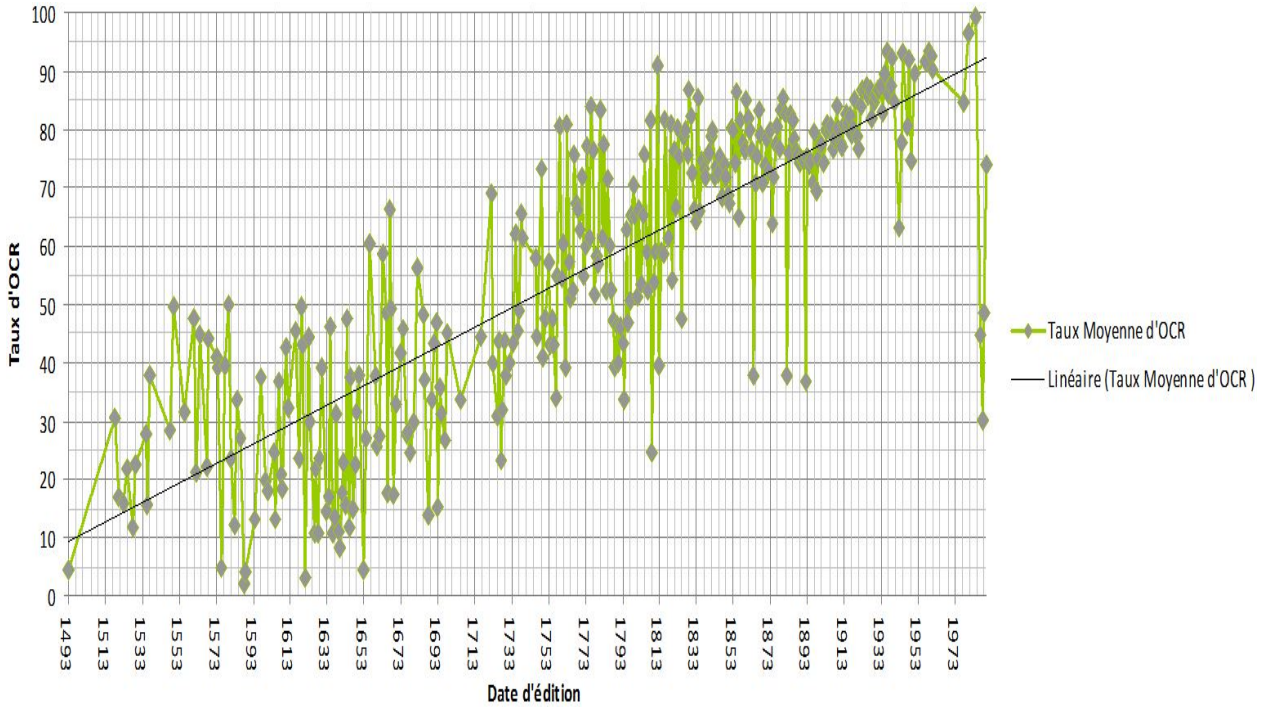
Bibliothèques numériques

- **Accès** aux documents patrimoniaux
- Outil d'**indexation**
- Système d'**interprétation** & de **recherche**

Images de documents patrimoniaux & OCR


A REVEREND I.SSIME
PRELAT, MESSIRE CHARLES
de Guillard, Eueſque de
Chartres.

MONSEIGNEVR, il y a ia long temps que ie vous euſſe preſenté les premiers fruicts de mes eſtudes & labeurs, ſi vne honte aucunemẽt ruſtique, & non paraduenture recuable ne m'en euſt par pluſieurs fois deſtourné, craignant vous empeſcher à la lecture de choſes ſi baſſes que ſont celles qui partent de mon entendement & ſcauoir, qui auez touſiours eu le maniement de grãdes charges & affaires en ce voyaume. Or à preſent



Résultats d'application d'un **OCR** sur des documents anciens [Cron (2012), BenSalah et al. (2013), Grana et al. (2014)]

Masque rouge : mot reconnu par l'OCR
Masque vert : élément graphique détecté par l'OCR

⇒**Performances faibles** en raison des **particularités** des documents anciens

[Cron (2012)] G. Cron, "Éléments ayant une influence sur la qualité et l'efficacité des OCRs," BnF, Tech. Rep., 2012
[BenSalah et al. (2013)] A. BenSalah, N. Ragot, and T. Paquet, "Adaptive detection of missed text areas in OCR outputs: application to the automatic assessment of OCR quality in mass digitization projects," in *DRR*, 2013
[Grana et al. (2014)] C. Grana, G. Serra, M. Manfredi, D. Coppi, and R. Cucchiara, "Layout analysis and content enrichment of digitized books," *MTA*, 2014

Particularités d'images de documents patrimoniaux

[Coustaty et al. (2011), Gallica]

[Coustaty et al. (2011)] M. Coustaty, R. Raveaux, and J. M. Ogier, "Historical document analysis: a review of French projects and open issues," in *EUSIPCO*, 2011

Objectifs

- Proposer une **solution complémentaire à l'OCR** pour l'indexation
- **S'affranchir**
  - **Limitations** des OCRs
  - **Connaissances *a priori*** (structure & contenu)

- **Sans aucune connaissance** sur la **structure des mises en page**
- **Sans connaissances *a priori*** sur les caractéristiques **typographiques & graphiques du contenu**

Hypothèses & contexte

- **Hypothèses**
  - **Redondance** du contenu des pages d'un même ouvrage
  - **Régularité** ou **homogénéité** des structures et contenus des pages d'un même ouvrage
- **Contexte**
  - **Réduire la complexité** du manque d'information à disposition
    $\Longrightarrow$ Approche **ouvrage** (*i.e.* redondance de l'information)

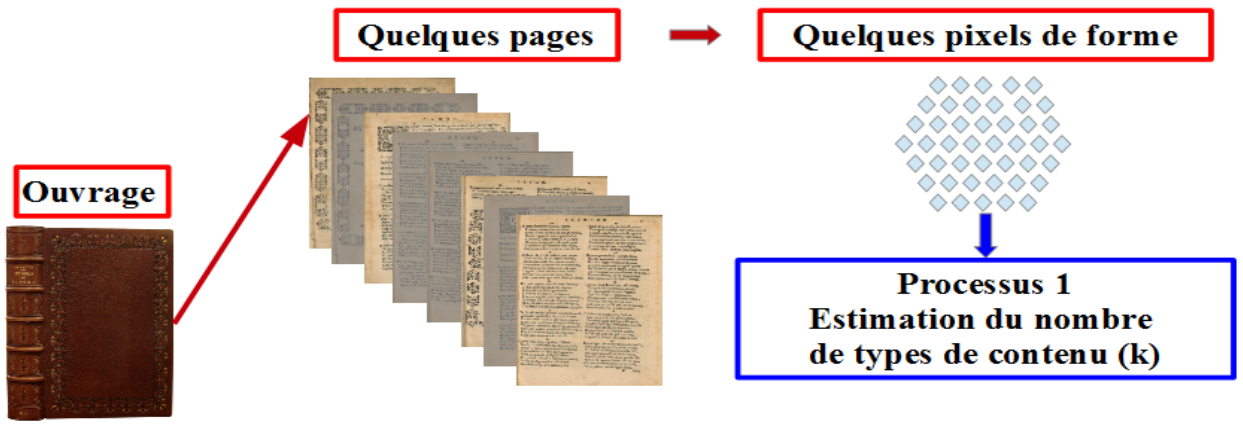
$\Rightarrow$ **Clés d'indexation : structure, redondance & régularité**

Plateforme d'étiquetage des pixels du contenu

- **Processus 1** : extraction et analyse d'information bas-niveau sur un extrait de l'ouvrage
- **Processus 2** : caractérisation de toutes les pages de l'ouvrage en exploitant les informations du processus 1

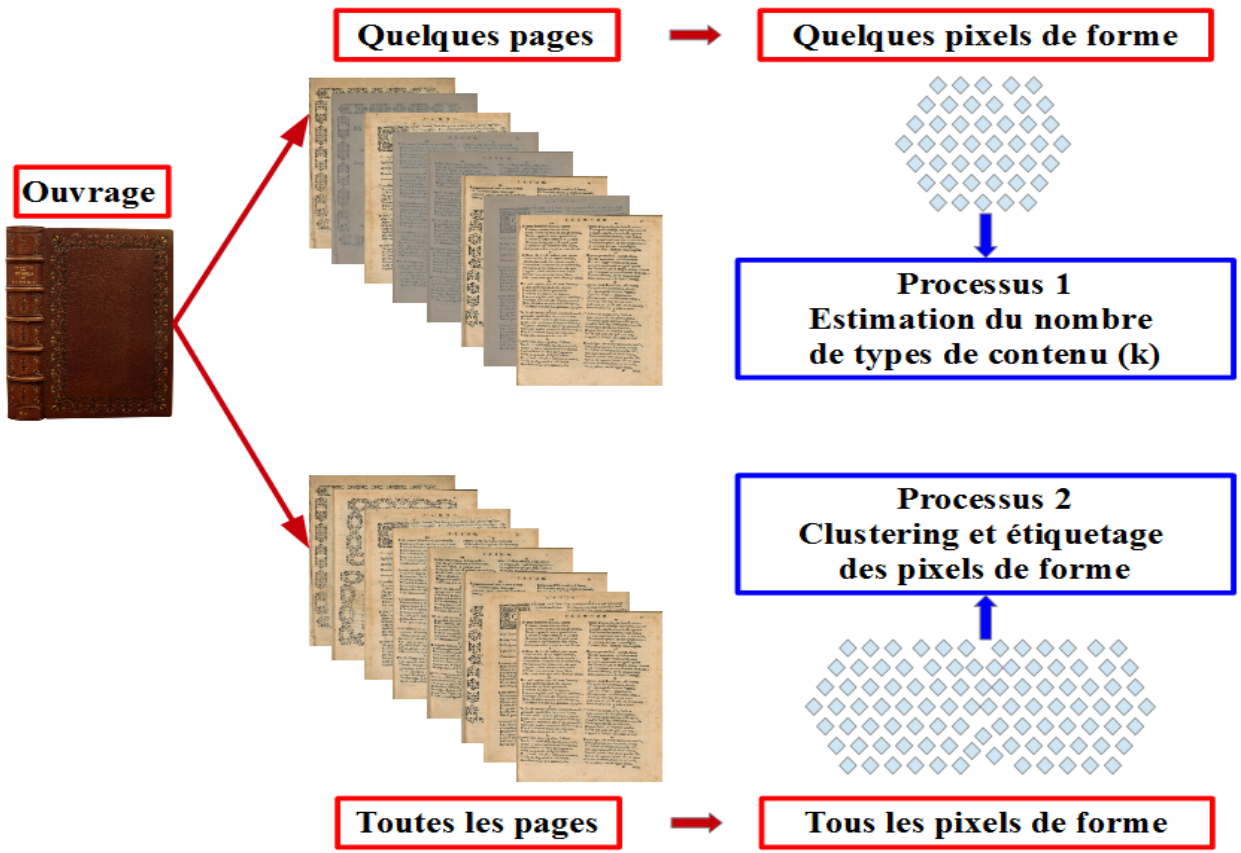
Plateforme d'étiquetage des pixels du contenu

- **Processus 1** : extraction et analyse d'information bas-niveau sur un extrait de l'ouvrage
- **Processus 2** : caractérisation de toutes les pages de l'ouvrage en exploitant les informations du processus 1



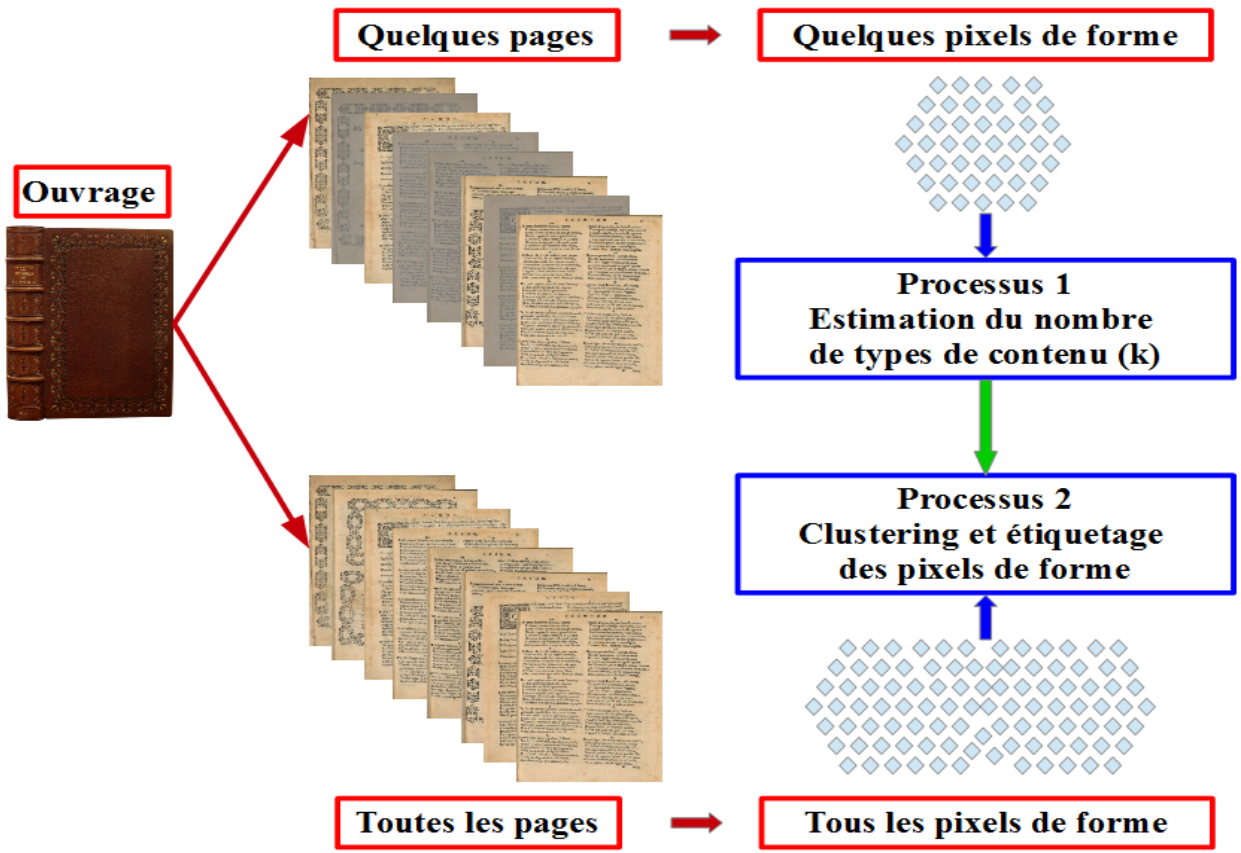
Plateforme d'étiquetage des pixels du contenu

- **Processus 1** : extraction et analyse d'information bas-niveau sur un extrait de l'ouvrage
- **Processus 2** : caractérisation de toutes les pages de l'ouvrage en exploitant les informations du processus 1



Plateforme d'étiquetage des pixels du contenu

- **Processus 1** : extraction et analyse d'information bas-niveau sur un extrait de l'ouvrage
- **Processus 2** : caractérisation de toutes les pages de l'ouvrage en exploitant les informations du processus 1



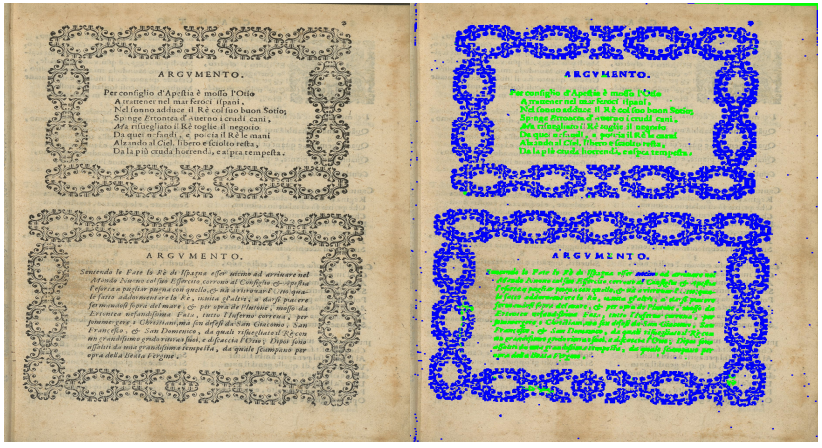
Plateforme d'étiquetage des pixels du contenu [Mehri et al. (2013), Mehri et al. (2015a)]

1. Extraction d'indices de texture
2. Estimation du nombre de types de contenu d'un ouvrage (1^{er} processus)
3. Clustering & étiquetage des pixels de forme (2^{nd} processus)

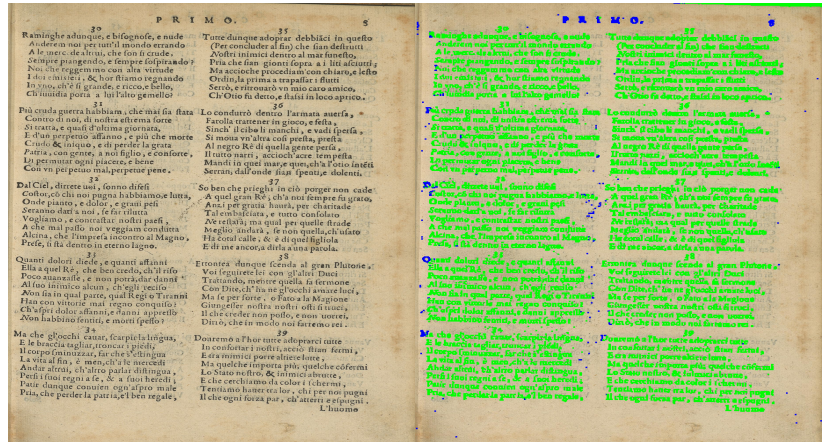
[Mehri et al. (2013)] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "A pixel labeling approach for historical digitized books," in *ICDAR*, 2013
[Mehri et al. (2015a)] M. Mehri, P. Gomez-Krämer, P. Héroux, A. Boucher, and R. Mullot, "A texture-based pixel labeling approach for historical books," *PAA*, 2015

Caractérisation des pages d'un ouvrage au niveau pixel

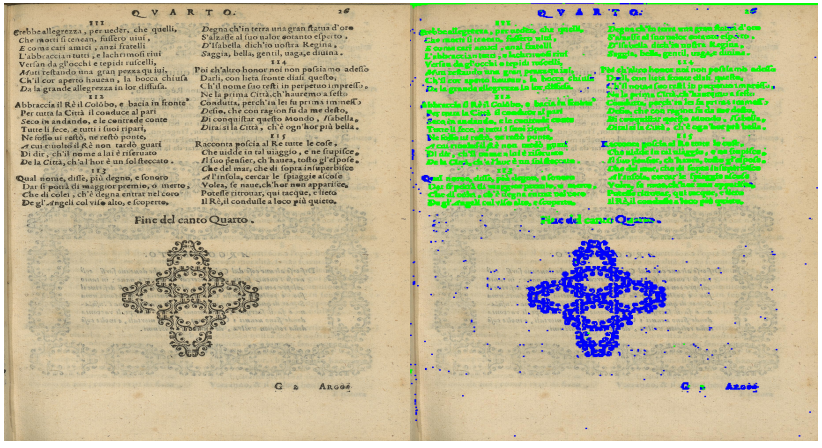
Résultats



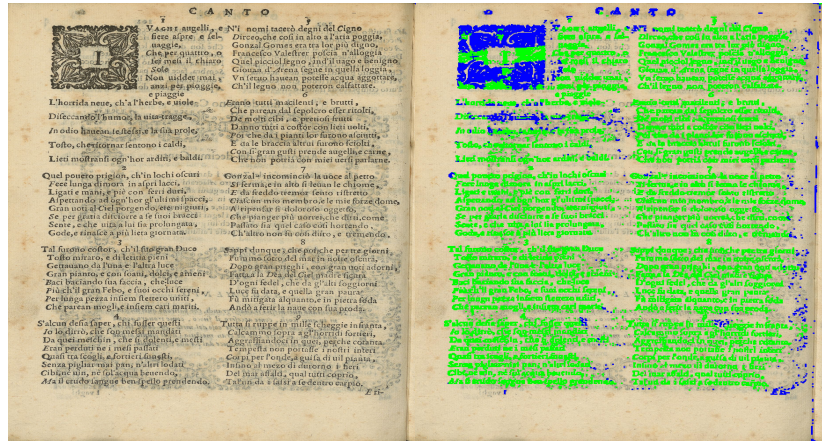
(a)



(b)



(c)



(d)

Objectifs

- **Représentation riche** & **holistique** du contenu & de la mise en page
- Applicable à une **grande variété** d'ouvrages anciens
- **Catégoriser** les pages d'un ouvrage en fonction de plusieurs critères

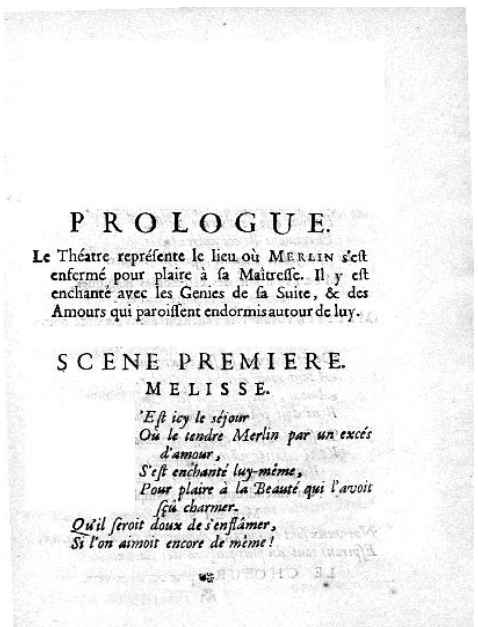
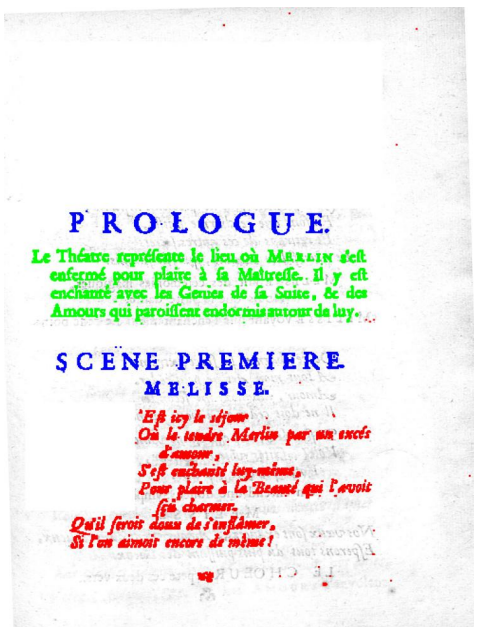
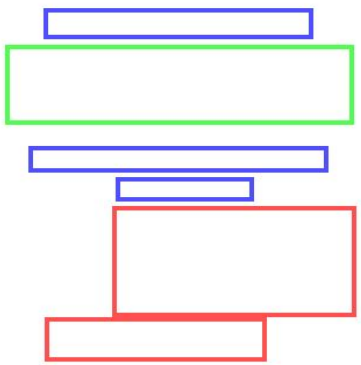


Image originale



Pixels étiquetés



Régions extraites

Caractérisation des pages d'un ouvrage

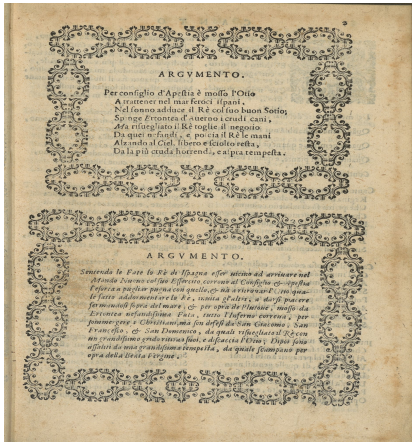
1. Extraction d'indices de texture
2. Estimation du nombre de types de contenu d'un ouvrage (1^{er} processus)
3. Clustering & étiquetage des pixels de forme (2^{nd} processus)
4. Extraction des régions homogènes [Mehri et al. (2015b)]
5. Génération de la signature structurelle par page [Mehri et al. (2015c)]

[Mehri et al. (2015b)] M. Mehri, P. Héroux, N. Sliti, P. Gomez-Krämer, N. E. B. Amara, and R. Mullot, "Extraction of homogeneous regions in historical document images," in *VISAPP*, 2015
[Mehri et al. (2015c)] M. Mehri, P. Héroux, J. Lerouge, P. Gomez-Krämer, and R. Mullot, "A structural signature based on texture for digitized historical book page categorization," in *ICDAR*, 2015

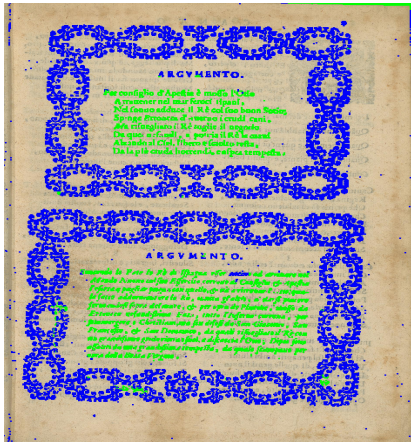
Caractérisation des pages d'un ouvrage

1. Extraction d'indices de texture
2. Estimation du nombre de types de contenu d'un ouvrage (1^{er} processus)
3. Clustering & étiquetage des pixels de forme (2^{nd} processus)
4. **Extraction des régions homogènes [Mehri et al. (2015b)]**
5. Génération de la signature structurelle par page [Mehri et al. (2015c)]

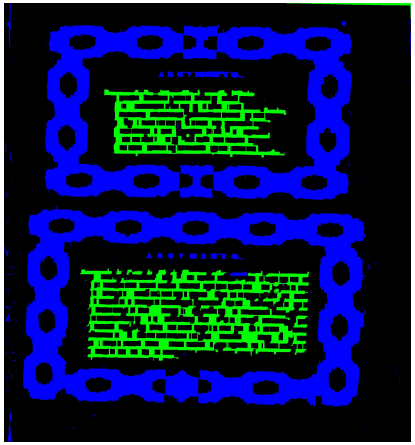
Extraction des régions homogènes



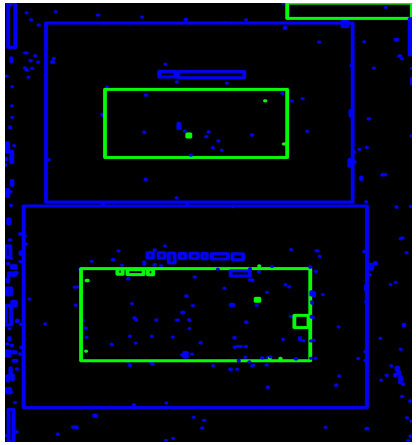
(a) Image originale



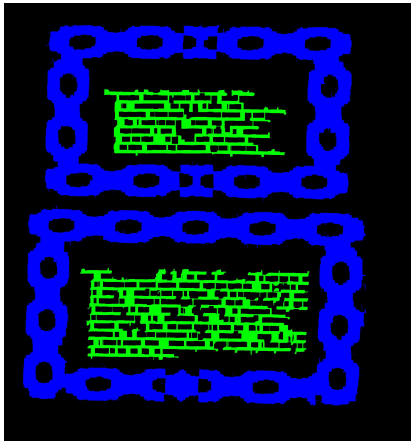
(b) Image étiquetée



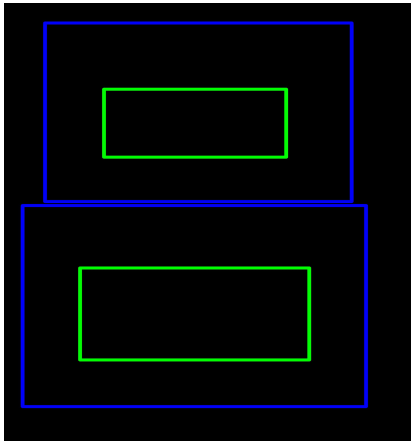
(c) Composantes connexes (CCs) étiquetées



(d) Régions homogènes extraites & étiquetées



(e) CCs sélectionnées



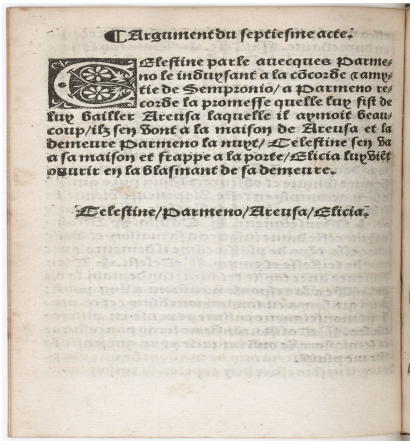
(f) Régions représentatives étiquetées

Caractérisation des pages d'un ouvrage

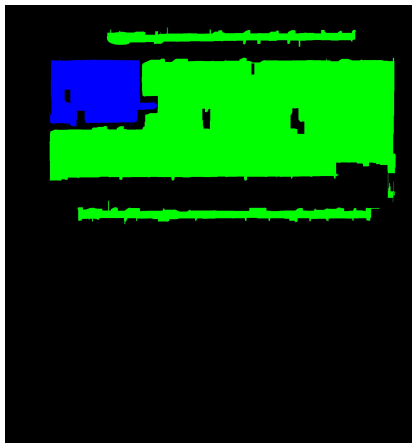
1. Extraction d'indices de texture
2. Estimation du nombre de types de contenu d'un ouvrage (1^{er} processus)
3. Clustering & étiquetage des pixels de forme (2^{nd} processus)
4. Extraction des régions homogènes [Mehri et al. (2015b)]
5. **Génération de la signature structurelle par page [Mehri et al. (2015c)]**

Génération de la signature structurelle par page

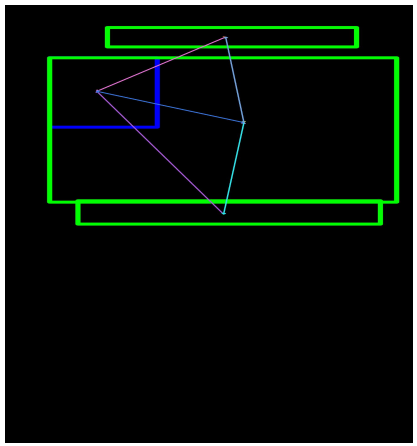
- **Représentation de la signature structurelle à base de graphe**



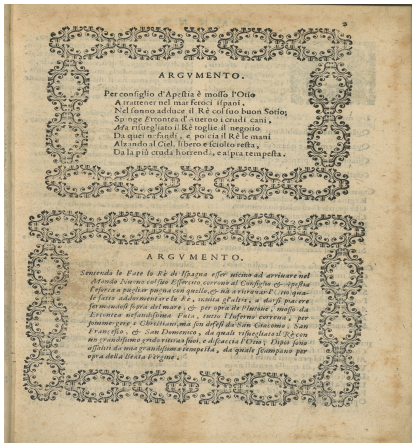
(a)Image originale



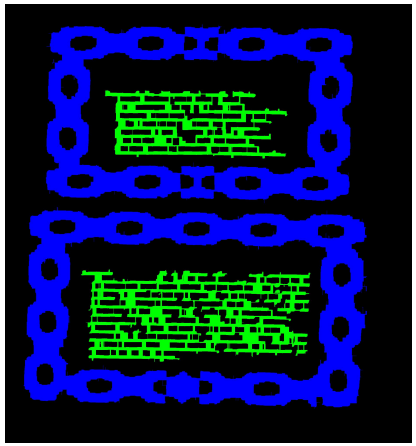
(b)Régions homogènes



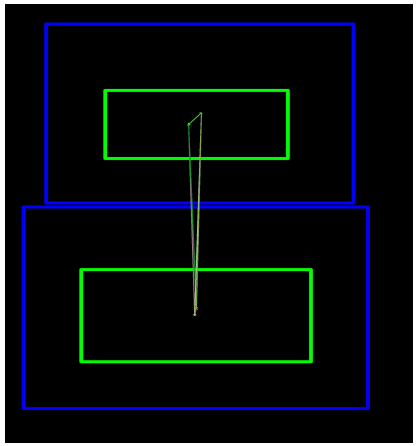
(c)Graphe



(d)Image originale



(e)Régions homogènes



(f)Graphe

Génération de la signature structurelle par page

- **Graphe**
  - **Attributs de nœud**
    - ✓ **192 indices de Gabor**
    - ✓ **46 indices de forme, géométriques & topologiques**
      - Positions spatiales du centroïde
      - Nombre de pixels de forme
      - Moyenne des niveaux de gris
      - Surface & périmètre du contour
      - Moments de Hu, spatiaux, centraux & centraux normalisés
      - *etc.*
  - **Attributs d'arc**
    - ✓ **Différences absolues** entre les 2 centroïdes
    - ✓ **Force d'attraction**

Génération de la signature structurelle par page

- **Force d'attraction**
  - Souligner les régions spatialement les plus **proches** & **représentatives**
  - Loi universelle de la **gravitation**
  - Nombre de pixels du **nœud destination**
  - Distance Euclidienne entre **deux nœuds**

Force d'attraction

$$F_e^{s,d} = \frac{N_{G_v^d}}{(ED_{G_v^{s,d}})^2} \quad (1)$$

Un arc entre deux nœuds est défini $\Leftrightarrow$ $\boldsymbol{F_e^{s,d} \geq Th_e}$.

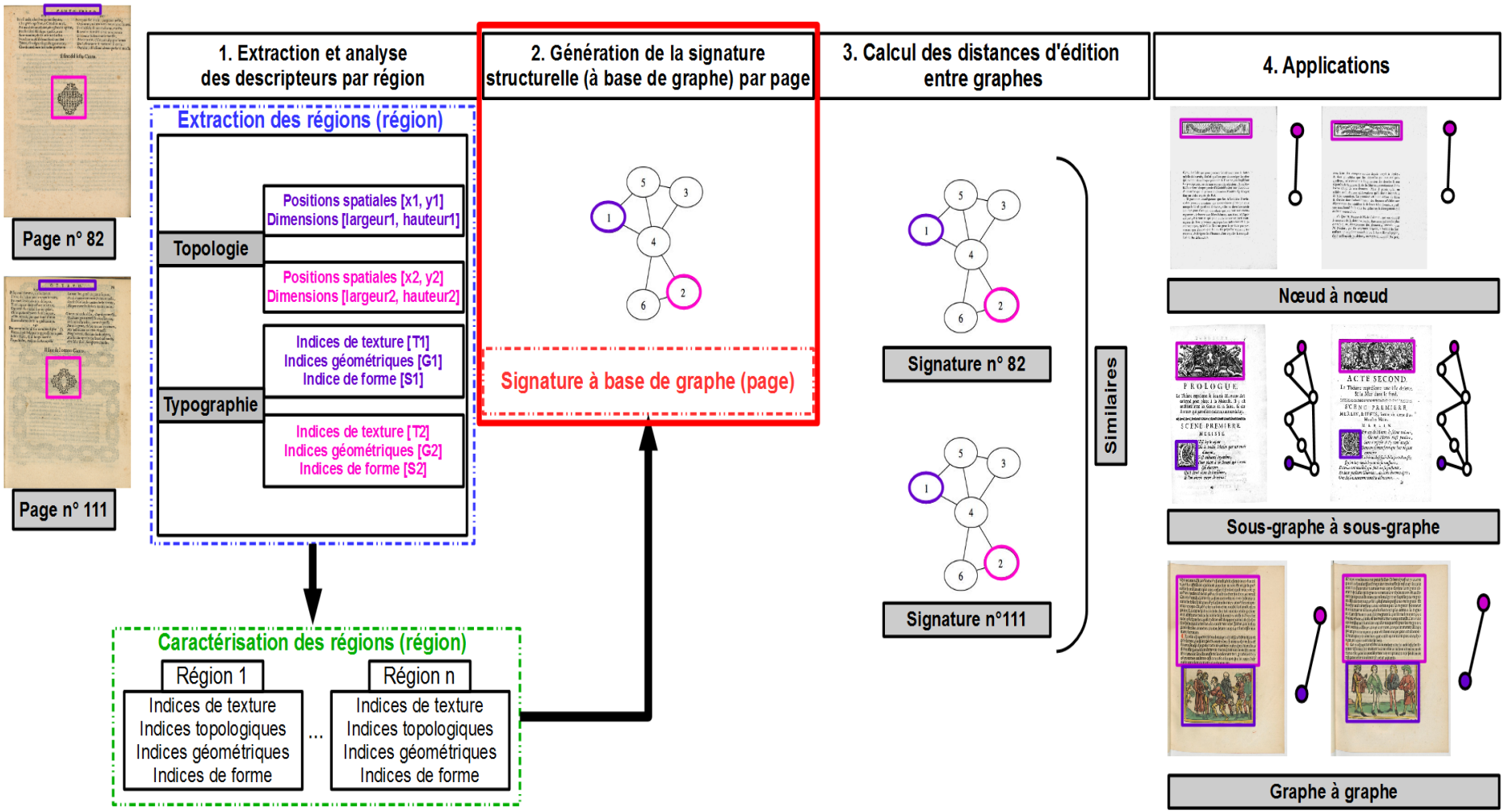
où

$\boldsymbol{ED_{G_v^{s,d}}}$: distance Euclidienne entre deux nœuds : source (G_v^s) & destination (G_v^d)

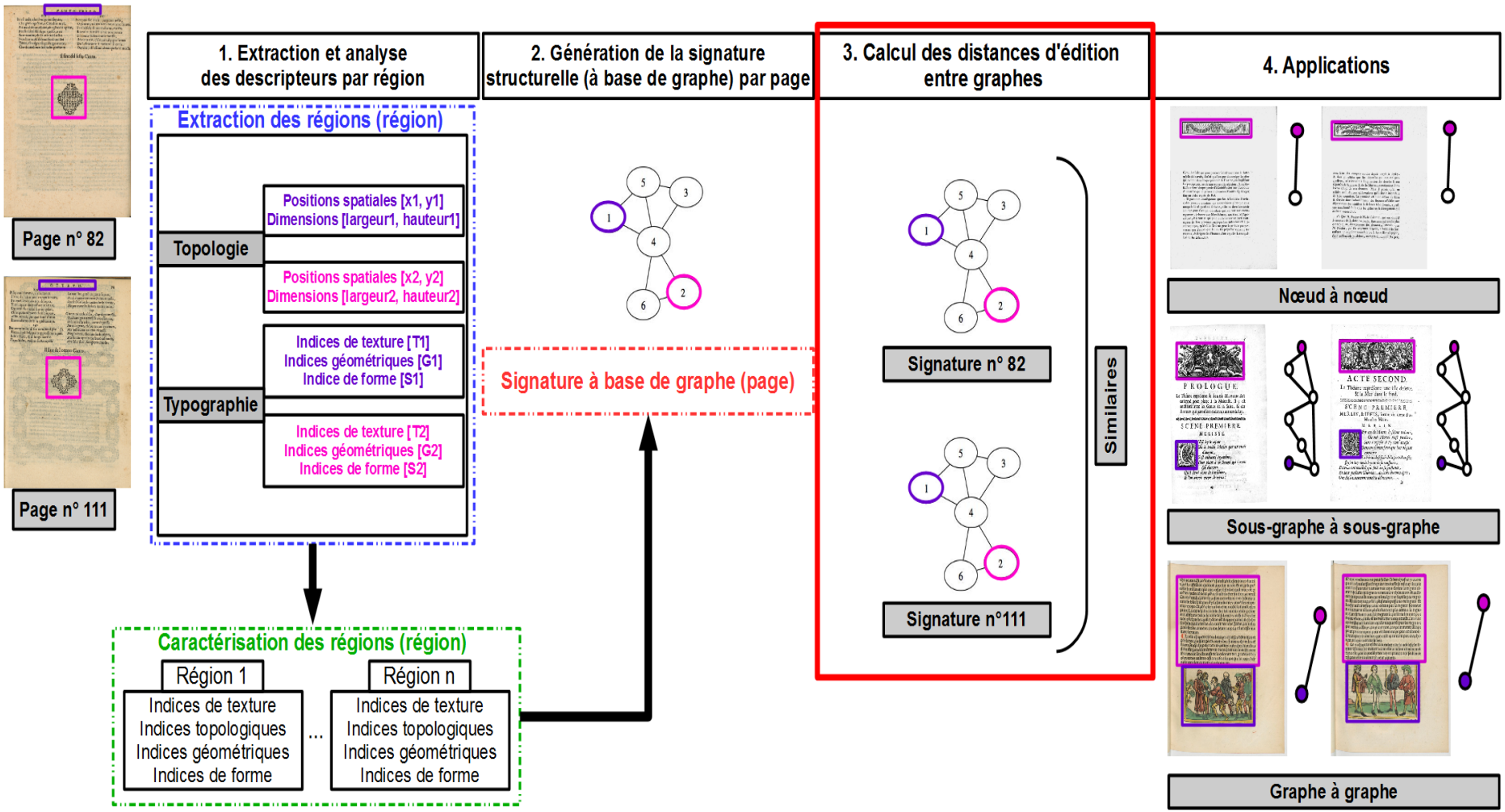
$\boldsymbol{N_{G_v^d}}$: nombre de pixels du nœud destination (G_v^d)

$\boldsymbol{Th_e}$: seuil de la force d'attraction

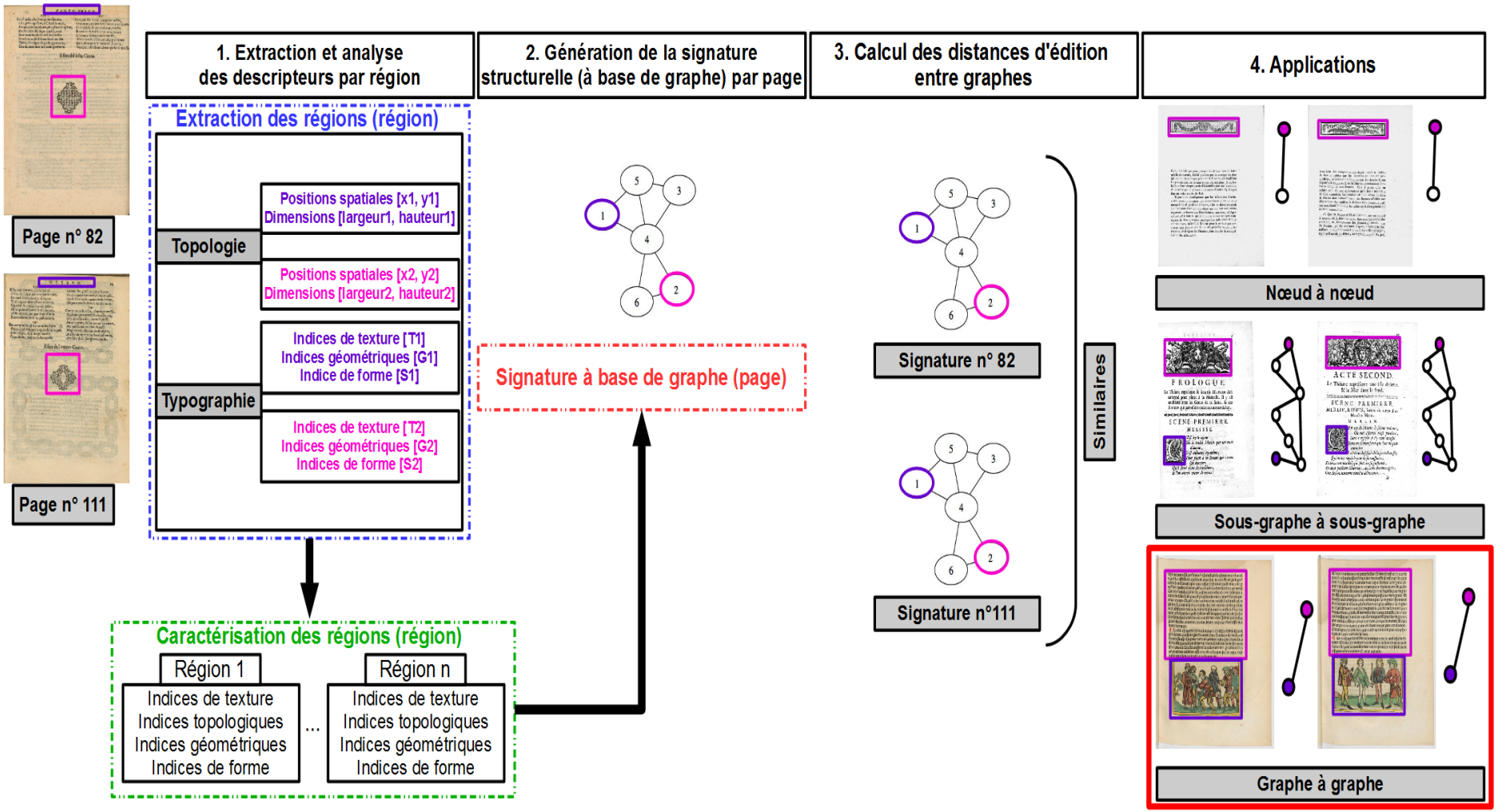
Applications potentielles



Applications potentielles



Applications potentielles



Applications potentielles

- **Distance d'édition entre graphes (GED)**
  - **Littérature**
    - ✓ Mesure de la **(dis)similarité entre les graphes**
    - ✓ Calcul du **coût minimum** de la séquence (e.g. insertion, suppression ou substitution de nœuds/arcs) [Bunke and Riesen (2012)]
  - **Expérimentations**
    - ✓ **Programmation linéaire binaire** [GEM++]
    - ✓ **Analyse statistique de la variation** des attributs
    - ✓ **Complexité algorithmique** de la GED (11 nœuds)

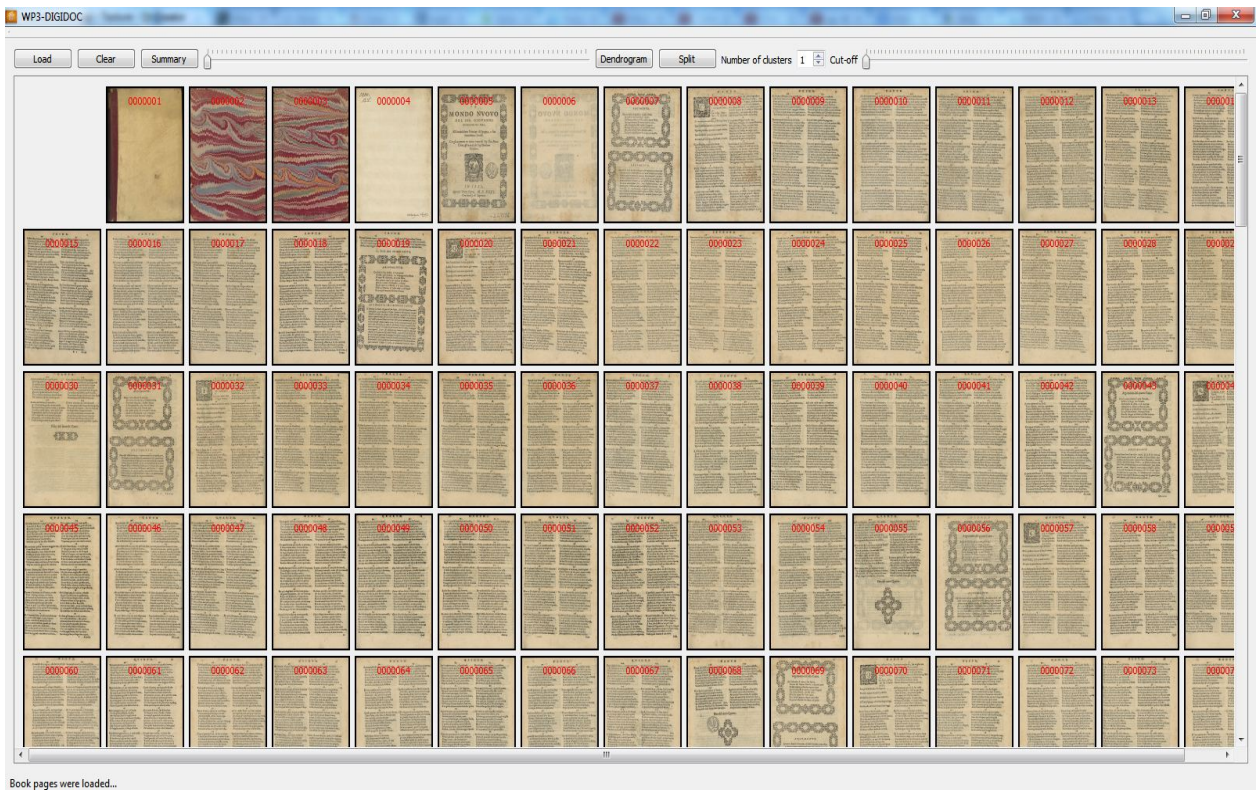
[Bunke and Riesen (2012)] H. Bunke and K. Riesen, "Towards the unification of structural and statistical pattern recognition," *PRL*, 2012
[GEM++] `http://litis-ilpiso.univ-rouen.fr/ILPIso/gem++.html`

Segmentation de flux des pages

- **Résumé d'un ouvrage**

Corpus expérimental

- Monographie imprimée de **Gallica-BnF** [Monographie]
- Intitulée « Il mondo nuovo, del sig. Giov. Giorgini da Jesi »
- Datée de 1596
- Écrite en italien
- Composée de **322 pages** numérisées en couleurs

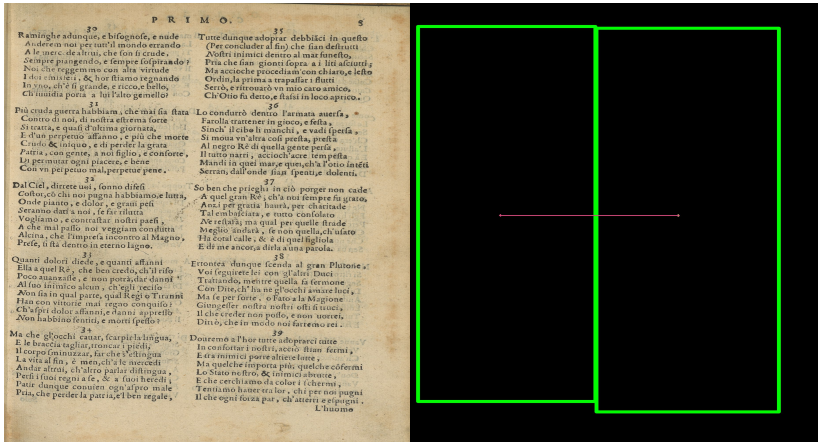


[Monographie] http://gallica.bnf.fr/ark:/12148/bpt6k132294p/f5.planchecontact.r=.langFR

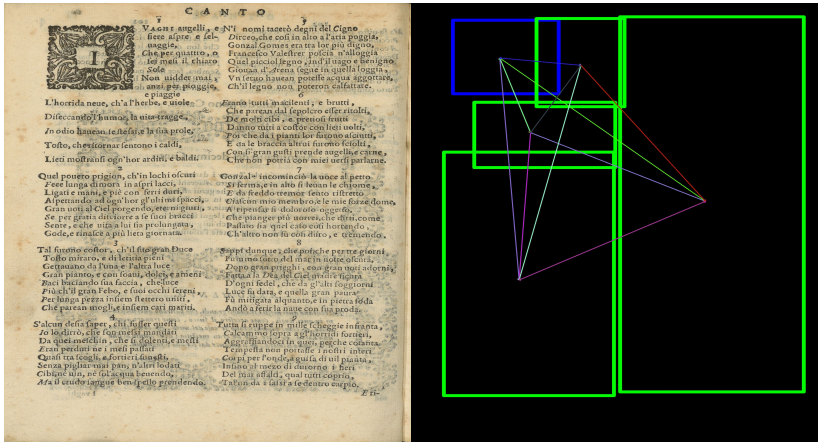
Protocole expérimental

- **Vérité terrain**
  - **102 paires** de pages successives considérées comme des **pages de transition**
- **Méthode d'évaluation**
  - **ROC**

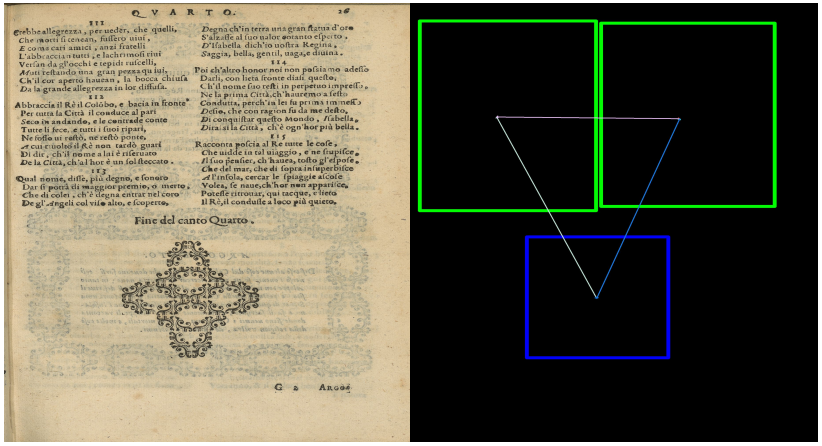
Caractérisation des pages d'un ouvrage



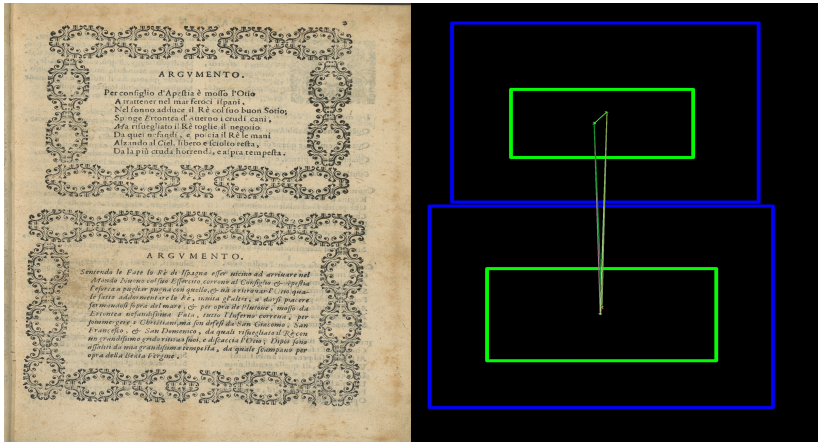
(a)



(b)

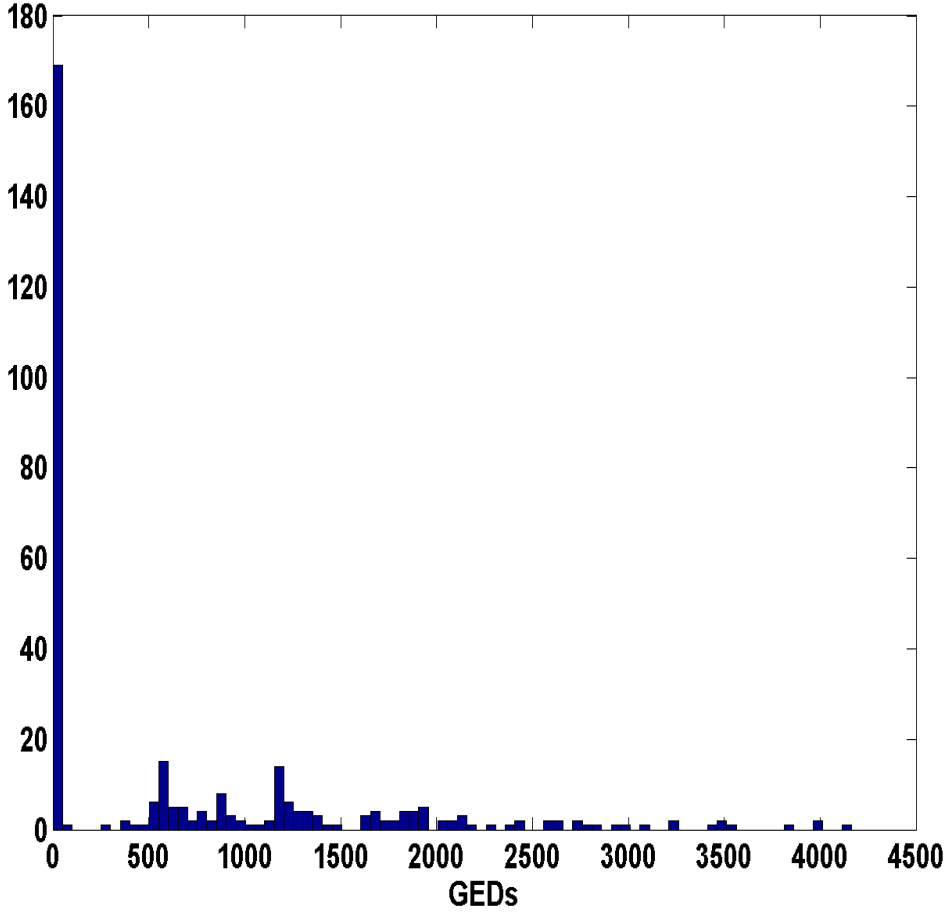


(c)

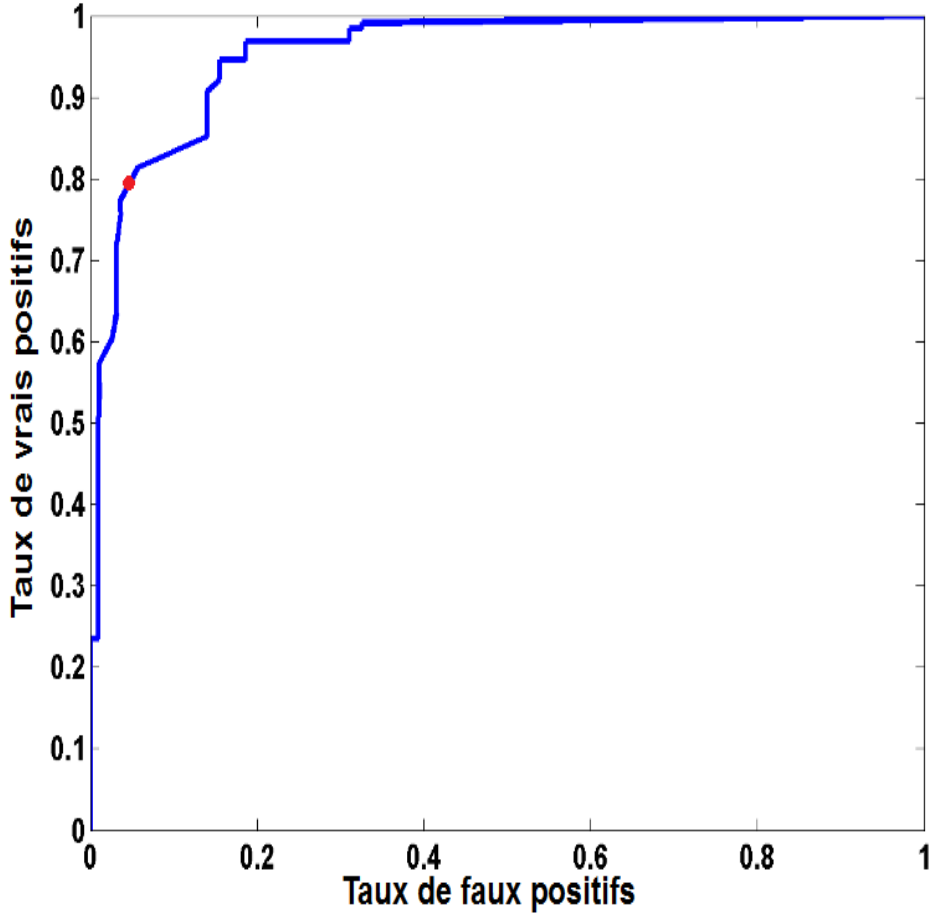


(d)

Segmentation de flux des pages

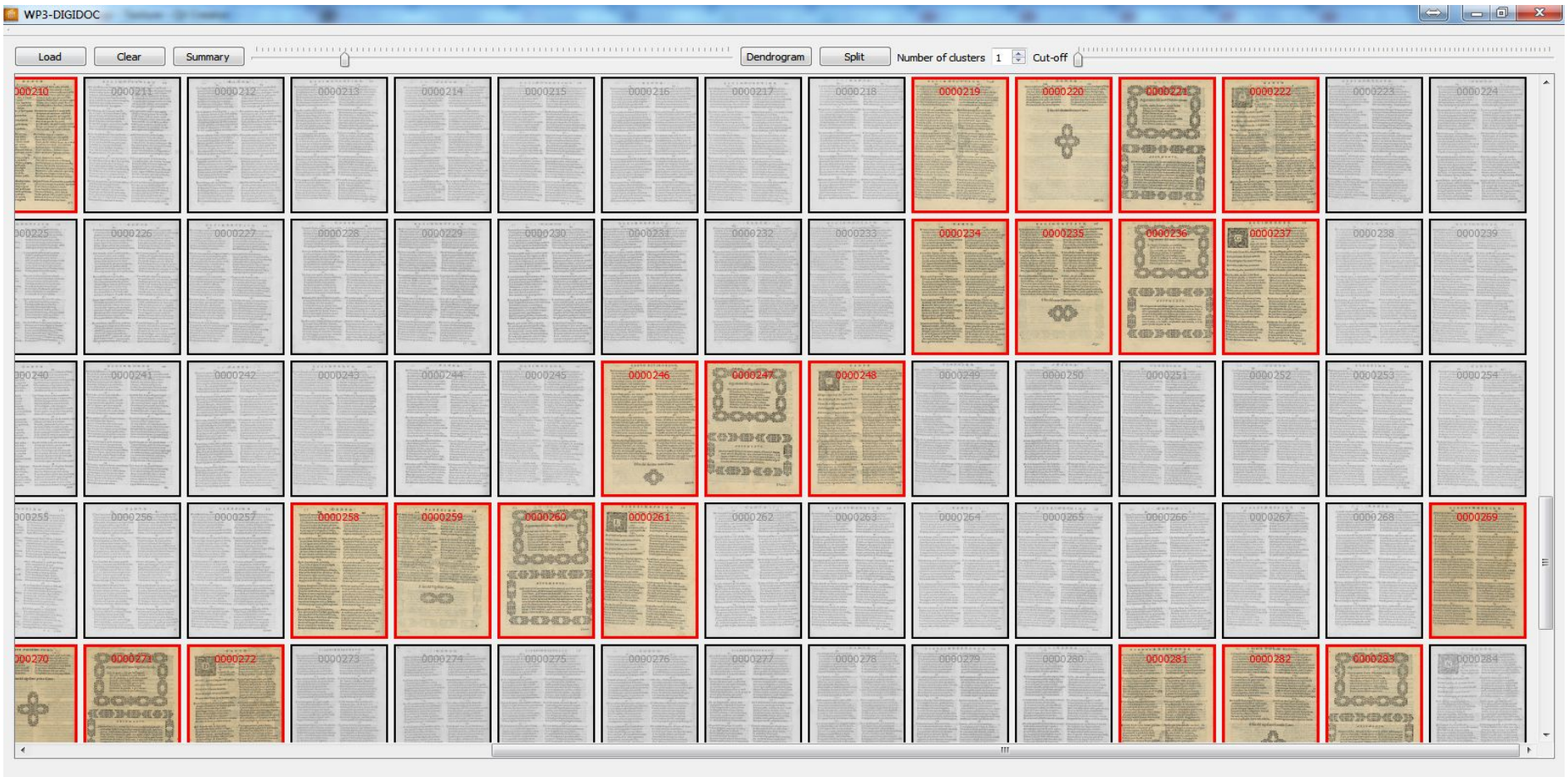


Histogramme



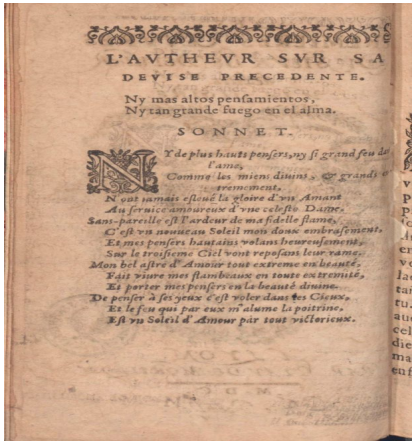
ROC

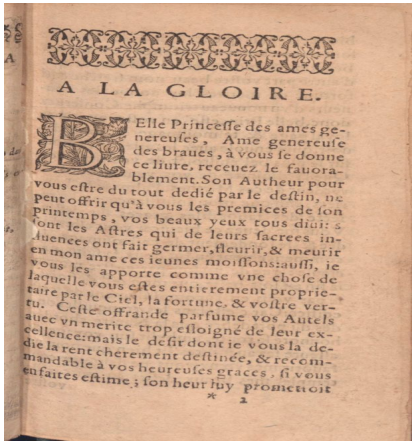
Segmentation de flux des pages

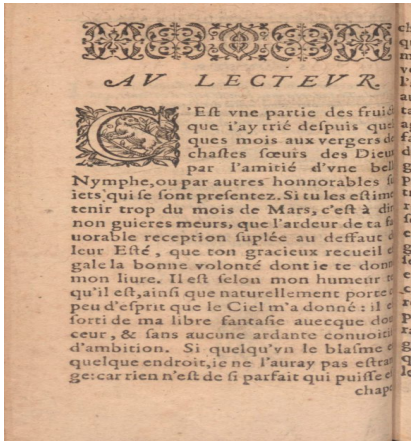


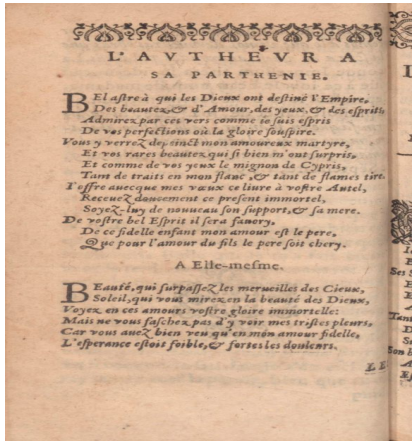
Exemples

- Détection d'une **défaillance** (e.g. pages floues, pliées ou inclinées)

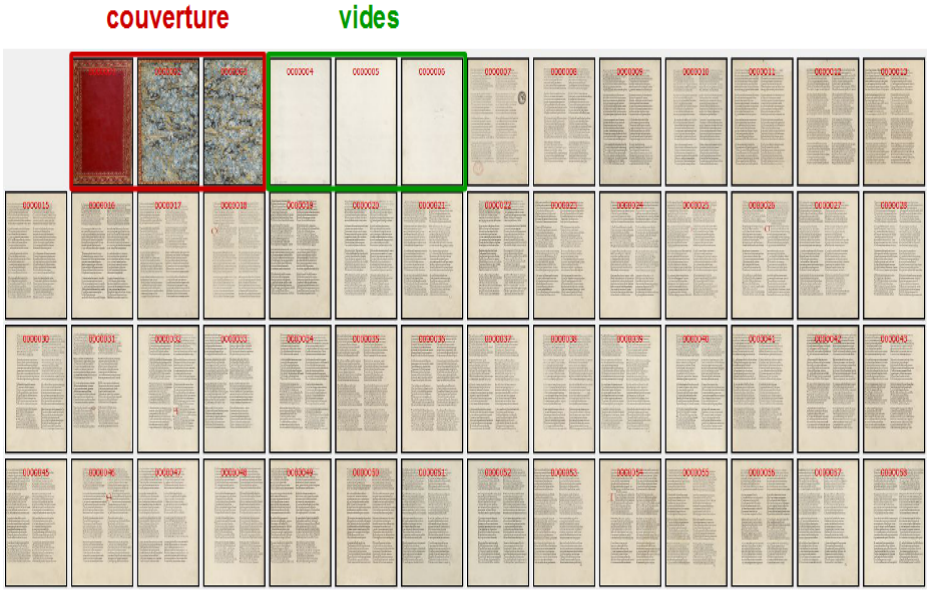

L'AVTHEVR SVR SA
DEVISE PRECEDENTE.
Ny mas altos penſamientos,
Ny tan grande fuego en el alma.
SONNET.


A LA GLOIRE.


AV LECTEVR.


L'AVTHEVR A
SA PARTHENIE.

- Regroupement des pages **de couverture** ou **vides**, *etc.*



Conclusions & perspectives

- **Conclusions**
  - **Sans connaissances *a priori*** sur la structure physique ou logique des documents (**structure** & **contenu**)
  - **Signature structurelle** à base de texture
    - ✓ **Résumé d'un ouvrage**
    - ✓ **Interface graphique utilisateur**
- **Perspectives**
  - **Nombreuses applications potentielles exploitant la signature**
    - ✓ **Pages similaires** à une **page requête**
    - ✓ Pages dont la structure contient un **groupe d'éléments particuliers**
    - ✓ *etc.*
  - **Exploitation des méta-données caractérisant de l'ouvrage**

*** Merci de votre attention ***

Questions ?

Annexe

Références (1)

G. Cron, "Éléments ayant une influence sur la qualité et l'efficacité des OCRs," BnF, Tech. Rep., 2012.

A. BenSalah, N. Ragot, and T. Paquet, "Adaptive detection of missed text areas in OCR outputs : application to the automatic assessment of OCR quality in mass digitization projects," in *DRR*, 2013.

C. Grana, G. Serra, M. Manfredi, D. Coppi, and R. Cucchiara, "Layout analysis and content enrichment of digitized books," *MTA*, 2014.

M. Coustaty, R. Raveaux, and J. M. Ogier, "Historical document analysis : a review of French projects and open issues," in *EUSIPCO*, 2011.

http://gallica.bnf.fr.

M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "A pixel labeling approach for historical digitized books," in *ICDAR*, 2013.

M. Mehri, P. Gomez-Krämer, P. Héroux, A. Boucher, and R. Mullot, "A texture-based pixel labeling approach for historical books," *PAA*, 2015.

M. Mehri, P. Héroux, N. Sliti, P. Gomez-Krämer, N. E. B. Amara, and R. Mullot, "Extraction of homogeneous regions in historical document images," in *VISAPP*, 2015.

M. Mehri, P. Héroux, J. Lerouge, P. Gomez-Krämer, and R. Mullot, "A structural signature based on texture for digitized historical book page categorization," in *ICDAR*, 2015.

H. Bunke and K. Riesen, "Towards the unification of structural and statistical pattern recognition," *PRL*, 2012.

http://litis-ilpiso.univ-rouen.fr/ILPIso/gem++.html.

http://gallica.bnf.fr/ark:/12148/bpt6k132294p/f5.planchecontact.r=.langFR.