Numérisation des documents patrimoniaux

1. Les structures

2. Applications: Recherche d'information, Navigation…

Après la numérisation, les éventuelles opérations de restauration: que faire des images?

Objectifs :

Rechercher et exploiter des *méta-données* et des spécificités propres à :

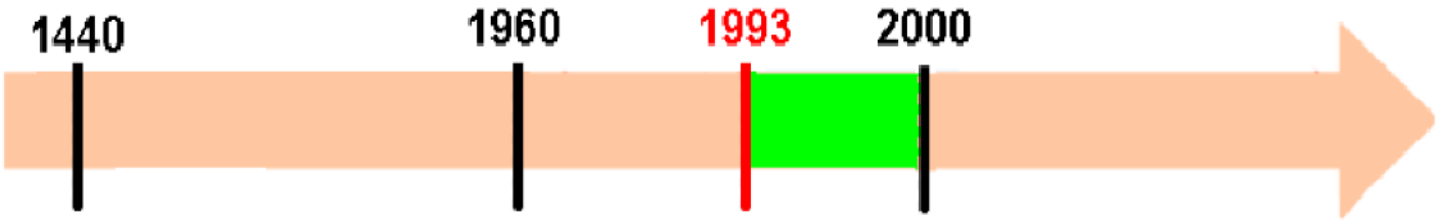
- ***La période de l'histoire***
- ***Le lieu où le document original a été produit***
- ***Les intentions d'usage, les besoins***

Proposer des solutions innovantes et adaptées au besoin de la reconnaissance de l'écrit et de l'analyse des contenus (*débruitage, restauration, caractérisation des écritures, indexation de corpus, recherche d'information fine…*)

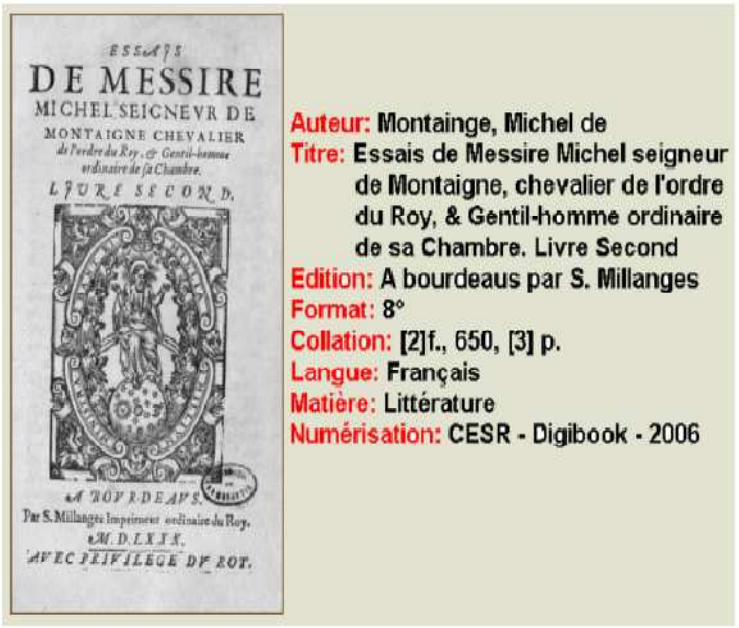
Mais quelles solutions…?

Quelles solutions?

Exploiter les documents en mode texte…



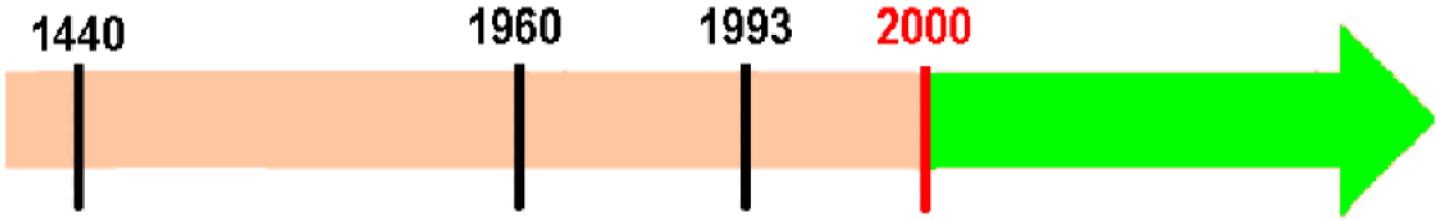
- Normes permettant de saisir des métadonnées
  - Text Encoding Initiative, Open Archive Initivative
  - IconClass
- Indispensable mais pas suffisant
  - Subjectivité
  - Ampleur de la tâche (voire impossible)



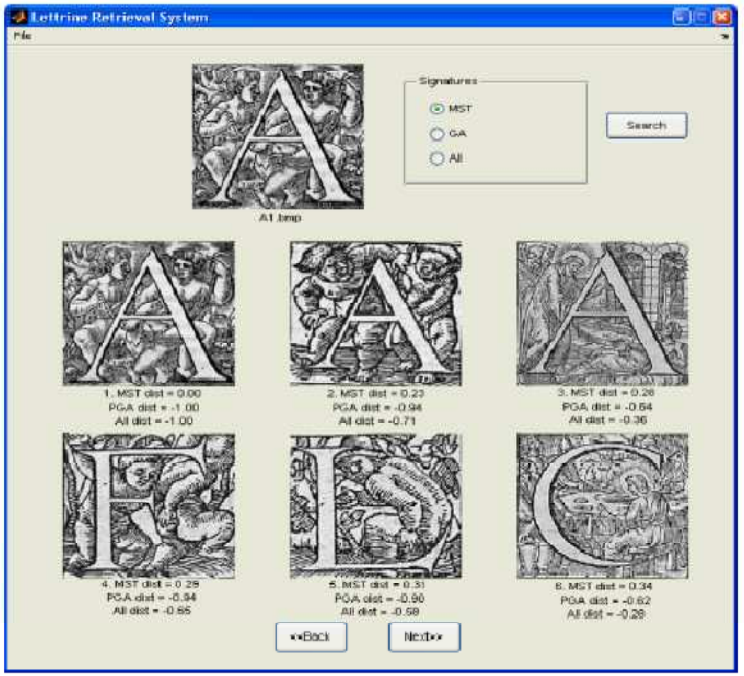
Bibliothèque numérique du CESR

Quelles solutions?

Exploiter le document en mode image…



- Reconnaissance de caractères dédiée aux documents anciens (Projet DEBORA [?])
- Indexation de la structure (Plate-Forme AGORA [?])
- Indexation d'illustrations de documents anciens pour une comparaison de dessins de traits [?]



Des outils et des techniques pour aider la reconnaissance et l'analyse

Objectif : Analyser,Reconnaître et Indexer les documents du patrimoine en adaptant les outils existants et en en créant de nouveaux

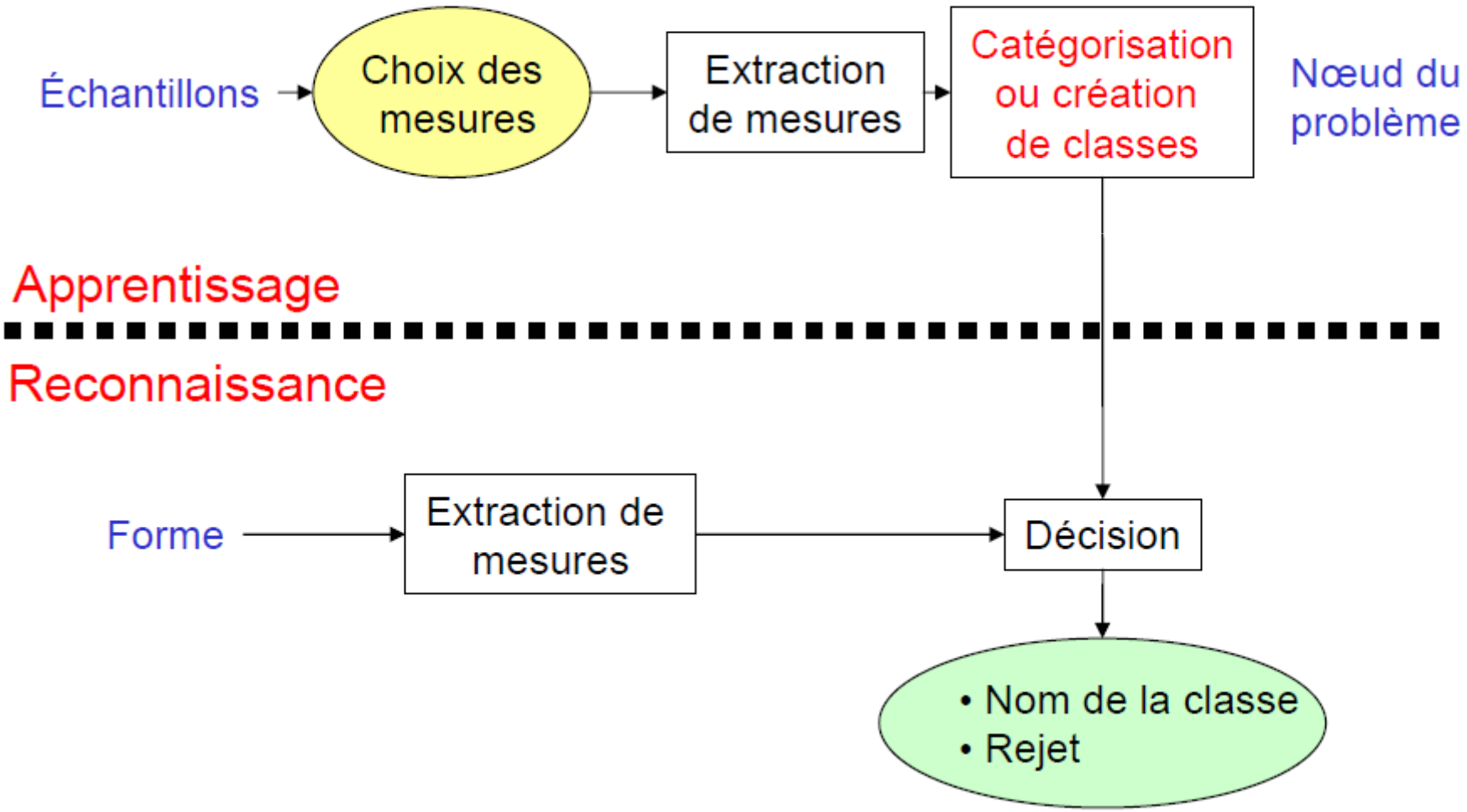
Quelques domaines d'applications existants:

- ▶Lecture de textes imprimés (caractères industriels, textes dans les vidéos, documents imprimés, pièces de monnaie…)
- ▶Reconnaissance de textes manuscrits on/off-line
- ▶Reconnaissance des structures des documents
- ▶Analyse de documents divers (Plans du cadastre, plans mécaniques..)
- ▶Analyse de documents anciens (manuscrits, imprimés)

Applications diverses - méthodes communes - difficultés différentes

Analyse automatique des images

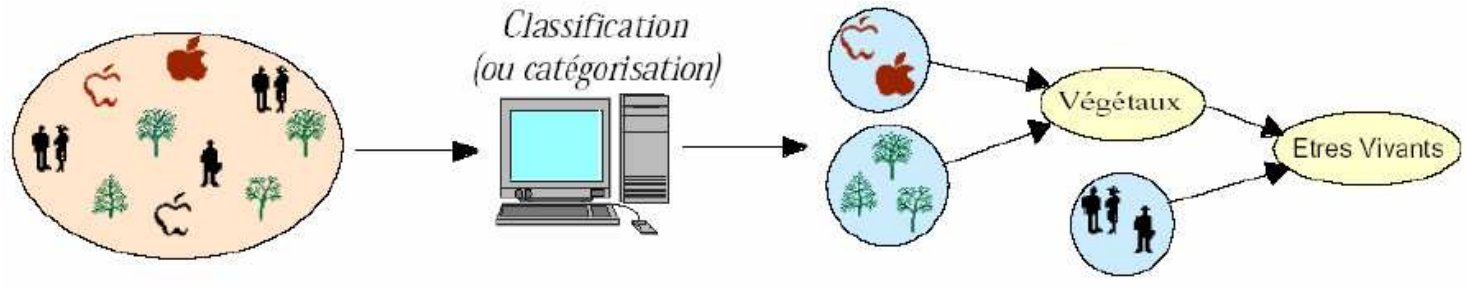
Schéma de reconnaissance



Analyse automatique des images

La catégorisation (classification au sens large)

- Définition
  - Opération ayant pour objectif d'organiser un ensemble d'observations en groupes **homogènes** et **contrastés**



Analyse automatique des images

La catégorisation (classification au sens large)

Il s'agit de ...

1. Résumer de manière pertinente un ensemble d'informations
2. Mettre en évidence des liens structurels entre les données

☛ Utilisation dans une démarche d'analyse exploratoire des données

- Mise en évidence de nouveaux concepts
- Inférence de valeurs inconnues à partir de la définition des classes

■ La classification peut être

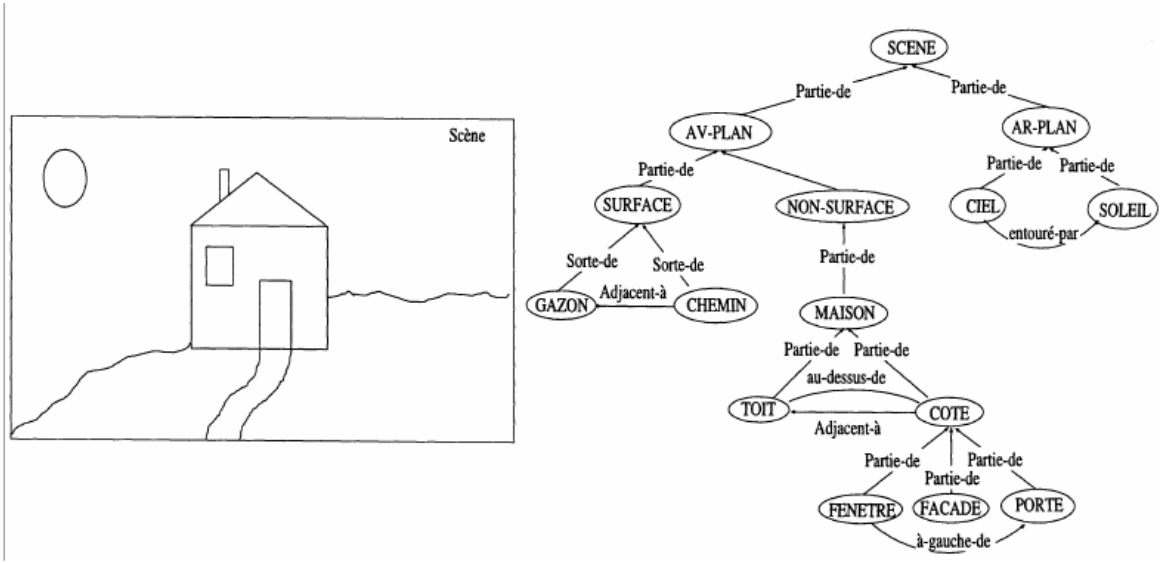
– Supervisée (anglais : classification)
  - Les catégories (ou classes) sont connues a priori
  - Elles ont en général un sens pour l'utilisateur

– Non supervisée (anglais : clustering)
  - Les classes sont fondées sur la structure propre de l'ensemble d'objets (e.g. proximités)
  - Leur signification est plus sujette à caution

Analyse automatique des images

La catégorisation (reconnaissance) structurelle

- Approche numérique
  - Inefficace sur des formes complexes où la décision :
    - Description d'une situation
    - Disposition mutuelle des objets
    - Taille relative
- Approche structurelle
  - Nécessité d'analyse en termes
    - Formes constituantes
    - Relations



Analyse automatique des images

La catégorisation (reconnaissance) structurelle

- Fonctionnement
  - Recherche et exploitation de règles de construction de formes à partir de leurs composants
  - Analogie avec l'analyse grammaticale
- Tendance
  - Hiérarchiser la structure de formes
  - Combiner récursivement l'assemblage de formes
- Point de vue
  - Chomsky pour analyser la structure du langage
- Puissance de l'approche
  - Choix des primitives
  - Choix des relations
- Analogie avec les langages
  - Choix des symboles $\in$ alphabet
  - Choix de règles de production de phrases

Analyse automatique des images : règles d'or

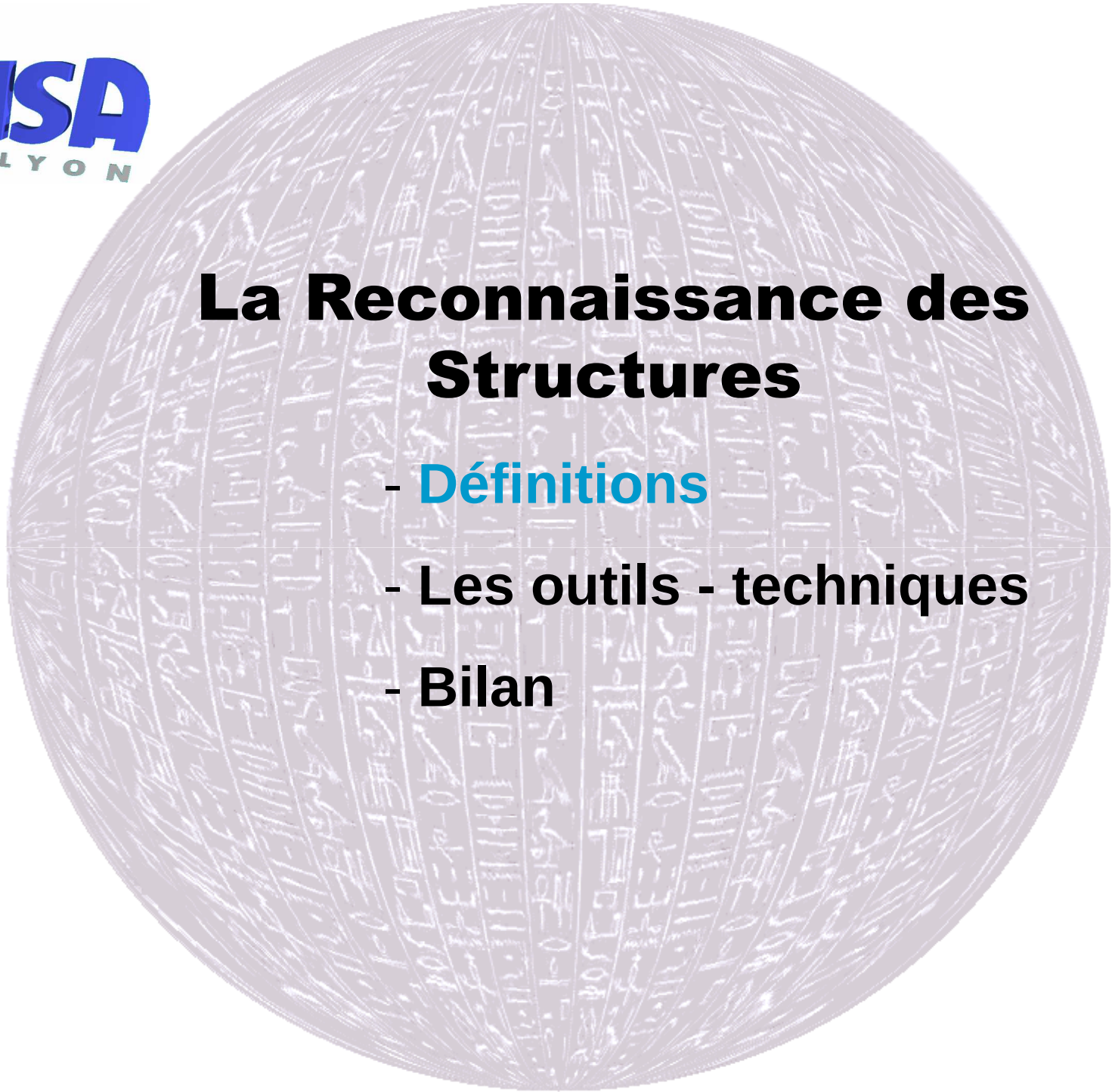
Les experts doivent fournir des critères sûrs et non ambigus des objets qu'ils veulent reconnaître

Cas 1: Si les objets à reconnaître ont des formes plus diverses que le nombre de formes que l'on peut modéliser alors le traitement automatique des images est impossible

Cas 2: Si l'on dispose de très peu d'observations pour l'apprentissage de la forme d'un objet alors il sera impossible de le reconnaître automatiquement

Exemple de traitement impossible : Trouver le mot « incipit » dans des manuscrits anciens quelque soit

- la césure,
- son alignement,
- son style d'écriture,
- sa taille, sa couleur et sa syntaxe…



La Reconnaissance des Structures

- Définitions
- Les outils - techniques
- Bilan

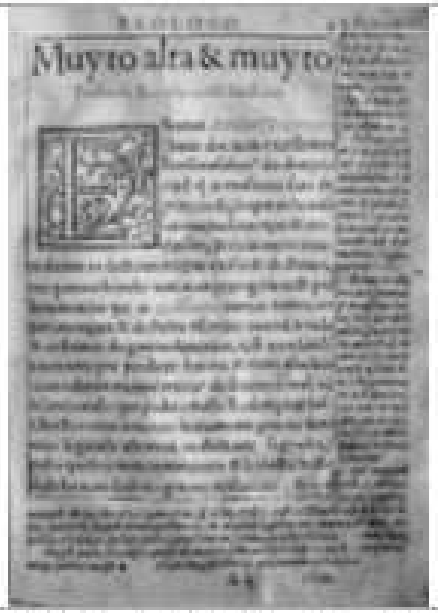
Structures: une grande diversité

1) Elaborer un modèle générique variable selon l'époque et la nature du contenu: une tâche difficile

2) Localiser les blocs d'information de l'image et reconnaître les contenus

3) Mise en place d'outils dédiés selon la nature des contenus

Manuscrit médiéval (source IRHT) XIII



*Livre Imprimé de la renaissance XVI*ème

*Imprimé du XVIII*ème

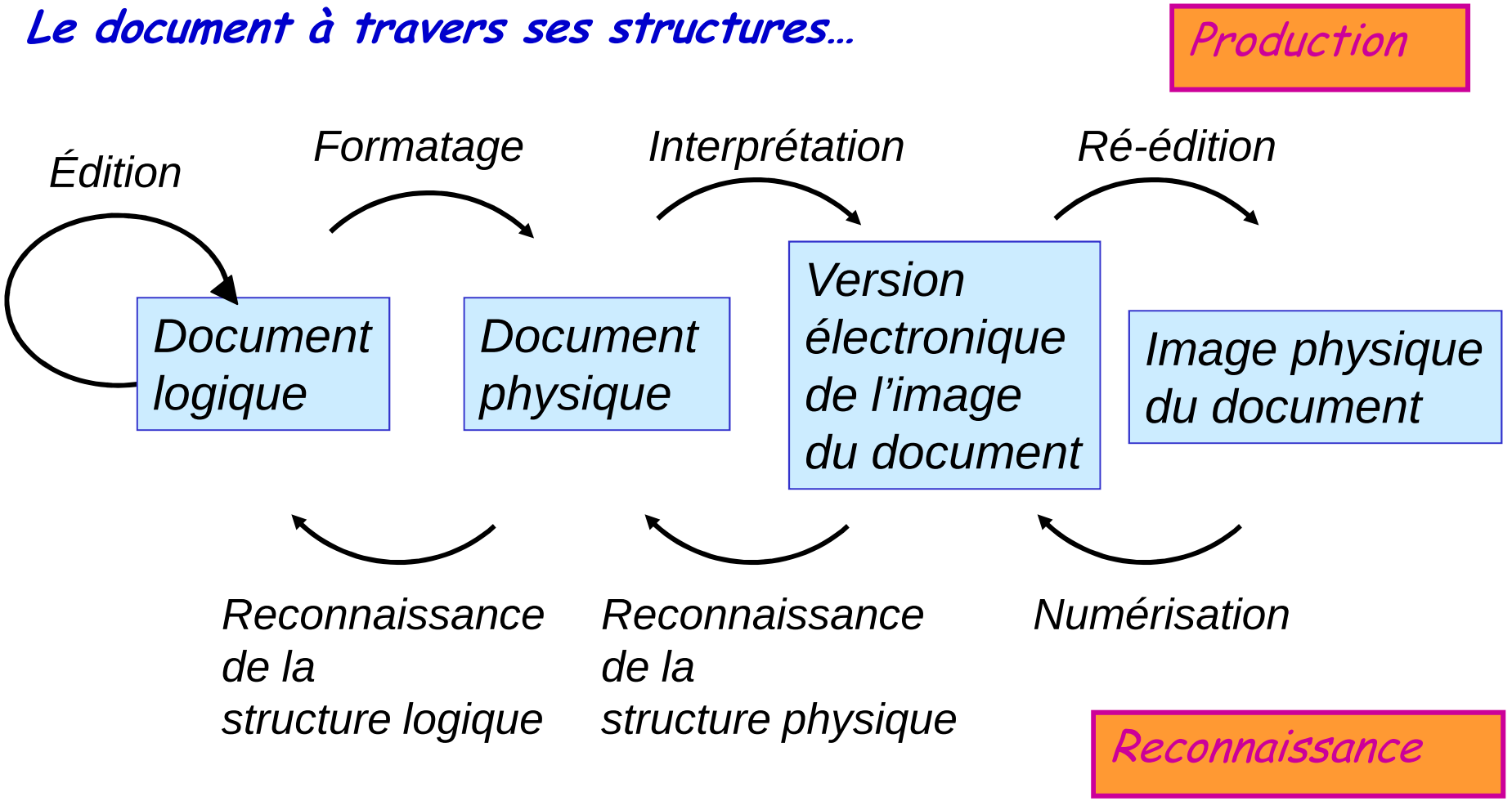
*Journal du XX*ème

Structures: une grande diversité

- Les classes de documents
  - Documents composites
    - Texte / non texte, structures du texte
  - Documents administratifs
    - Structures à motifs et à mots clés
  - Documents anciens
    - Lettrines, lignes et mots
  - Documents postaux
    - Repérage de mots clés

Les différentes structures des documents réguliers

Le document à travers ses structures…



Les différentes structures des documents réguliers

- Structure physique
  - Très régulière
  - Composée de blocs rectangulaires
    - Les lignes sont horizontales
  - Arrangée suivant une certaine hiérarchie
    - Page >> colonne >> région >> bloc >> ligne >> caractère
- Structure logique
  - Très régulière et hiérarchique également
    - Article >> section >> paragraphe >> phrase >> mot
  - La correspondance avec la structure logique est très étroite

Les différentes structures des documents réguliers

A quoi sert la structure physique?

- *Analyser, reconnaître et archiver les documents imprimés de toute nature (périodiques, livres, revues techniques, lettres commerciales,listing...)*

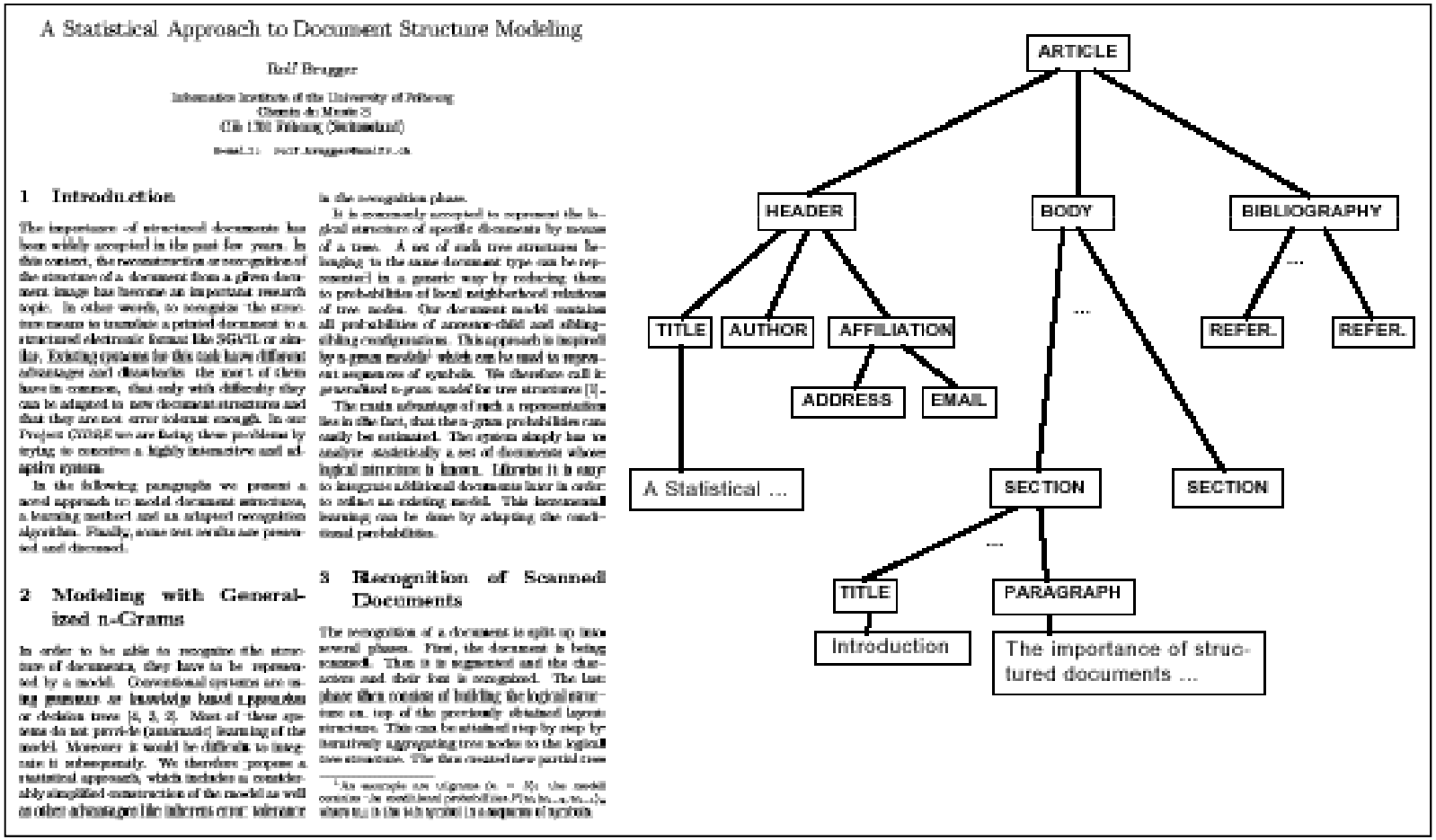
- *Cela dépend du type de document concerné... à la base de nombreuses applications.*

On peut citer en particulier les besoins en:

Stockage
Recherche (indexation)
Manipulation
Re-édition
Génération de vues différentes (reformatage)

Les différentes structures des documents réguliers

Lien entre structure physique et logique?



A Statistical Approach to Document Structure Modeling

Ralf Brugger

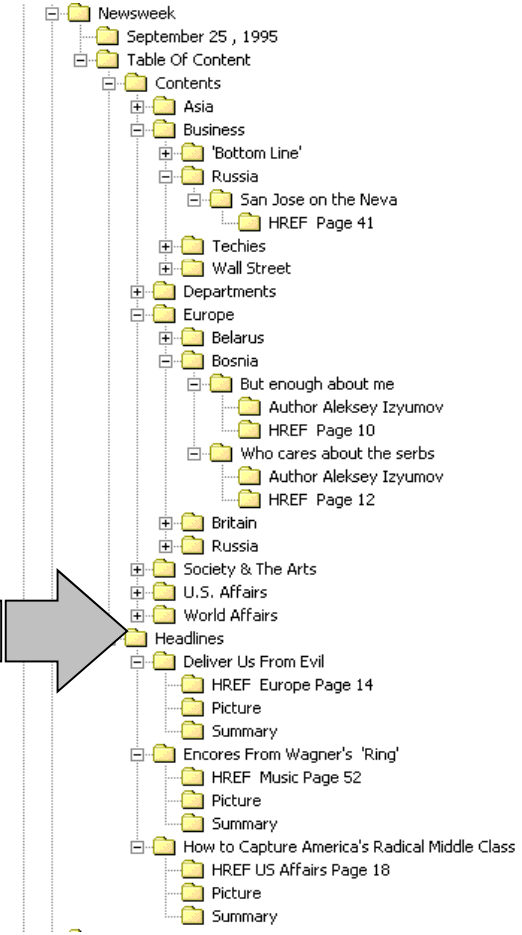
1 Introduction

2 Modeling with Generalized n-Grams

3 Recognition of Scanned Documents

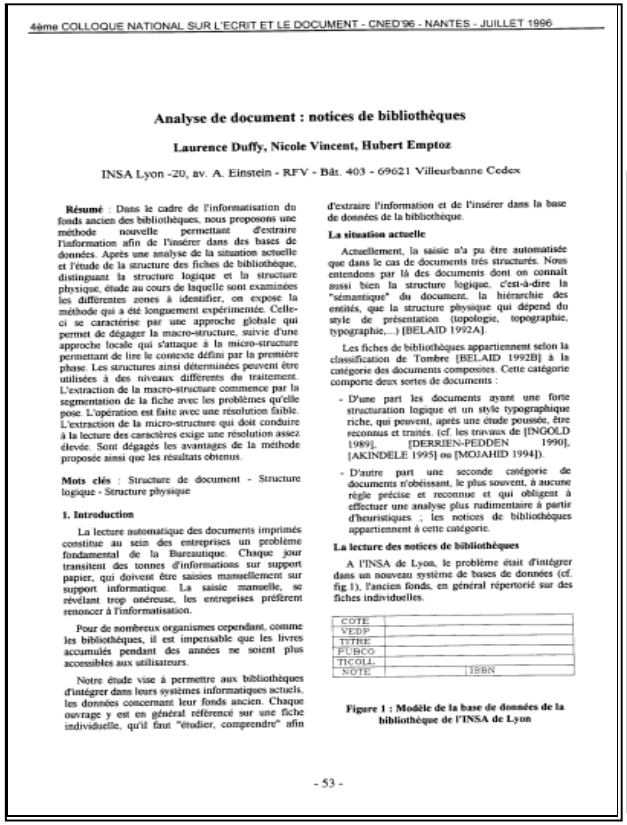
Physique => Logique=> Encodage

Trouver la fonction de chaque bloc de texte ▶ **XML**



Structure PHYSIQUE d'un document (1)

la structure physique décrit l'organisation du document, en termes d'objets graphiques *(caractères, mots, lignes, blocs, paragraphes, colonnes, images, typographie)* et des relations entre ces objets *(décomposition hiérarchique, positions absolues et relatives dans la page)*.



4ème COLLOQUE NATIONAL SUR L'ECRIT ET LE DOCUMENT - CNED'96 - NANTES - JUILLET 1996

Analyse de document : notices de bibliothèques

Laurence Duffy, Nicole Vincent, Hubert Emptoz

INSA Lyon -20, av. A. Einstein - RFV - Bât. 403 - 69621 Villeurbanne Cedex

Résumé : Dans le cadre de l'informatisation du fonds ancien des bibliothèques, nous proposons une méthode nouvelle permettant d'extraire l'information afin de l'insérer dans des bases de données. Après une analyse de la situation actuelle et l'étude de la structure des fiches de bibliothèque, distinguant la structure logique et la structure physique, étude au cours de laquelle sont examinées les différentes zones à identifier, on expose la méthode qui a été longuement expérimentée. Celle-ci se caractérise par une approche globale qui permet de dégager la macro-structure, suivie d'une approche locale qui s'attaque à la micro-structure permettant de lire le contexte défini par la première phase. Les structures ainsi déterminées peuvent être utilisées à des niveaux différents du traitement. L'extraction de la macro-structure commence par la segmentation de la fiche avec les problèmes qu'elle pose. L'opération est faite avec une résolution faible. L'extraction de la micro-structure qui doit conduire à la lecture des caractères exige une résolution assez élevée. Sont dégagés les avantages de la méthode proposée ainsi que les résultats obtenus.

Mots clés : Structure de document - Structure logique - Structure physique

1. Introduction

La lecture automatique des documents imprimés constitue au sein des entreprises un problème fondamental de la Bureautique. Chaque jour transitent des tonnes d'informations sur support papier, qui doivent être saisies manuellement sur support informatique. La saisie manuelle, se révélant trop onéreuse, les entreprises préfèrent renoncer à l'informatisation.

Pour de nombreux organismes cependant, comme les bibliothèques, il est impensable que les livres accumulés pendant des années ne soient plus accessibles aux utilisateurs.

Notre étude vise à permettre aux bibliothèques d'intégrer dans leurs systèmes informatiques actuels, les données concernant leur fonds ancien. Chaque ouvrage y est en général référencé sur une fiche individuelle, qu'il faut "étudier, comprendre" afin d'extraire l'information et de l'insérer dans la base de données de la bibliothèque.

La situation actuelle

Actuellement, la saisie n'a pu être automatisée que dans le cas de documents très structurés. Nous entendons par là des documents dont on connaît aussi bien la structure logique, c'est-à-dire la "sémantique" du document, la hiérarchie des entités, que la structure physique qui dépend du style de présentation (topologie, topographie, typographie,...) [BELAID 1992A].

Les fiches de bibliothèques appartiennent selon la classification de Tombre [BELAID 1992B] à la catégorie des documents composites. Cette catégorie comporte deux sortes de documents :

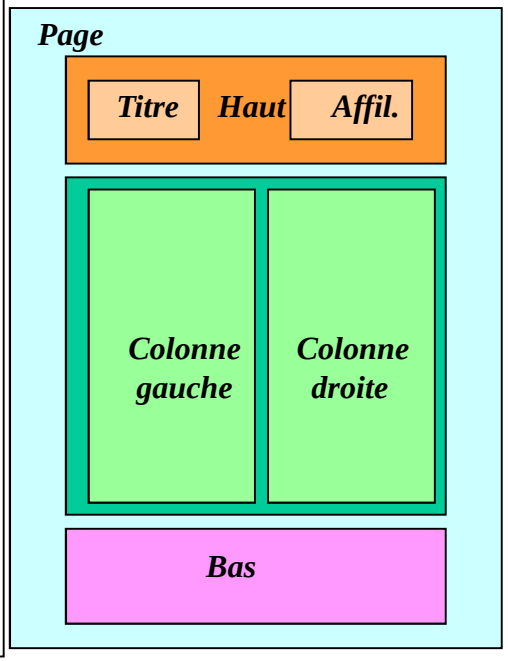
- D'une part les documents ayant une forte structuration logique et un style typographique riche, qui peuvent, après une étude poussée, être reconnus et traités. (cf. les travaux de [INGOLD 1989], [DERRIEN-PEDDEN 1990], [AKINDELE 1995] ou [MOJAHID 1994]).
- D'autre part une seconde catégorie de documents n'obéissant, le plus souvent, à aucune règle précise et reconnue et qui obligent à effectuer une analyse plus rudimentaire à partir d'heuristiques ; les notices de bibliothèques appartiennent à cette catégorie.

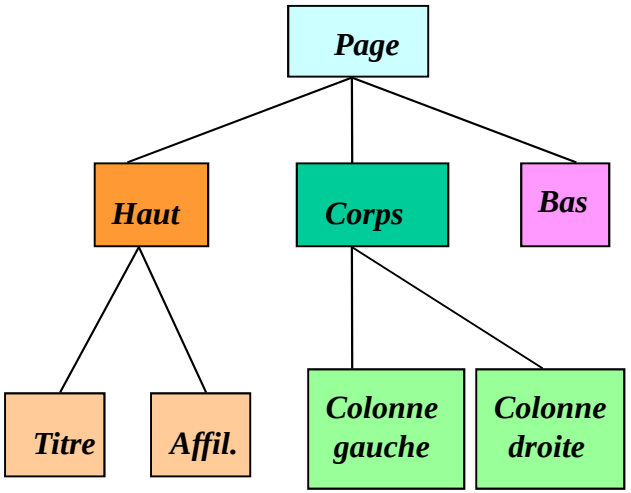
La lecture des notices de bibliothèques

A l'INSA de Lyon, le problème était d'intégrer dans un nouveau système de bases de données (cf. fig 1), l'ancien fonds, en général répertorié sur des fiches individuelles.

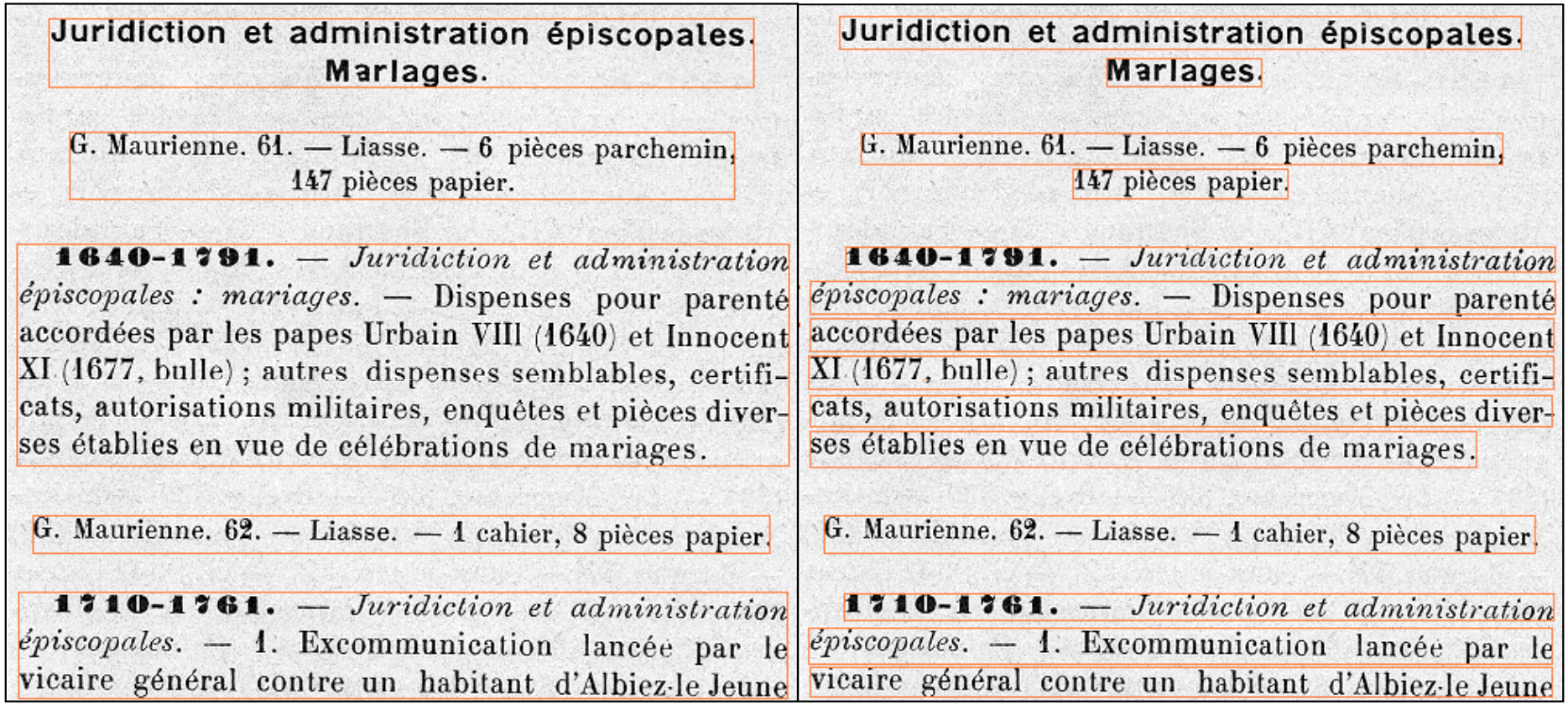
COTE	
VEDP	
TITRE	
PUBCO	
TICOLL	
NOTE	ISBN

Figure 1 : Modèle de la base de données de la bibliothèque de l'INSA de Lyon





Structure PHYSIQUE d'un document (2)



Juridiction et administration épiscopales. Mariages.

G. Maurienne. 61. — Liasse. — 6 pièces parchemin, 147 pièces papier.

1640-1791. — *Juridiction et administration épiscopales : mariages.* — Dispenses pour parenté accordées par les papes Urbain VIII (1640) et Innocent XI (1677, bulle) ; autres dispenses semblables, certificats, autorisations militaires, enquêtes et pièces diverses établies en vue de célébrations de mariages.

G. Maurienne. 62. — Liasse. — 1 cahier, 8 pièces papier.

1710-1761. — *Juridiction et administration épiscopales.* — 1. Excommunication lancée par le vicaire général contre un habitant d'Albiez-le Jeune

Juridiction et administration épiscopales. Mariages.

G. Maurienne. 61. — Liasse. — 6 pièces parchemin, 147 pièces papier.

1640-1791. — *Juridiction et administration épiscopales : mariages.* — Dispenses pour parenté accordées par les papes Urbain VIII (1640) et Innocent XI (1677, bulle) ; autres dispenses semblables, certificats, autorisations militaires, enquêtes et pièces diverses établies en vue de célébrations de mariages.

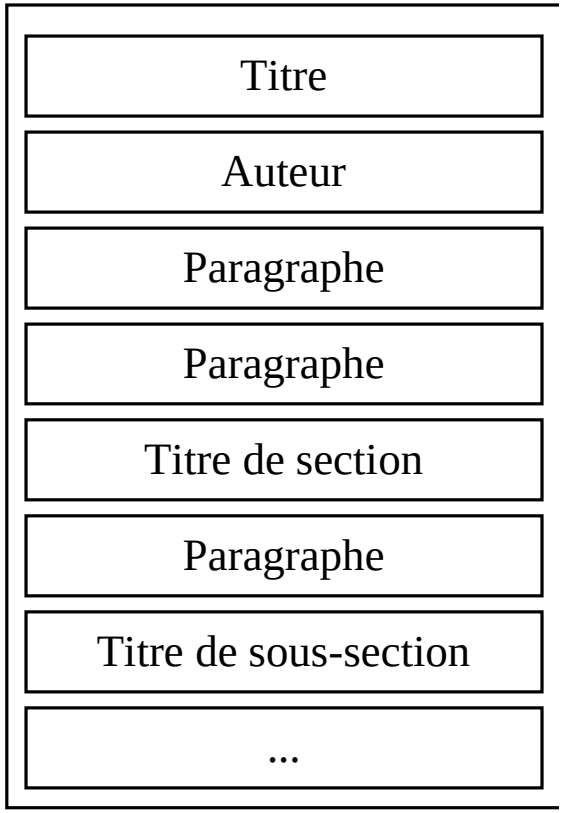
G. Maurienne. 62. — Liasse. — 1 cahier, 8 pièces papier.

1710-1761. — *Juridiction et administration épiscopales.* — 1. Excommunication lancée par le vicaire général contre un habitant d'Albiez-le Jeune

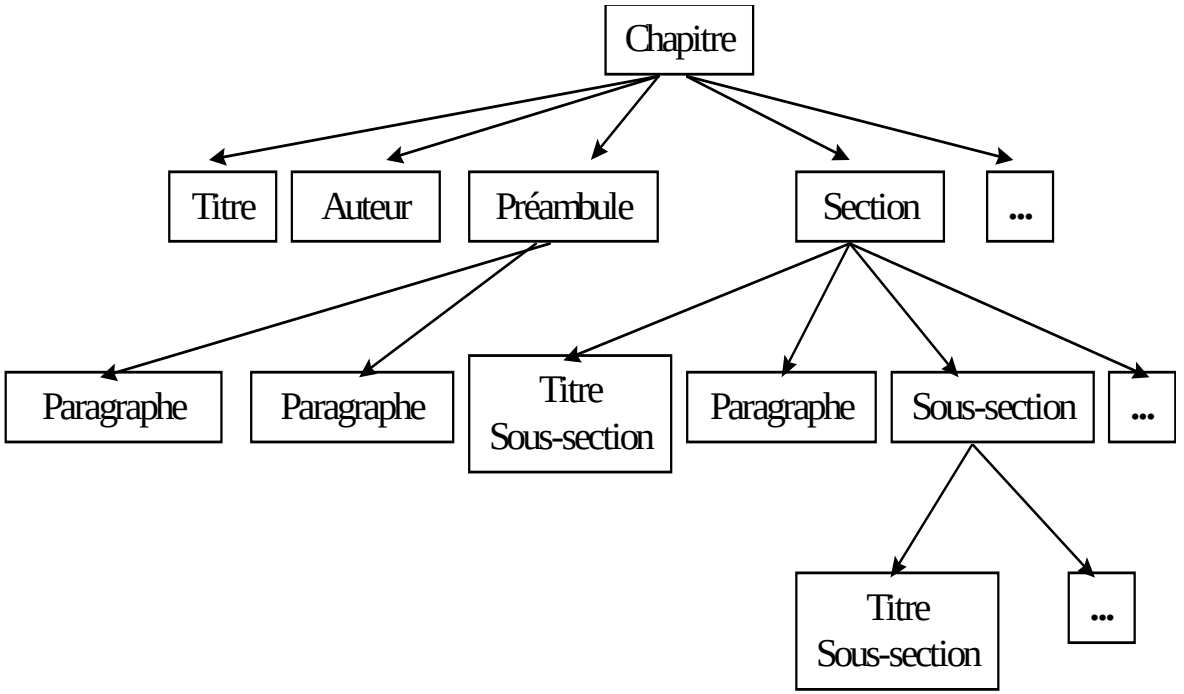
Pour une **localisation** des informations dans la page

Structure LOGIQUE d'un document (1)

La structure logique décompose le document en éléments d'information caractérisés par le rôle qu'ils jouent dans le document (titre de chapitre, chapitre, paragraphe). Elle spécifie les relations (syntaxique et sémantique) entre ces éléments appelés aussi entités logiques

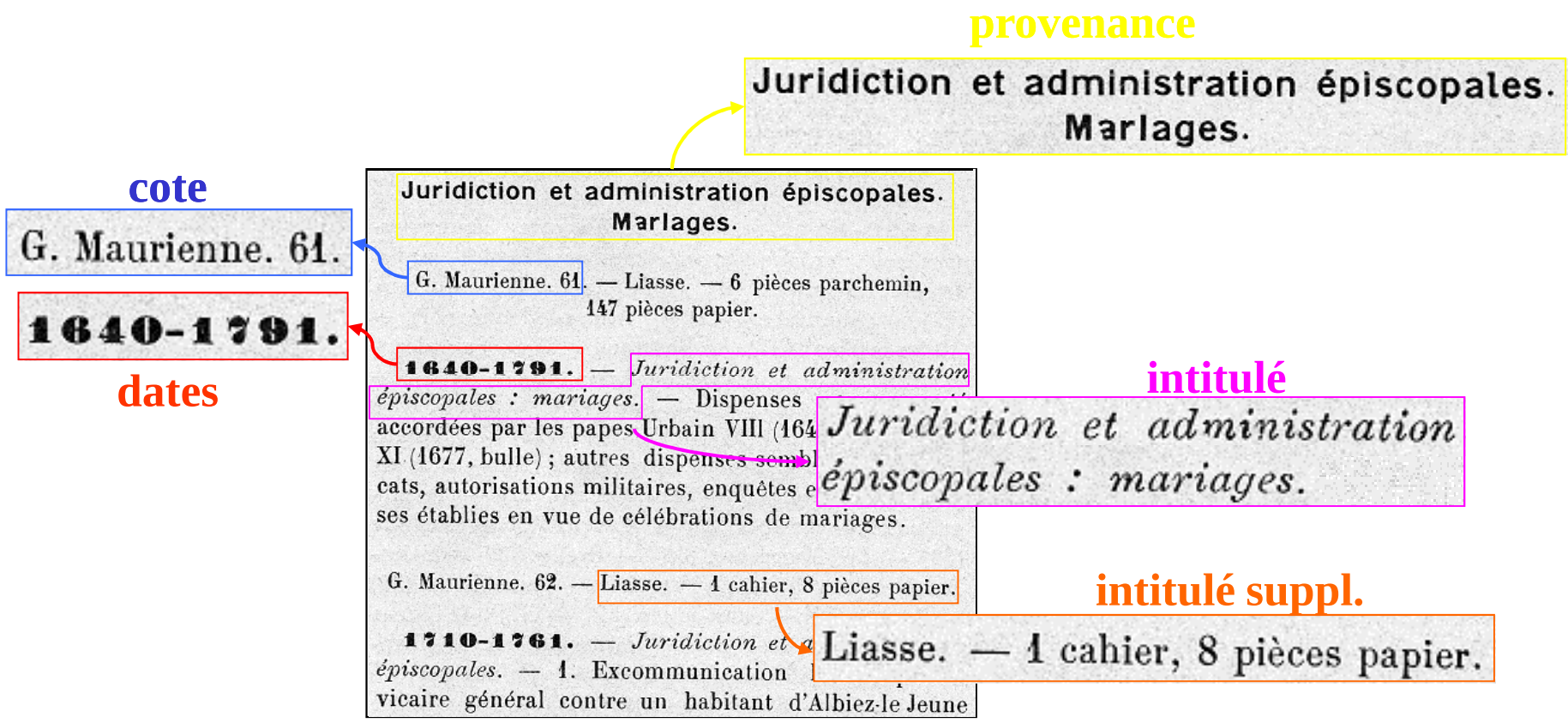


Exemple de structure logique de liste



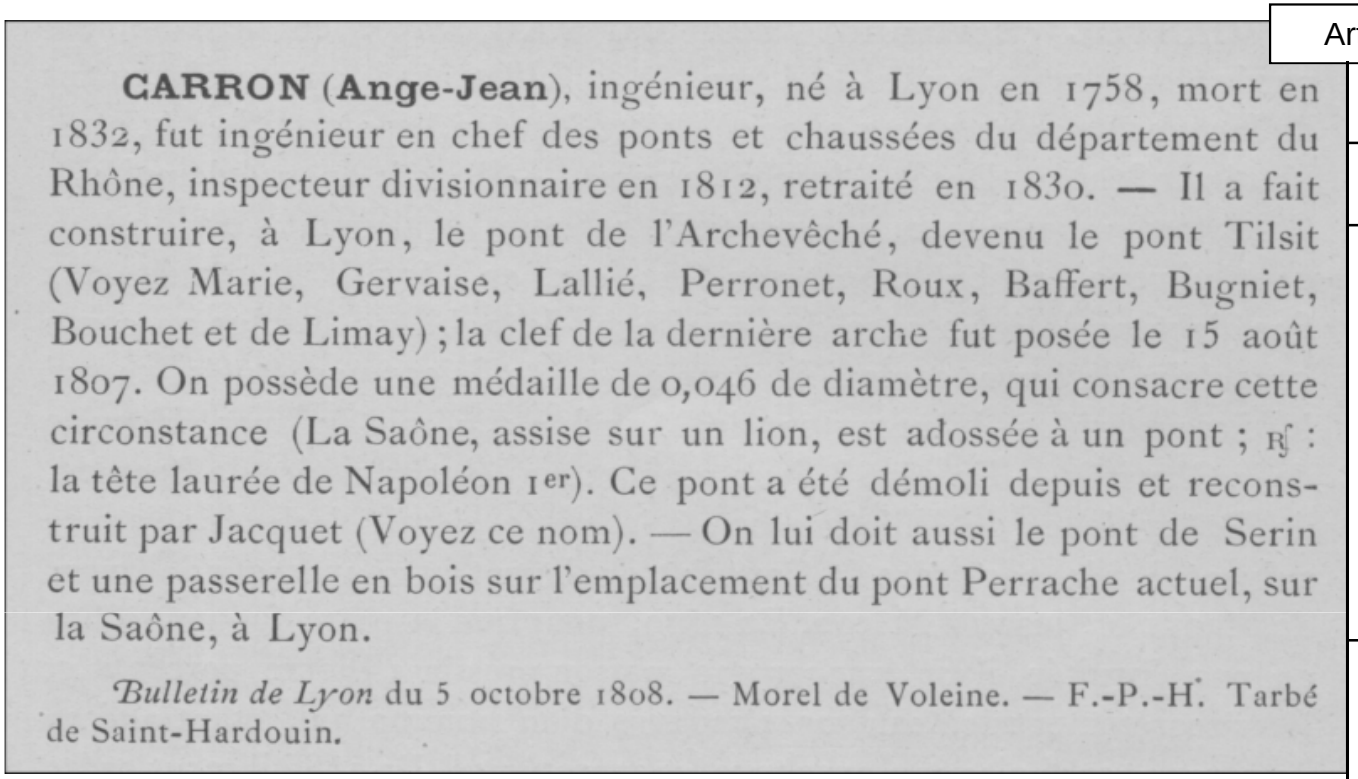
Exemple de structure logique en arbre

Structure LOGIQUE d'un document (2)



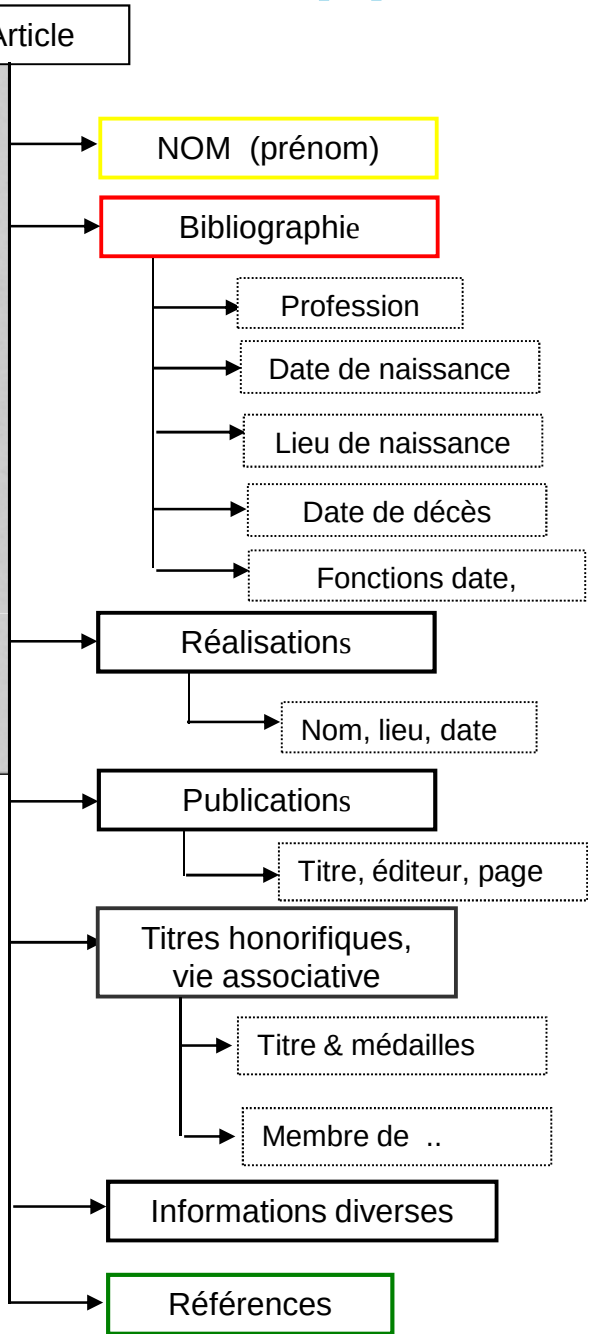
Pour une **identification** des zones informatives

Structure LOGIQUE d'un document (3)



CARRON (**Ange-Jean**), ingénieur, né à Lyon en 1758, mort en 1832, fut ingénieur en chef des ponts et chaussées du département du Rhône, inspecteur divisionnaire en 1812, retraité en 1830. — Il a fait construire, à Lyon, le pont de l'Archevêché, devenu le pont Tilsit (Voyez Marie, Gervaise, Lallié, Perronet, Roux, Baffert, Bugniet, Bouchet et de Limay) ; la clef de la dernière arche fut posée le 15 août 1807. On possède une médaille de 0,046 de diamètre, qui consacre cette circonstance (La Saône, assise sur un lion, est adossée à un pont ; ℞ : la tête laurée de Napoléon 1er). Ce pont a été démoli depuis et reconstruit par Jacquet (Voyez ce nom). — On lui doit aussi le pont de Serin et une passerelle en bois sur l'emplacement du pont Perrache actuel, sur la Saône, à Lyon.

Bulletin de Lyon du 5 octobre 1808. — Morel de Voleine. — F.-P.-H. Tarbé de Saint-Hardouin.



Études possibles

1) Recherches sur le nombre d'adhérents d'une association à une date donnée,
2) Étudier l'évolution du nombre de personnes par catégorie socioprofessionnelle
3) Étudier les rapports possibles entre deux personnes

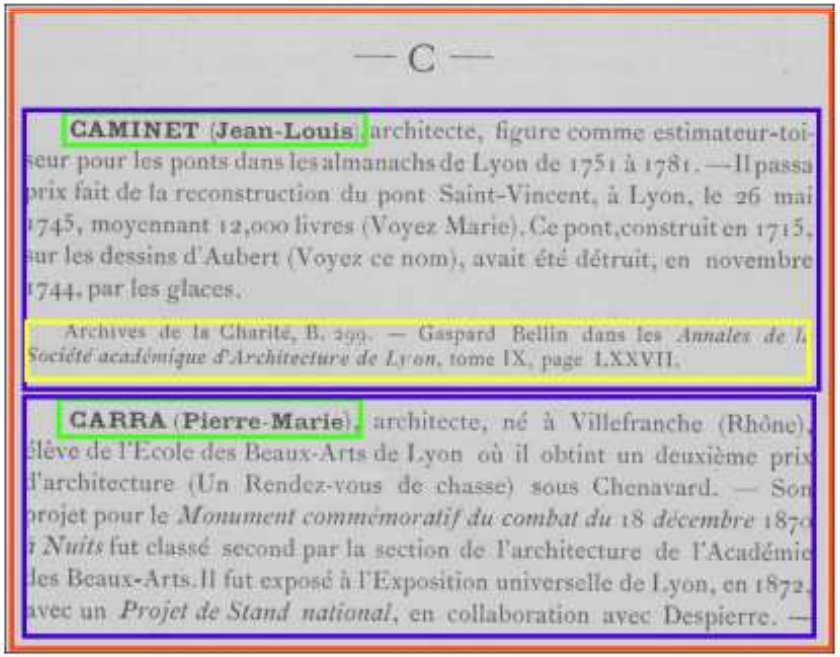
Vers une structure de l'ouvrage…

Structure d'un ouvrage (préface, chapitres, tables des matières)

Structure d'une page (texte, bas de page, paragraphes, sections)

Structure fine du texte (Fonction de chaque mot)

▶ **Vers XML, format de représentation des éléments de structure**

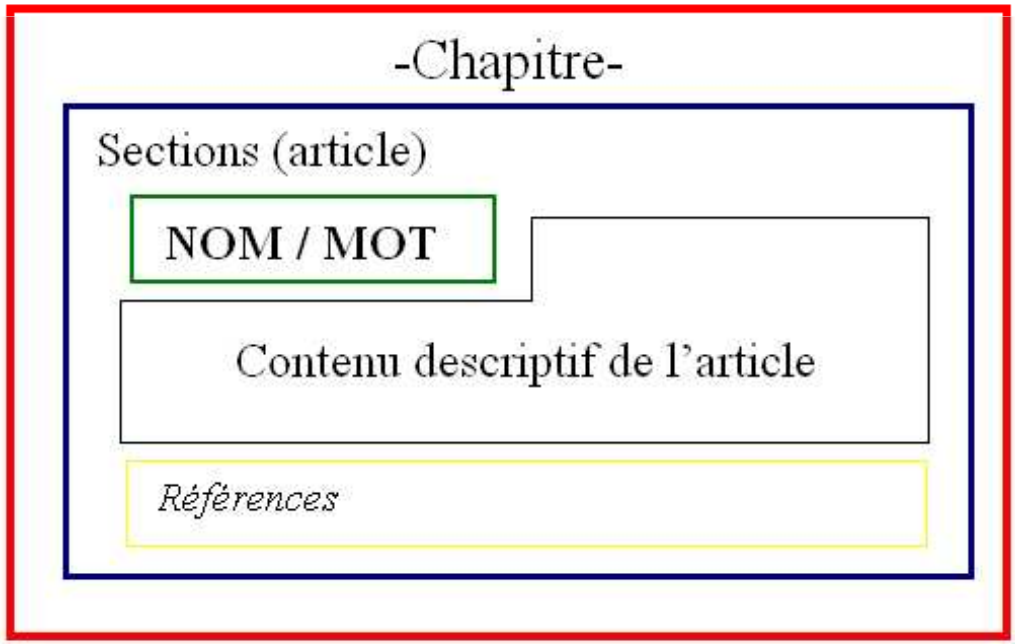


— C —

CAMINET (Jean-Louis), architecte, figure comme estimateur-toiseur pour les ponts dans les almanachs de Lyon de 1751 à 1781. — Il passa prix fait de la reconstruction du pont Saint-Vincent, à Lyon, le 26 mai 1745, moyennant 12,000 livres (Voyez Marie). Ce pont, construit en 1715, sur les dessins d'Aubert (Voyez ce nom), avait été détruit, en novembre 1744, par les glaces.

Archives de la Charité, B. 299. — Gaspard Bellin dans les *Annales de la Société académique d'Architecture de Lyon*, tome IX, page LXXVII.

CARRA (Pierre-Marie), architecte, né à Villefranche (Rhône), élève de l'École des Beaux-Arts de Lyon où il obtint un deuxième prix d'architecture (Un Rendez-vous de chasse) sous Chenavard. — Son projet pour le *Monument commémoratif du combat du 18 décembre 1870* [illegible] *Nuits* fut classé second par la section de l'architecture de l'Académie des Beaux-Arts. Il fut exposé à l'Exposition universelle de Lyon, en 1872, avec un *Projet de Stand national*, en collaboration avec Despierre. —



Les structures: pour quels usages?

▶ un outil de comparaison des modèles de documents

Exemple de description logique d'un document:

> syntaxe et grammaire

+Position de l'information
+Nature de l'information
+relation avec les autres informations
+ordonnancement
+fiabilité (0=facultative)

Titre 2eme colonne gras sans sérif taille 18 suivi de rien à gauche

↓ **1**

Sous-Titre 2eme colonne, aligné sur le précédent, style gras sans sérif taille 14

↓ **1**

Texte à gauche du précédent,caractère avec sérif taille 12

0,1

nom de l'auteur à la fin du texte, caractère avec sérif italique taille 12

0,1

Page aligné sur *nom de l'auteur* ou **Texte** style gras sans sérif taille 14 (chiffres<100)

1

<10

n

Special Report
Hostile Takeover
by Dorinda Elliott and Melinda Liu 10
Israel: One Threat It Never Prepared For 14

Europe
Bosnia: A Painful Joy *by Rod Nordland* 18
Spain: Farewell to Felipe? 20
Russia: Sex Tiger Seeks Genius 21

World Affairs
South Africa: Fortress Mentality
by Joseph Contreras 22
Iraq: Enemies Like These 24
Egypt: Digging Up Dangerous History 24
Mexico: The Cantina Comeback 25
'Dispatches' 25

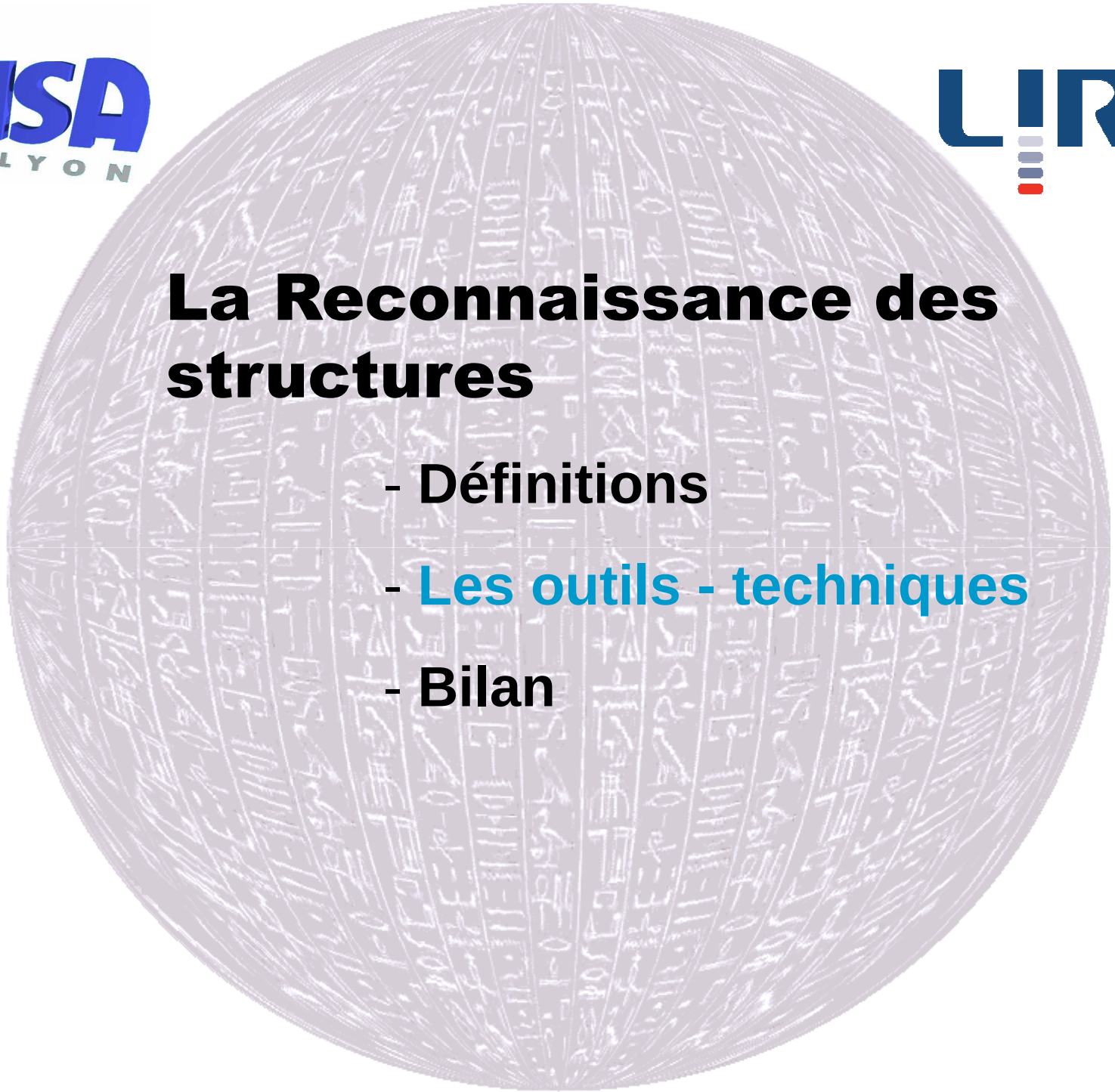
Asia
U.S. Forces: Battle of Okinawa II
by Tony Emerson 26
Japan: The Old Way Won't Work 28
India: The Gods Are Thirsty 29
Natural Disasters: The Waters of Death 30

U.S. Affairs
Reagan: The Long Goodbye *by Eleanor Clift* 32
Alzheimer's Terrible Toll 36
Congress: The Alaskan Assault 37

Business
Wall Street: Big Deals, Big Talk
by Allan Sloan 38
Time's Uneasy Pieces
by Johnnie L. Roberts 40
Turner Wants to Be Big 42
IBM: Private Banking 43
'Bottom Line' 43
Nike: Just Doing It 44

Society & The Arts
Olympics: Bounds for Glory *by Mark Starr* 46
Books: Advice for the Expectant 53
Science: A New King of the Hill 53
Music: California Dreamin' *by Katrine Ames* 54

Departments
Periscope 3
Cyberscope 4
Perspectives 9
Newsmakers 31
Transition 31
Interview: William Knipe 56



La Reconnaissance des structures

- Définitions
- Les outils - techniques
- Bilan

Au départ, la segmentation

Une bonne segmentation est déjà une étape importante pour la reconnaissance !!!!

Une segmentation est un traitement **irréversible** car c'est le résultat d'une **interprétation** suivant un **critère** et une **méthode**.

Les images non segmentées conservent toute l'information

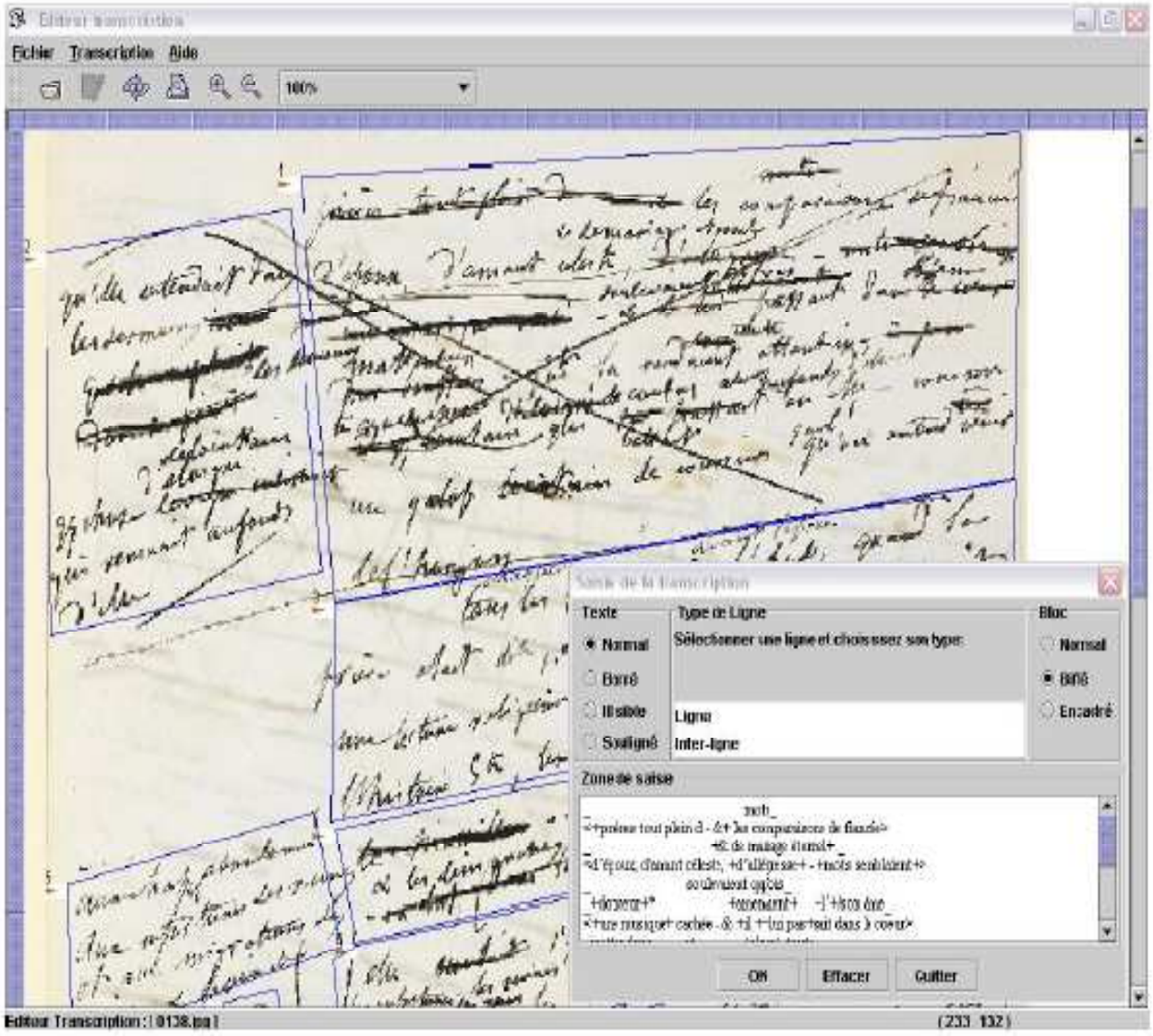
Toute segmentation est le produit d'un choix d'une méthode et de ses paramètres !

Plusieurs méthodes de segmentation :

- suivant la couleur (seuillage)
- à partir d'analyse de formes
- à partir de connaissances

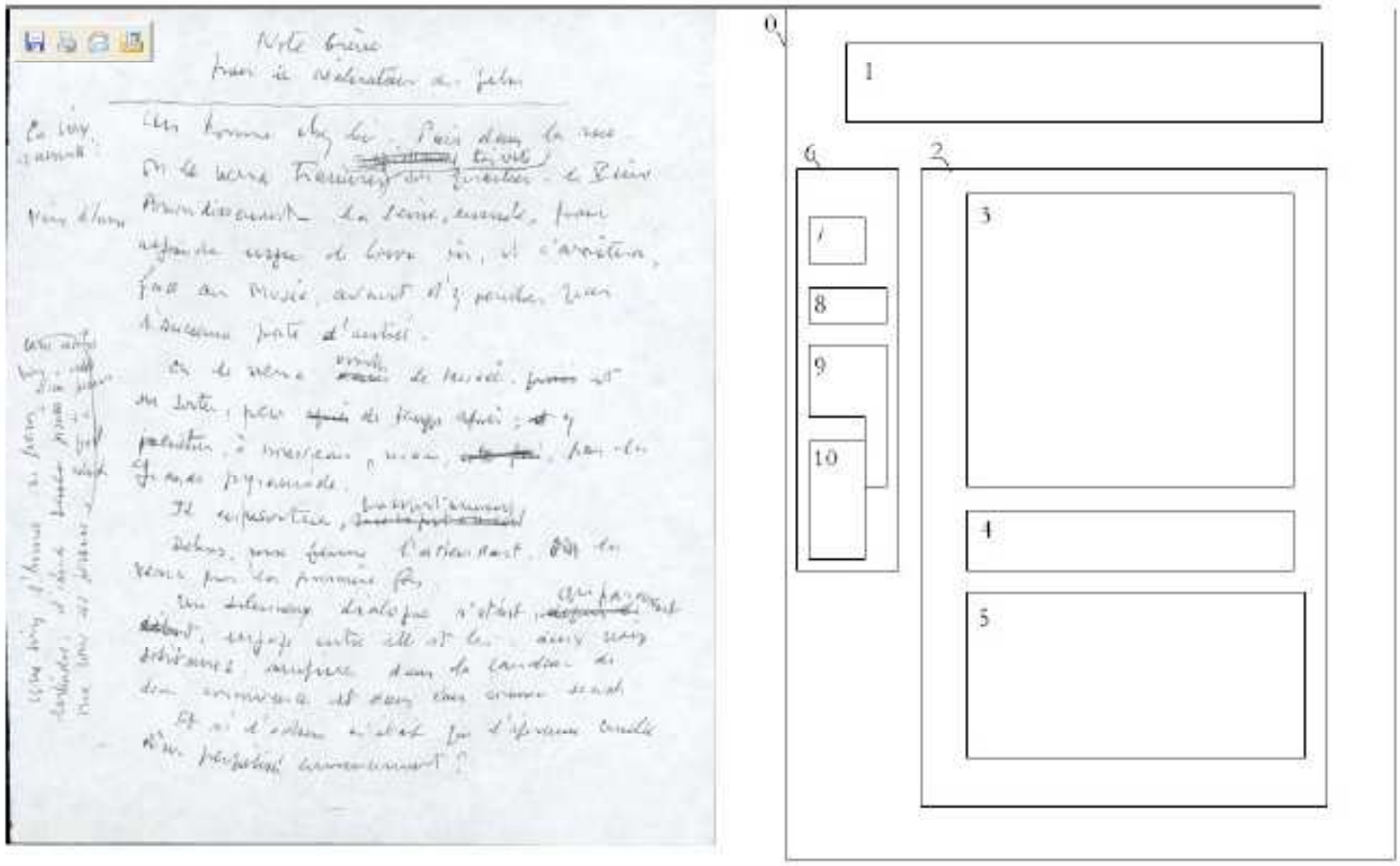
Cette opération n'est pas toujours réalisable, dépend de la qualité initiale de l'image

Les cas improbables



Automatiser ce que l'expert doit faire « à la main ». Interface EMMA, 2005.

Les découpages idéaux



Découpage idéal d'un feuillet manuscrit en régions. Fekete, 2004

Des méthodes pour une segmentation physique

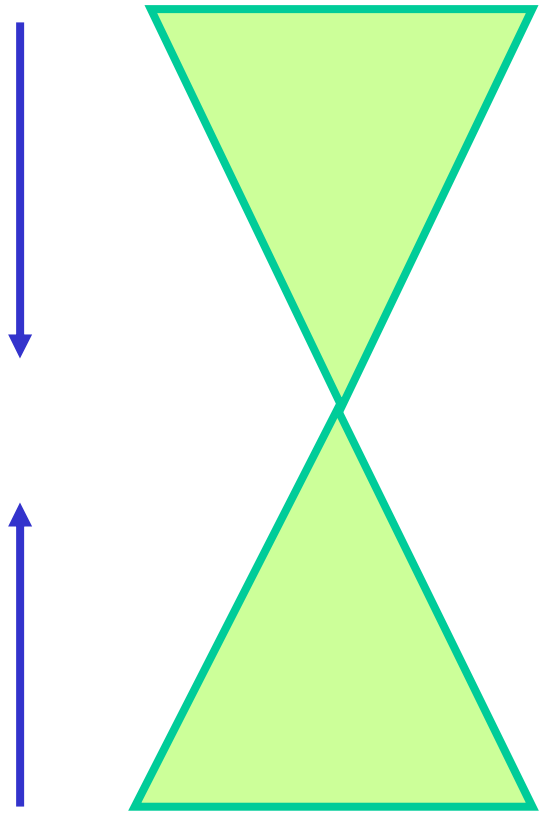
- Cas des documents réguliers composites
- Cas des courriers d'entreprises
- Cas des documents du patrimoines et des manuscrits irréguliers

La structure physique du document

La structure physique : les méthodes d'analyse

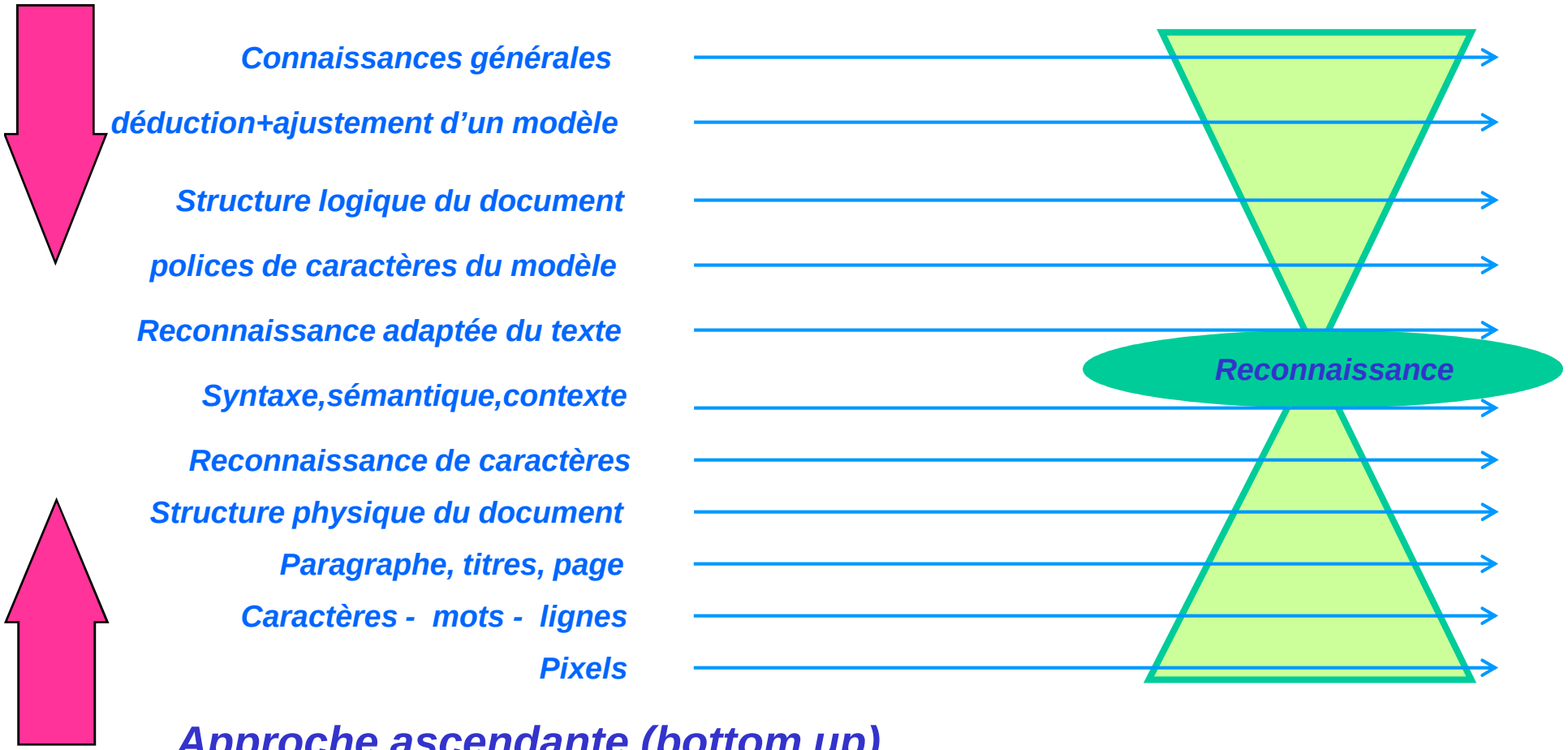
- Les méthodes ne sont pas concurrentes
- Elles sont complémentaires
- On distingue plusieurs classes

  - Descendantes
  - Ascendantes
  - Mixtes
  - Les variantes



La structure physique du document: Approches

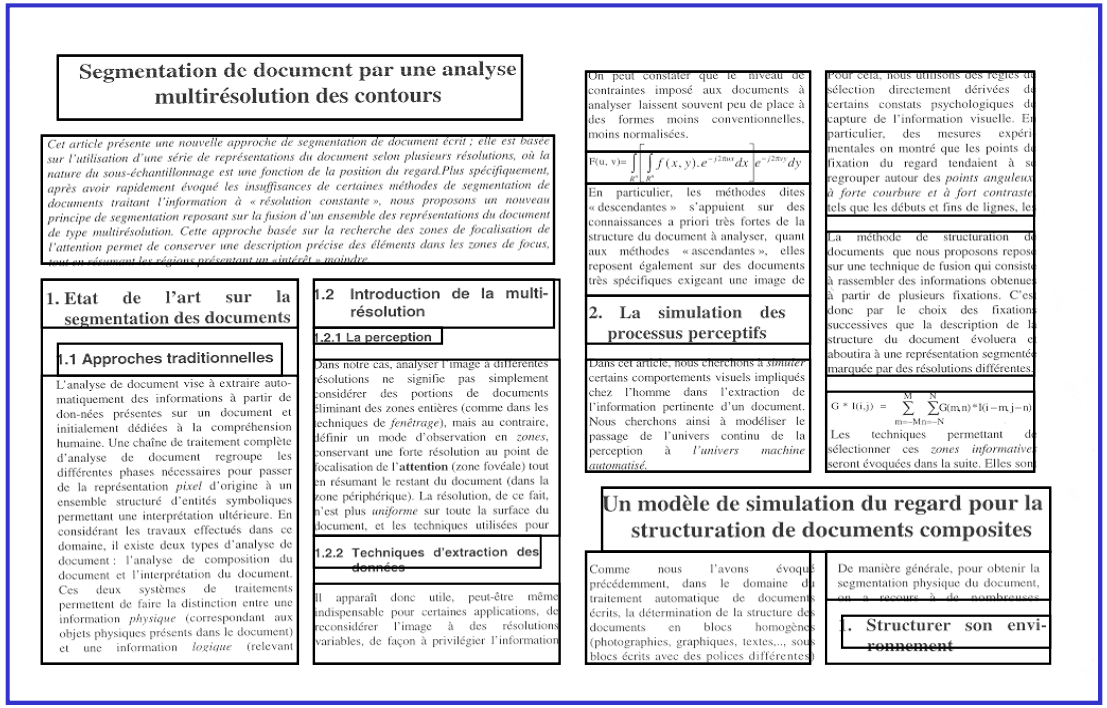
Approche descendante (top down)



Approche ascendante (bottom up)

Approches descendantes (top-down)

Fondées sur le découpage de l'image en zones de grande taille découpées ensuite en petites zones par analyse de propriétés spécifiques en relation avec la nature du document traité.



Segmentation de document par une analyse multirésolution des contours

Cet article présente une nouvelle approche de segmentation de document écrit : elle est basée sur l'utilisation d'une série de représentations du document selon plusieurs résolutions, où la nature du sous-échantillonnage est une fonction de la position du regard.Plus spécifiquement, après avoir rapidement évoqué les insuffisances de certaines méthodes de segmentation de documents traitant l'information à « résolution constante », nous proposons un nouveau principe de segmentation reposant sur la fusion d'un ensemble des représentations du document de type multirésolution. Cette approche basée sur la recherche des zones de focalisation de l'attention permet de conserver une description précise des éléments dans les zones de focus, tout en résumant les régions présentant un intérêt « moindre ».

1. Etat de l'art sur la segmentation des documents

1.1 Approches traditionnelles

L'analyse de document vise à extraire automatiquement des informations à partir de données présentes sur un document et initialement dédiées à la compréhension humaine. Une chaîne de traitement complète d'analyse de document regroupe les différentes phases nécessaires pour passer de la représentation *pixel* d'origine à un ensemble structuré d'entités symboliques permettant une interprétation ultérieure. En considérant les travaux effectués dans ce domaine, il existe deux types d'analyse de document : l'analyse de composition du document et l'interprétation du document. Ces deux systèmes de traitements permettent de faire la distinction entre une information *physique* (correspondant aux objets physiques présents dans le document) et une information *logique* (relevant

1.2 Introduction de la multi-résolution

1.2.1 La perception

Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents éliminant des zones entières (comme dans les techniques de *fenêtrage*), mais au contraire, définir un mode d'observation en *zones*, conservant une forte résolution au point de focalisation de l'**attention** (zone fovéale) tout en résumant le restant du document (dans la zone périphérique). La résolution, de ce fait, n'est plus *uniforme* sur toute la surface du document, et les techniques utilisées pour

1.2.2 Techniques d'extraction des données

Il apparaît donc utile, peut-être même indispensable pour certaines applications, de reconsidérer l'image à des résolutions variables, de façon à privilégier l'information

On peut constater que le niveau de contraintes imposé aux documents à analyser laissent souvent peu de place à des formes moins conventionnelles, moins normalisées.

$$F(u,v)=\int_{R^n}\left[\int_{R^m} f(x,y).e^{-j2\pi ux}dx\right]e^{-j2\pi vy}dy$$

En particulier, les méthodes dites « descendantes » s'appuient sur des connaissances a priori très fortes de la structure du document à analyser, quant aux méthodes « ascendantes », elles reposent également sur des documents très spécifiques exigeant une image de

2. La simulation des processus perceptifs

Dans cet article, nous cherchons à *simuler* certains comportements visuels impliqués chez l'homme dans l'extraction de l'information pertinente d'un document. Nous cherchons ainsi à modéliser le passage de l'univers continu de la perception à *l'univers machine automatisé.*

Pour cela, nous utilisons des règles de sélection directement dérivées de certains constats psychologiques de capture de l'information visuelle. En particulier, des mesures expérimentales on montré que les points de fixation du regard tendaient à se regrouper autour des *points anguleux à forte courbure et à fort contraste* tels que les débuts et fins de lignes, les

La méthode de structuration de documents que nous proposons repose sur une technique de fusion qui consiste à rassembler des informations obtenues à partir de plusieurs fixations. C'est donc par le choix des fixations successives que la description de la structure du document évoluera et aboutira à une représentation segmentée marquée par des résolutions différentes.

$$G * I(i,j) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} G(m,n) * I(i-m, j-n)$$

Les techniques permettant de sélectionner ces *zones informatives* seront évoquées dans la suite. Elles sont

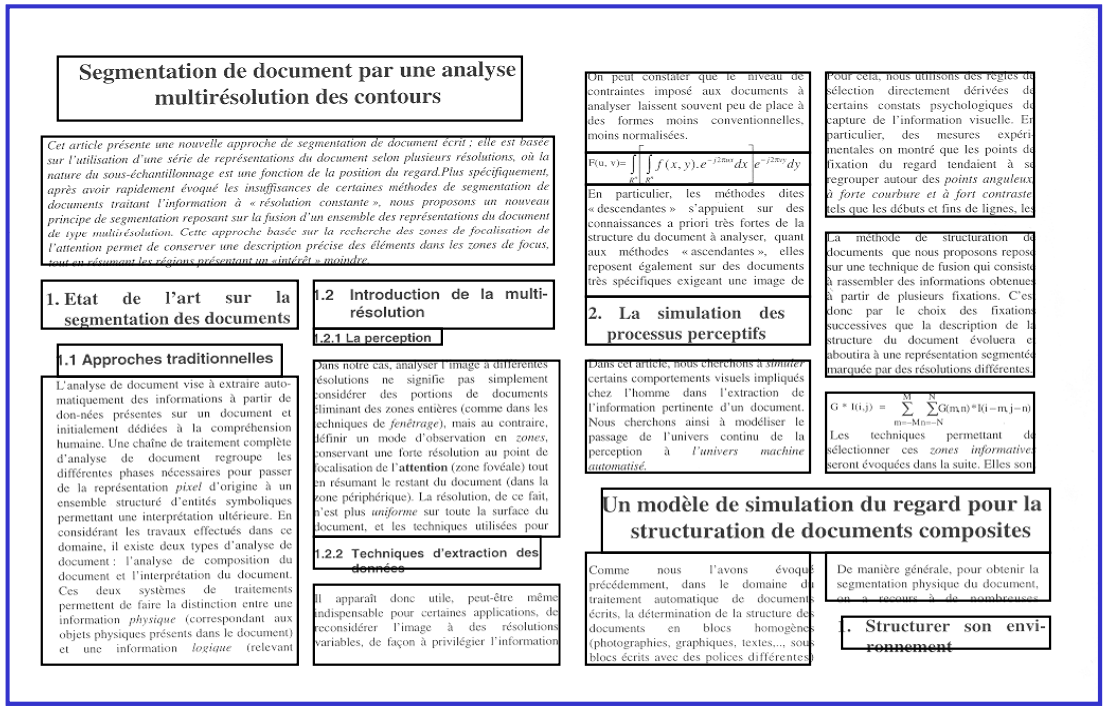
Un modèle de simulation du regard pour la structuration de documents composites

Comme nous l'avons évoqué précédemment, dans le domaine du traitement automatique de documents écrits, la détermination de la structure des documents en blocs homogènes (photographies, graphiques, textes,.., sous blocs écrits avec des polices différentes)

De manière générale, pour obtenir la segmentation physique du document, on a recours à de nombreuses

1. Structurer son environnement

Approches descendantes (top-down)



Segmentation de document par une analyse multirésolution des contours

Cet article présente une nouvelle approche de segmentation de document écrit ; elle est basée sur l'utilisation d'une série de représentations du document selon plusieurs résolutions, où la nature du sous-échantillonnage est une fonction de la position du regard.Plus spécifiquement, après avoir rapidement évoqué les insuffisances de certaines méthodes de segmentation de documents traitant l'information à « résolution constante », nous proposons un nouveau principe de segmentation reposant sur la fusion d'un ensemble des représentations du document de type multirésolution. Cette approche basée sur la recherche des zones de focalisation de l'attention permet de conserver une description précise des éléments dans les zones de focus, tout en résumant les régions présentant un «intérêt» moindre.

1. Etat de l'art sur la segmentation des documents

1.1 Approches traditionnelles

L'analyse de document vise à extraire automatiquement des informations à partir de données présentes sur un document et initialement dédiées à la compréhension humaine. Une chaîne de traitement complète d'analyse de document regroupe les différentes phases nécessaires pour passer de la représentation *pixel* d'origine à un ensemble structuré d'entités symboliques permettant une interprétation ultérieure. En considérant les travaux effectués dans ce domaine, il existe deux types d'analyse de document : l'analyse de composition du document et l'interprétation du document. Ces deux systèmes de traitements permettent de faire la distinction entre une information *physique* (correspondant aux objets physiques présents dans le document) et une information *logique* (relevant

1.2 Introduction de la multi-résolution

1.2.1 La perception

Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents éliminant des zones entières (comme dans les techniques de *fenêtrage*), mais au contraire, définir un mode d'observation en *zones*, conservant une forte résolution au point de focalisation de l'**attention** (zone fovéale) tout en résumant le restant du document (dans la zone périphérique). La résolution, de ce fait, n'est plus *uniforme* sur toute la surface du document, et les techniques utilisées pour

1.2.2 Techniques d'extraction des données

Il apparaît donc utile, peut-être même indispensable pour certaines applications, de reconsidérer l'image à des résolutions variables, de façon à privilégier l'information

On peut constater que le niveau de contraintes imposé aux documents à analyser laissent souvent peu de place à des formes moins conventionnelles, moins normalisées.

$$F(u,v)=\int_{\mathbb{R}}\left[\int_{\mathbb{R}} f(x,y).e^{-j2\pi ux}dx\right]e^{-j2\pi vy}dy$$

En particulier, les méthodes dites « descendantes » s'appuient sur des connaissances a priori très fortes de la structure du document à analyser, quant aux méthodes « ascendantes », elles reposent également sur des documents très spécifiques exigeant une image de

2. La simulation des processus perceptifs

Dans cet article, nous cherchons à *simuler* certains comportements visuels impliqués chez l'homme dans l'extraction de l'information pertinente d'un document. Nous cherchons ainsi à modéliser le passage de l'univers continu de la perception à *l'univers machine automatisé*.

Pour cela, nous utilisons des règles de sélection directement dérivées de certains constats psychologiques de capture de l'information visuelle. En particulier, des mesures expérimentales on montré que les points de fixation du regard tendaient à se regrouper autour des *points anguleux à forte courbure et à fort contraste* tels que les débuts et fins de lignes, les

La méthode de structuration de documents que nous proposons repose sur une technique de fusion qui consiste à rassembler des informations obtenues à partir de plusieurs fixations. C'est donc par le choix des fixations successives que la description de la structure du document évoluera et aboutira à une représentation segmentée marquée par des résolutions différentes.

$$G * I(i,j) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} G(m,n) * I(i-m,j-n)$$

Les techniques permettant de sélectionner ces *zones informatives* seront évoquées dans la suite. Elles son

Un modèle de simulation du regard pour la structuration de documents composites

Comme nous l'avons évoqué précédemment, dans le domaine du traitement automatique de documents écrits, la détermination de la structure des documents en blocs homogènes (photographies, graphiques, textes,... sous blocs écrits avec des polices différentes)

De manière générale, pour obtenir la segmentation physique du document, on a recours à de nombreuses

1. Structurer son environnement

Propriétés spécifiques:

- *Mise en page*
- *Nombre de colonnes*
- *Présence de titres*
- *Adjacence de blocs de textes*
- *Présence de graphique*
- *De tableau, de formules…*

Approches descendantes (top-down)

***Principe générique* : objectifs**

Automatiser la localisation des régions d 'intérêts

- *Analyse de structures complexes (documents composites : pages de journaux)*
- *Analyse de formulaires complexes à base de tables*
- *Analyse de mails*
- *Analyse de documents administratifs*
-...

Approches descendantes (top-down)

- *Connaissances **a priori** de la structure du document à analyser (connaissances de haut niveau).*

- *S'appliquent essentiellement à des documents **très spécifiques** et **très hiérarchisés** (documents administratifs et scientifiques).*

***Trois variantes de cette classe de méthodes** :*

Techniques de projection (XY-Tree)
Technique de lissage directionnel
Technique du pavage du fond de l'image

Approches descendantes (top-down)

La méthode R-XYC (Recursive X-Y Cuts) [Nagy 1986]

Hypothèse *: les éléments de structure (colonnes, paragraphes, figures) peuvent être contenus dans des blocs rectangulaires.*

Basée sur les profils de projections horizontale et verticale.

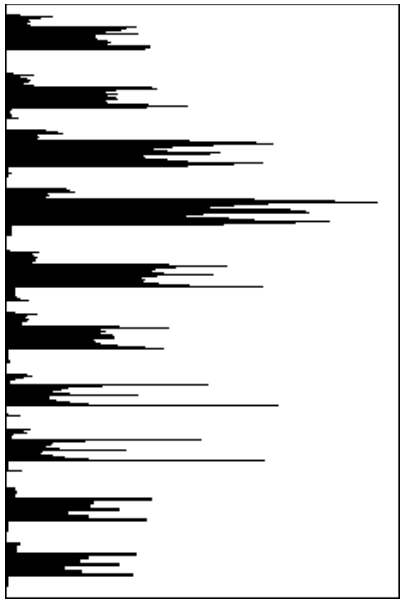
- *Les colonnes de texte sont localisées par profil de projection verticale,*
- *Les coupures entre paragraphes sont obtenues par profil de projection horizontale.*
- *Les lignes à l'intérieur des paragraphes peuvent ensuite être obtenues par projections horizontales (récursives).*

Approches descendantes (top-down)

La méthode R-XYC

Exemple de profil horizontal sur un bloc de texte

Voilà du texte en Times tout bête.
Maintenant, de l'italique pour faire le test.
Et si on essayait avec de l'Arial, maintenant?
Voila l'Arial italique qui se présente bien différemment.
On peut également essayer le Book Antiqua.
L'italique de cette chose n'est pas la même.
Le fameux courrier ne doit pas être négligé.
L'italique va être critique ici, je le sens.
Le sans sérif est catastrophique.
L'italique est encore bien pire!



Approches descendantes (top-down)

La méthode R-XYC

1. *Projection récursive de l'image alternativement en x et en y.*
2. *Analyse des histogrammes de projection.*
3. *Se baser sur des seuils évolutifs suivant le degré de récursivité.*
4. *Puis définir les points de découpage :*
   - *Détection de pics locaux à partir des profils horizontaux (PH)*
   - *Localisation des espaces interlignes (PH) et intercolonnes (PV).*
   - *Les points de découpage dans les deux axes correspondent aux espaces les plus larges localisés sur les profils de projection. (détection des grandes zones de transition entre deux paragraphes ou deux colonnes.)*

Approches descendantes (top-down)

La méthode R-XYC :

Pour la localisation des espaces interlignes et intercolonnes

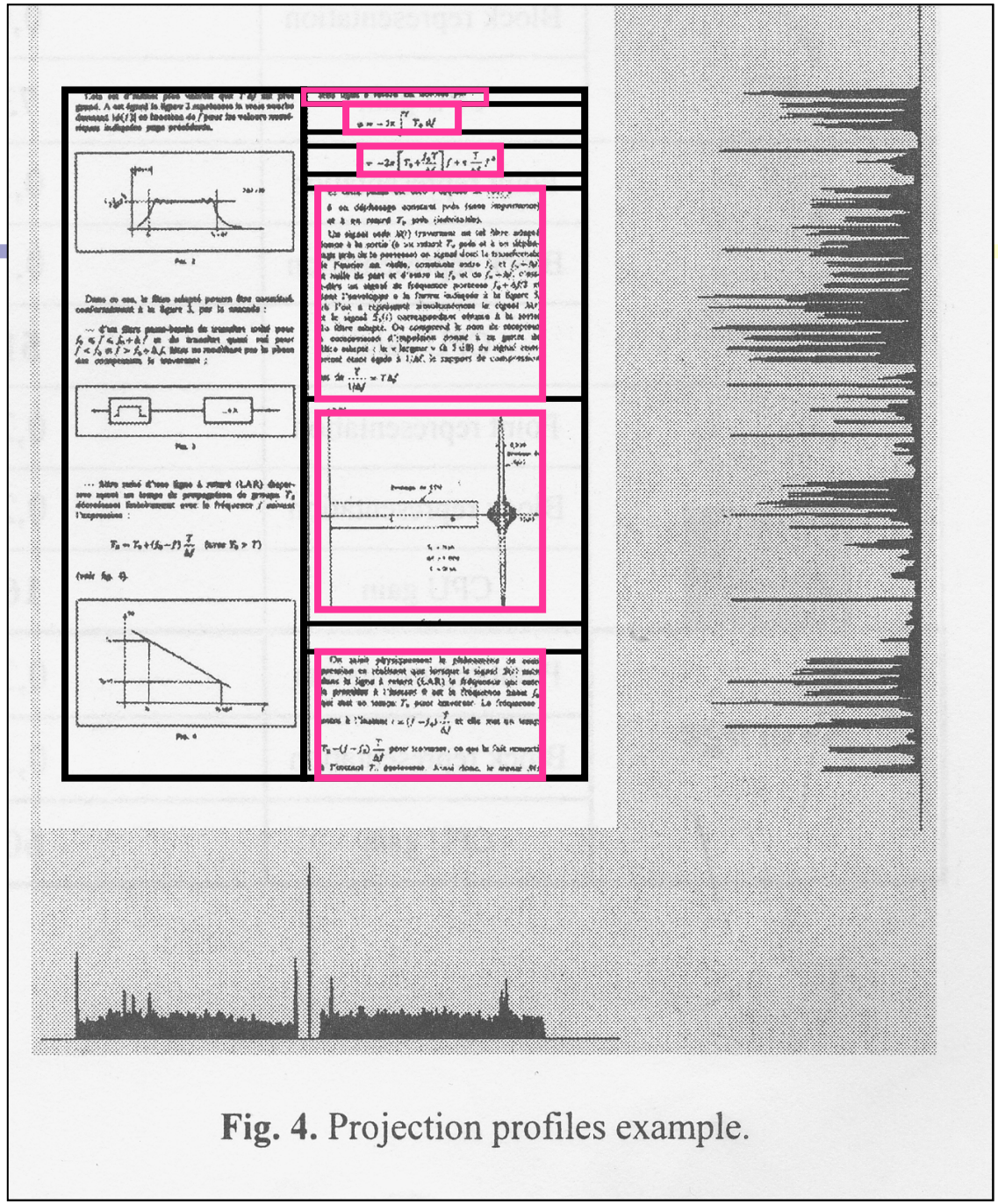
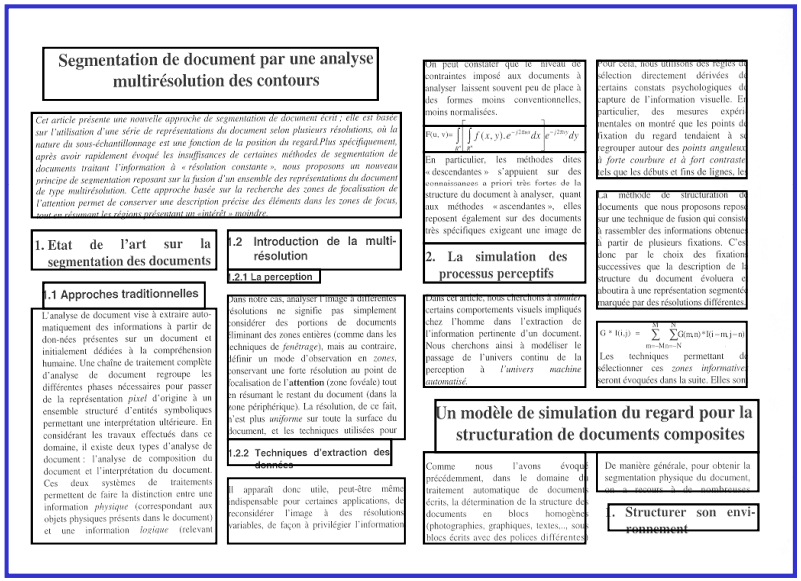


Fig. 4. Projection profiles example.

Approches descendantes (top-down)

La représentation X-Y Tree



Segmentation de document par une analyse multirésolution des contours

Cet article présente une nouvelle approche de segmentation de document écrit ; elle est basée sur l'utilisation d'une série de représentations du document selon plusieurs résolutions, où la nature du sous-échantillonnage est une fonction de la position du regard.Plus spécifiquement, après avoir rapidement évoqué les insuffisances de certaines méthodes de segmentation de documents traitant l'information à « résolution constante », nous proposons un nouveau principe de segmentation reposant sur la fusion d'un ensemble des représentations du document de type multirésolution. Cette approche basée sur la recherche des zones de focalisation de l'attention permet de conserver une description précise des éléments dans les zones de focus, tout en résumant les régions présentant un «intérêt » moindre.

1. Etat de l'art sur la segmentation des documents

1.1 Approches traditionnelles

L'analyse de document vise à extraire automatiquement des informations à partir de documents présentes sur un document et initialement dédiées à la compréhension humaine. Une chaîne de traitement complète d'analyse de document regroupe les différentes phases nécessaires pour passer de la représentation *pixel* d'origine à un ensemble structuré d'entités symboliques permettant une interprétation ultérieure. En considérant les travaux effectués dans ce domaine, il existe deux types d'analyse de document : l'analyse de composition du document et l'interprétation du document. Ces deux systèmes de traitements permettent de faire la distinction entre une information *physique* (correspondant aux objets physiques présents dans le document) et une information *logique* (relevant

1.2 Introduction de la multi-résolution

1.2.1 La perception

Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents éliminant des zones entières (comme dans les techniques de *fenêtrage*), mais au contraire, définir un mode d'observation en *zones*, conservant une forte résolution au point de focalisation de l'**attention** (zone fovéale) tout en résumant le restant du document (dans la zone périphérique). La résolution, de ce fait, n'est plus *uniforme* sur toute la surface du document, et les techniques utilisées pour

1.2.2 Techniques d'extraction des données

Il apparaît donc utile, peut-être même indispensable pour certaines applications, de reconsidérer l'image à des résolutions variables, de façon à privilégier l'information

On peut constater que le niveau de contraintes imposé aux documents à analyser laissent souvent peu de place à des formes moins conventionnelles, moins normalisées.

$$F(u,v)=\int_{x}\left[\int_{y} f(x,y).e^{-j2\pi ux}dx\right]e^{-j2\pi vy}dy$$

En particulier, les méthodes dites « descendantes » s'appuient sur des connaissances a priori très fortes de la structure du document à analyser, quant aux méthodes « ascendantes », elles reposent également sur des documents très spécifiques exigeant une image de

2. La simulation des processus perceptifs

Dans cet article, nous cherchons à *simuler* certains comportements visuels impliqués chez l'homme dans l'extraction de l'information pertinente d'un document. Nous cherchons ainsi à modéliser le passage de l'univers continu de la perception à *l'univers machine automatisé*.

Un modèle de simulation du regard pour la structuration de documents composites

Comme nous l'avons évoqué précédemment, dans le domaine du traitement automatique de documents écrits, la détermination de la structure des documents en blocs homogènes (photographies, graphiques, textes.., sous blocs écrits avec des polices différentes

Pour cela, nous utilisons des règles de sélection directement dérivées de certains constats psychologiques de capture de l'information visuelle. En particulier, des mesures expérimentales on montré que les points de fixation du regard tendaient à se regrouper autour des *points anguleux, à forte courbure et à fort contraste*, tels que les débuts et fins de lignes, les

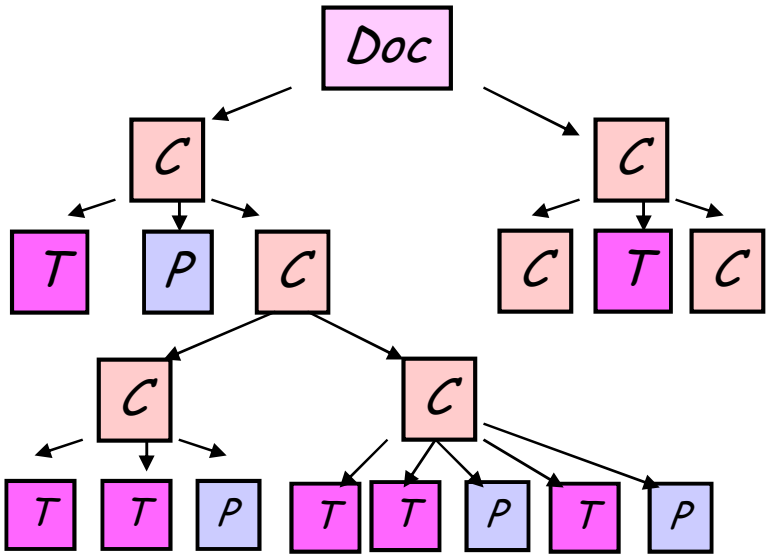
La méthode de structuration de documents que nous proposons repose sur une technique de fusion qui consiste à rassembler des informations obtenues à partir de plusieurs fixations. C'est donc par le choix des fixations successives que la description de la structure du document évoluera et aboutira à une représentation segmentée marquée par des résolutions différentes

$$G*I(i,j)=\sum_{m=-M}^{M}\sum_{n=-N}^{N}G(m,n)*I(i-m,j-n)$$

Les techniques permettant de sélectionner ces *zones informatives* seront évoquées dans la suite. Elles son

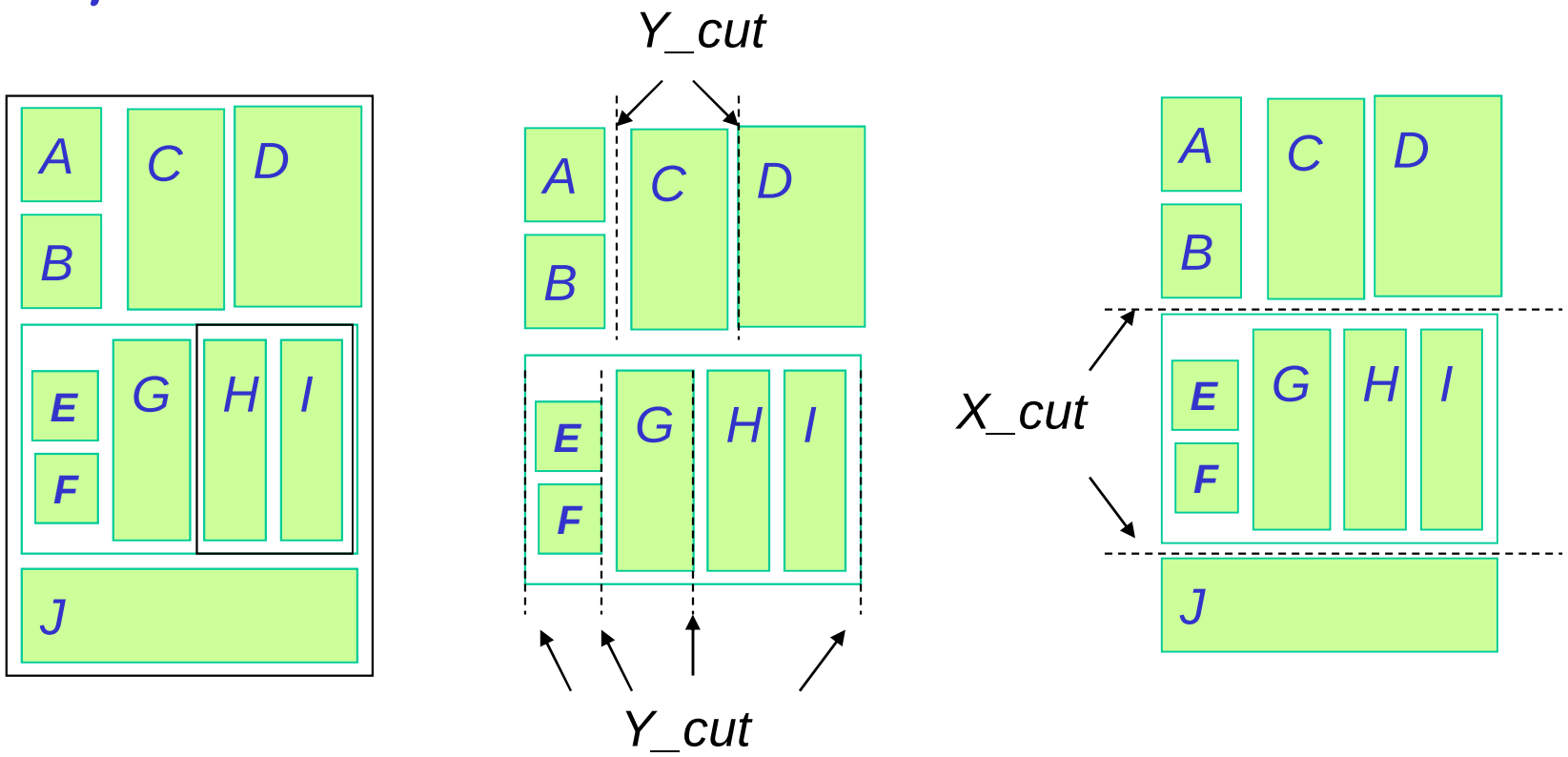
De manière générale, pour obtenir la segmentation physique du document, on a recours à de nombreuses

1. Structurer son environnement



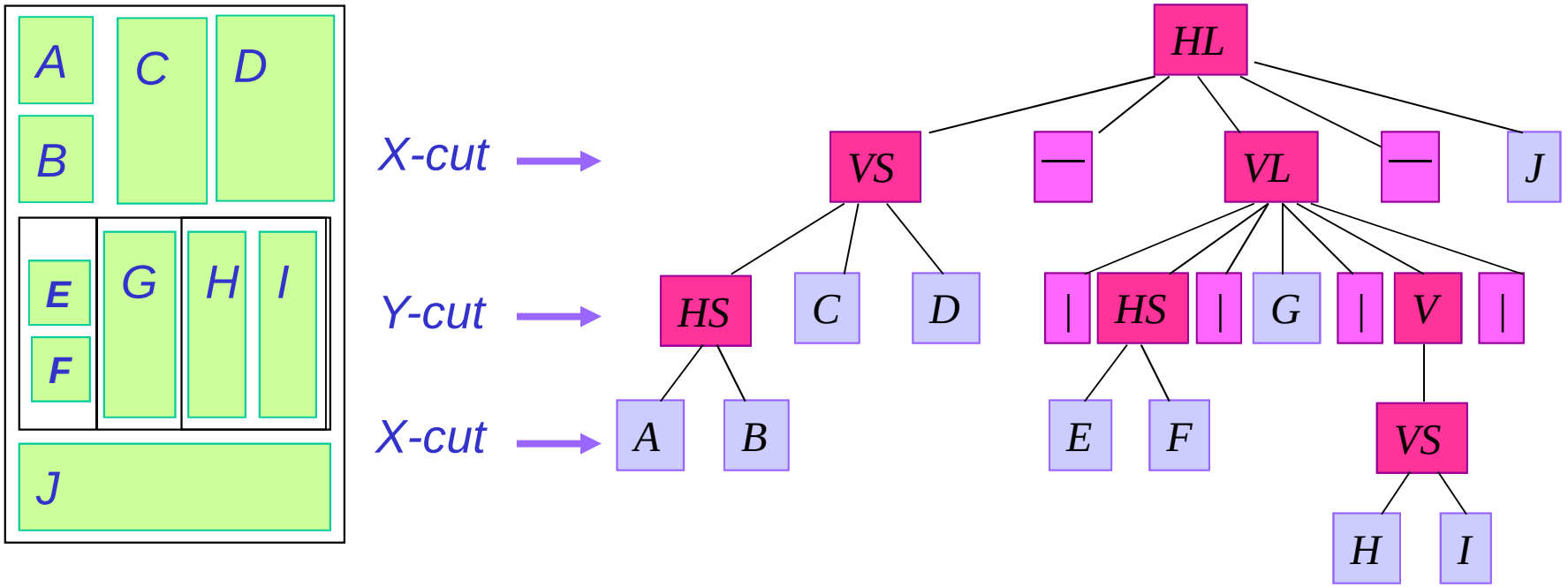
Approches descendantes (top-down)

La représentation X-Y Tree



Approches descendantes (top-down)

La représentation X-Y Tree



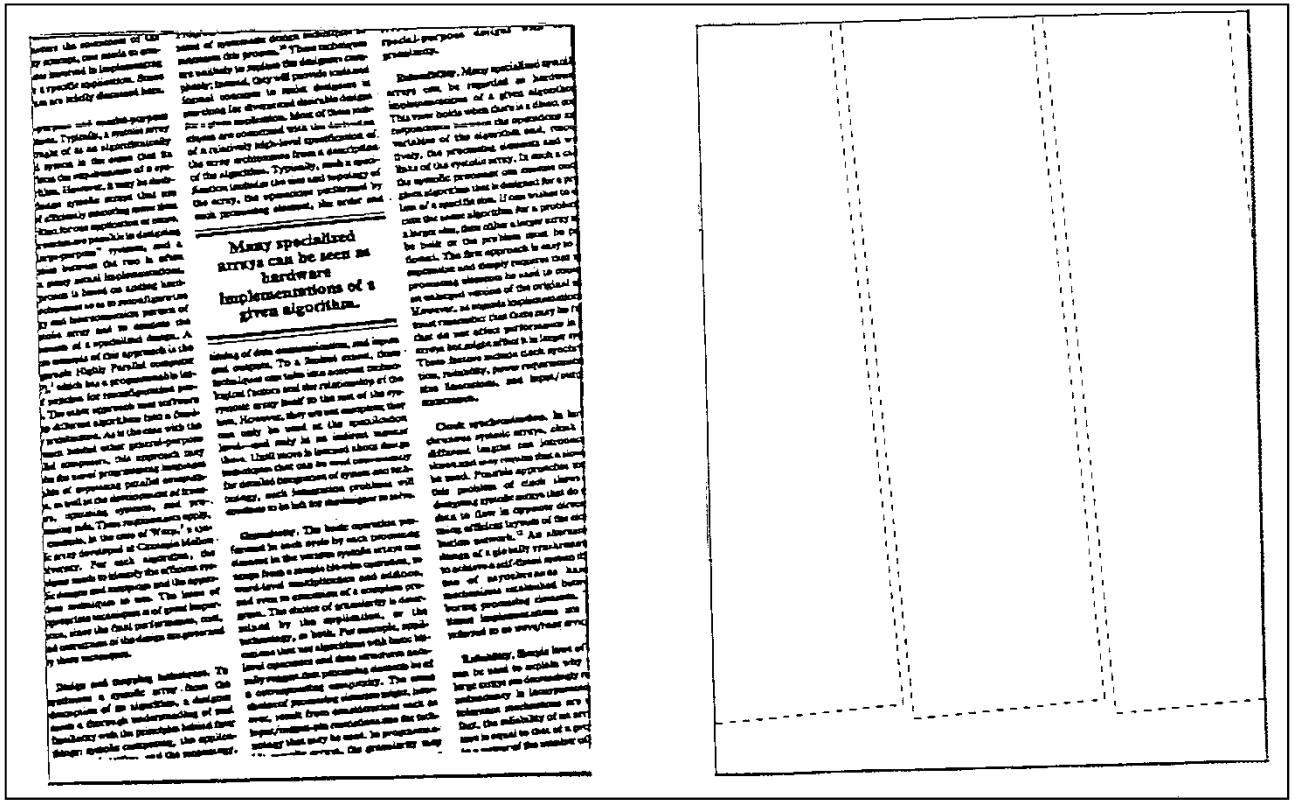
HL VL : horizontal and vertical lines

HS VS : horizontal and vertical spaces

V : vacuous cut

Approches descendantes (top-down)

L'analyse du fond par la recherche de plages blanches [Pavlidis 1991]



Approches descendantes (top-down)

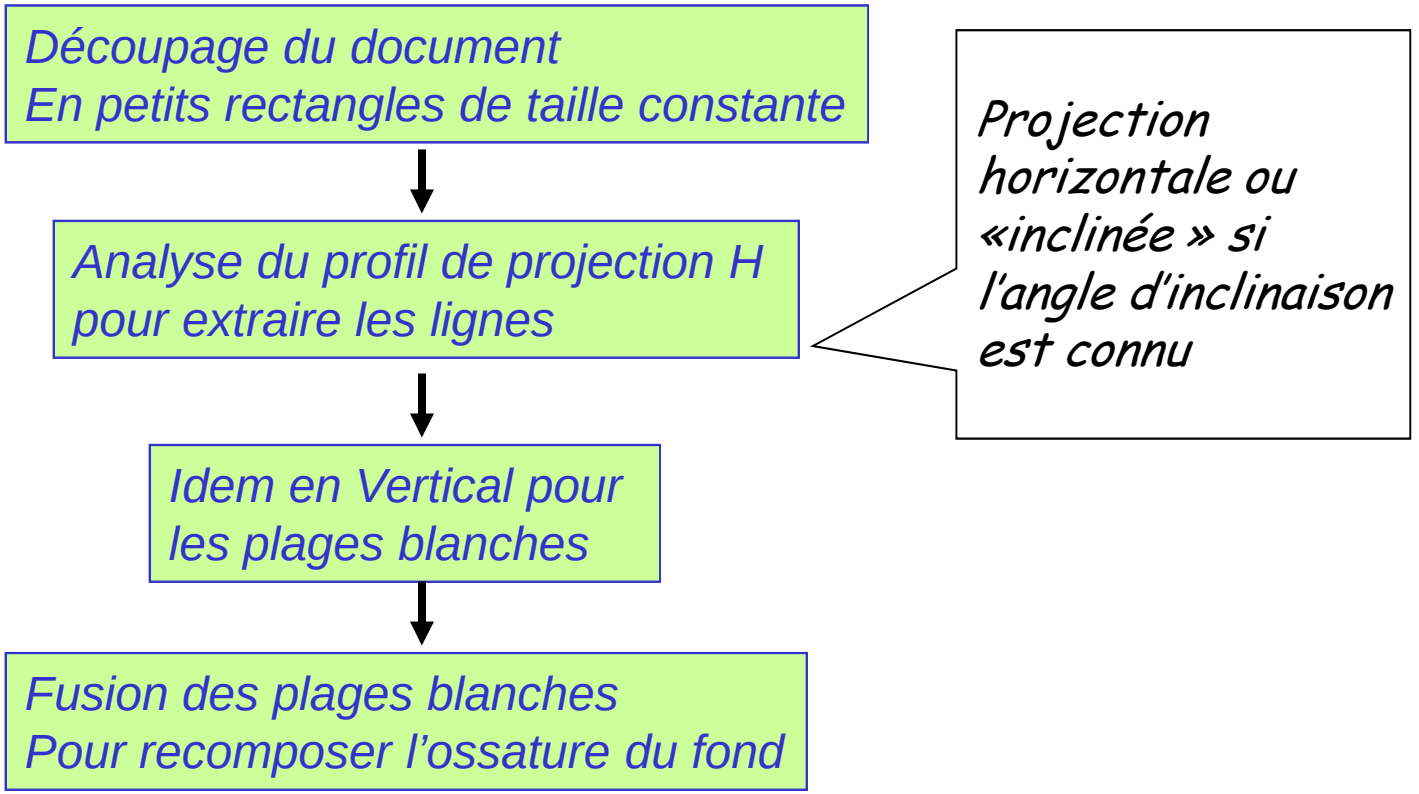
La recherche de plages blanches

- *Rechercher dans l'image des séquences homogènes de plages blanches pour récupérer la structure des blocs en ligne et en colonne.*
- *Les plages blanches constituent un tramage du fond qui permet de récupérer la structure en lignes et en colonnes.*
- *Les colonnes sont reconstruites à partir des profils « verticaux » fusionnés.*

- *Le résultat = une **partition de la page** en colonnes (séparées les unes des autres).*

Approches descendantes (top-down)

La recherche de plages blanches



Approches descendantes (top-down)

La recherche de plages blanches

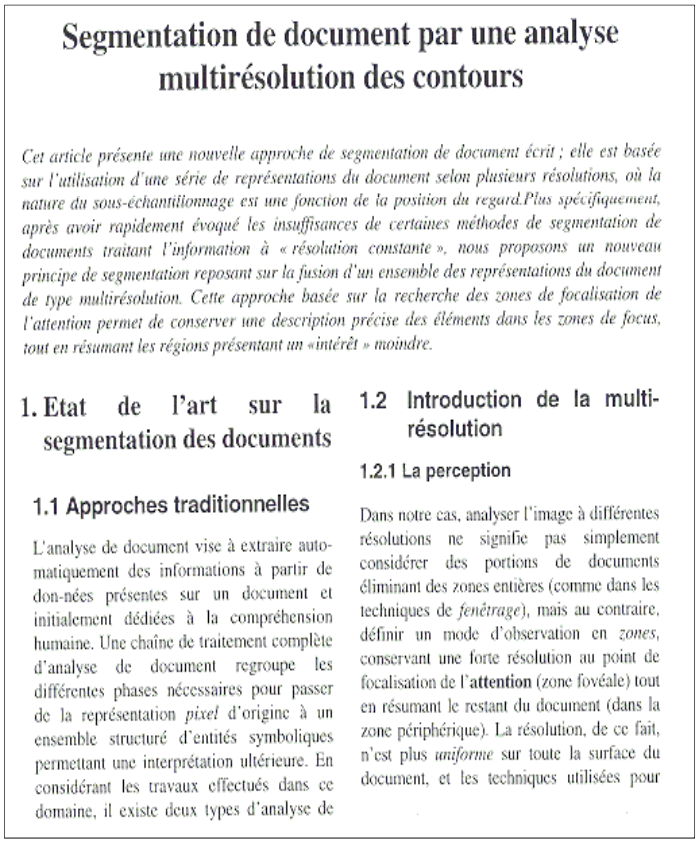
- *Bons résultats sur des images dont le texte est incliné. Connaissance a priori de l'angle de l'inclinaison.*

- *Pas des résultats toujours très informatifs : l'inclinaison des lignes génèrent un découpage en petits blocs qui n'ont pas toujours de relation entre eux.*

- *Méthode de partitionnement du fond largement reprise : segmentation par rectangulation basée sur la recherche des espaces séparateurs de texte (espaces inter-mots et inter-caractères), Baird en 1992 et Boukined en 1995.*

Approches descendantes (top-down)

Les approches multirésolution [Déforges 1997]



Segmentation de document par une analyse multirésolution des contours

Cet article présente une nouvelle approche de segmentation de document écrit ; elle est basée sur l'utilisation d'une série de représentations du document selon plusieurs résolutions, où la nature du sous-échantillonnage est une fonction de la position du regard.Plus spécifiquement, après avoir rapidement évoqué les insuffisances de certaines méthodes de segmentation de documents traitant l'information à « résolution constante », nous proposons un nouveau principe de segmentation reposant sur la fusion d'un ensemble des représentations du document de type multirésolution. Cette approche basée sur la recherche des zones de focalisation de l'attention permet de conserver une description précise des éléments dans les zones de focus, tout en résumant les régions présentant un «intérêt » moindre.

1. Etat de l'art sur la segmentation des documents

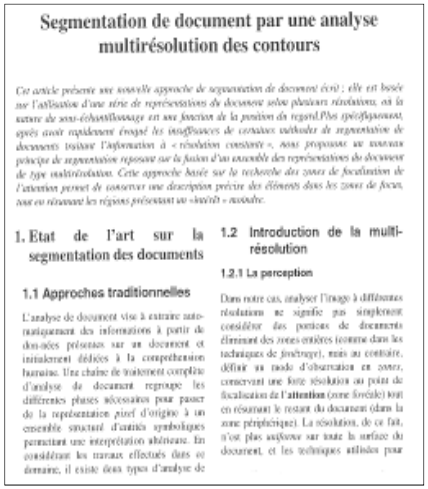
1.1 Approches traditionnelles

L'analyse de document vise à extraire automatiquement des informations à partir de données présentes sur un document et initialement dédiées à la compréhension humaine. Une chaîne de traitement complète d'analyse de document regroupe les différentes phases nécessaires pour passer de la représentation *pixel* d'origine à un ensemble structuré d'entités symboliques permettant une interprétation ultérieure. En considérant les travaux effectués dans ce domaine, il existe deux types d'analyse de

1.2 Introduction de la multi-résolution

1.2.1 La perception

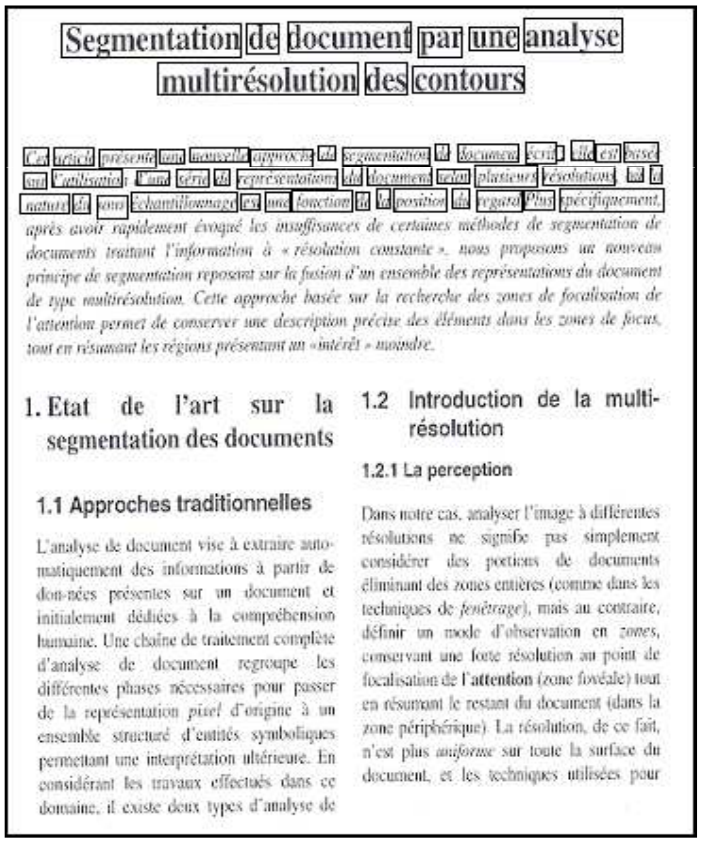
Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents éliminant des zones entières (comme dans les techniques de *fenêtrage*), mais au contraire, définir un mode d'observation en *zones*, conservant une forte résolution au point de focalisation de l'**attention** (zone fovéale) tout en résumant le restant du document (dans la zone périphérique). La résolution, de ce fait, n'est plus *uniforme* sur toute la surface du document, et les techniques utilisées pour



Approches descendantes (top-down)

Les approches multirésolution

Haute et basse résolution :



Segmentation de document par une analyse multirésolution des contours

Cet article présente une nouvelle approche de segmentation de document écrit ; elle est basée sur l'utilisation d'une série de représentations du document selon plusieurs résolutions, où la notion du sous-échantillonnage est une fonction de la position du regard. Plus spécifiquement, après avoir rapidement évoqué les insuffisances de certaines méthodes de segmentation de documents traitant l'information à « résolution constante », nous proposons un nouveau principe de segmentation reposant sur la fusion d'un ensemble des représentations du document de type multirésolution. Cette approche basée sur la recherche des zones de focalisation de l'attention permet de conserver une description précise des éléments dans les zones de focus, tout en résumant les régions présentant un «intérêt » moindre.

1. Etat de l'art sur la segmentation des documents

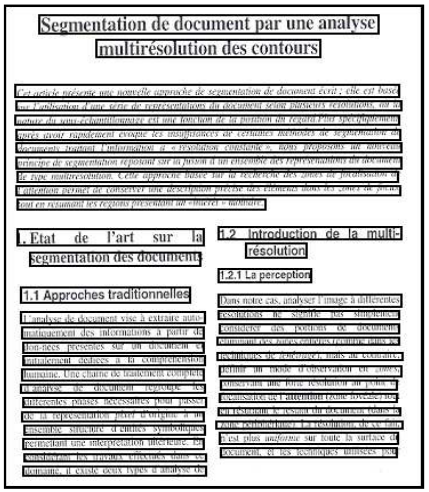
1.1 Approches traditionnelles

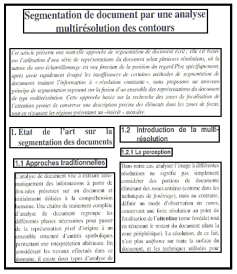
L'analyse de document vise à extraire automatiquement des informations à partir de données présentes sur un document et initialement dédiées à la compréhension humaine. Une chaîne de traitement complète d'analyse de document regroupe les différentes phases nécessaires pour passer de la représentation *pixel* d'origine à un ensemble structuré d'entités symboliques permettant une interprétation ultérieure. En considérant les travaux effectués dans ce domaine, il existe deux types d'analyse de

1.2 Introduction de la multi-résolution

1.2.1 La perception

Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents éliminant des zones entières (comme dans les techniques de *fenêtrage*), mais au contraire, définir un mode d'observation en *zones*, conservant une forte résolution au point de focalisation de l'**attention** (zone fovéale) tout en résumant le restant du document (dans la zone périphérique). La résolution, de ce fait, n'est plus *uniforme* sur toute la surface du document, et les techniques utilisées pour





⇒ *diminution des détails*

⇒ *fusion des composantes connexes en mots*

⇒ *des mots en lignes*

⇒ *des lignes en blocs*

Approches descendantes (top-down)

Bilan des méthodes top-down

- *Méthodes sont assez rapides car traitent les données dans leur* ***globalité.***

- *Elles ne nécessitent* ***pas une très grande précision*** *de résolution (résolution liée à la précision de la numérisation) : # des techniques d'OCR qui traitent les images à plus de 300 dpi.*

- *Nécessitent de* ***connaître la structure a priori*** *du document (documents très hiérarchisés) : environnement fortement normalisé, pas envisageable sur des blocs polygonaux.*

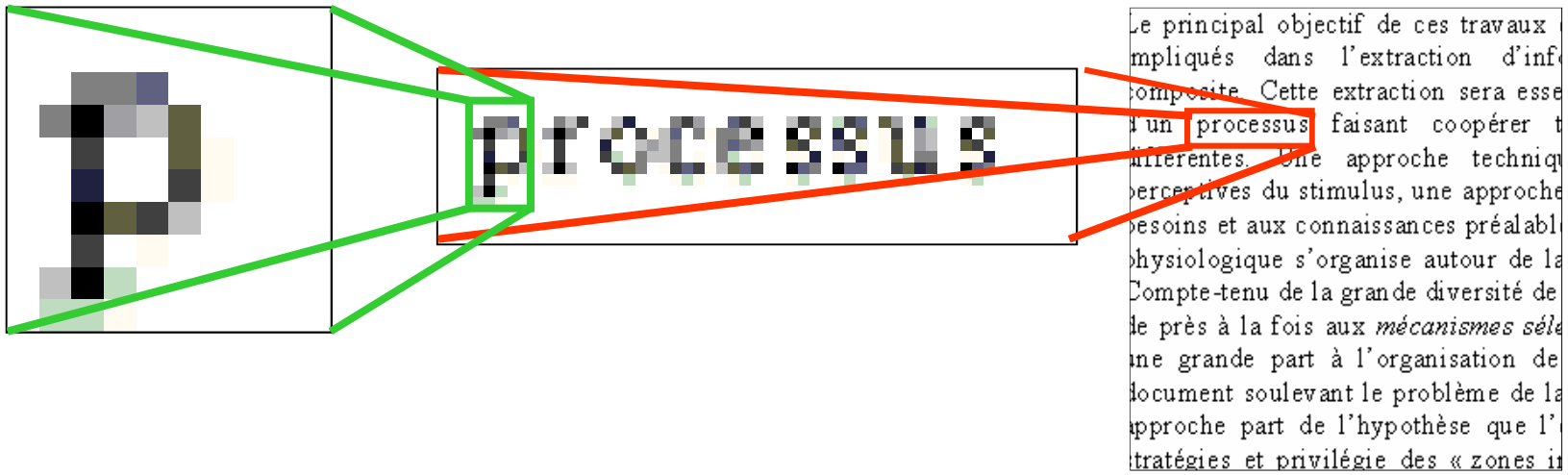
Les approches ascendantes (bottom-up)

Approches ascendantes (bottom-up)

Fondées sur l'analyse de composantes connexes

Principe : fusionner les morceaux jusqu'à l'assemblage complet de la page du document.

Exemple :fusionner les caractères en mots, puis les mots en lignes et enfin les lignes en paragraphes de texte ou en colonnes.



Des méthodes pour une segmentation physique

Méthodes de segmentation dirigée par les données de l'image

Caractères = tous les objets de taille moyenne régulièrement alignés

Image originale

Segmentation des objets

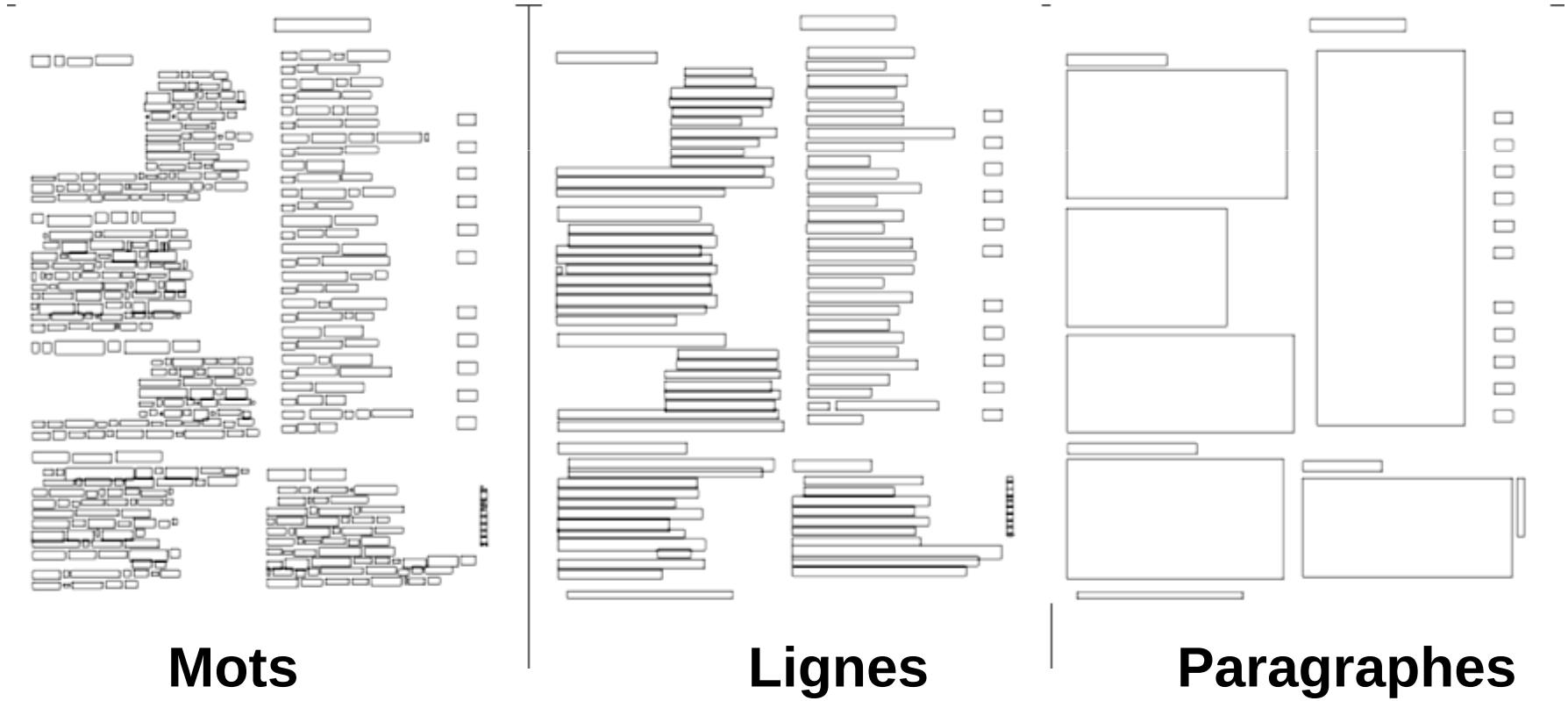
Regroupement des objets en caractères

Des méthodes pour une segmentation physique

Mots = Regroupements de caractères faiblement espacés

Lignes = Regroupement de mots adjacents faiblement espacés

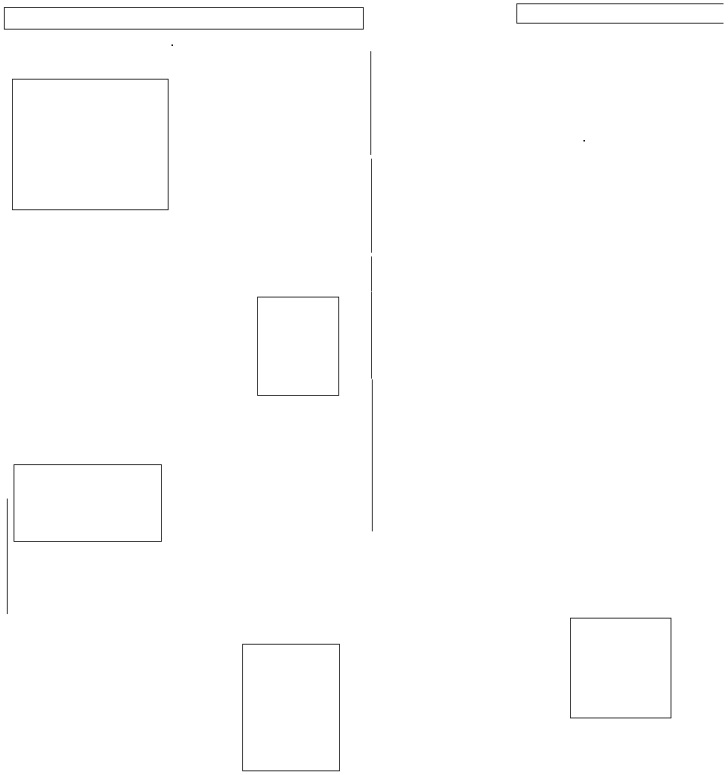
Paragraphes = Regroupement de lignes de faible interligne



Mots **Lignes** **Paragraphes**

Des méthodes pour une segmentation physique

Illustrations = Objets non classés comme du texte



Illustrations

SOMMAIRE

Fini et sans limites

Héritier du terme grec *apeiron* (de *peras* qui signifie limite), le mot latin *infinitum* et son dérivé français « in-fini » ont longtemps été considérés comme le synonyme d'« illimité ». On peut cependant envisager, sans contradiction aucune, des espaces finis ne possédant aucune limite. C'est notamment le cas de la surface finie d'une sphère ou d'un tore (espace à deux dimensions) qui ne comporte aucune frontière. *(La clôture du monde. Page 10.)*

Le passage du fini à l'infini

En arithmétique, la démonstration d'un théorème général, applicable à tous les nombres, requiert en principe une infinité de vérifications : montrer que le théorème est vrai pour 1, 2, 3 et ainsi de suite. Aussi loin que nous allions, nous ne parviendrons jamais à établir la vérité du théorème général. Seule la méthode dite de la récurrence permet, selon le mot du mathématicien Poincaré, de « passer du fini à l'infini ». *(Un, deux, trois... l'infini ! Page 70.)*

A la conquête de l'espace infini

Pour la première fois dans l'histoire de l'art, les peintres de la Renaissance italienne donnent une représentation figurée de l'infini sous la forme de « ligne d'horizon » ou de « point de fuite ». L'infini se voit ainsi accorder sur les tableaux une trace aussi réelle que n'importe quel autre point de l'espace. *(L'infini en perspective. Page 32.)*

L'infini comme méthode

Loin de toujours compliquer la vie du physicien, l'infini peut au contraire la faciliter. En physique statistique, par exemple, c'est seulement par le biais d'un passage à l'infini, soit en faisant tendre le nombre de constituants élémentaires (atomes, molécules) d'un système vers l'infini, que le physicien parvient à retrouver les discontinuités de certaines propriétés physiques (densité, capacité calorifique) qui caractérisent un changement d'état, par exemple la transformation de l'eau en glace. *(L'infini en pratique. Page 40.)*

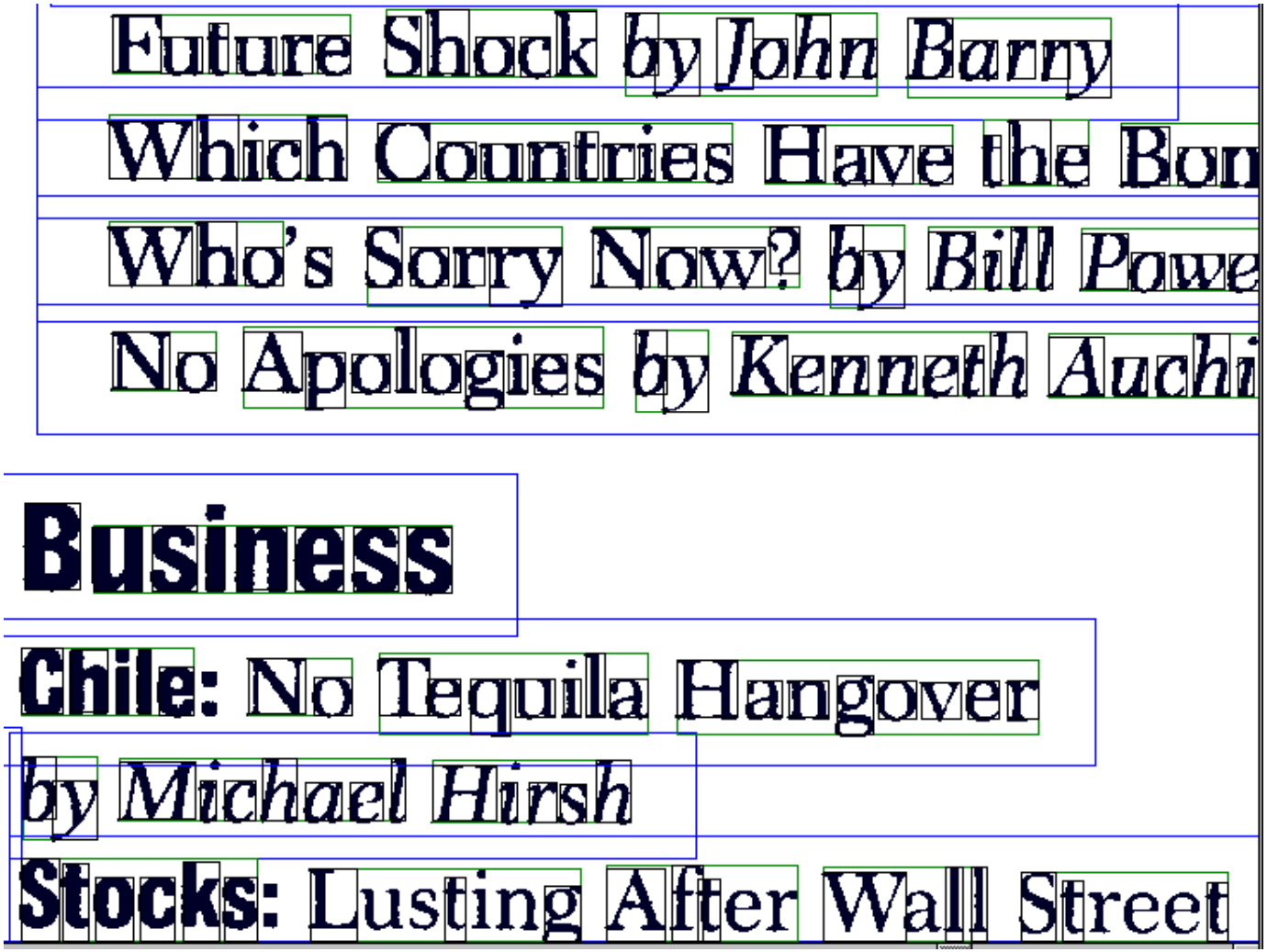
Illusion d'infini

L'étude des « variantes » topologiques d'espace à trois dimensions susceptibles de permettre la construction de modèles pertinents du monde réel révèle la possibilité de multiples illusions d'infinis. Si nous vivions par exemple dans un espace hyperbolique compact, nous aurions l'impression d'être dans un espace cellulaire pavé à l'infini par des dodécaèdres déformés par des illusions d'optique. *(Notre univers est-il chiffonné ? Page 12.)*

Superposition avec l'image

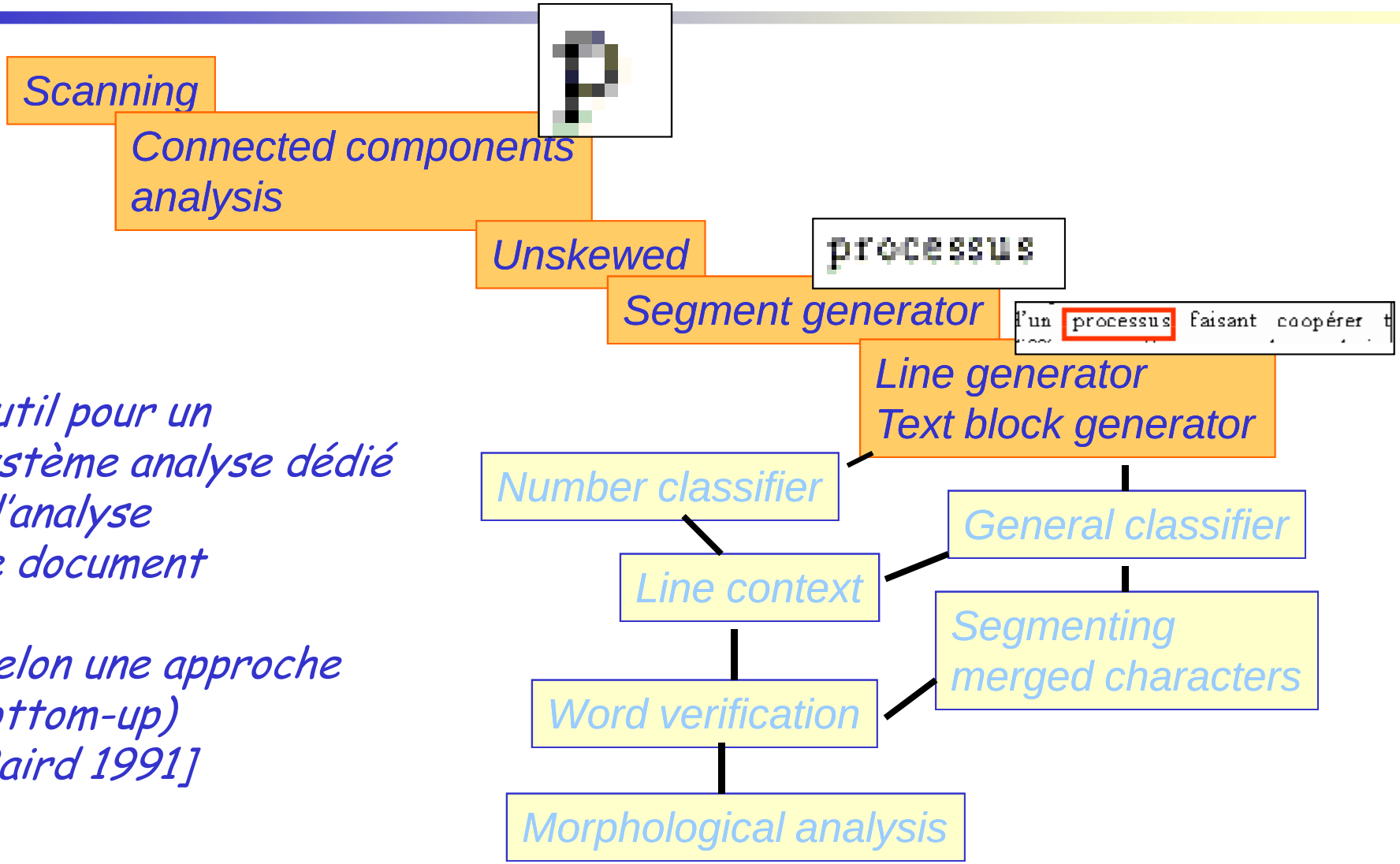
Structure physique hiérarchique du texte

(Colonnes>paragraphes>Lignes>mots>caractères>accents et ponctuation…)



Résultat de la Segmentation de la structure physique

Approches ascendantes (bottom-up)



Outil pour un système analyse dédié à l'analyse de document

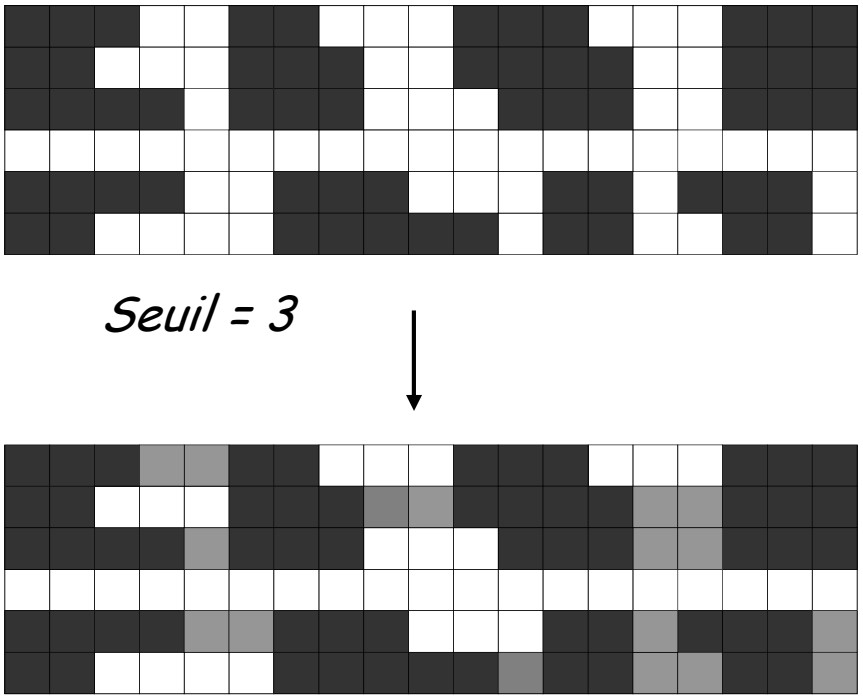
(selon une approche bottom-up) [Baird 1991]

Approches ascendantes (bottom-up)

La technique de lissage directionnel

La technique du RLSA

(Run-Length Smoothing Algorithm, Wong 82)



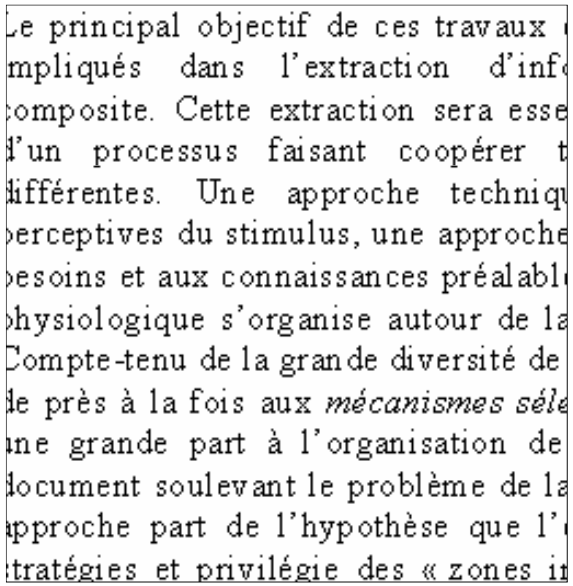
Pixel objet

Pixel bouché

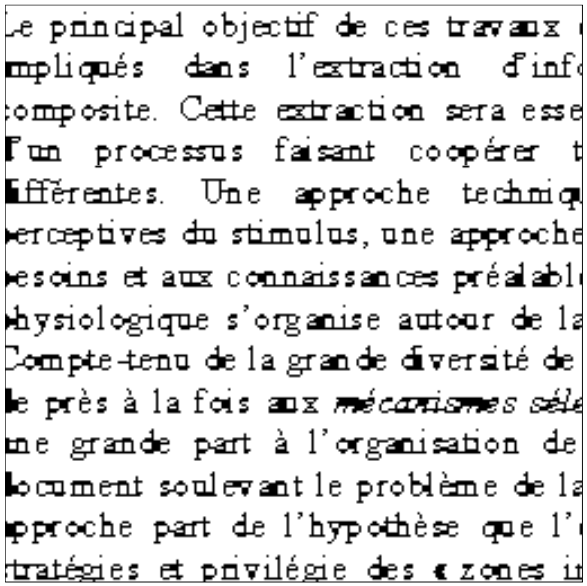
Pixel blanc

Approches ascendantes (bottom-up)

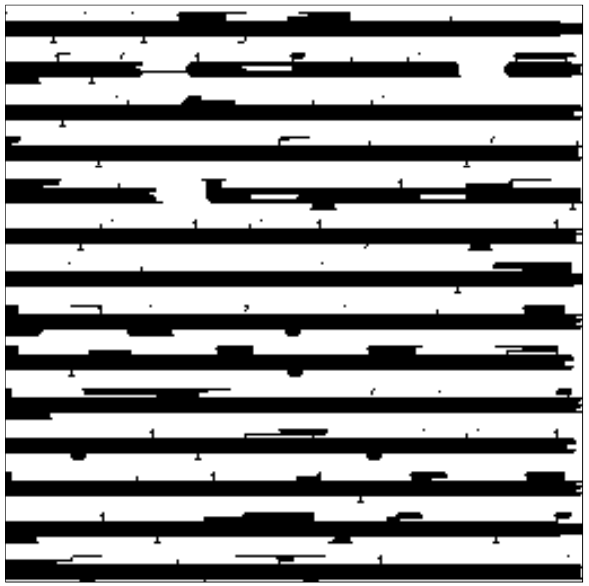
La technique du RLSA



Image



Seuil =3



Seuil =20

Approches ascendantes (bottom-up)

La technique du RLSA


DES CINÉASTES EN COLÈRE

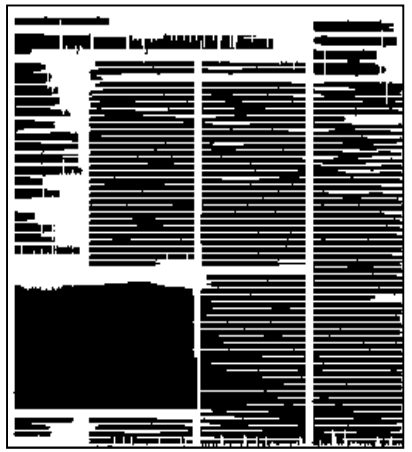
Ségolène Royal contre les professionnels du cinéma

Bertrand Tavernier : « Nous ne serons pas les sans-papiers du service public »

Bouchage Horizontal

Bouchage Vertical

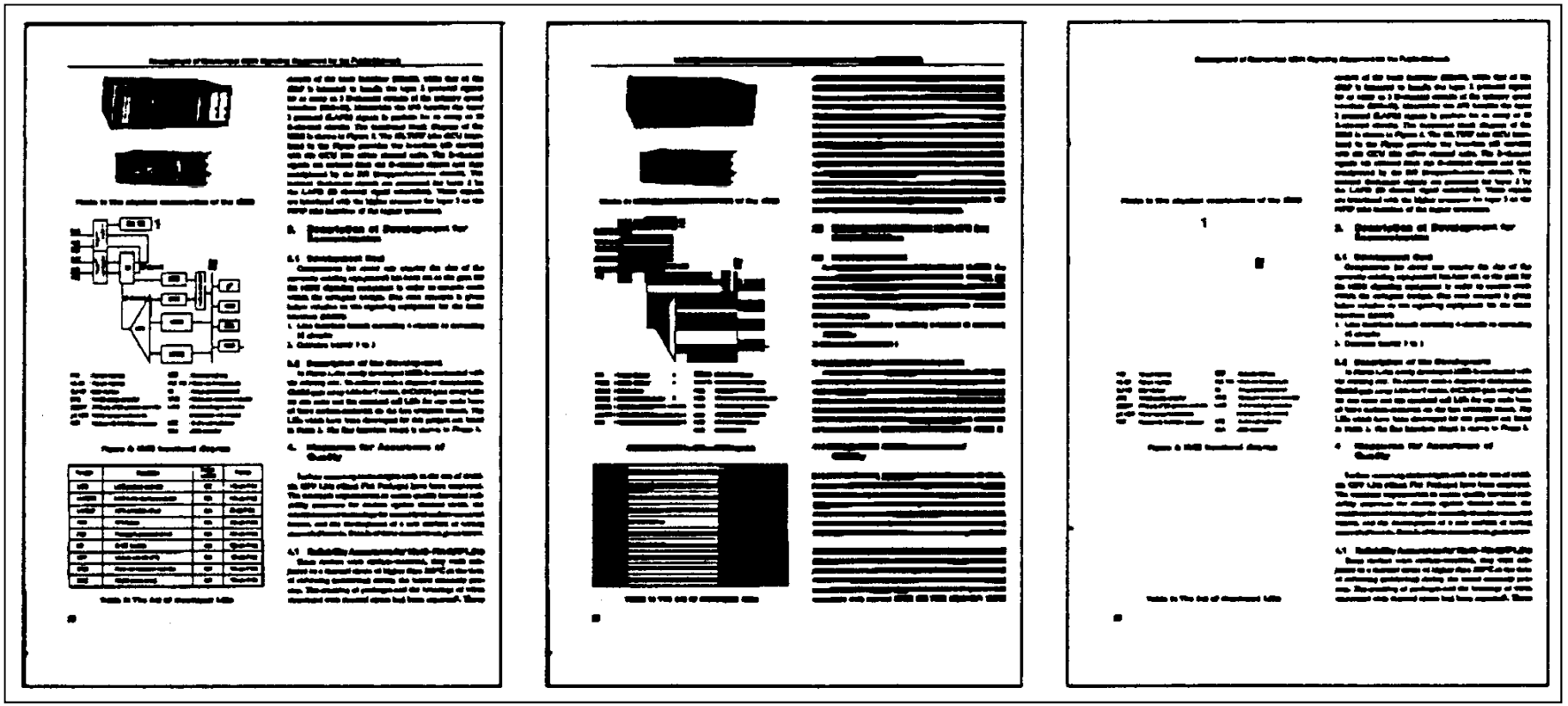
ET logique



Seuil = 20 pixels

Approches ascendantes (bottom-up)

Les variantes à la technique du RLSA



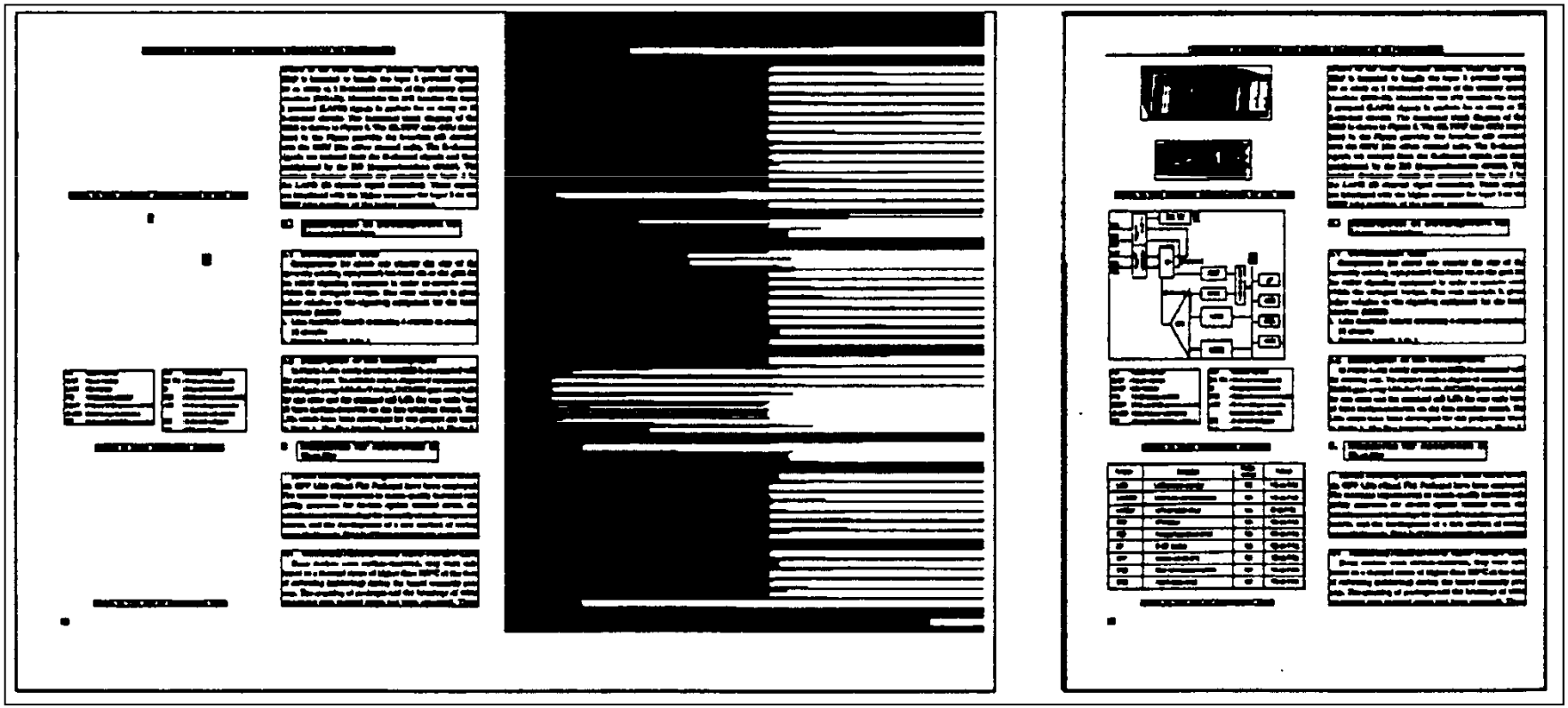
1. Image échantillonnée

2.Image des blocs

3. Image des caractères

Approches ascendantes (bottom-up)

Les variantes à la technique du RLSA

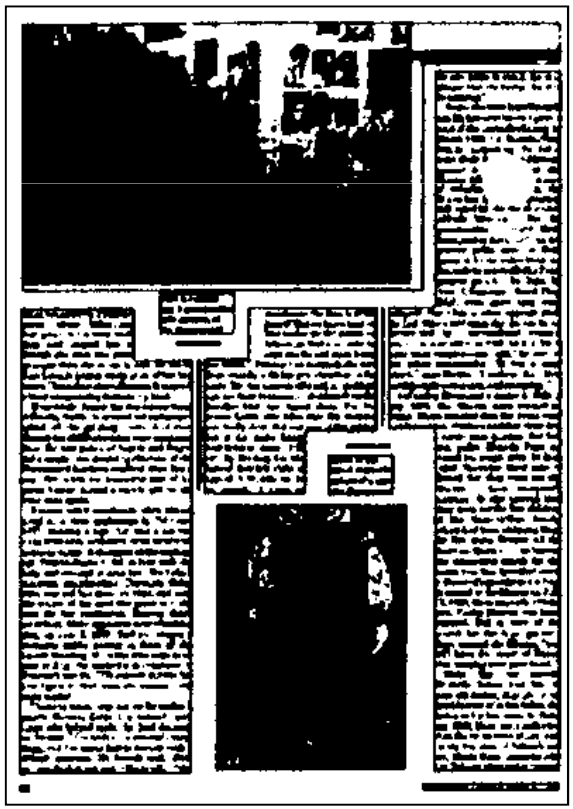


4. Profils de projection

5. Résultats

Approches ascendantes (bottom-up)

Suivi de segments pour l'extraction de blocs polygonaux



Approches ascendantes (bottom-up)

Bilan

+++

- *Très grande variété des approches*
- *Méthodes plus spécifiques répondant mieux à un problème particulier.*
- *Plus grande souplesse : permet en particulier de traiter des documents fortement inclinés*

- *Connaissance a priori des typographies utilisées (travail sur les cc)*
- *Une très grande précision dans la résolution des images est requise : manipulation de gros volumes de données.*

Les approches mixtes (à la fois ascendantes et descendantes) constituent de nouvelles pistes.

Approches mixtes

Méthodes hybrides (combinaison de méthodes mi-ascendantes, mi-descendantes) plus efficaces pour l'analyse du fond (par les espaces blancs). Plus grande robustesse des résultats

Étape descendante

Recherche de séparateurs
Entre les paragraphes
(plages blanches ou lignes noires)

Extraction des régions

Étape ascendante

Fusion des composantes connexes
dans la direction orthogonale
aux marqueurs du 1.

Mise en évidence des régions de texte
Des images et des graphiques

Exemple de méthode (Okamoto 1993) :

Approches mixtes

A Timetable: Out by Dec. 31

Résultats de la segmentation par la méthode d'Okamoto dans [Ok93].

Conclusion sur la reconnaissance des structures

Des travaux émergents sur les documents anciens

Fonctionne sur des documents bien définis dont les structures sont cohérentes et régulières (journaux, revues..)

Systèmes réalisés et paramétrés en laboratoire pour un fonctionnement limité à certains documents

Les systèmes à apprentissage apportent un peu de souplesse au détriment d'un travail supplémentaire

Difficile de transférer la technologie aux sociétés privées.

La Recherche avance.. doucement, les entreprises privés et les usagers attendent …

APPLICATION DE LA SEGMENTATION

1. Segmentation par PROJECTIONS récursives

The plane formed by $\tilde{v}_{i,j}$ and the focal point of the camera must include $\hat{v}_{i,j}$. Let this plane be designated by its normal $n_{i,j}$.

$$n_{i,j} = \tilde{p}_{i,j} \times \tilde{p}_{i,j+1} \quad (1)$$

Since $n_{i,j}$ is perpendicular to $\hat{v}_{i,j}$

$$n_{i,j} \cdot \hat{v}_{i,j} = 0 \quad (2)$$

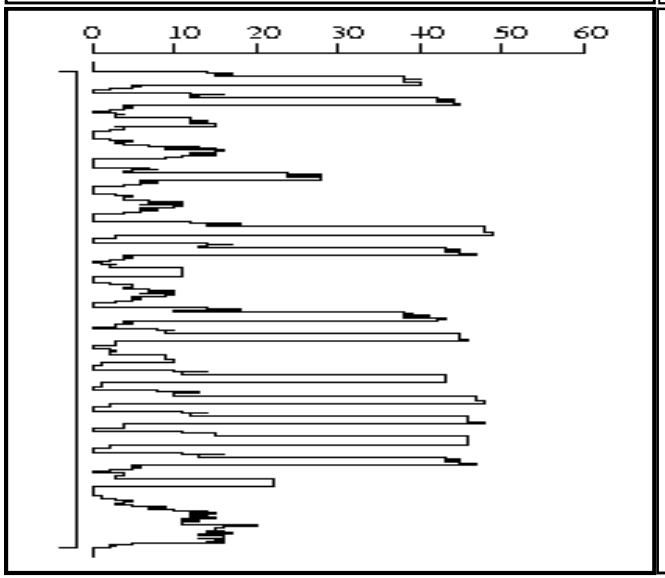
In the case of purely translational motion, the direction of $\hat{v}_{i,j}$ is constant for all i. Therefore, Equation 2 can be rewritten as

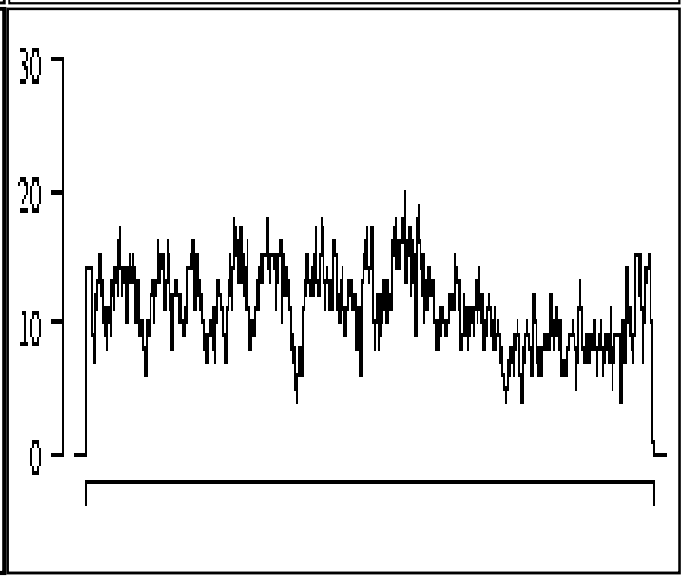
$$n_{i,j} \cdot \hat{v}_j = 0 \quad (3)$$

where $\hat{v}_j = \hat{v}_{i,j}$ for all i. This equation is linear with three unknowns, and can be solved using a least squares technique.

An error measure is used to evaluate the validity of the local translation approximation. The error measure we use is the average, taken over the local neighborhood, of the angle between each flow vector plane and the local translation. Using the normals $n_{i,j}$ from Equation 1, the error measure is defined as

$$\frac{1}{m}\sum_{i=1}^{m} \left| \sin^{-1}\left(\frac{n_{i,j} \cdot \hat{v}_j}{\|n_{i,j}\| \|\hat{v}_j\|} \right) \right| \quad (4)$$





2. Segmentation par BINARISATION

Binarisation adaptative dans chaque ligne

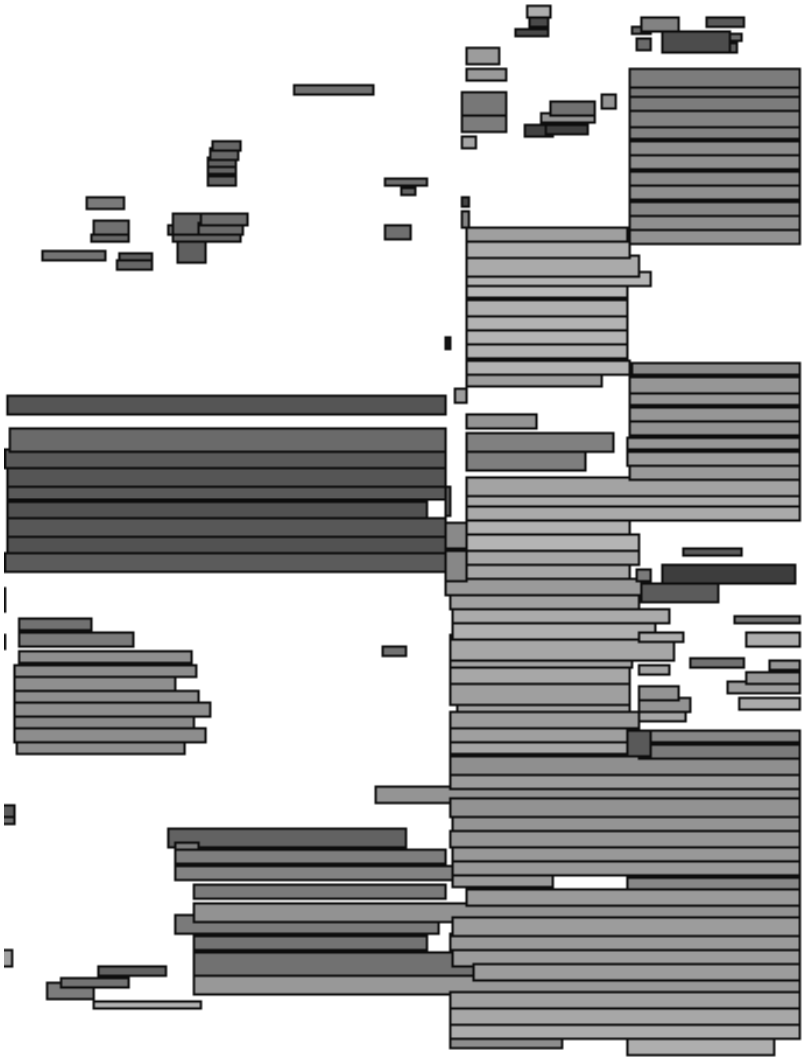
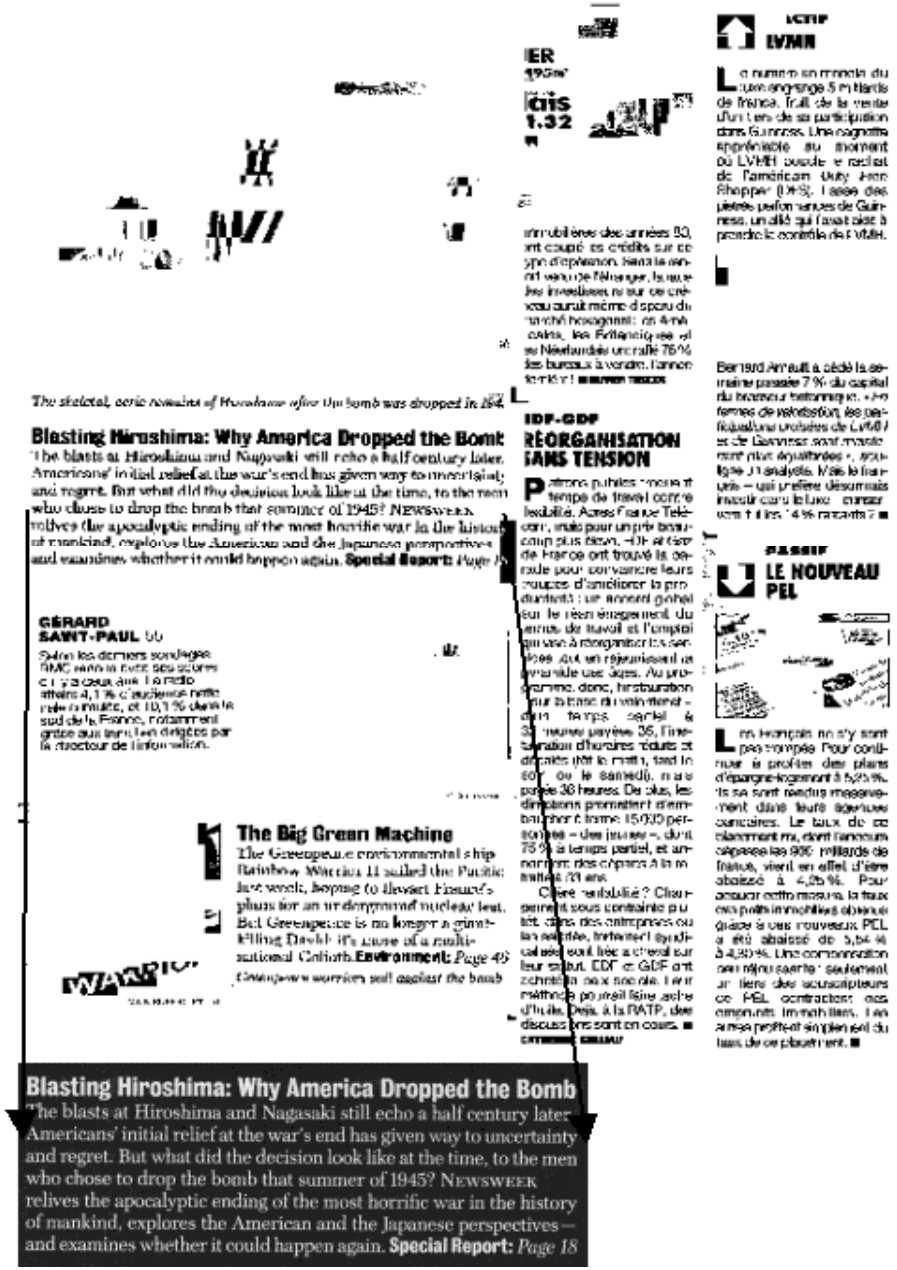
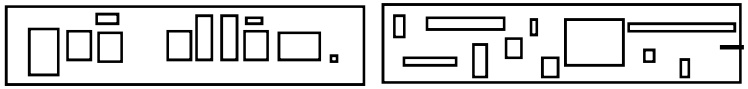


Image binaire obtenue



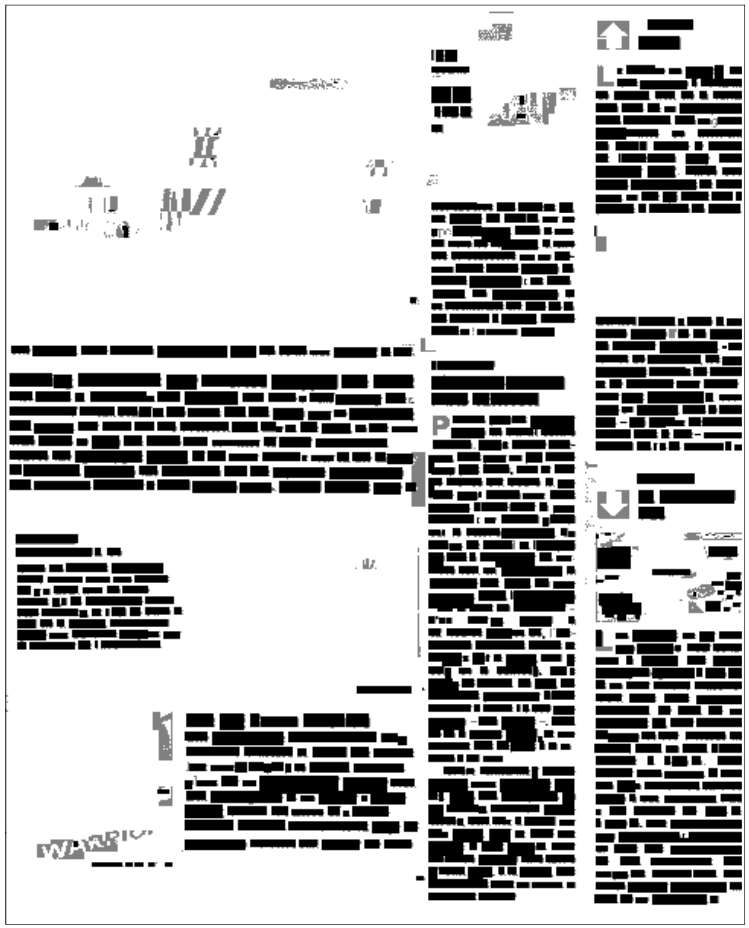
2. Segmentation par BINARISATION

Séparation texte/graphique : cohérence des alignements



Texte **graphique**

Texte/graphique



Zones graphiques rejetées

Zones encore reconnues comme texte

Résultat de l'analyse automatique

3. Séparation texte/graphique par texture

Par filtrage horizontal et morphologie mathématique sur l'image binaire

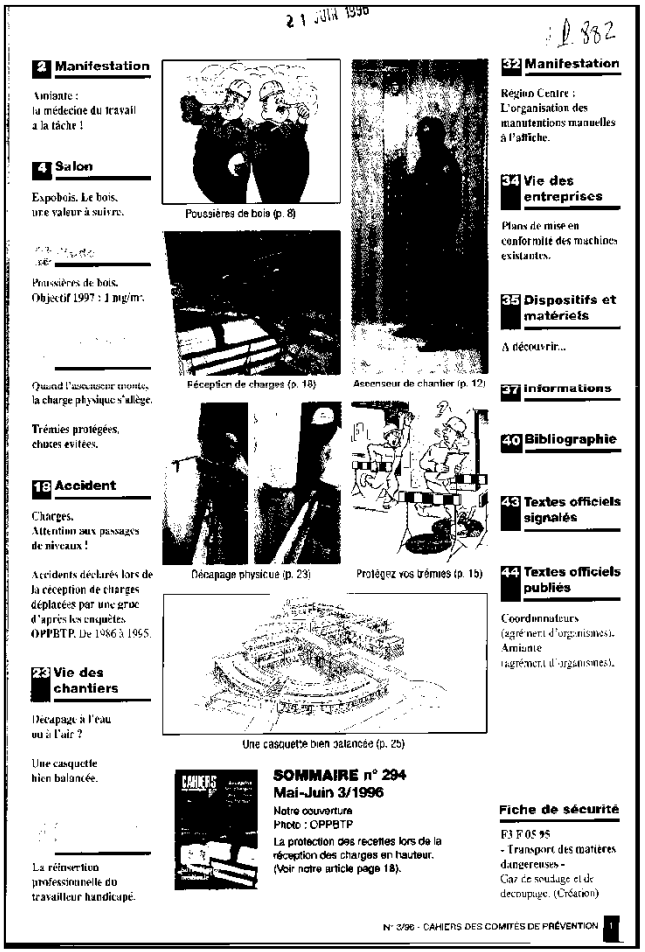
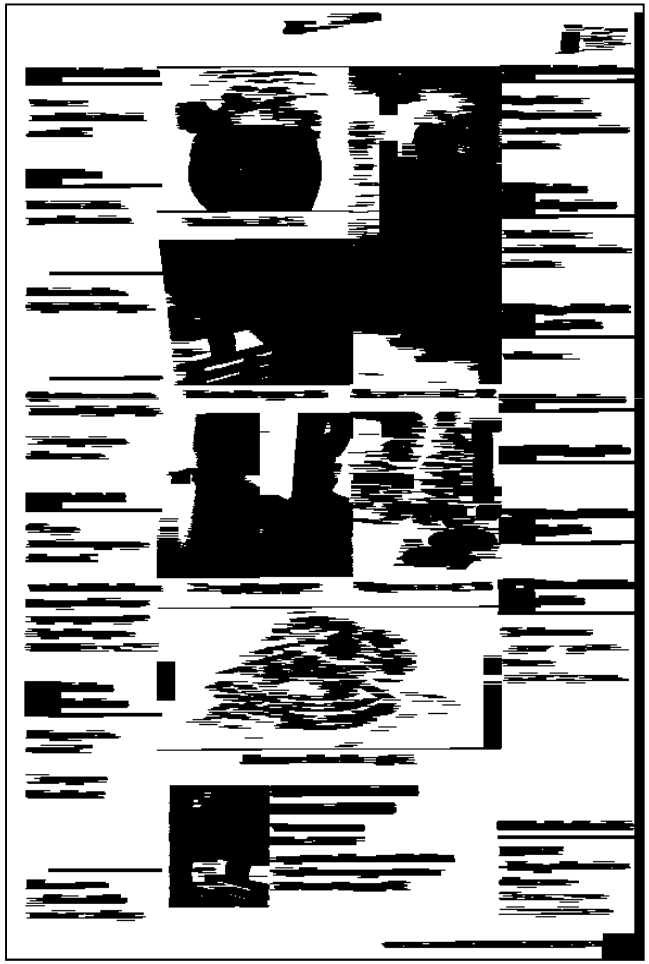
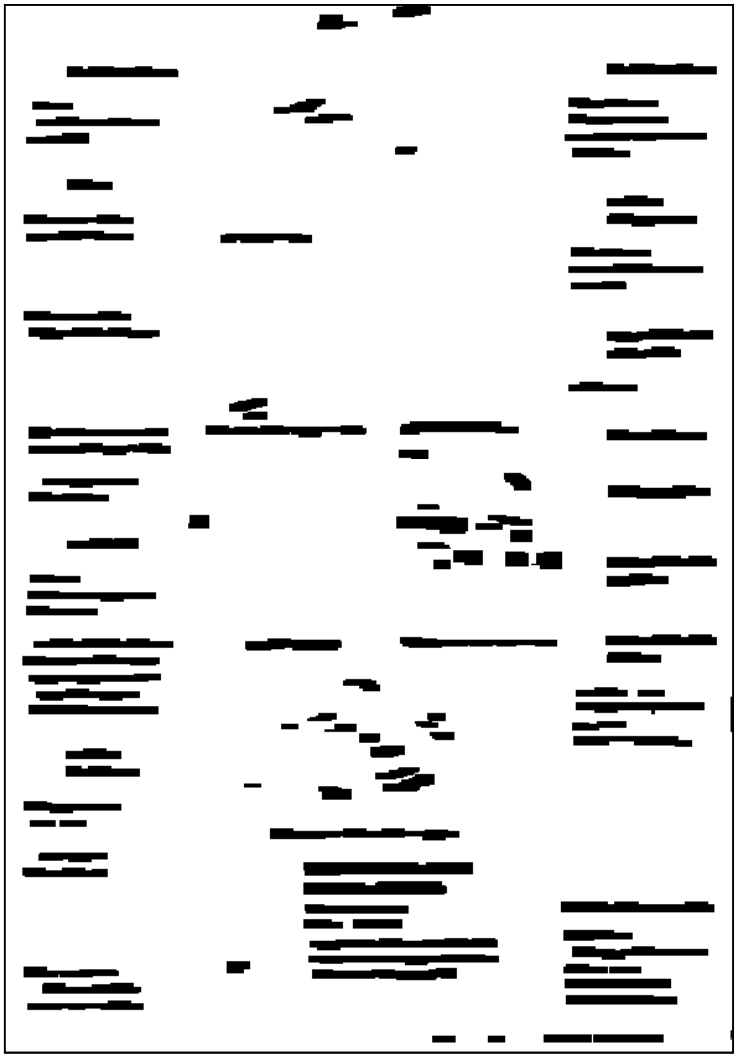


Image binaire



Résultat d'un filtrage horizontal

3. Séparation texte/graphique par texture



Détection des alignements par morphologie

Résultat de la segmentation

3. Séparation texte/graphique par texture

Par filtrage des variations d'intensité lumineuse de l'image (nuances de gris)

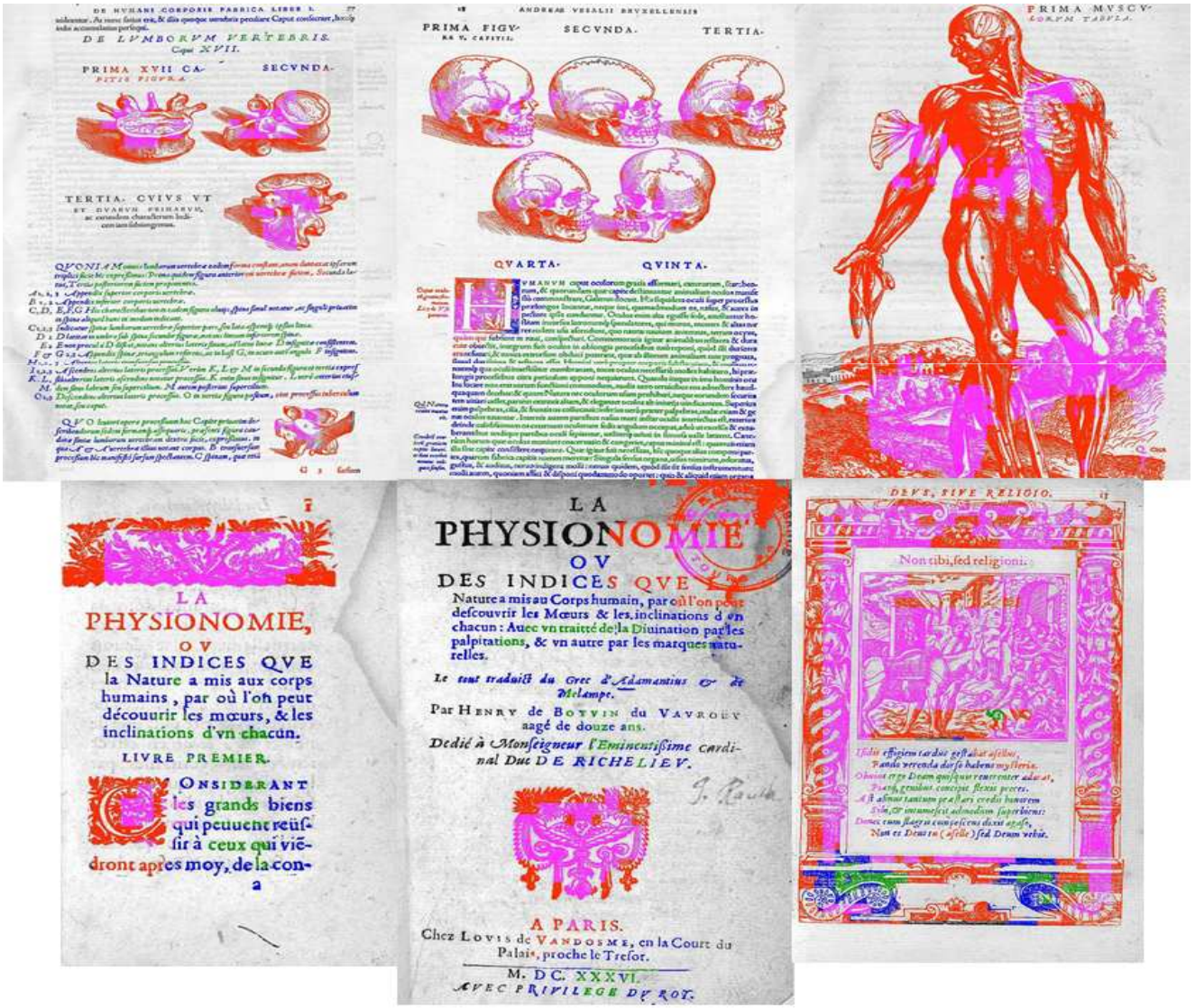
- Faible complexité de calcul, vitesse élevée de traitement (10'')
- La taille du voisinage est définie par l'échelle d'analyse.
- Le résultat du filtre est normalisé par la surface du voisinage puis comparé à une seuil global.

3. Séparation texte/graphique par texture

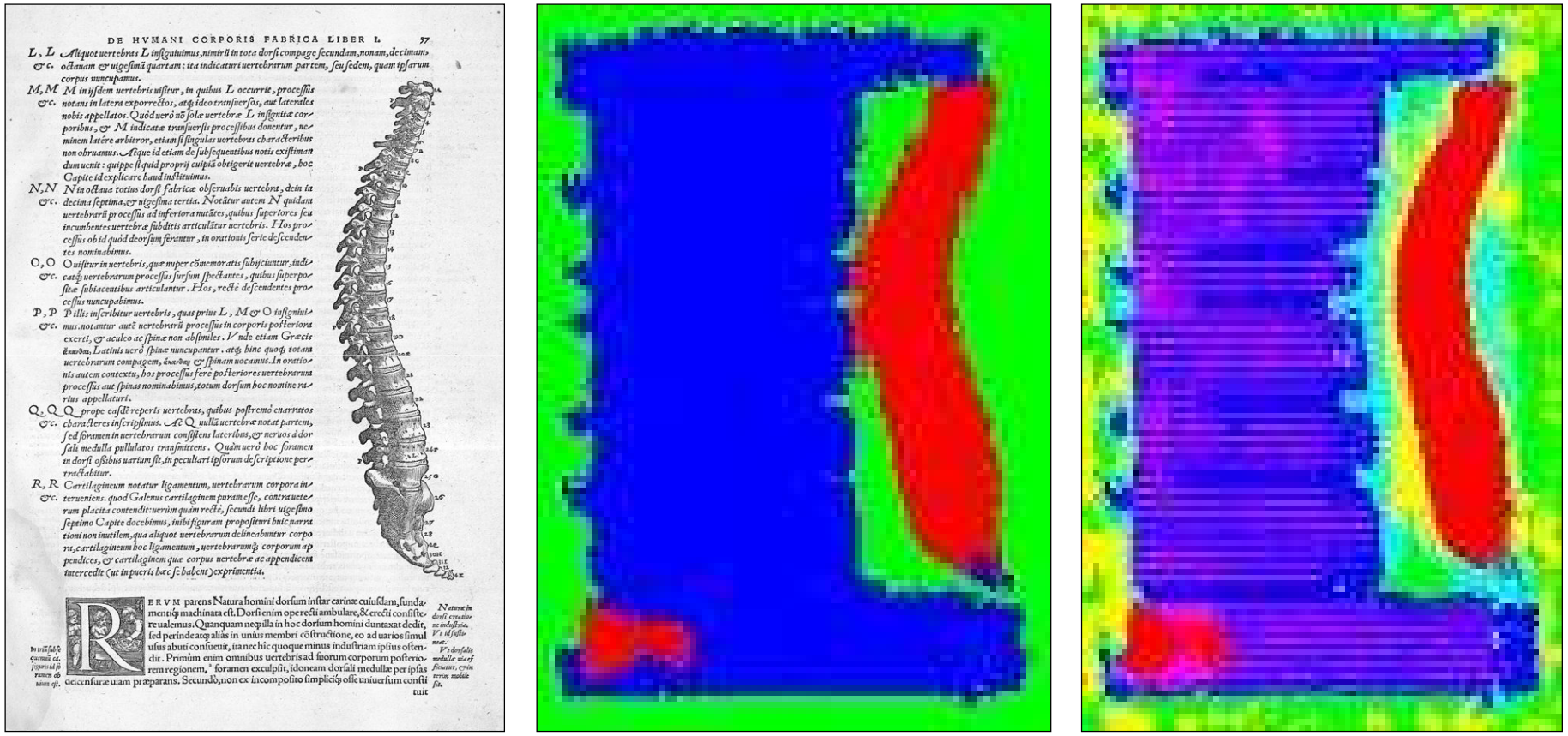
Par une analyse systématique sans a priori
(nuances de gris)

Détection et interprétation des orientations

(thèse Nicholas Journet – La Rochelle 2006)

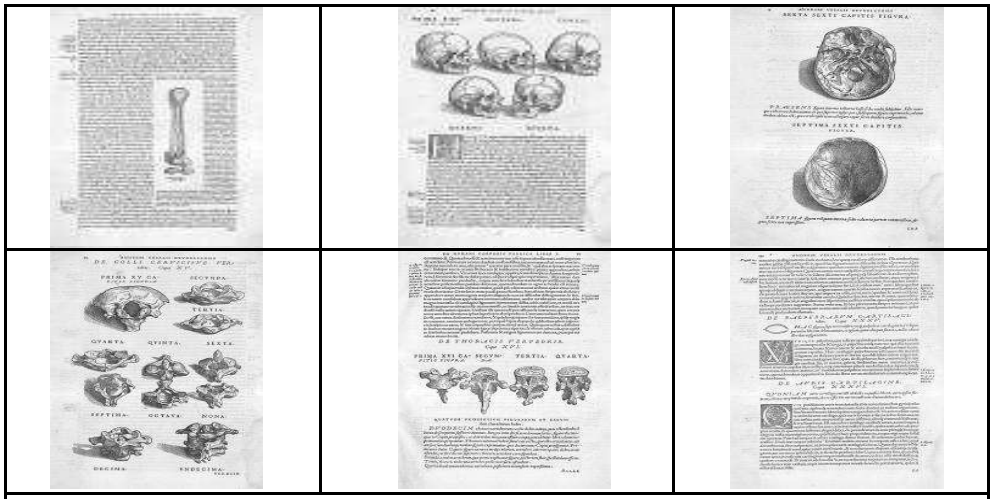


3. Séparation texte/graphique par texture

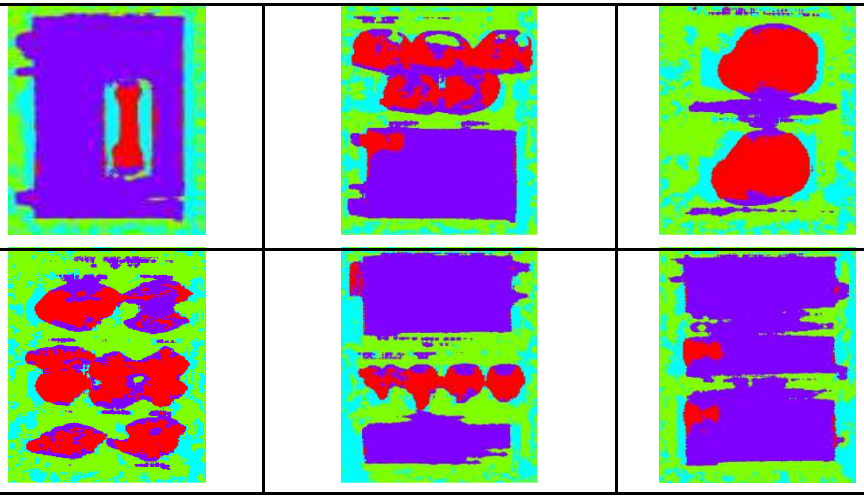


Détection et interprétation de critères de texture par classification: orientations, densité, complexité - Nicholas Journet – La Rochelle 2006

3. Séparation texte/graphique par texture



Tested pages (from the same book)



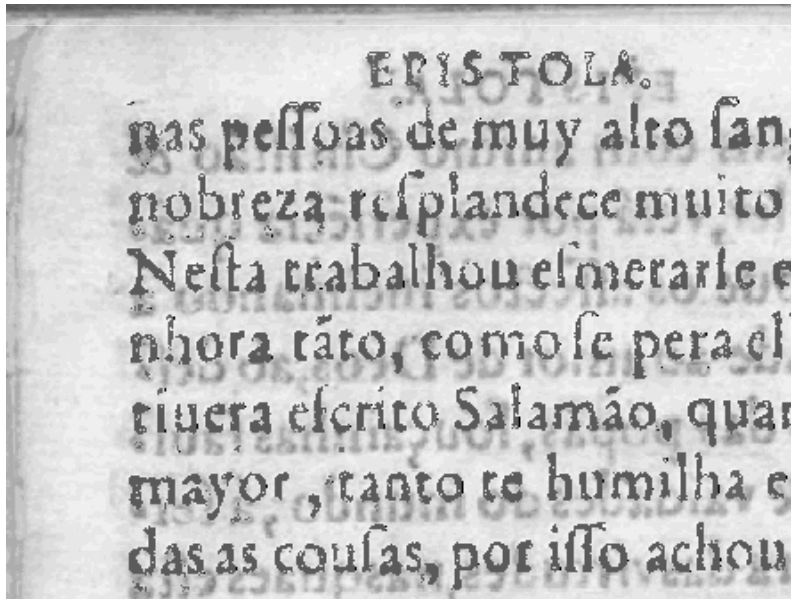
Results of a 4 classes classification
(simulation of text/drawings separation)

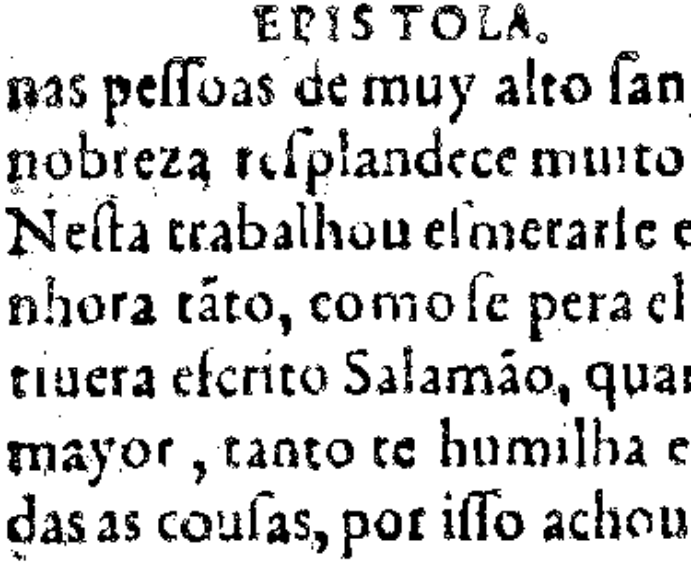
4. Séparation par redondance des formes

- **Critère d'extraction : le taux de redondance des formes de caractères**

Une segmentation inadaptée produit des caractères collés ou l'apparition du verso sur le recto ou des caractères cassés.

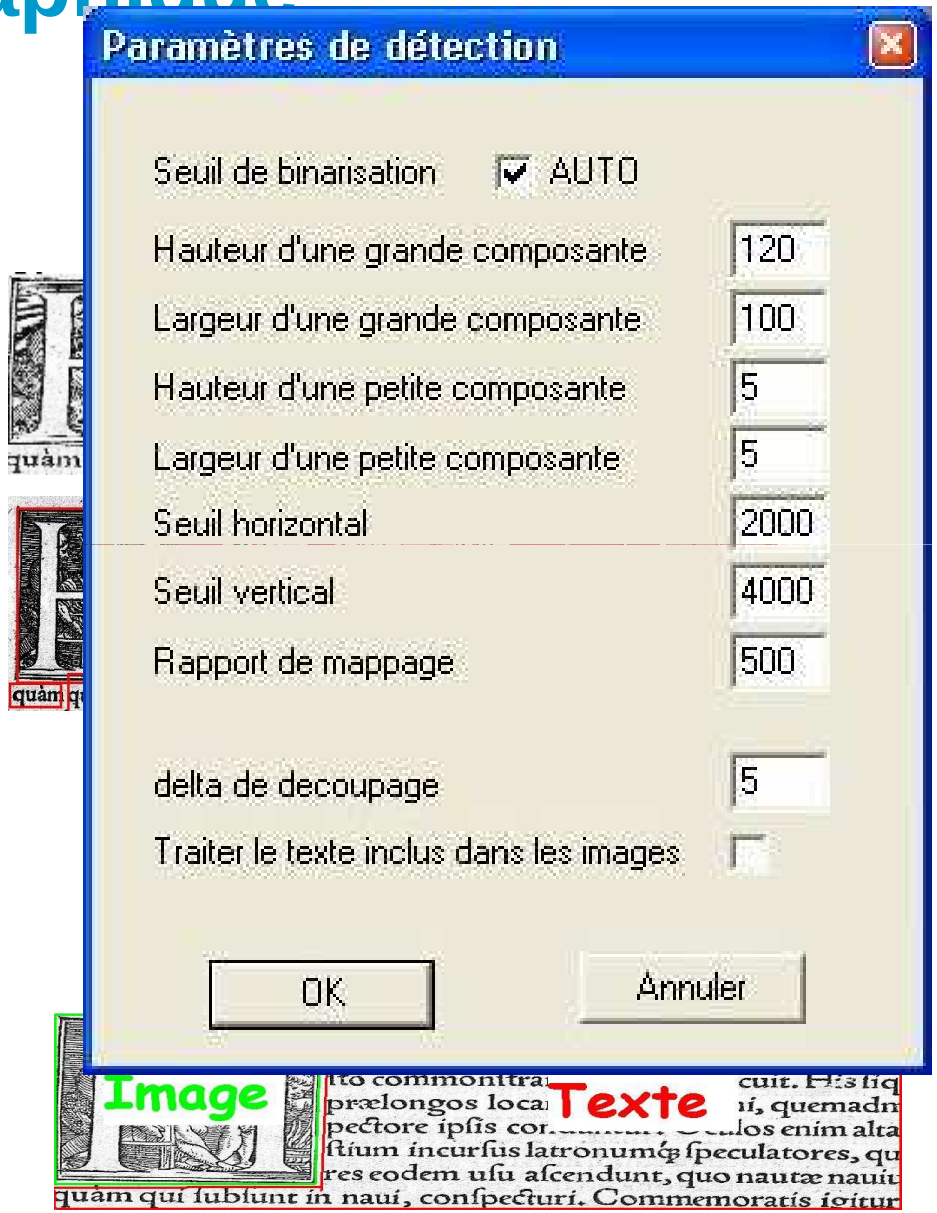
Une segmentation optimale correspond à un taux maximal de redondance des formes de caractères !





Bilan autour de la séparation texte/graphique

- Méthodes (morpho, filtrage direct., fréquentiel, connexités…)
- Connaissances a priori de l'échelle (taille des composantes de base, des lignes)
- Paramétrage coûteux
- Méthodes spécialisées
- Quelles connaissances pour une intervention utilisateur?



ATTENTION A L'USINE A GAZ

Conclusion sur la reconnaissance des structures

►Des travaux émergents pour les documents anciens (% 50 ans sur l'OCR)

► Encore au stade d'expérimentation: fonctionne bien sur des documents bien définis, peu de généricité (pas de modèle unique)

► Les systèmes à apprentissage apportent un peu de souplesse au détriment d'un travail supplémentaire

► Vers des approches plus interactives (implication de l'usager) et nécessitant moins de paramètres et de connaissances techniques

► L'analyse d'image offre aux SHS des nouvelles opportunités scientifiques *(expertise automatique des éléments typographiques de la Renaissance, reconnaissance des scripteurs, authentification de manuscrits, datation des écritures pour la paléographie, indexation…)*

Applications

- Les courriers d'entreprise
- Inventaires sommaires, Archives
- Documents mathématiques
- Manuscrits Médiévaux
- Manuscrits contemporains
- Imprimés de la Renaissance

Les courriers d’entreprises

Le tri automatique des documents : *problématique et besoins récents*

Débit = 13 lettres / seconde

Le tri automatique des documents : *problématique et besoins récents*

Familles de documents et de courriers d'entreprises

- *Courrier postal*
- *Courrier interne manuscrit*
- *Courrier interne dactylographique*
- *Formulaire*
- *Planus*
- *Carte bleue*
- *Listing A3*
- *Listing A4*
- *NPAI*
- *Chèque circulant*
- *Pochettes DVD, etc.*

Le tri automatique des documents : *problématique et besoins récents*

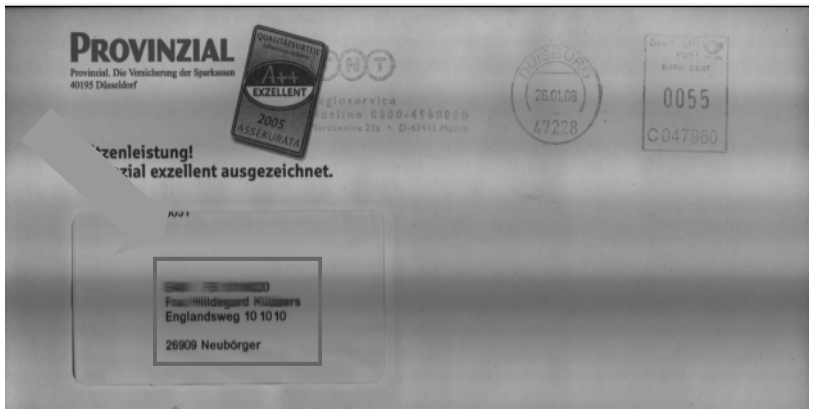
Deux types de documents

1 *Documents structurés (formulaires…)*

Documents non structurés (courrier postal…) **2**

Localisation de la zone d'intérêt

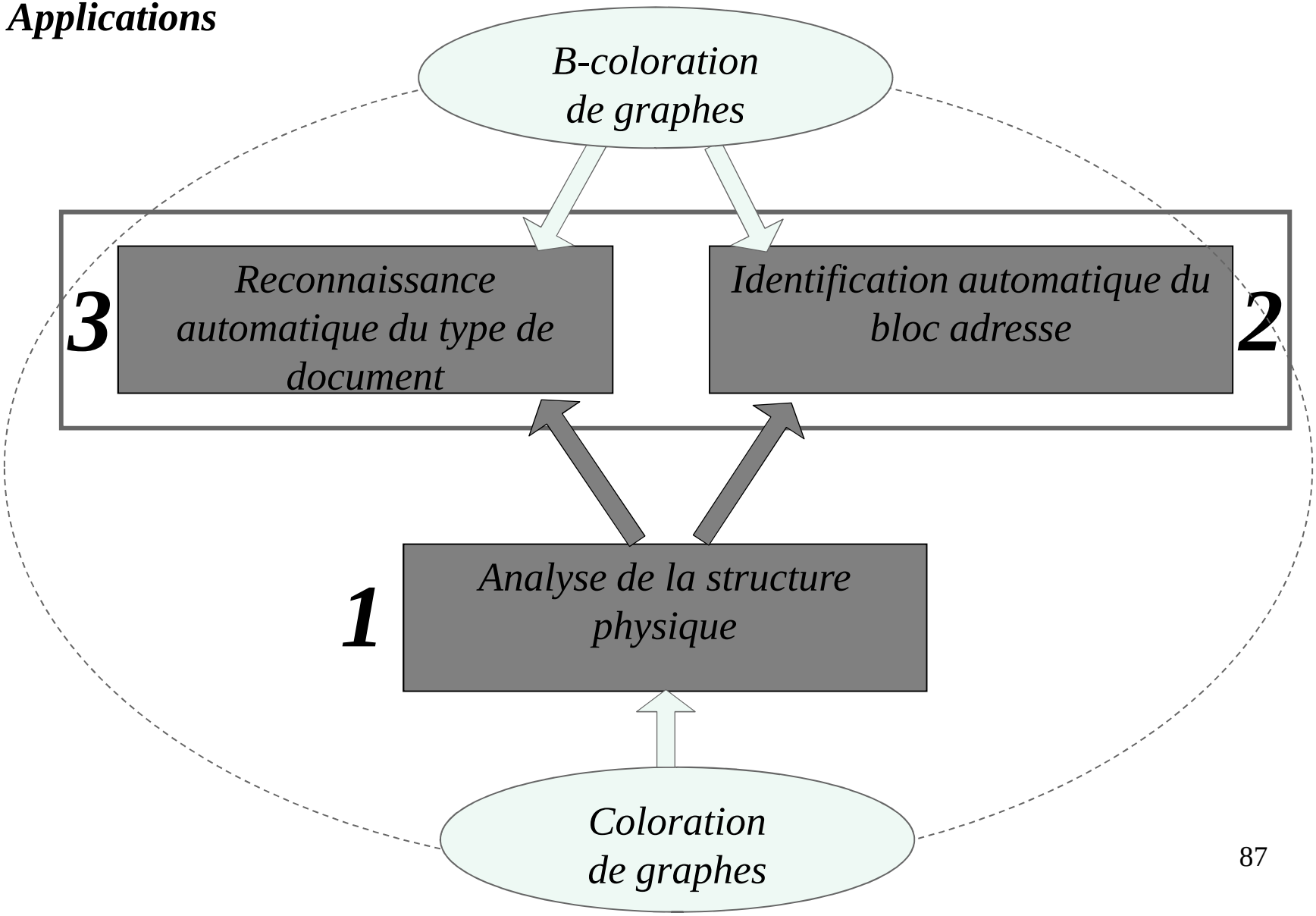
Lecture de code ACI manuscrit



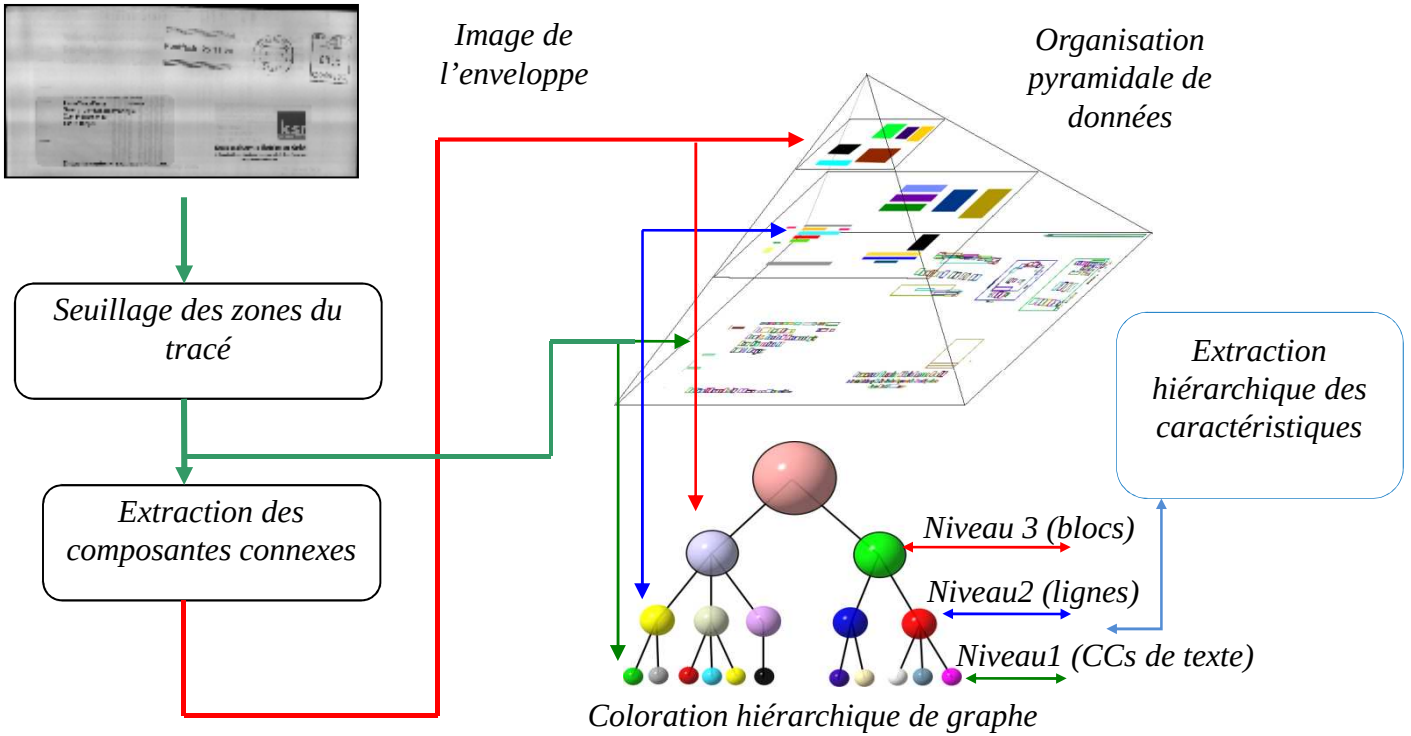
Lecture de l'adresse de destination

Des approches conventionnelles aux solutions industrielles

Applications



❑ *Diagramme fonctionnel de notre méthode*



1. Segmentation « brute » par binarisation et détection des CCs

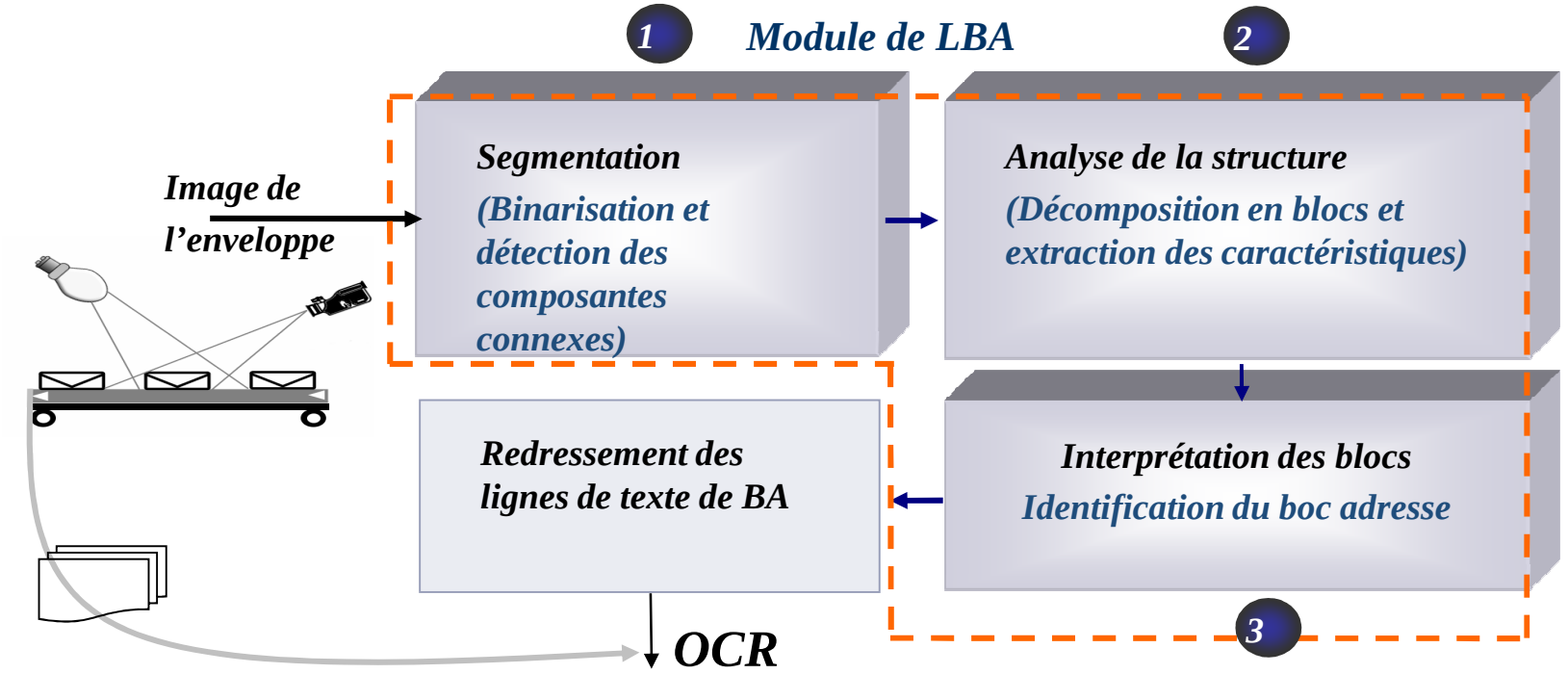
2. Extraction de la structuration par coloration hiérarchique de graphe

Grande variabilité de la zone d'adresse

- *Adresse de taille variable.*
- *Adresse manuscrite de styles variables ou imprimée de différentes polices de mises en forme variable (taille, espacement…).*
- *Technologies d'impression différentes : machines à écrire, imprimante matricielle, laser, à jet d'encre…etc.*
- *Adresse dégradée (faible contraste, luminance non uniforme, pli…)*
- *Adresse sur une étiquette adhésive, sur ou sous un film plastique, ou visible par une fenêtre transparente.*

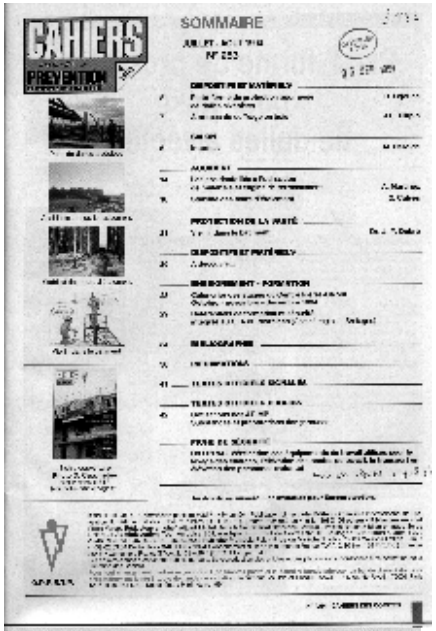
❑ *Étapes de conception du système de LBA*

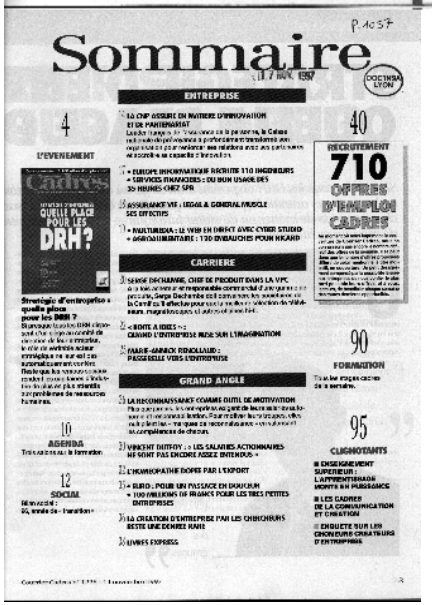
1. *Analyse hiérarchique de la structure physique (coloration hiérarchique de graphe)*
2. *Apprentissage pour la LBA (b-coloration de graphe)*
3. *Reconnaissance (identification) du bloc adresse parmi plusieurs candidats*



Inventaires sommaires

Reconnaissance de la structure des sommaires de revues





Variabilité des types de sommaires

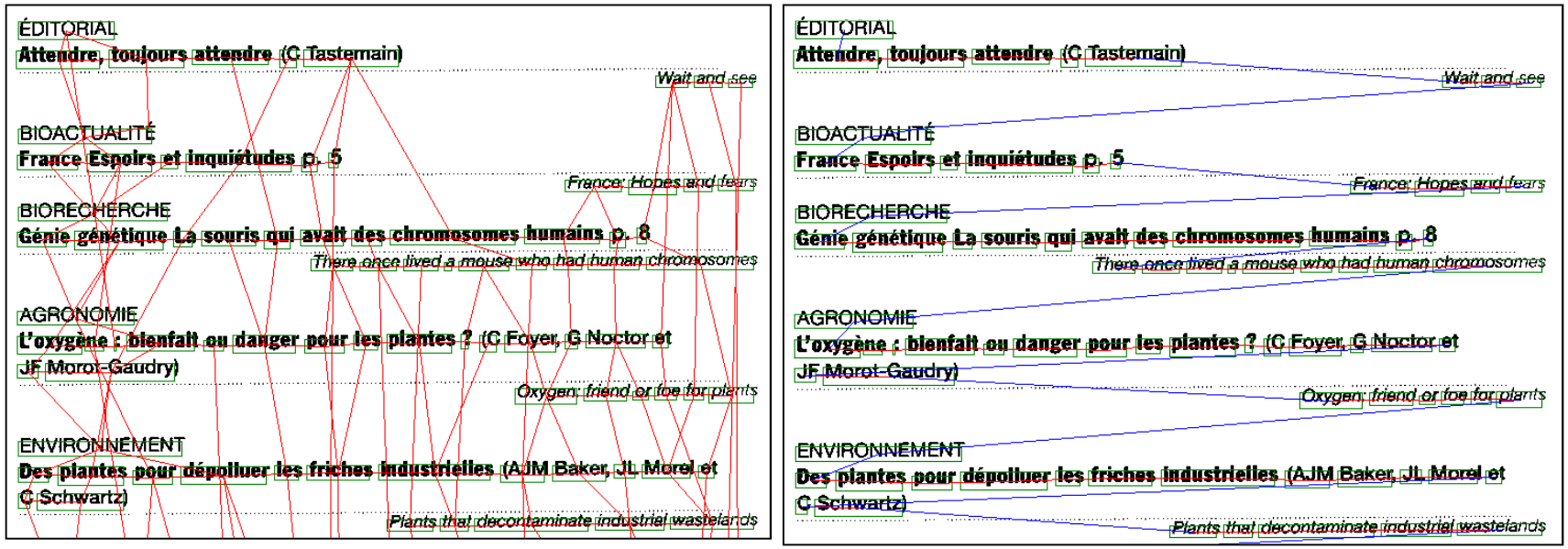
Reconnaissance des structures par apprentissage

La diversité des modèles de sommaires nous ont amené a réaliser un apprentissage pour chaque revue

Conservation de la mise en page

Chaque revue conserve une mise en page homogène :

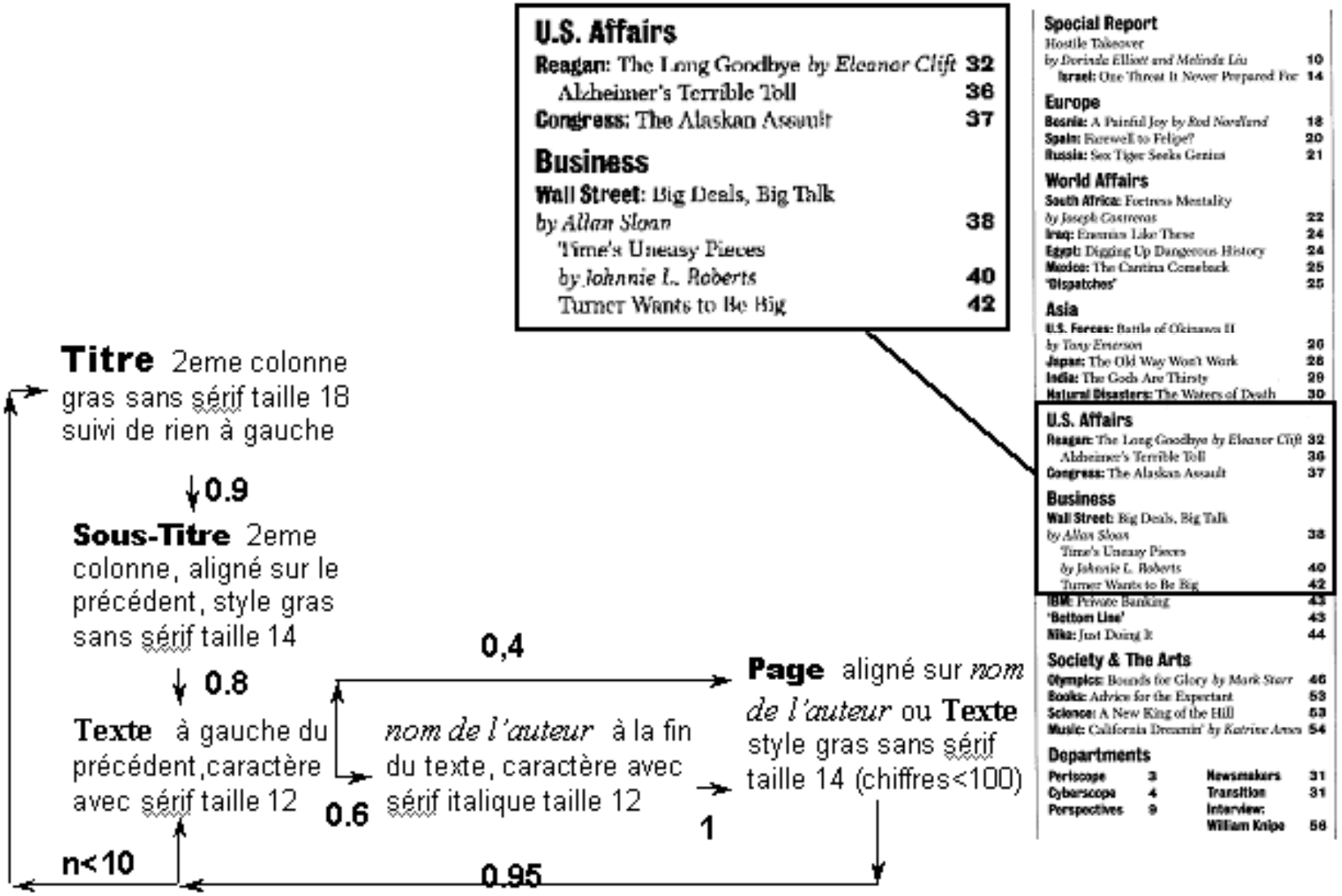
➡ *apprentissage des relations spatiales entre les blocs de texte*



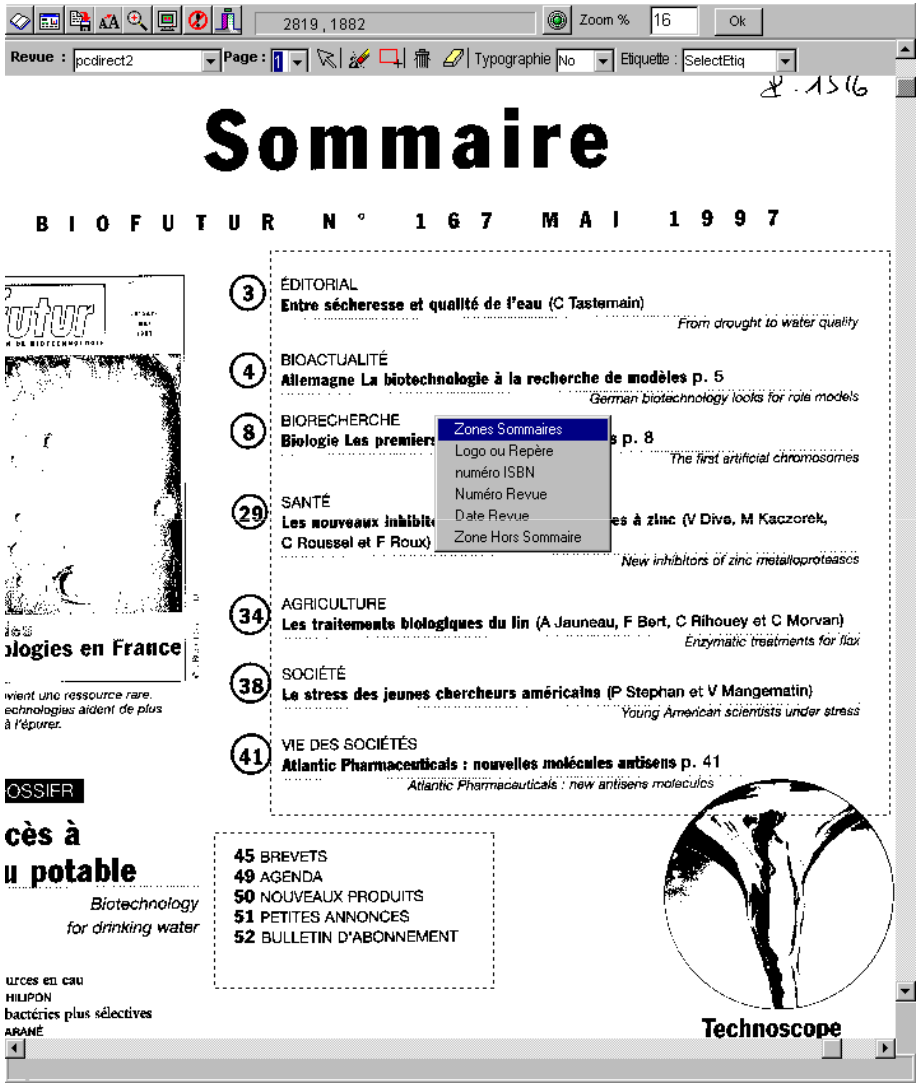
Modèle d'enchaînement flexible de la structure d'un sommaire

Modèle de décodage tolérant aux erreurs de segmentation

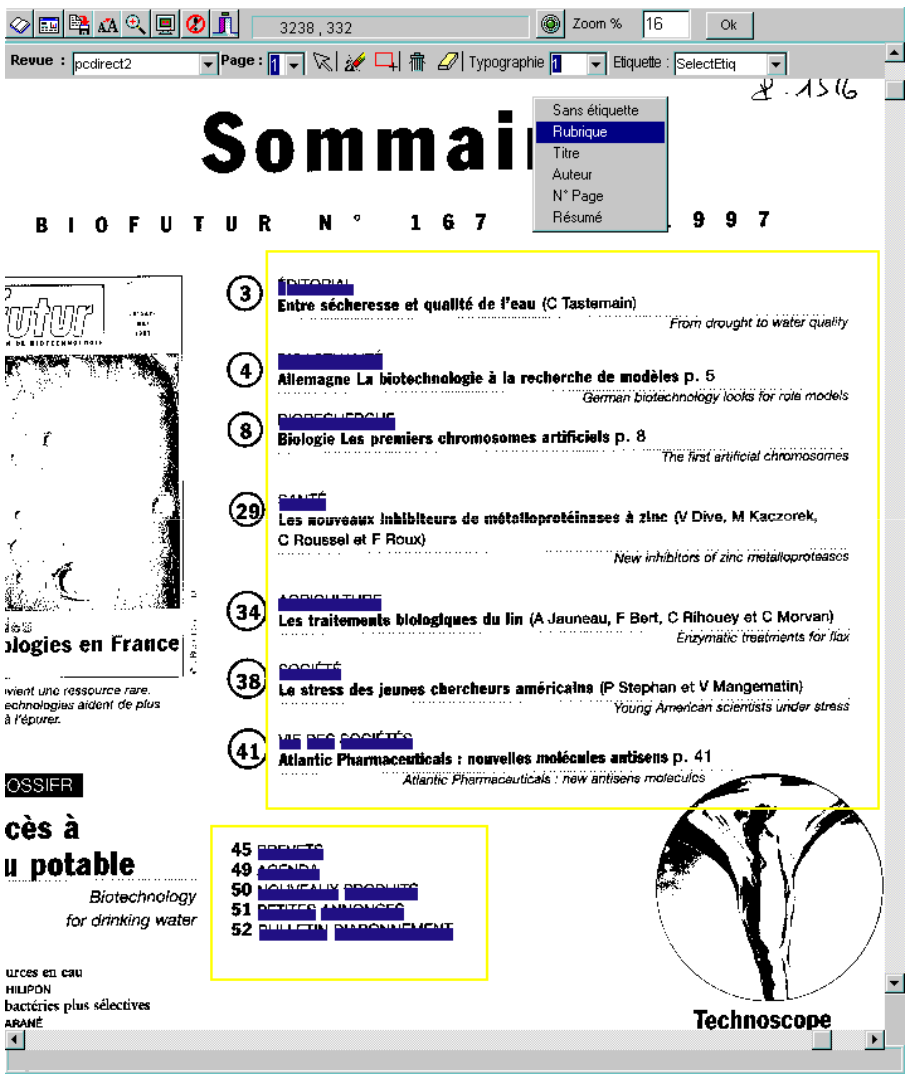
Utilisation quasi-récurrente du même style typographique par revue



Apprentissage par l'utilisateur

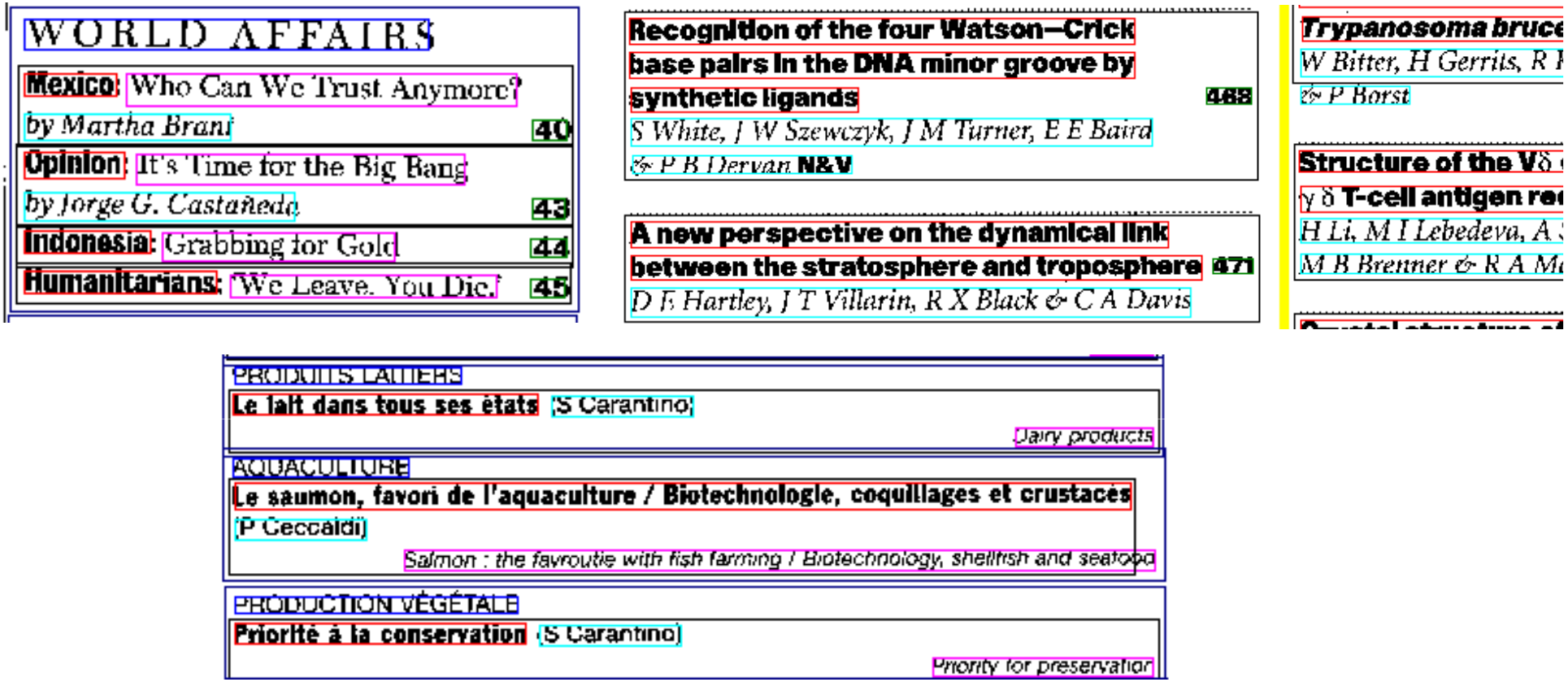


Apprentissage manuel des zones



Apprentissage des titres

Résultats de la reconnaissance



Conversion automatique des documents numérisés en fichier ré-éditable et structuré type XML

Les cas des documents du patrimoine… une ouverture nécessaire aux SHS

- Recherche d'une image de traits par l'exemple *(rechercher une lettrine, une décoration particulière, une forme de caractère…)*
- Classification automatique des éléments typographiques de la Renaissance *(comparaison des casses, des ornements et lettrines)*
- Segmentation et classification d'ouvrages de la Renaissance pour la recherche d'informations
- Classification spatio-temporelle des écritures manuscrites médiévales pour la paléographie *en cours* …
- Authentification de scripteurs dans les manuscrits d'auteurs
- Recherche de similarités dans les manuscrits (compression)

Quelques exemples de travaux réalisés et en cours…

Manuscrits réguliers anciens Latins

(période du Moyen Age)

Les documents à travers les périodes

6-8

Old / New Roman scripts

Pre-carolingian scripts

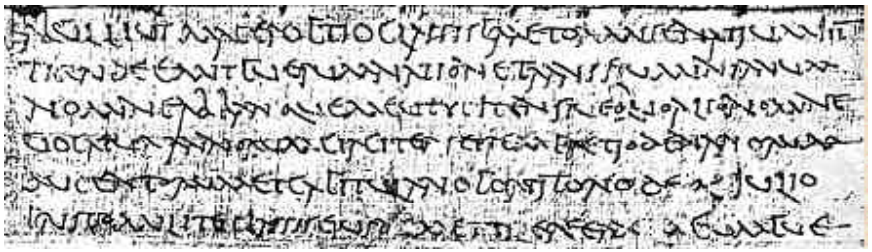
Caroline

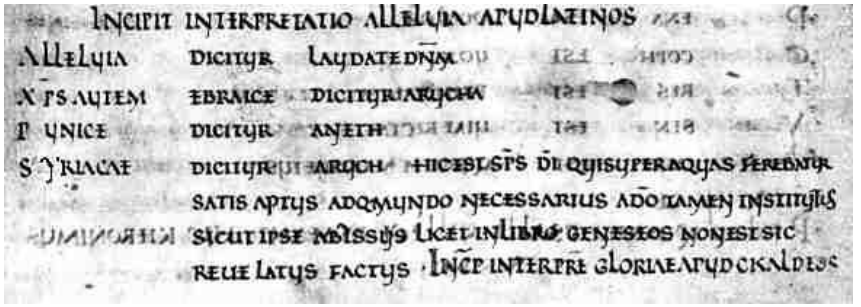
Gothic

- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

Humanistic scripts

Contemporary scripts





1. Carved inscription (MMD)
2. Old roman on Papyrus (BL)
3. Rustic capital (BL)

Les documents à travers les périodes

8-9

Old / New Roman scripts

Pre-carolingian scripts

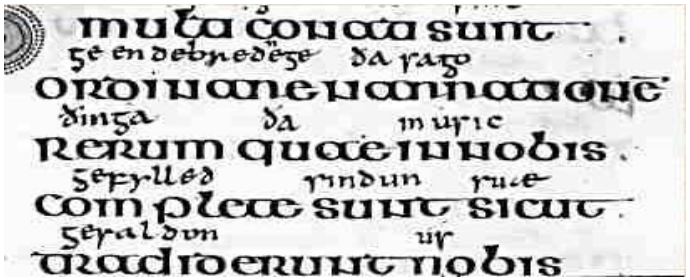
Caroline

Gothic

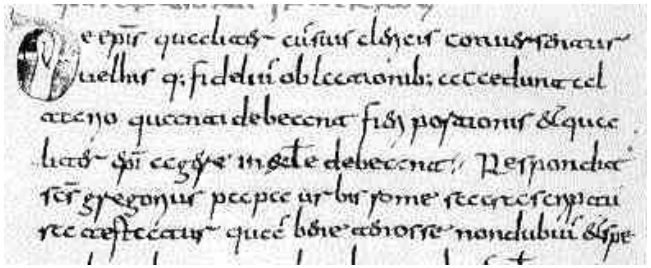
- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

Humanistic scripts

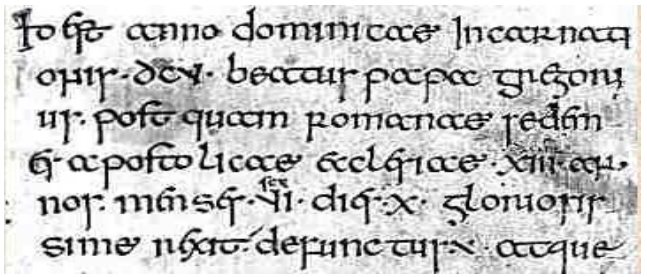
Contemporary scripts



8th



8th



9th

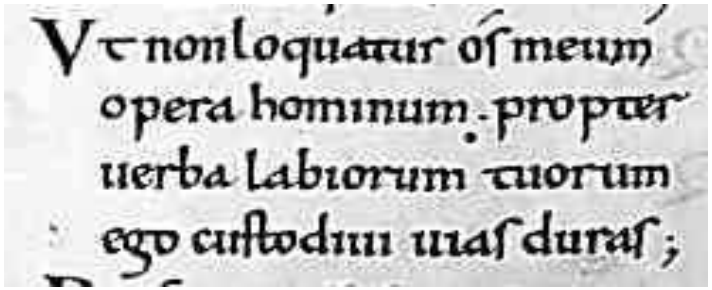
1. *Uncial Post-Roman (BL)*
2. *Merovingian (BL)*
3. *Pre-carolingian (BL)*

Les documents à travers les périodes

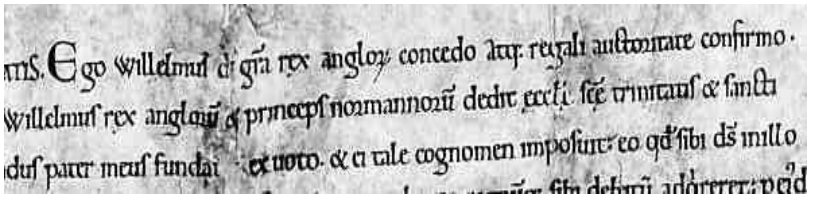
10-12

Old / New Roman scripts
Pre-carolingian scripts
Caroline
Gothic
- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

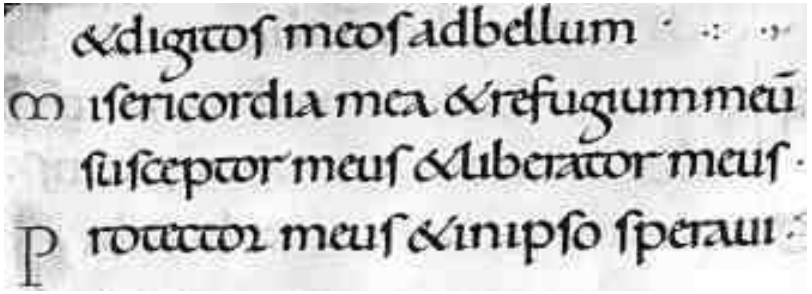
Humanistic scripts
Contemporary scripts



10th



11th



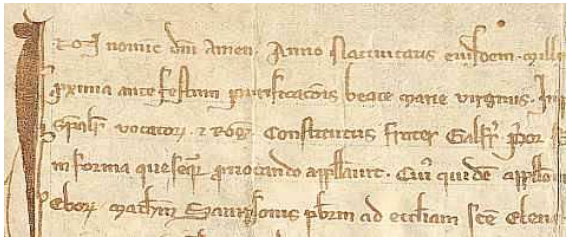
11th

Les documents à travers les périodes

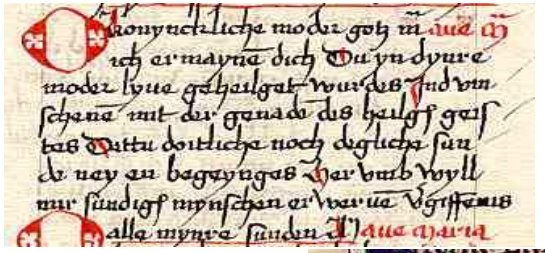
12-15

Old / New Roman scripts
Pre-carolingian scripts
Caroline
Gothic
- **Protogothic**
- **Textura**
- **Rotunda**
- **Cursiva**
- **Bastarda**

Humanistic scripts
Contemporary scripts



13th



15th

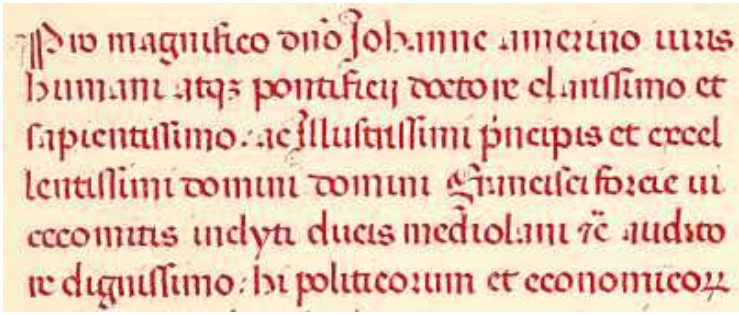
1. *Calligraphic (Private coll.)*
2. *Cursive Gothic (Private coll.)*
3. *Textura (NDL)*
4. *Batarda (Private coll.)*

Les documents à travers les périodes

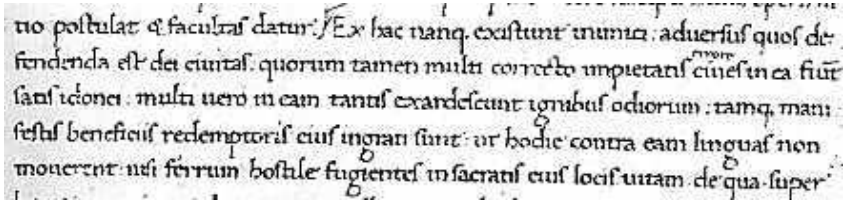
Old / New Roman scripts
Pre-carolingian scripts
Caroline
Gothic
- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

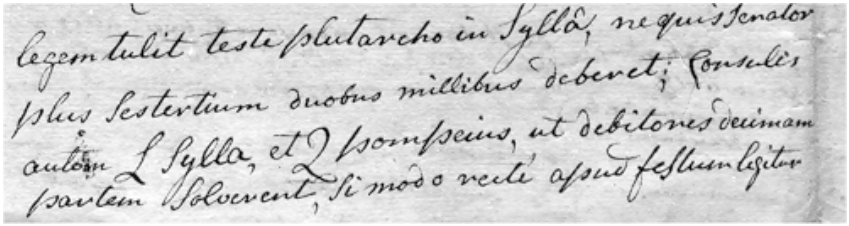
15-19

Humanistic scripts
Author's scripts



15th

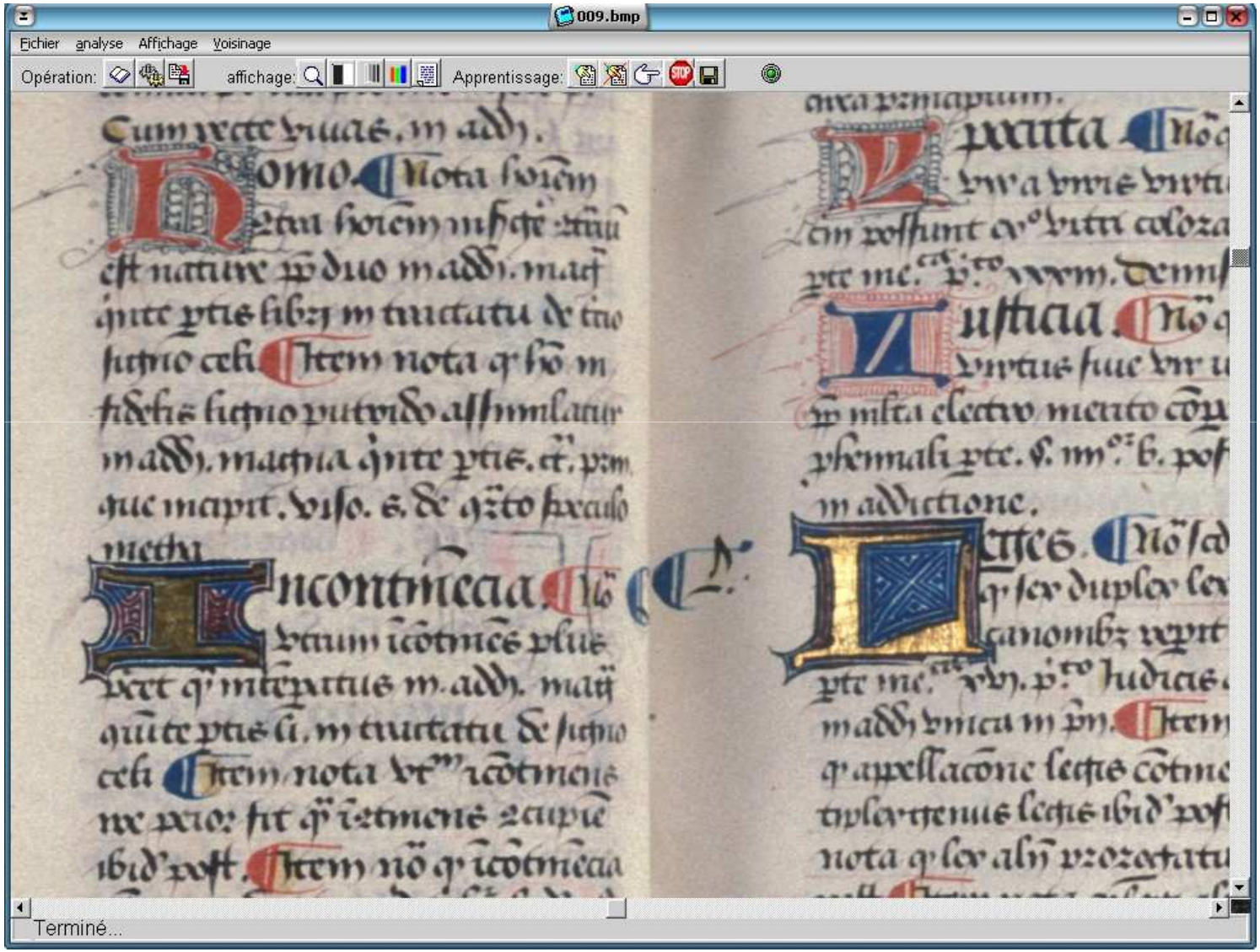




18th

1&2.Humanistic (BL)
3.Montesquieu's collection (BNF)

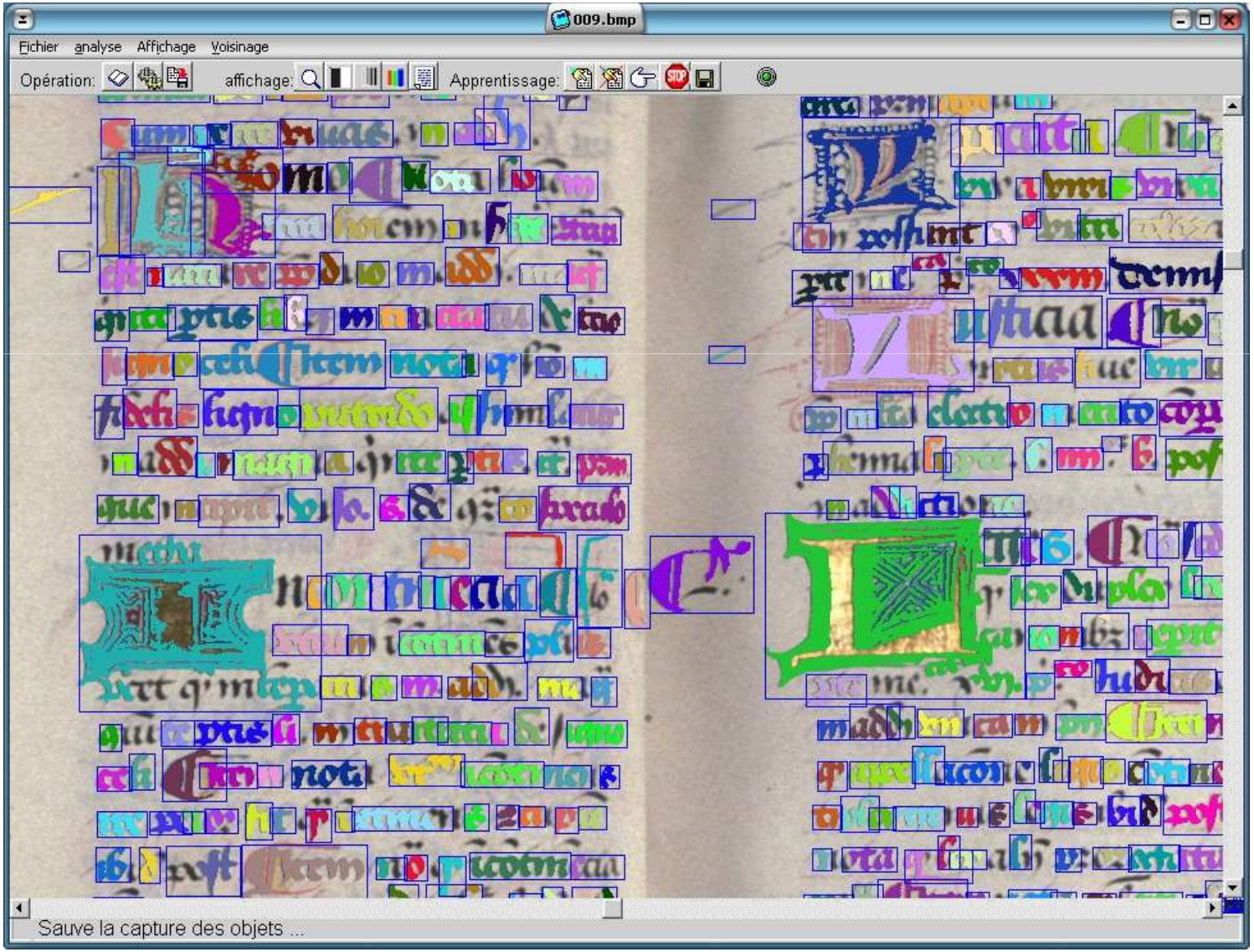
Segmentation et capture des objets



Segmentation et capture des objets

Mesures et caractérisation des objets

Caractéristiques de Formes, géométrie, et couleur…



Extraction des informations sur les relations spatiales entre les objets

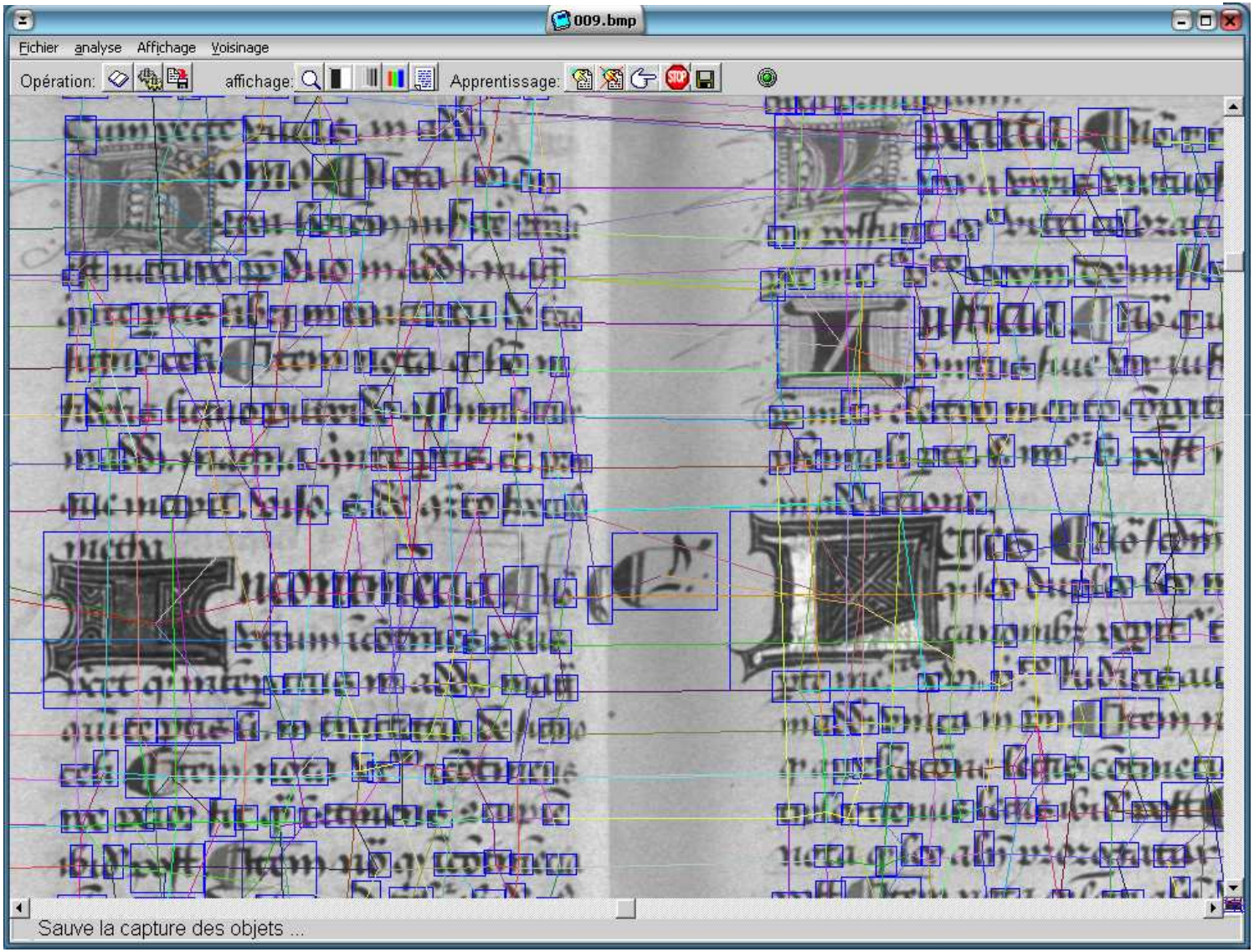
Exemple…

- Centre de gravité Cx et Cy.
- Surface [S] (en pixels)
- Périmètre [P] : mesuré comme la somme des distances entre pixels (somme de 1 et $\sqrt{2}$)

- Moyenne des niveaux de gris (ou couleurs)
- Variance des niveaux de gris (ou couleurs)
  - *Conseil : utiliser un autre espace couleur que RGB pour coder les couleurs (par exemple HLS), car celui-ci est non-linéaire fausse les résultats.*

- Facteur de forme : $Ff = \frac{4\pi S}{P^2}$ *Droite : Ff = 0* *Cercle : Ff = 1*
- **...**

Extraction des informations sur les relations spatiales entre les objets



Extraction des informations sur les relations spatiales entre les objets

Choix des caractéristiques

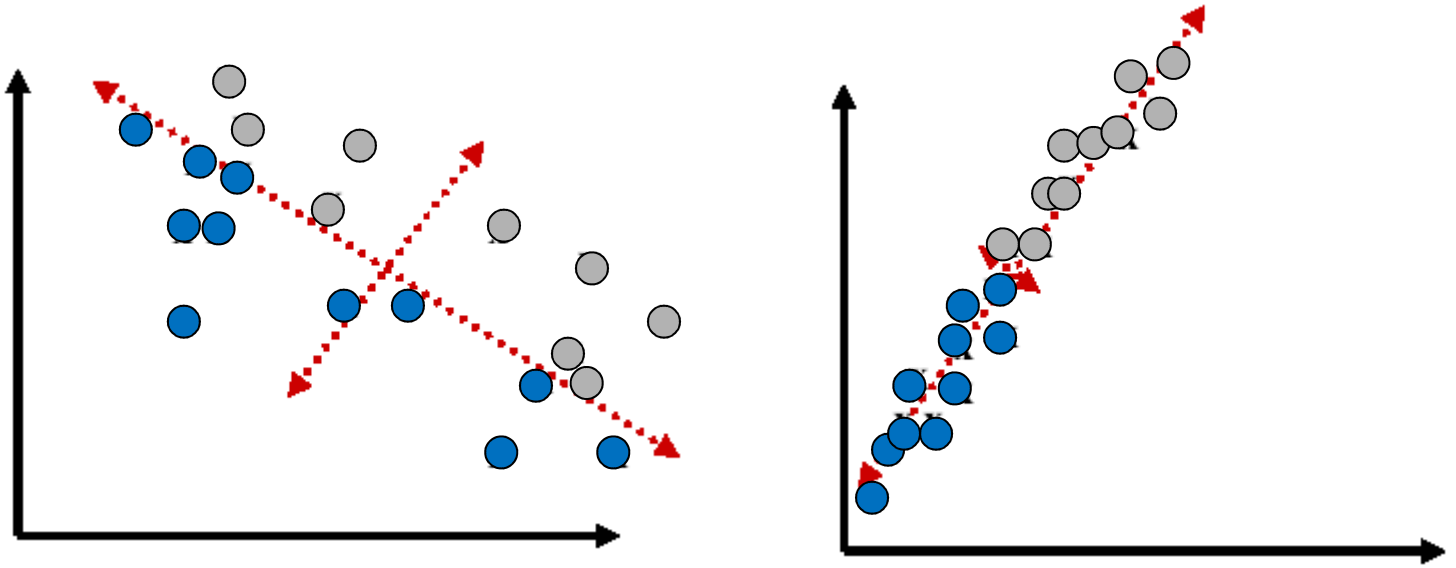
- Le principal problème est le bon choix des caractéristiques permettant de différencier les différentes classes d'objets.
  - *Classe de texte, de graphique, de lettrines…*

- Les caractéristiques identifiées par les experts d'un domaine (exemple : paléographes) ne sont pas forcément celles qui sont reconnaissables dans les images.

- Une méthode qui permet d'identifier les caractéristiques importantes est l'Analyse en Composantes Principales (ACP).

Analyse en Composantes Principales

- A partir d'un nombre élevé de caractéristiques, le but de cette méthode est de réduire les calculs à un petit nombre significatif de caractéristiques.
  - *Elimine la redondance entre les caractéristiques et les caractéristiques non-significatives.*



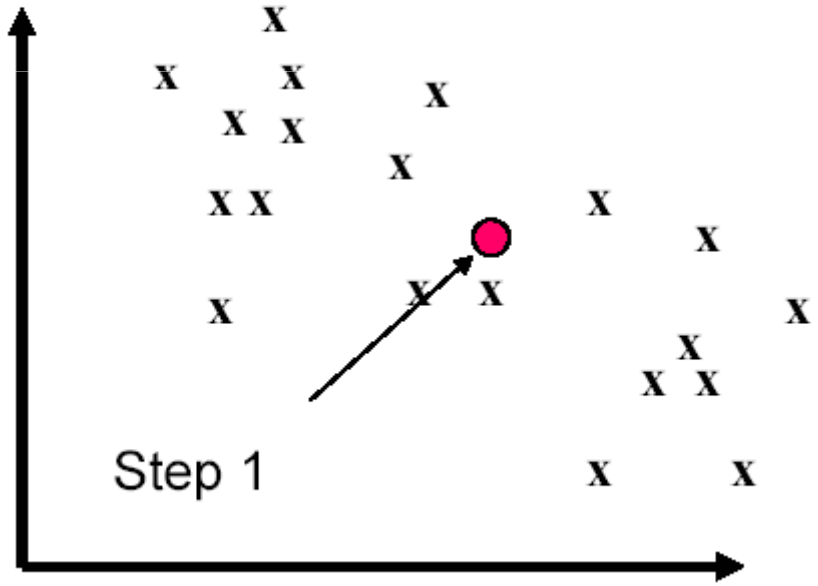
Analyse en Composantes Principales

- **Etape 1**
  Calculer la moyenne de chaque vecteur de caractéristiques.
- **Etape 2**
  Soustraire la moyenne de chaque vecteur de caractéristiques.
- **Etape 3**
  Calculer la matrice des covariances. $\Psi = \frac{1}{N_i - 1} \sum_{k=1}^{N} (x_k - \bar{x})(x_k - \bar{x})^T$
- **Etape 4**
  Calculer les valeurs et vecteurs propres de la matrice de covariance.
- **Etape 5**
  Ne conserver que les valeurs propres (+ vecteurs) les plus grandes.
- **Etape 6**
  Projeter les données dans ce nouvel espace propre.

Analyse en Composantes Principales

- **Etape 1**

  Calculer la moyenne de chaque vecteur de caractéristiques.

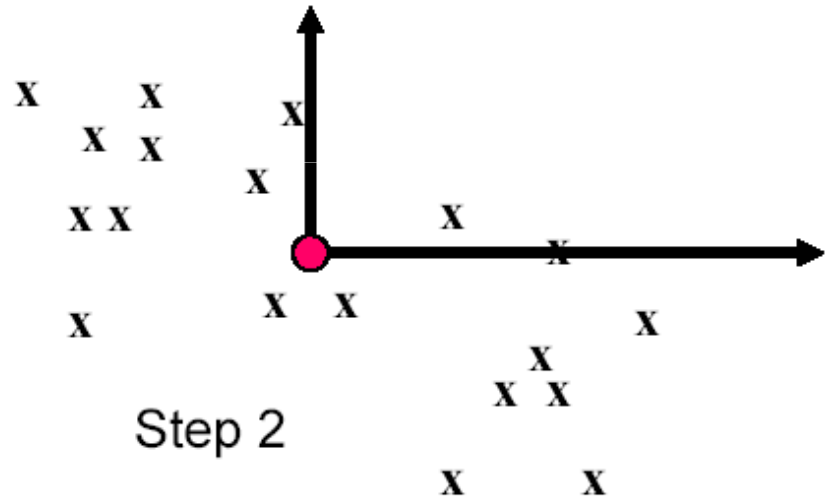


Source : cis.poly.edu/cs664

Analyse en Composantes Principales

- **Etape 2**

  Soustraire la moyenne de chaque vecteur de caractéristiques.



- **Etape 3**

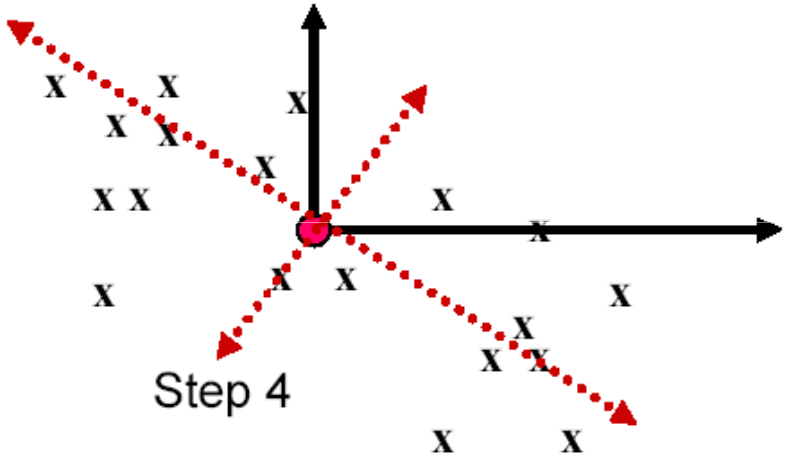
  Calculer la matrice des covariances.

$$\Psi = \frac{1}{N_i - 1} \sum_{k=1}^{N} (x_k - \bar{x})(x_k - \bar{x})^T$$

Analyse en Composantes Principales

- **Etape 4**

  Calculer les valeurs et vecteurs propres de la matrice de covariance.

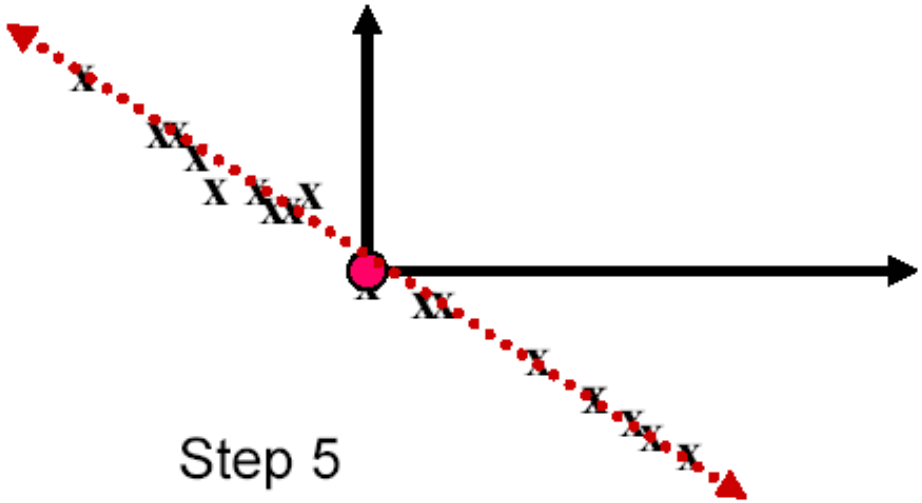


- Les vecteurs propres de la matrice de covariance représentent une base orthonormée d'axes principaux (significatifs) de l'ensemble des données.
- Chaque valeur propre exprime l'importance du vecteur propre associé : plus la valeur propre est grande, plus le vecteur propre associé est significatif

Analyse en Composantes Principales

- **Etape 5**

  Ne conserver que les valeurs propres (+ vecteurs) les plus grandes.



Step 5

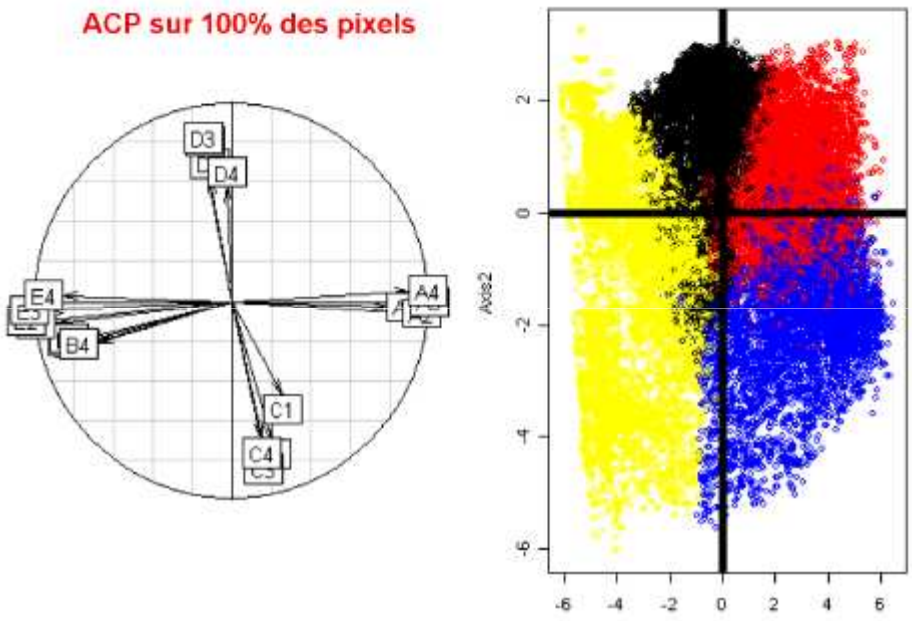
- **Etape 6**

  Projetter les données dans ce nouvel espace propre.

Analyse en Composantes Principales

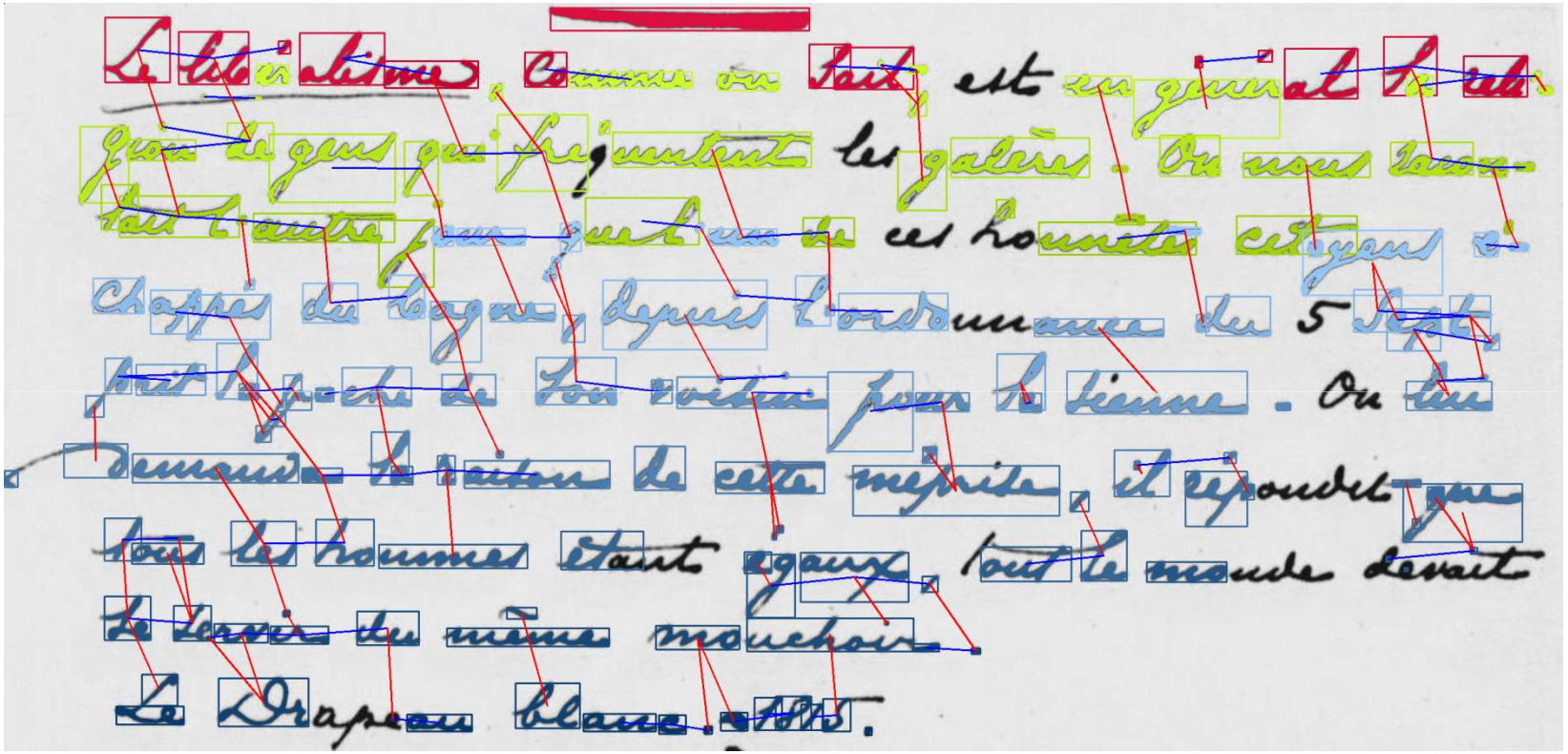
- L’Analyse en Composantes Principales sert à optimiser les caractéristiques utilisées pour la reconnaissance.
- Une fois l’entraînement (ou apprentissage) des données terminé, pour une nouvelle forme à reconnaître :
  - On calcule les caractéristiques de cette forme
  - On projète les caractéristiques dans le nouvel espace propre (espace de l’ACP)
  - On calcule la distance avec chacune des classes possibles pour trouver la bonne classe d’appartenance.

Exemple : classification des objets d'un document



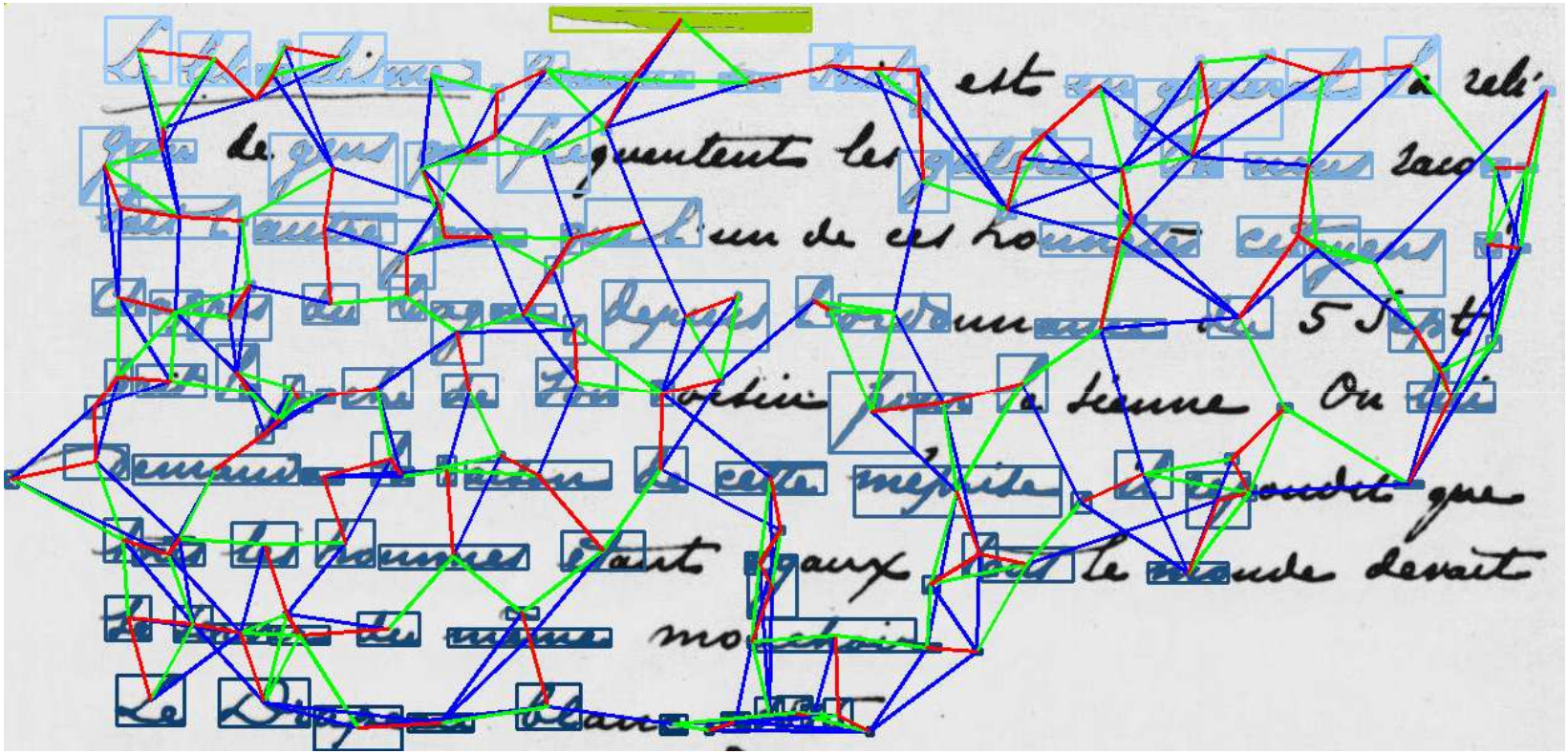
Etude des corrélations entre dimensions
Caractéristiques (cercle des corrélations)
- *contribution*
- *inter-dépendances*

Segmentation et capture des objets



Recherche de voisinage et construction d'un graphe planaire
Transformée et Hough, critère d'alignement et d'orientation

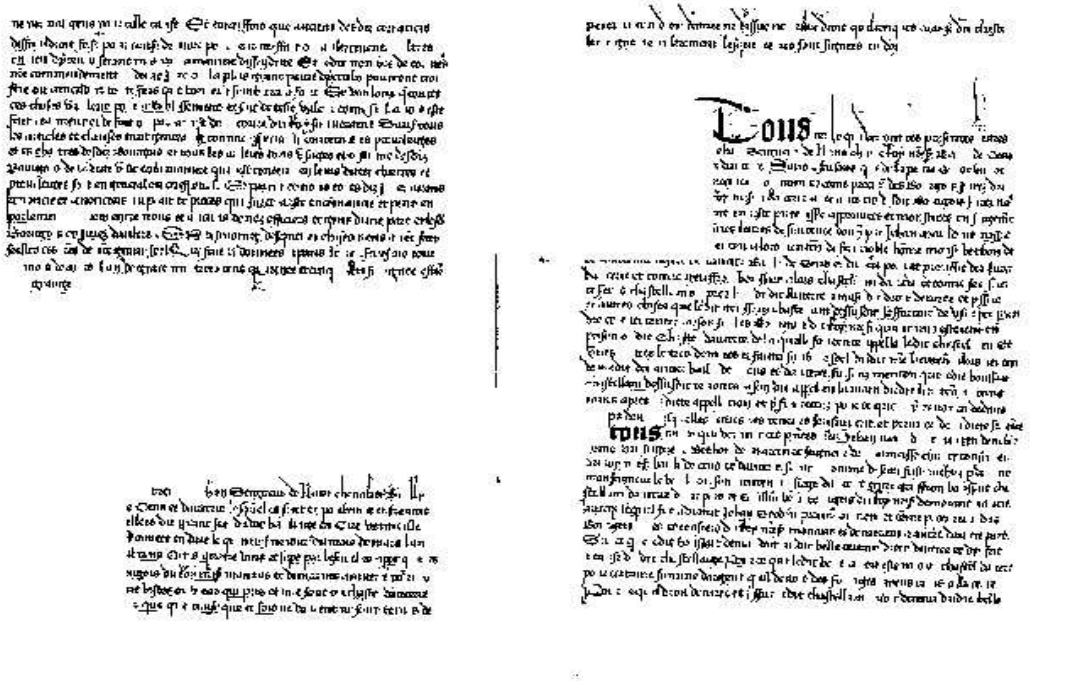
Segmentation et capture des objets



Segmentation en type d’écriture

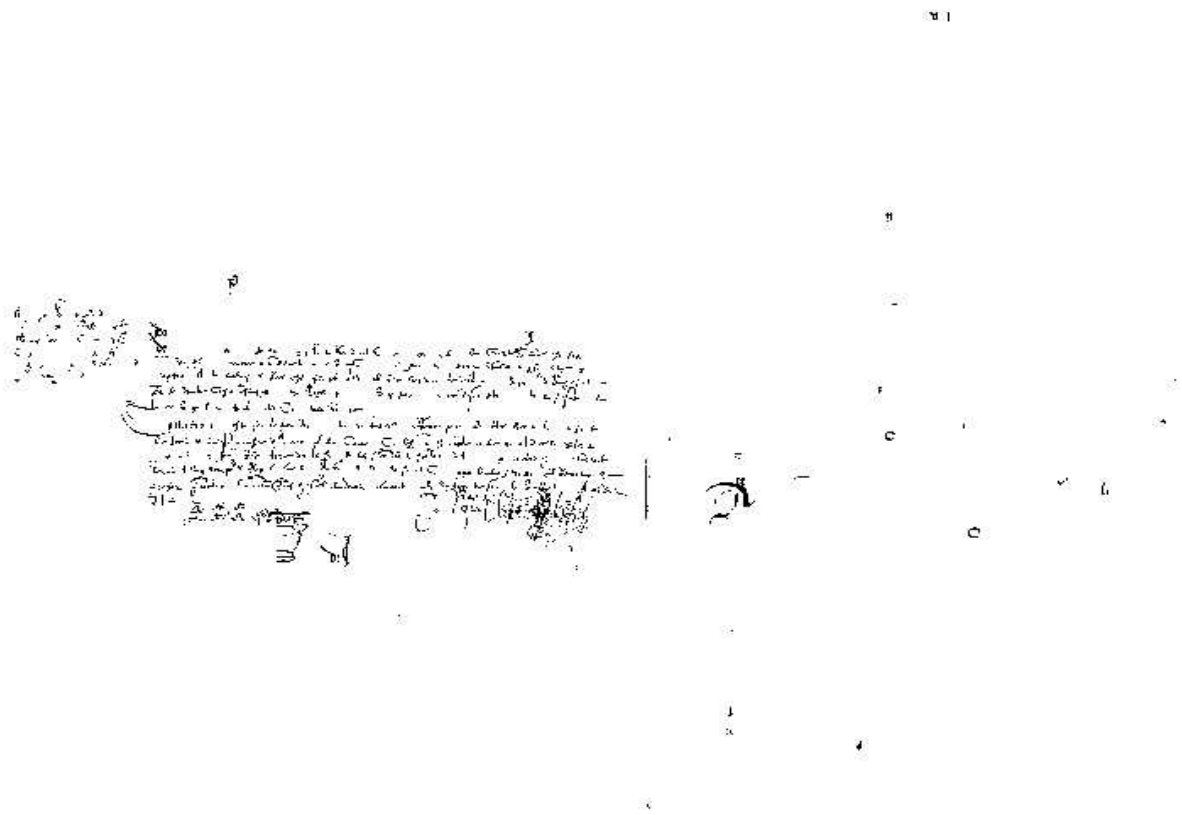
Image originale

Segmentation en type d’écriture



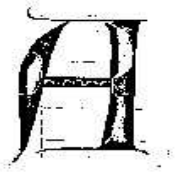
Classe n°1 : texte (Ecriture du 1 er copiste)

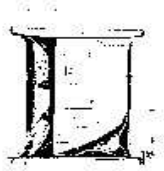
Segmentation en type d’écriture



Classe n°2 : Texte (Ecriture du 2 nd copiste)

Extraction automatique des lettrines



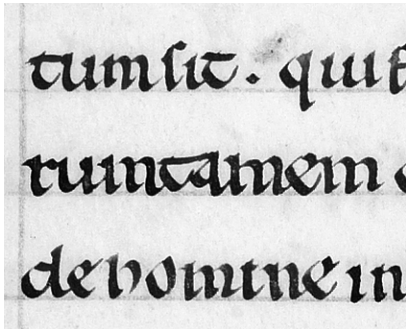


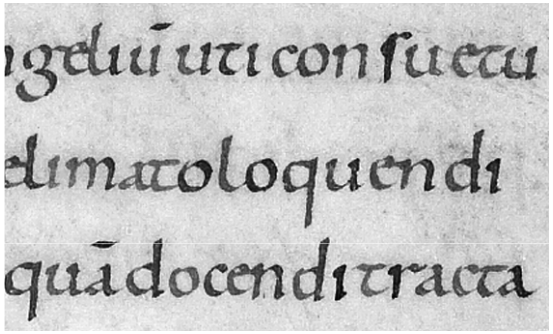
Classe n°3 : Lettrines de couleur

Paléographie sur manuscrits réguliers

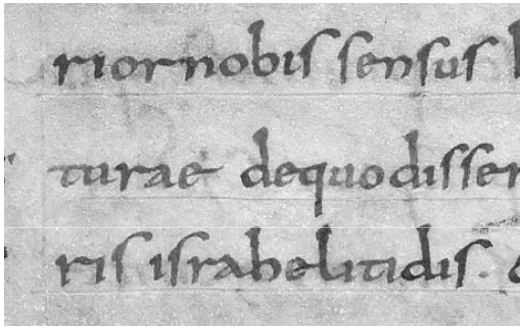
Requête : Trouver le style d'écriture la plus proche

Hypothèse de date : XII

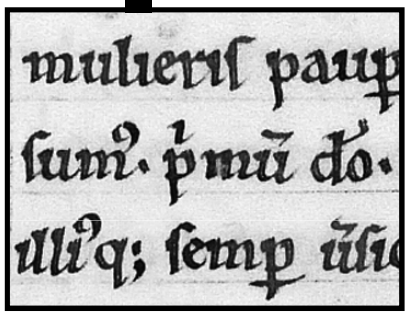




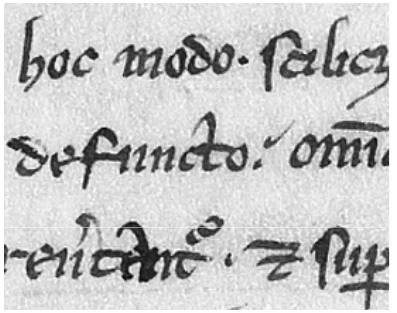
Corbie 850



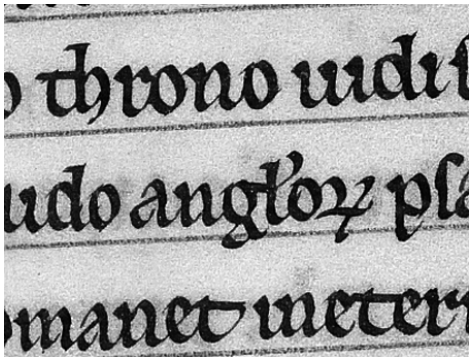
IX



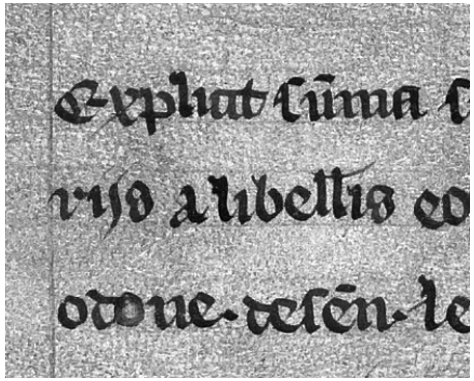
Vauclaire 1150



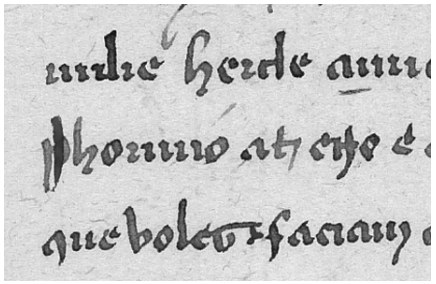
Laon 1262



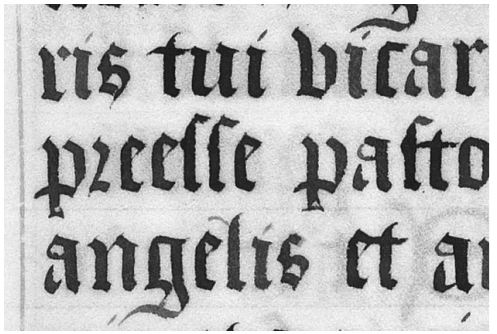
1222 Cistercien



1325

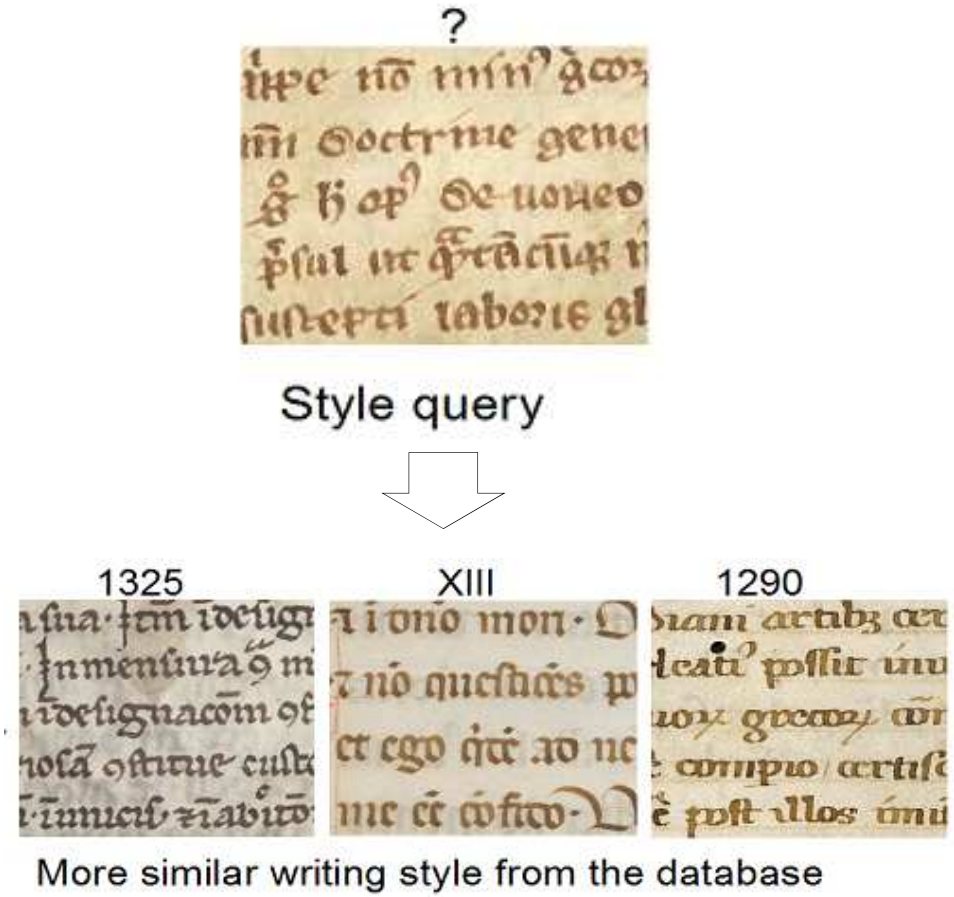


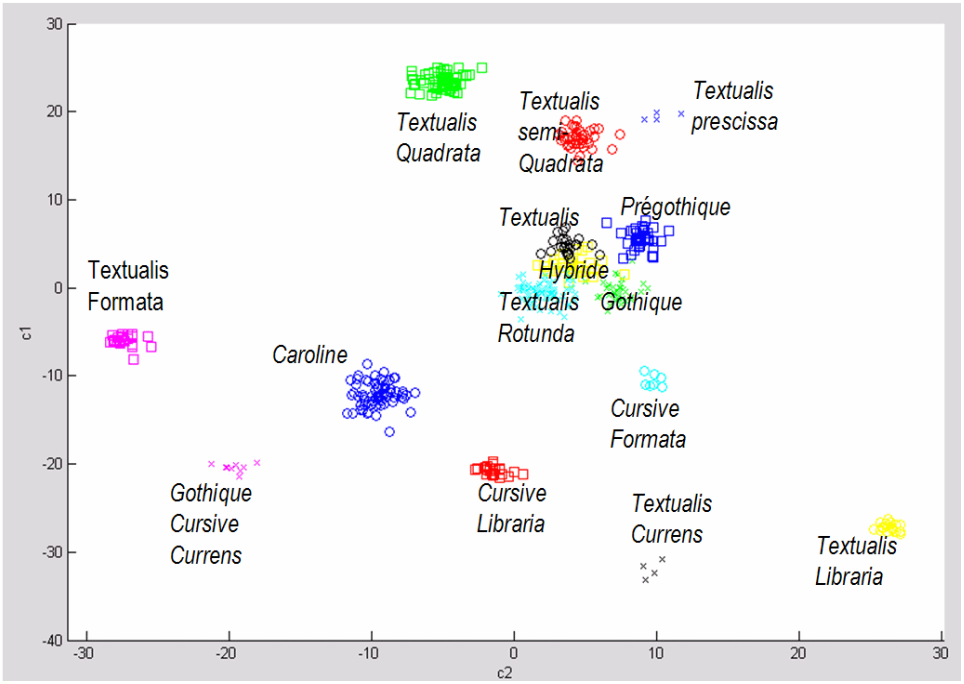
1401-1450



1535

Paléographie sur manuscrits réguliers

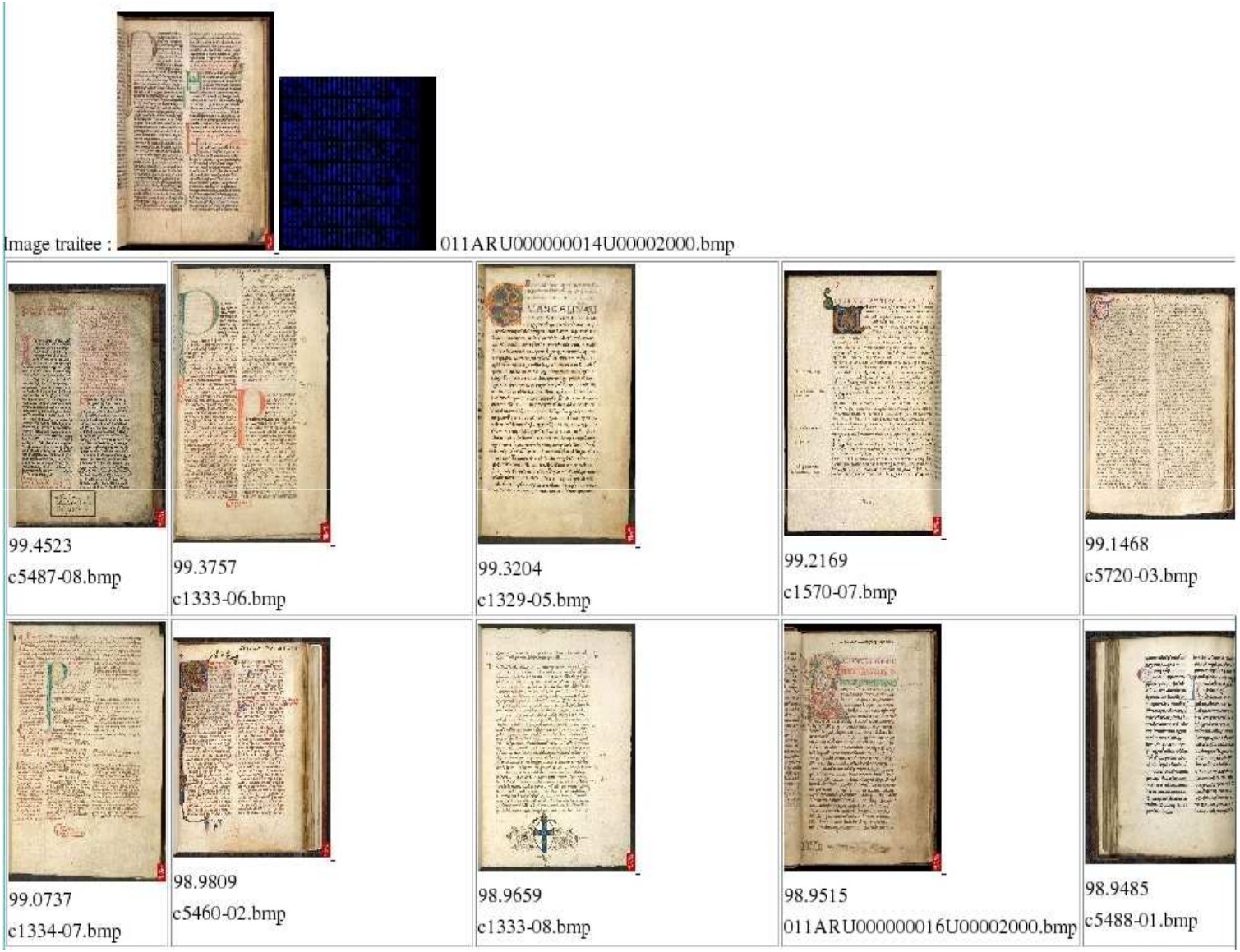




Linear discriminant analysis

Example of style query and most similar responses from the database.

Paléographie sur manuscrits réguliers

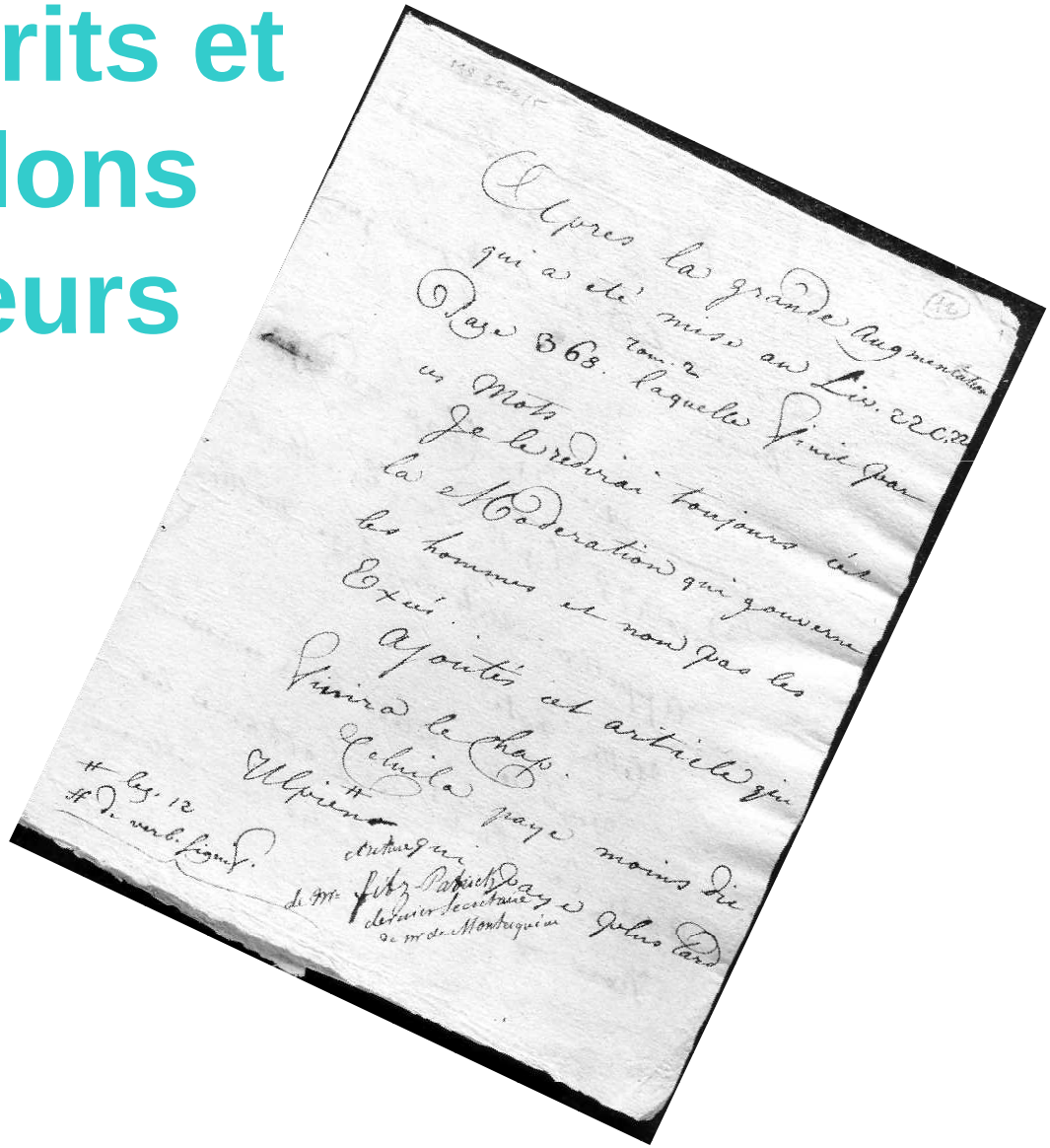


Example of style query and most similar responses from the database.

Paléographie sur manuscrits réguliers

Example of style query and most similar responses from the database.

Manuscrits et brouillons d'auteurs



Analyse des manuscrits contemporains

- **Exploitation scientifique des fonds manuscrits d'auteurs**
  - Rendre accessible aux chercheurs des textes de grande fragilité
  - Proposer à l'étude des textes dont l'authenticité reste à prouver
  - Préserver et valoriser les contenus

- **La problématique du brouillon d'auteur (Montesquieu- 1750)**
  - Rendre compte de la complexité des mise en pages
  - Souligner l'intervention de différentes mains
  - Proposer une trace historique de l'œuvre

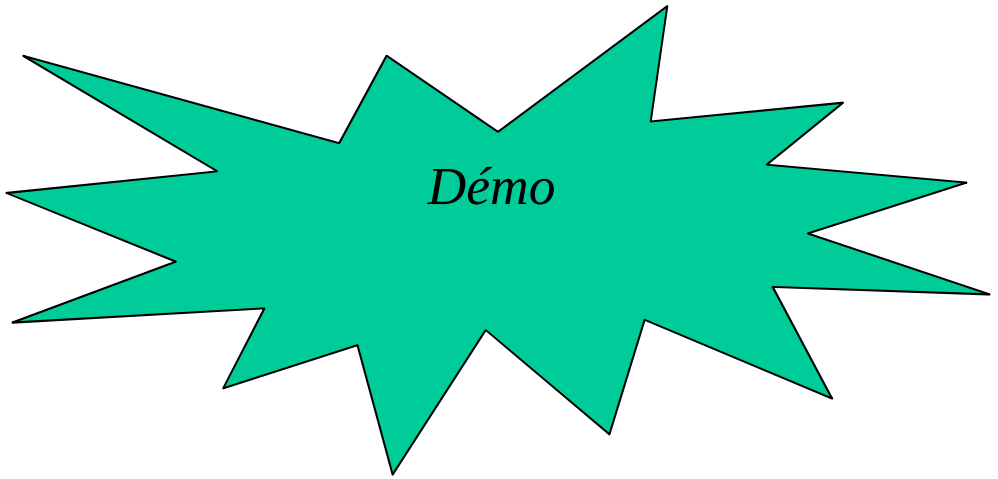
- **La problématique du texte inachevé: Bouvard et Pécuchet (Flaubert- 1881)**
  - Editer en ligne, sous une forme technologique innovante, un ensemble patrimonial cohérent, d'importance scientifique et culturelle reconnue
  - Fournir une interface d'interrogation multiple performante et différents types de consultation des documents :
    - une consultation linéaire selon l'organisation patrimoniale, le classement chronologique,
    - des parcours thématiques guidés par champs scientifiques distincts (histoire, médecine, philosophie…).
    - Une circulation entre fragments et une comparaison des notes prises par Flaubert avec les textes dont elles sont issues.

Mise en ligne du corpus Bouvard et Pécuchet de Flaubert, 2007

Analyse des manuscrits contemporains fortement bruités

http://dossiers-flaubert.ish-lyon.cnrs.fr/login.php

veglin + kexect3

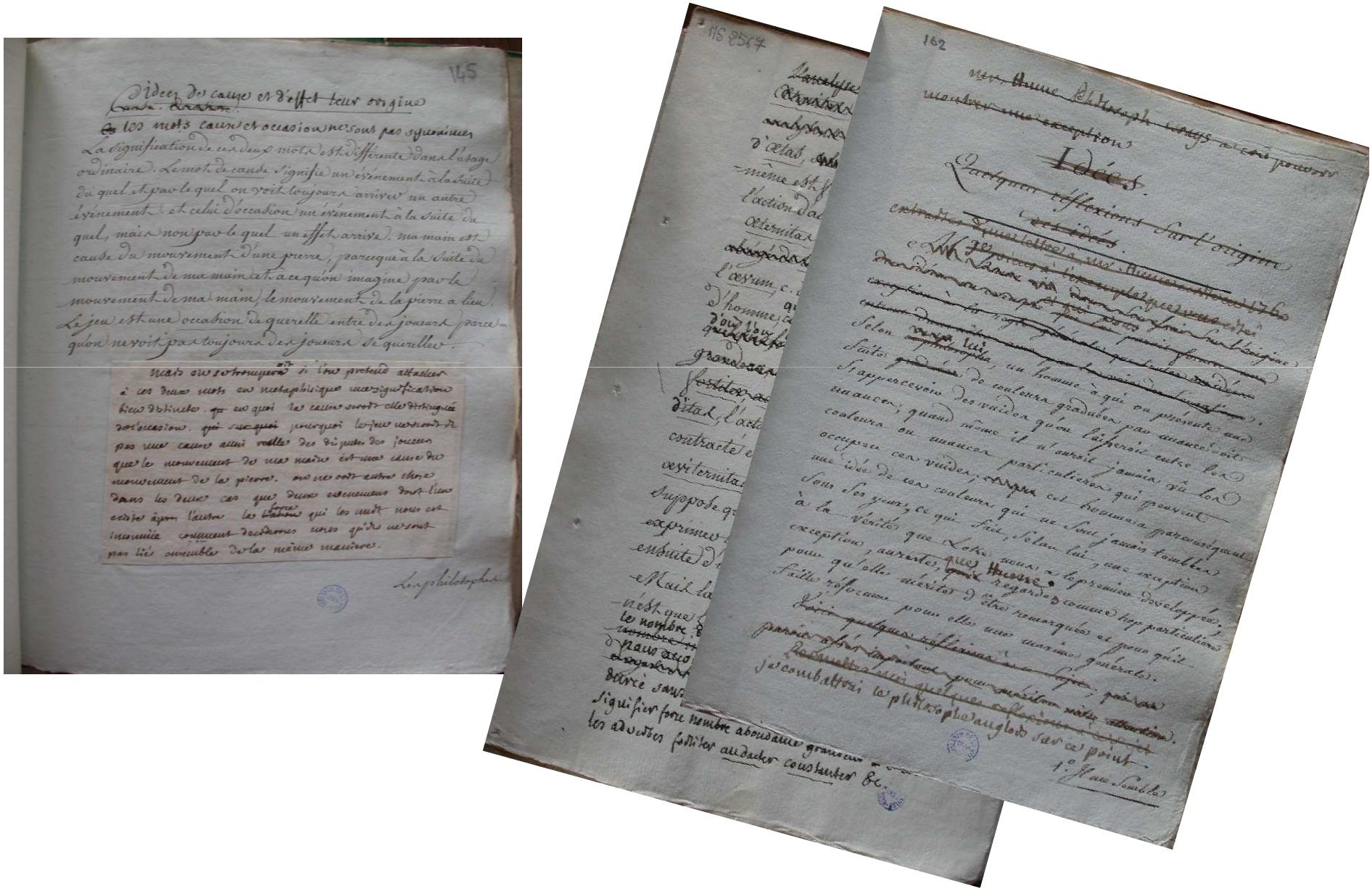


Analyse des manuscrits contemporains

Quelques exemples…

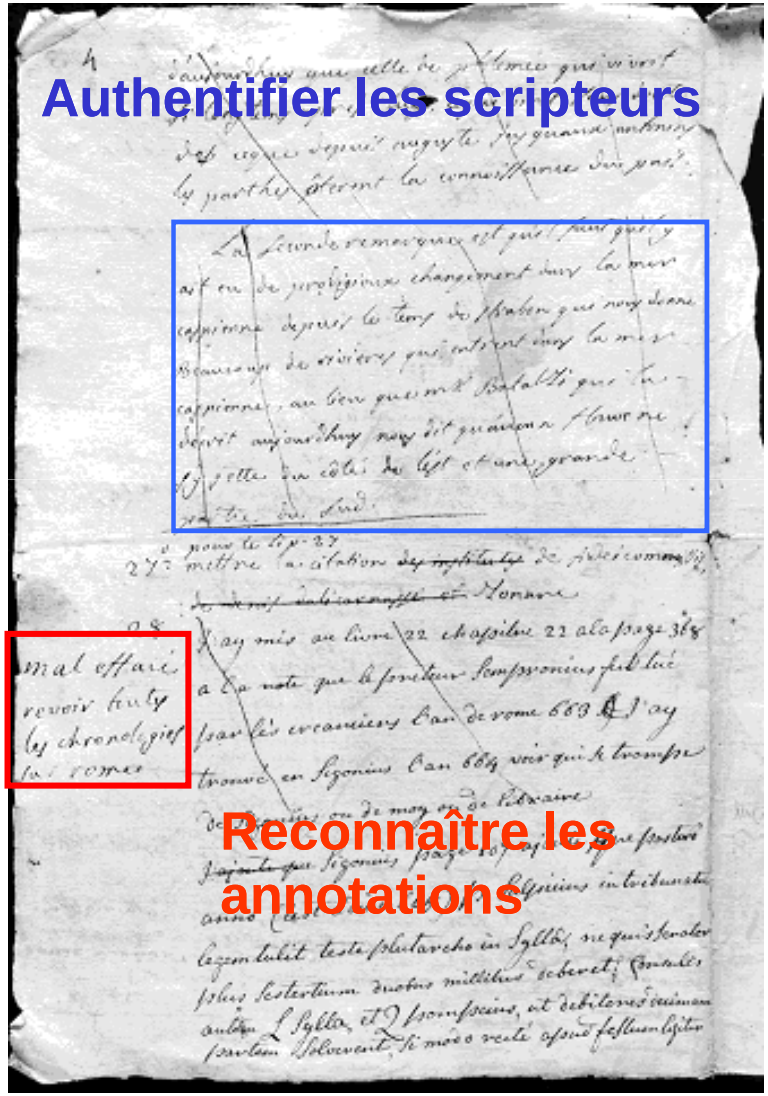
Analyse des manuscrits contemporains

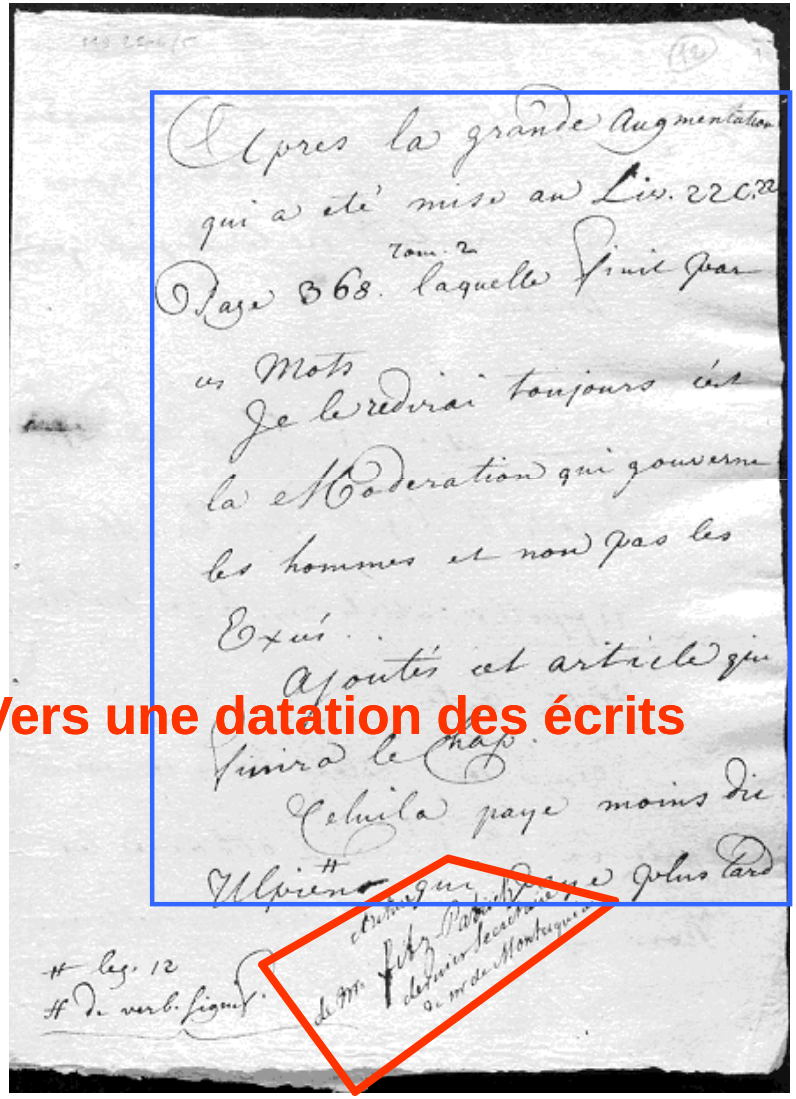
Quelques exemples, Economiste de la fin 18ème (BM Lyon)



Analyse des manuscrits contemporains

Découpage idéal

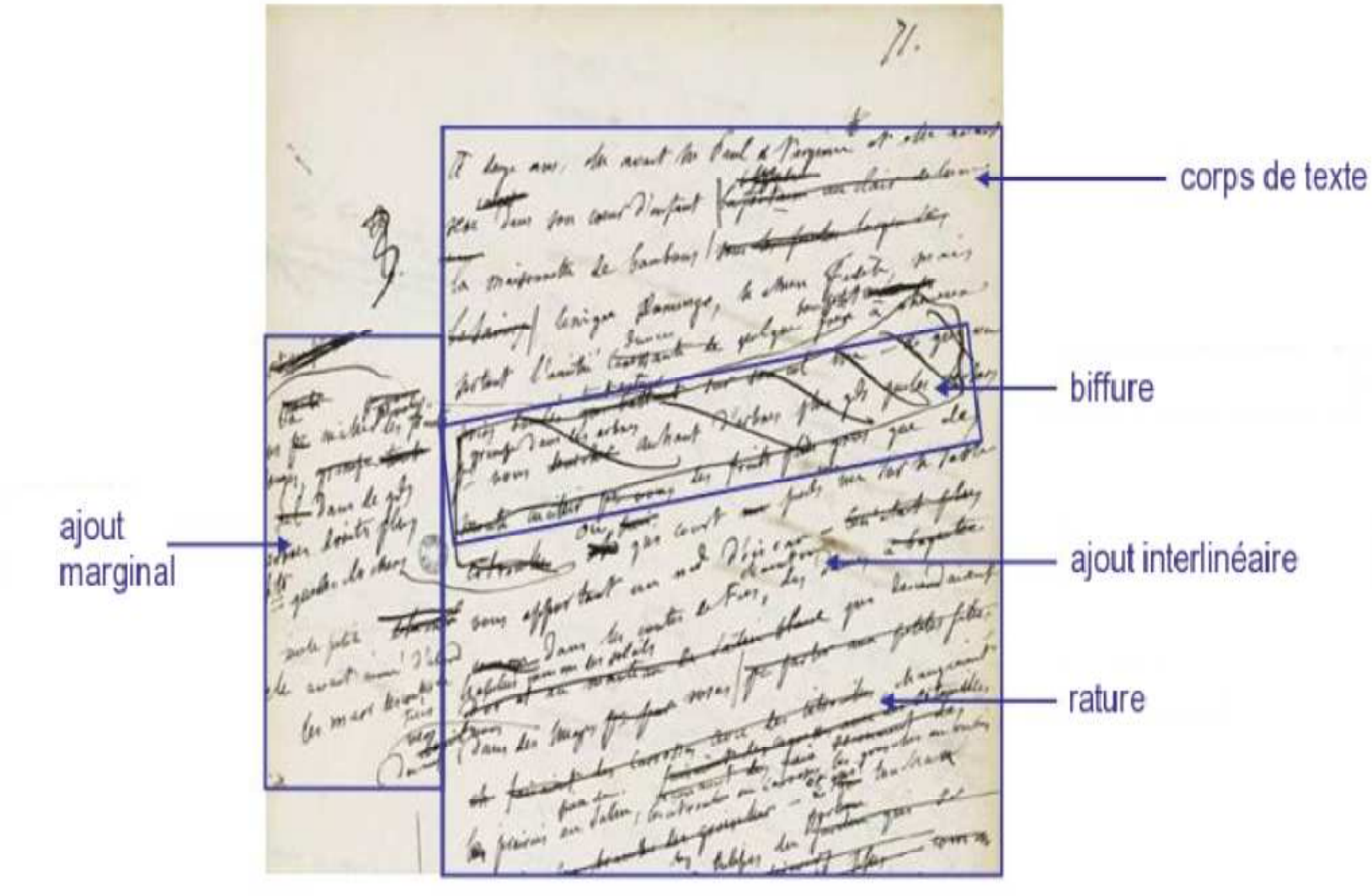




Montesquieu – Extrait de l'Esprit des Lois (1750)

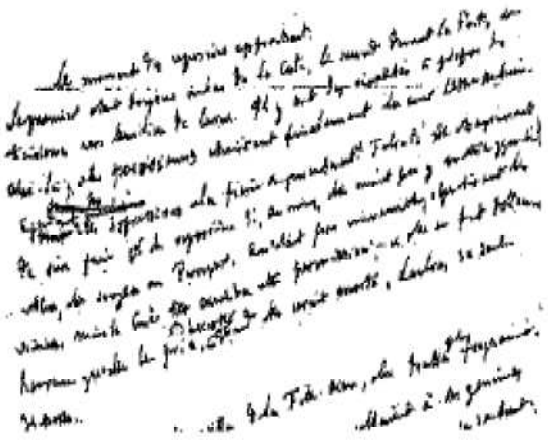
Analyse des manuscrits contemporains

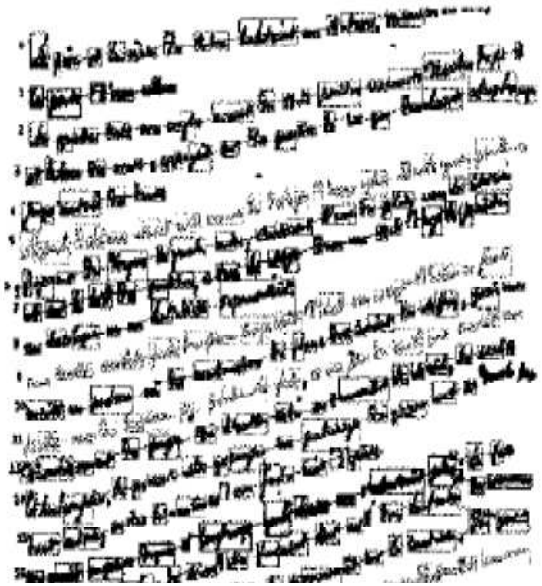
Découpage idéal



G. Flaubert – Madame Bovary

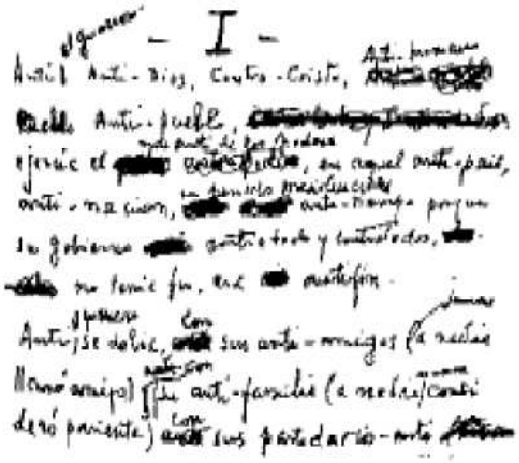
Analyse des manuscrits contemporains

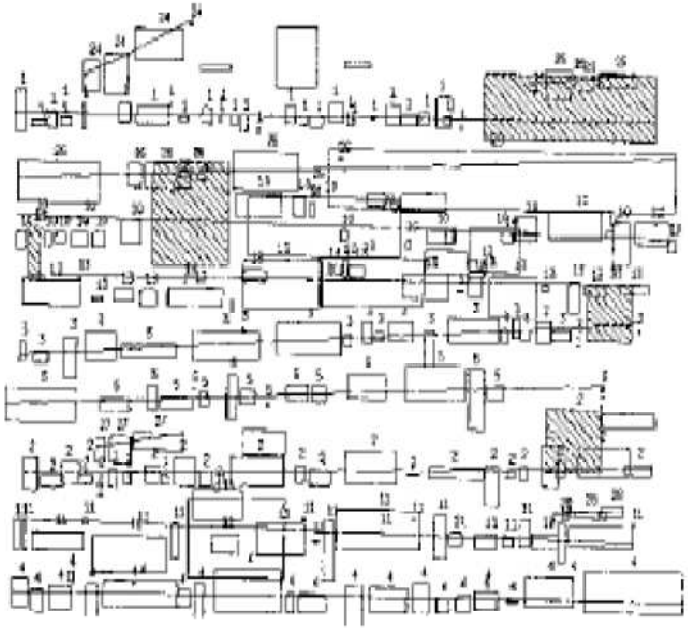


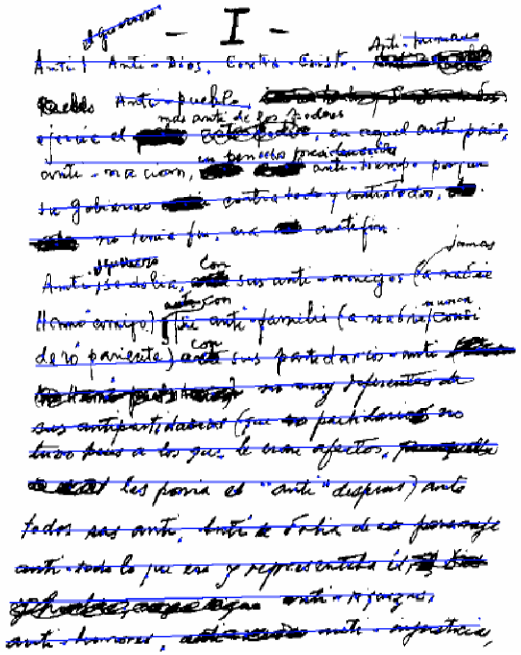


Au niveau des lignes et du bloc de texte

Séparation des lignes. Likforman, 2000

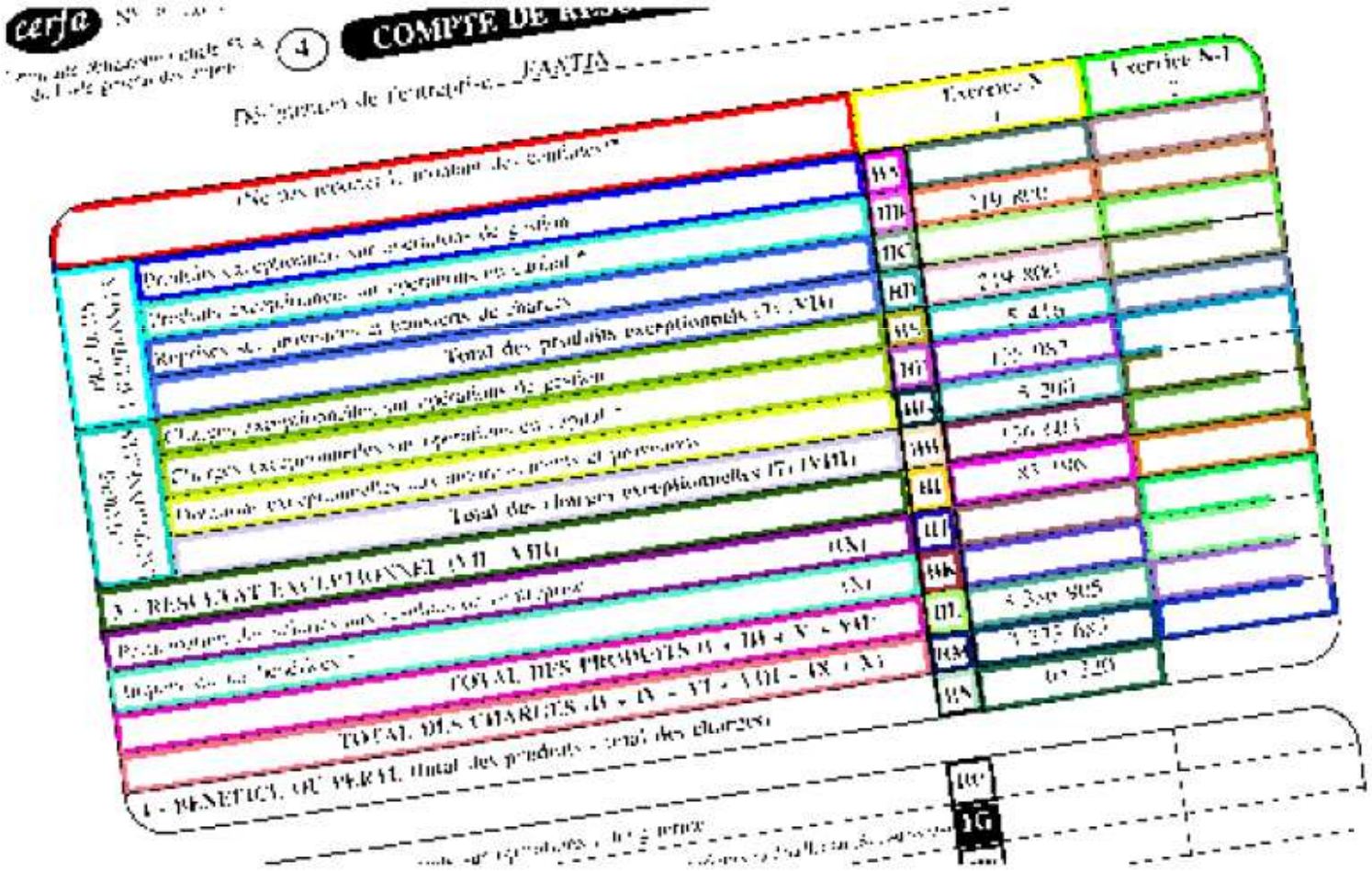






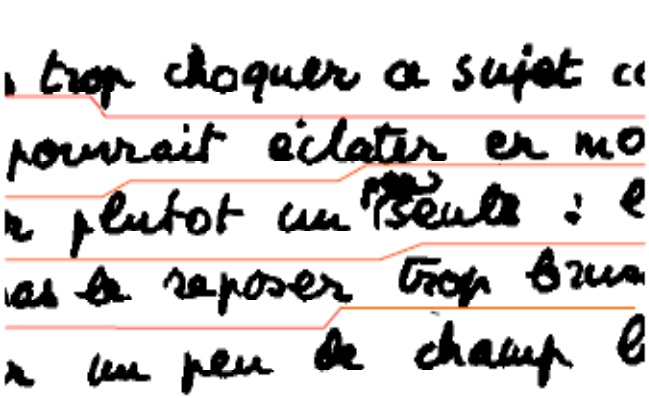
Analyse des manuscrits contemporains

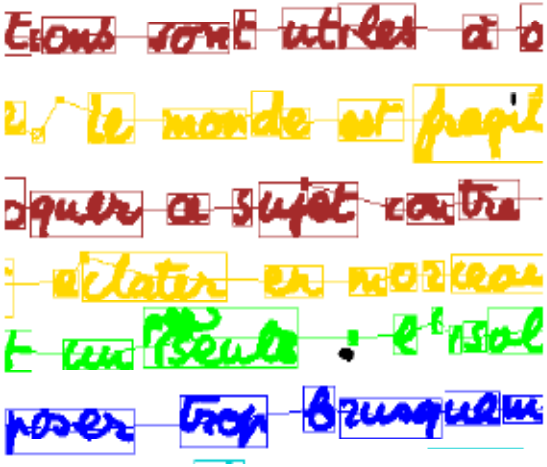
Hough + Détection de cellules

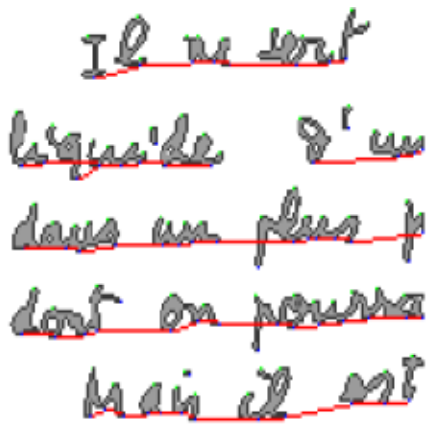


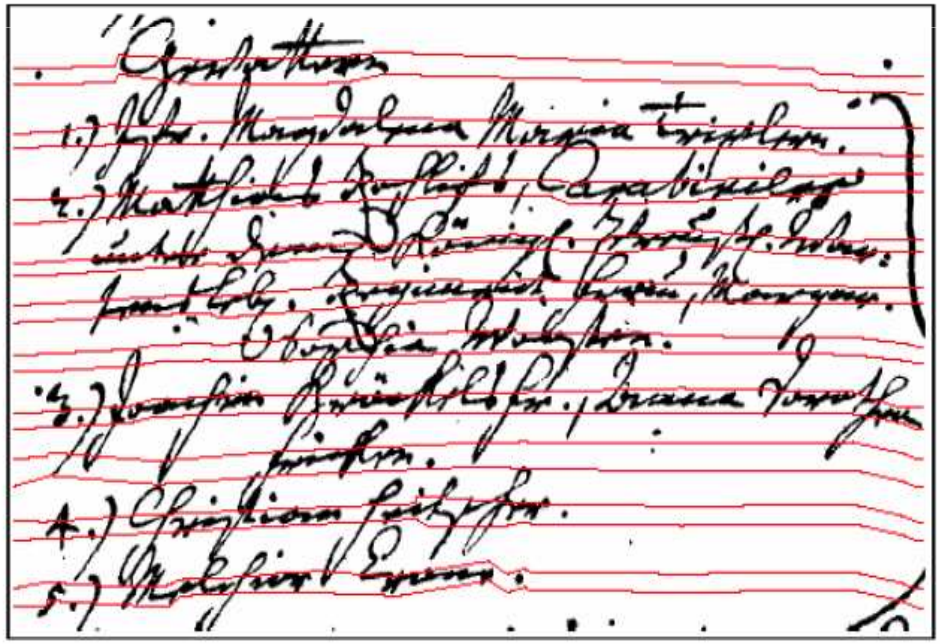
Analyse des manuscrits contemporains

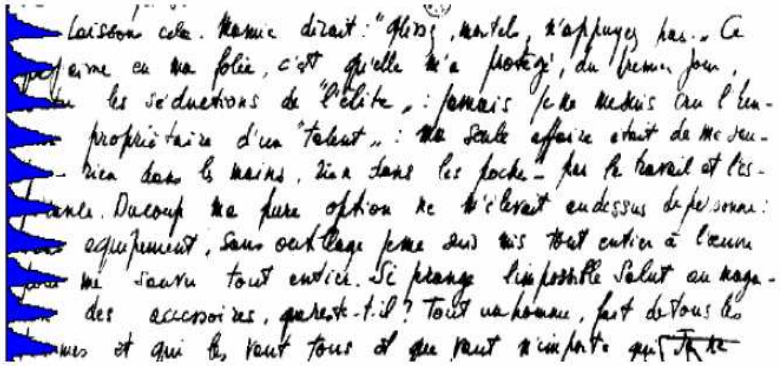
Au niveau des lignes et du bloc de texte

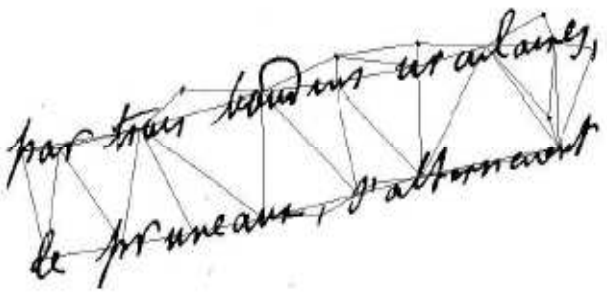






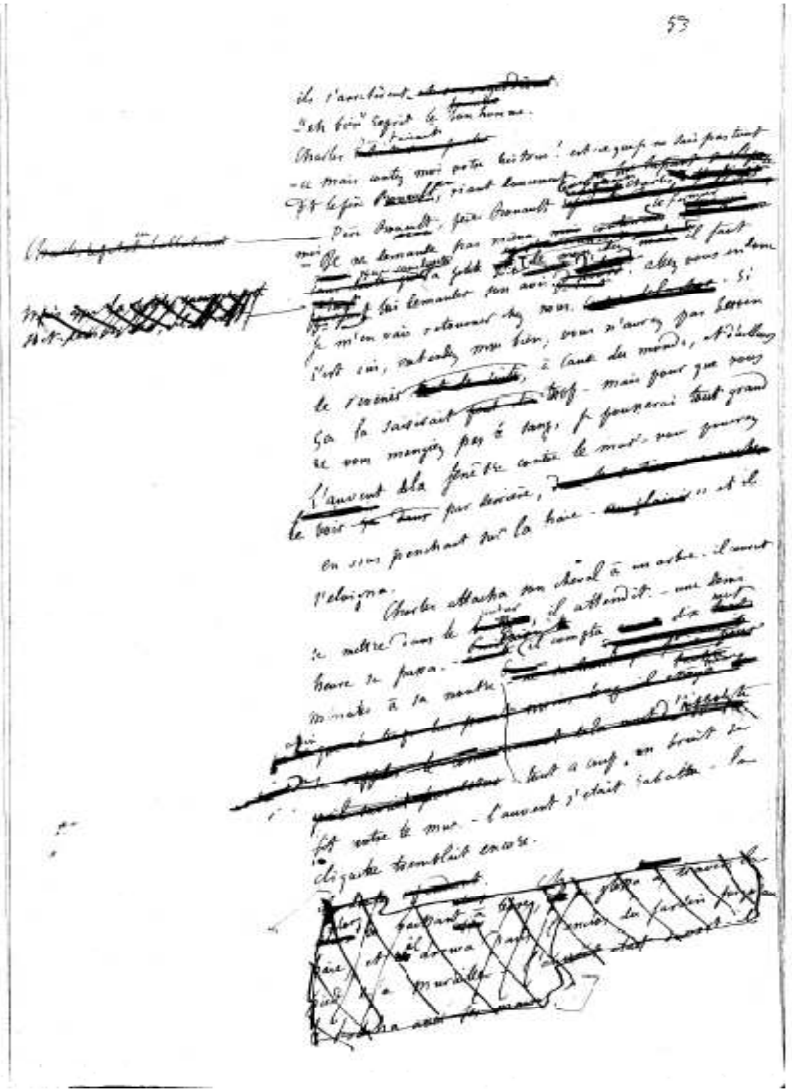


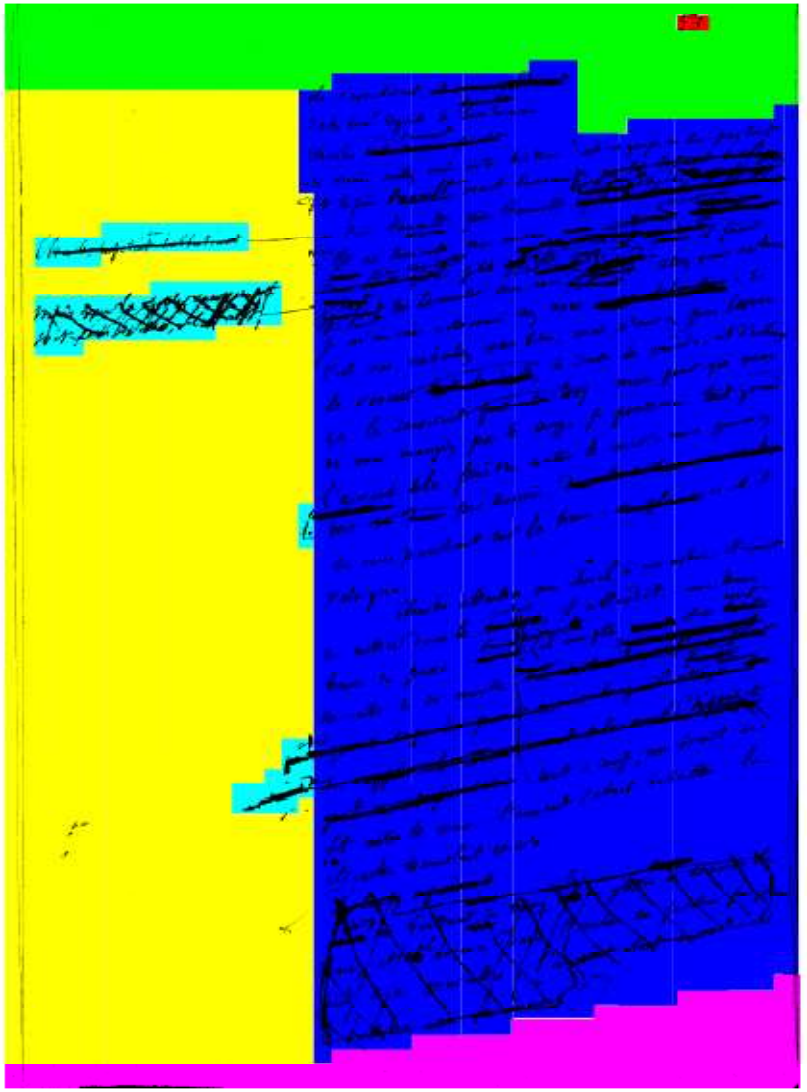




Analyse des manuscrits contemporains

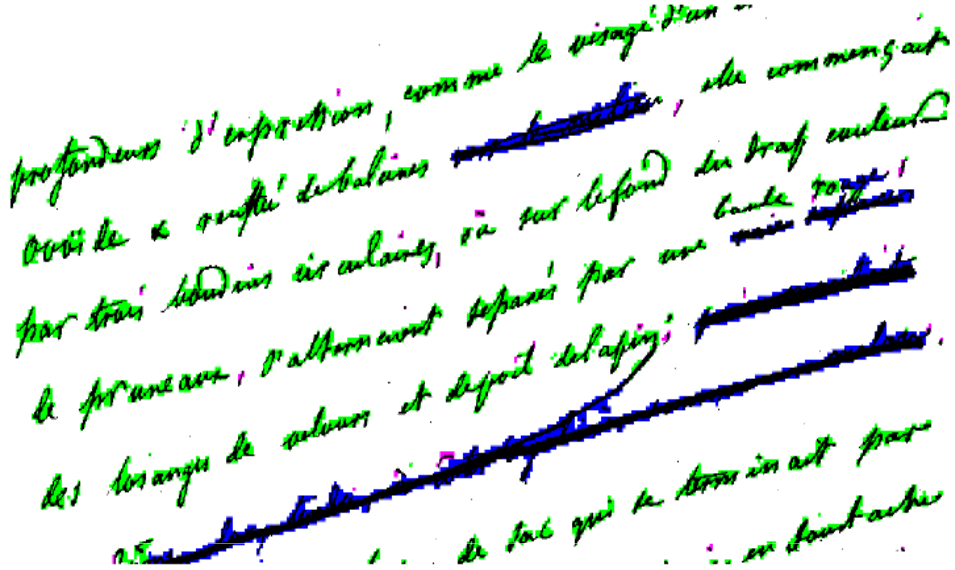
Au niveau des lignes et du bloc de texte



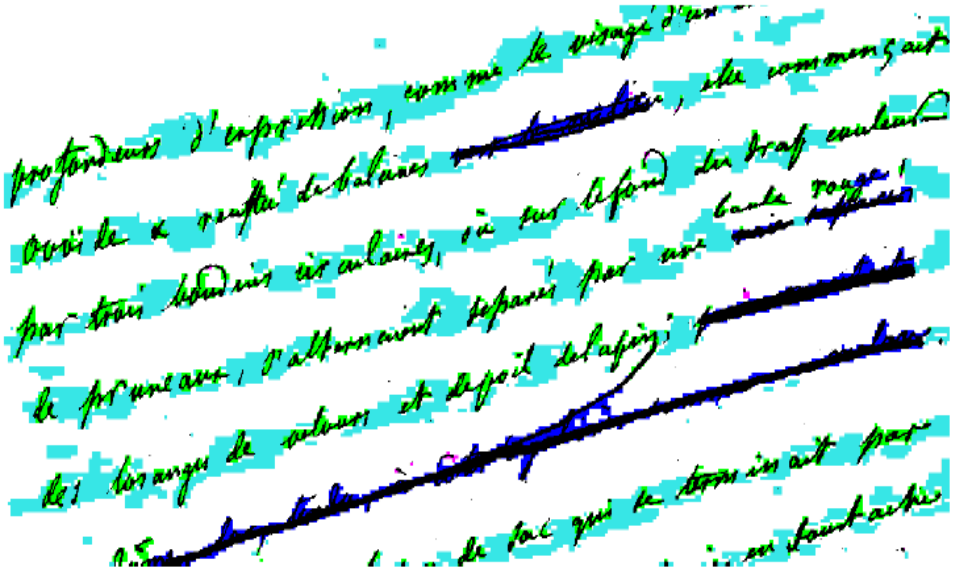


Application à l'analyse des manuscrits

Au niveau des lignes et du bloc de texte



(a)

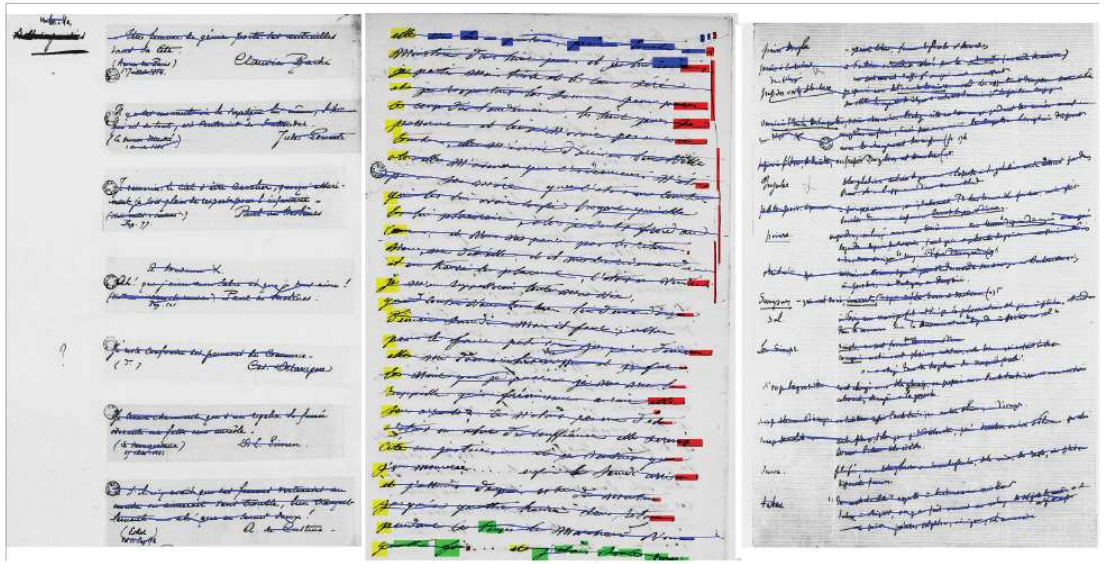


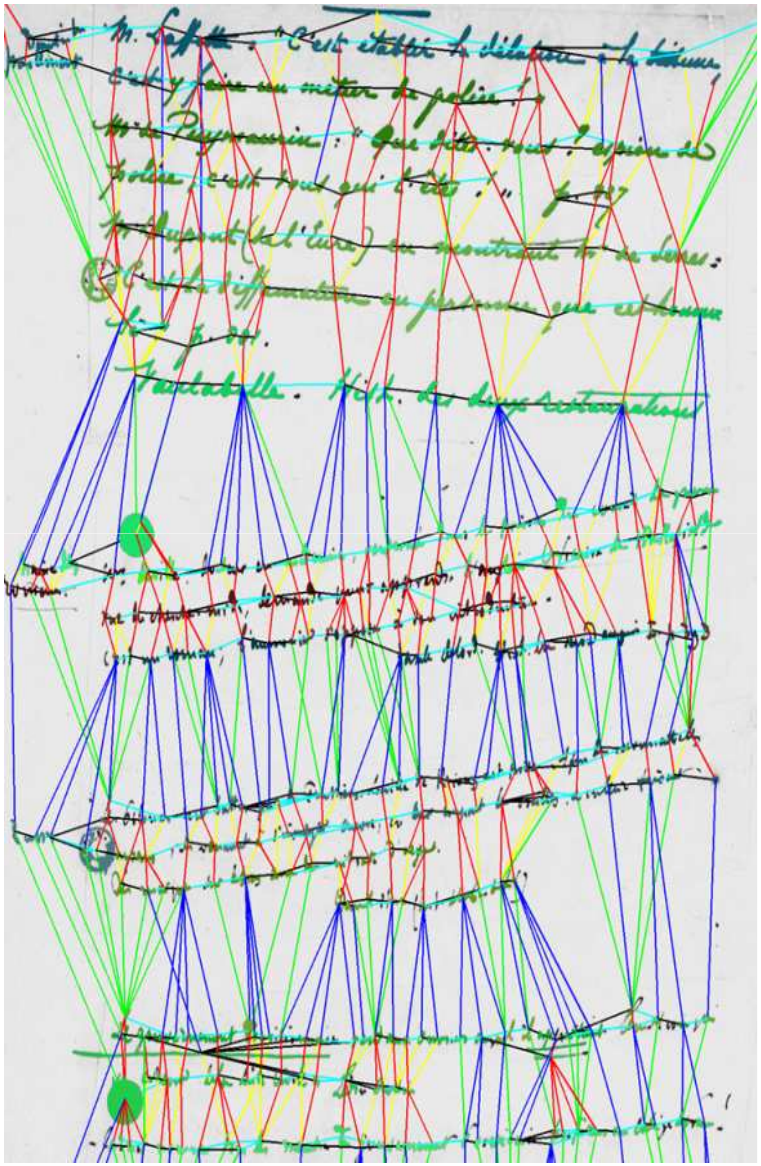
(b)

Application à l'analyse des manuscrits de Flaubert

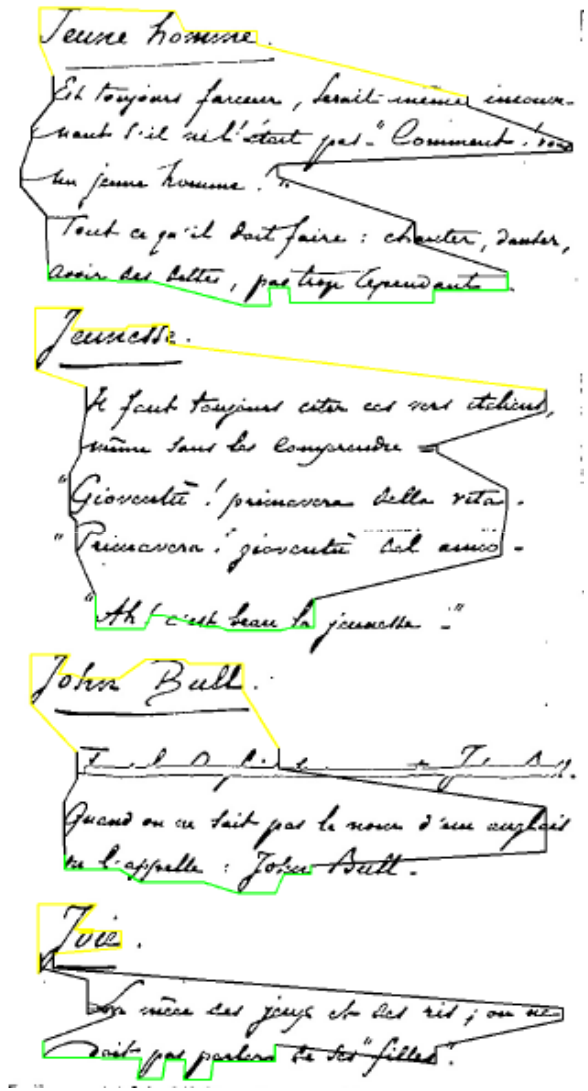
Extraction des lignes

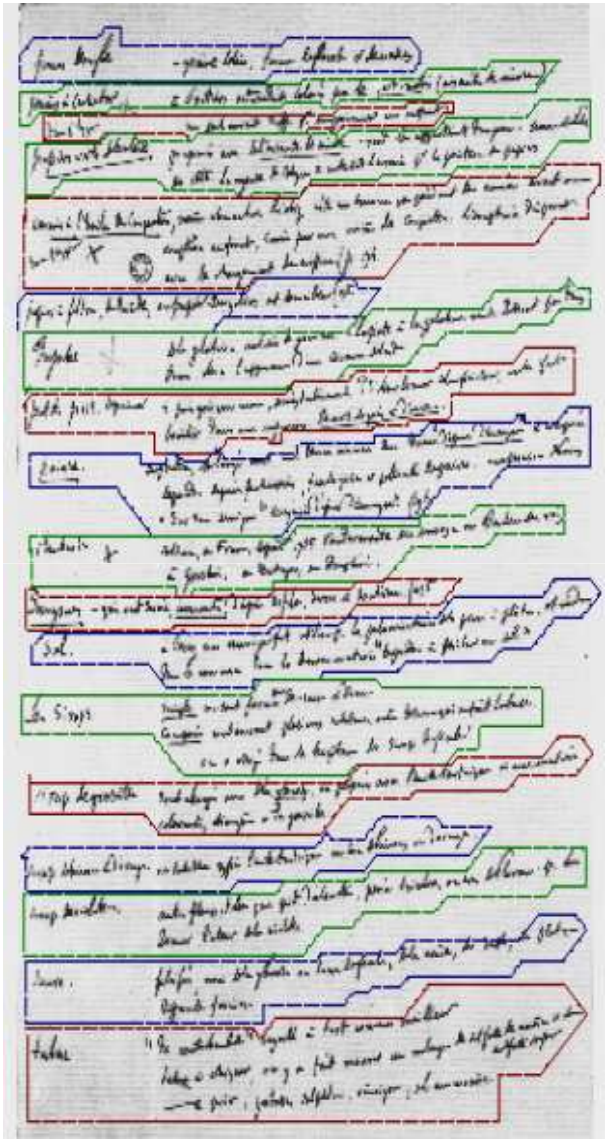
- Extraction des composantes connexes
- Utilisation de la transformée de Hough
- Construction d'un graphe géométrique décrivant les relations spatiales entres composantes
- Segmentation du graphe en n sous-graphes correspondant aux lignes

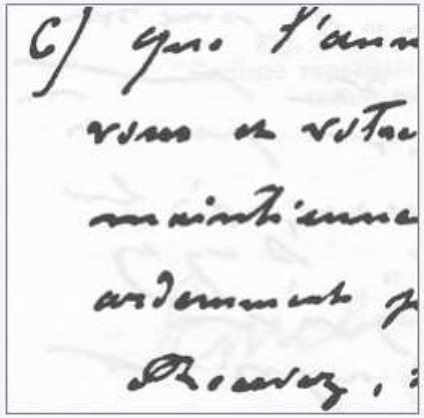


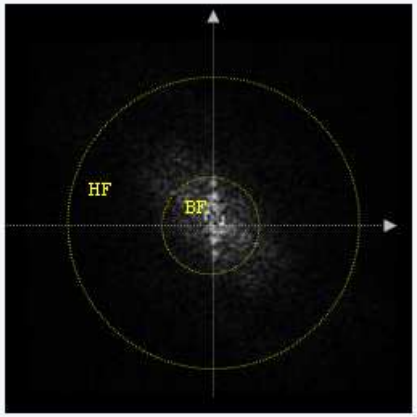


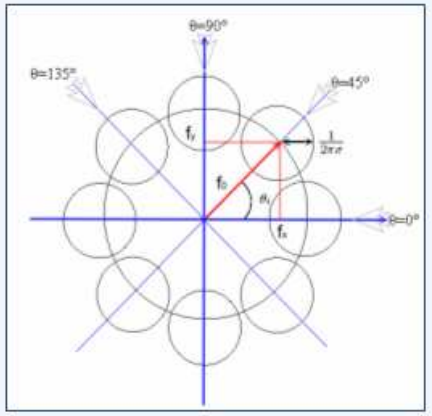
Application à l’analyse des manuscrits de Flaubert

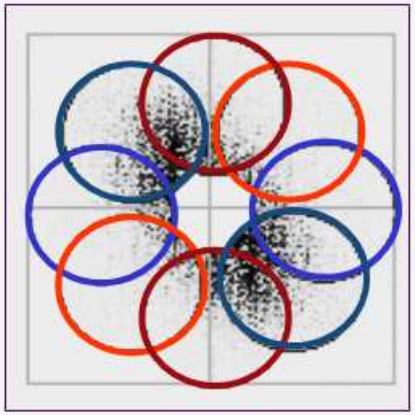




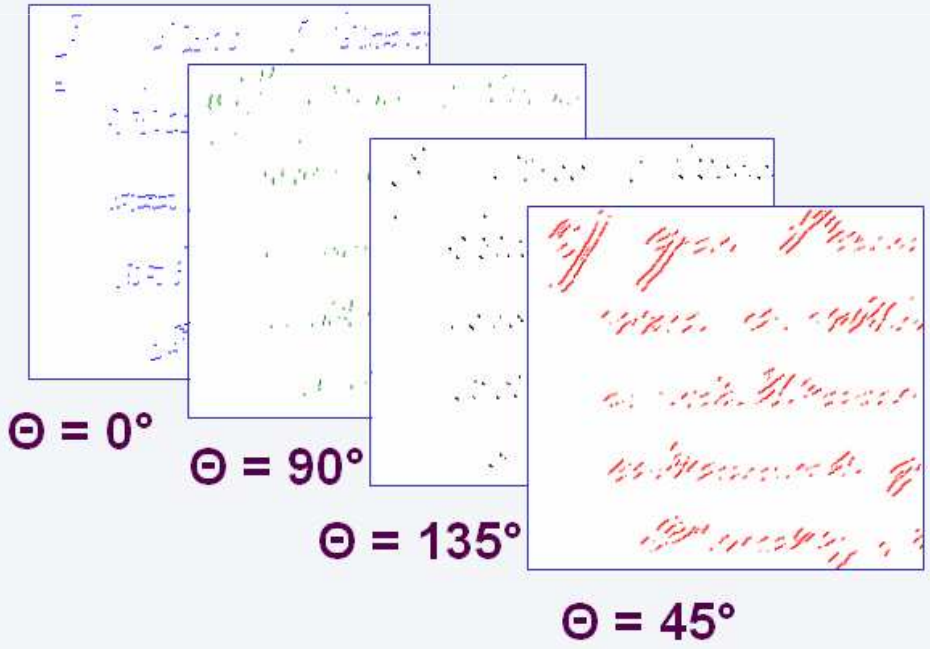


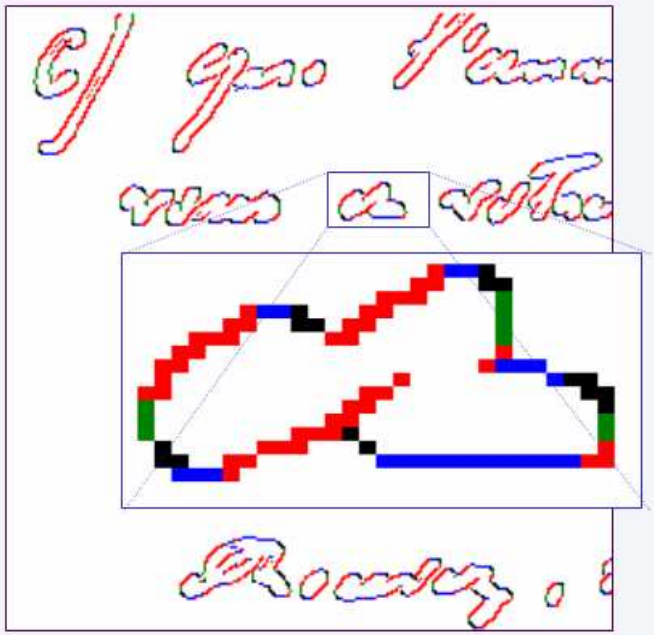






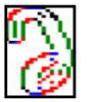
Gabor filtering – outlines marking and directional maps decomposition

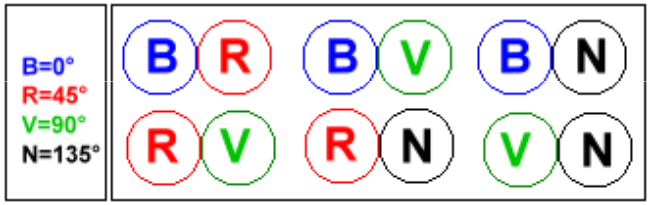




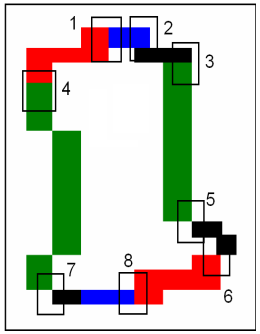
Forme recherchée

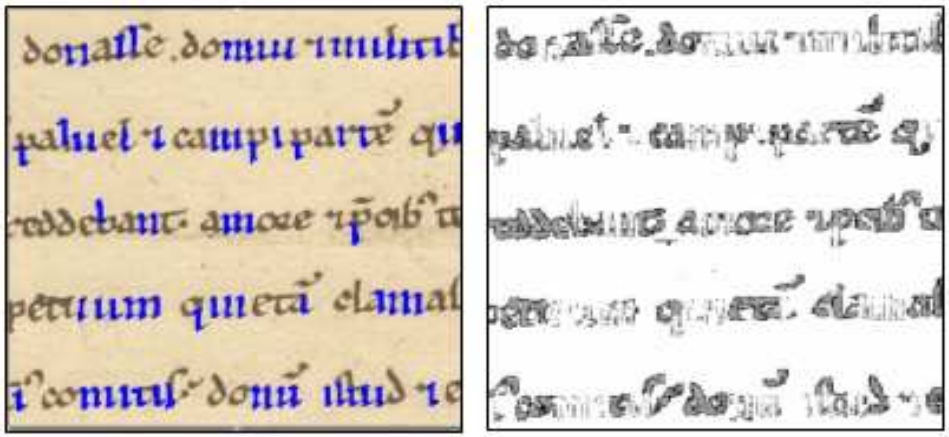
A partir de sa version codée

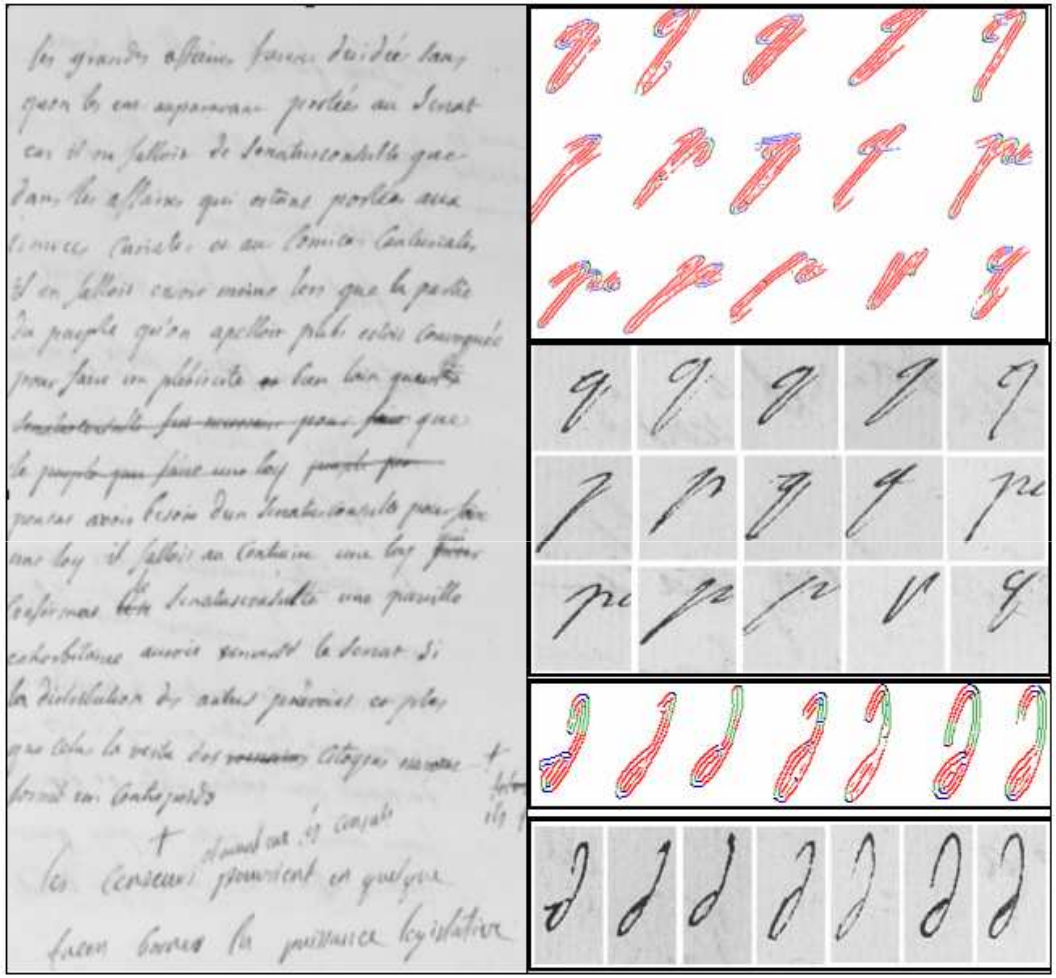




Junctions coding

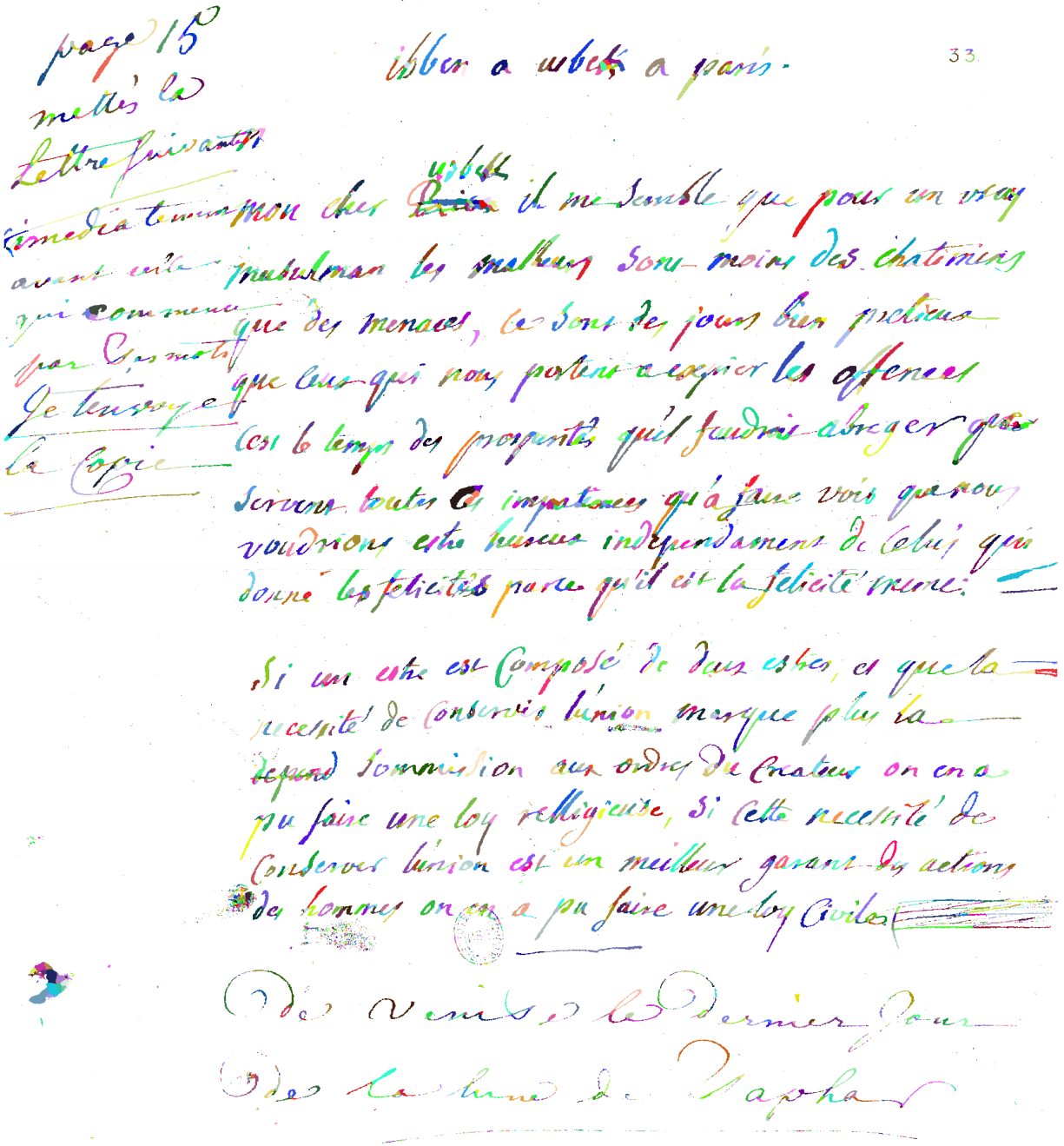


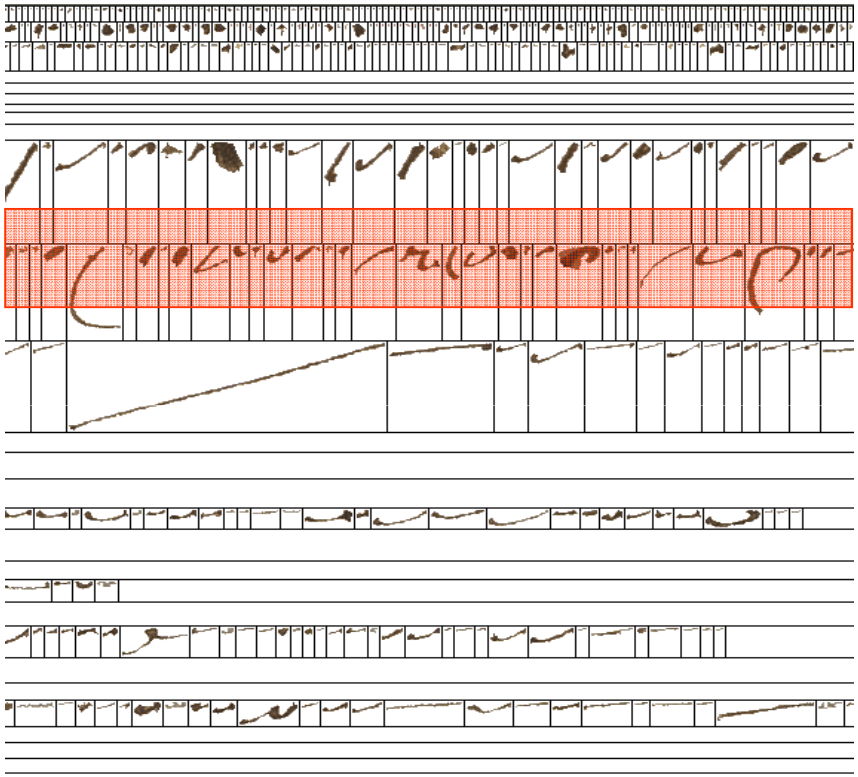




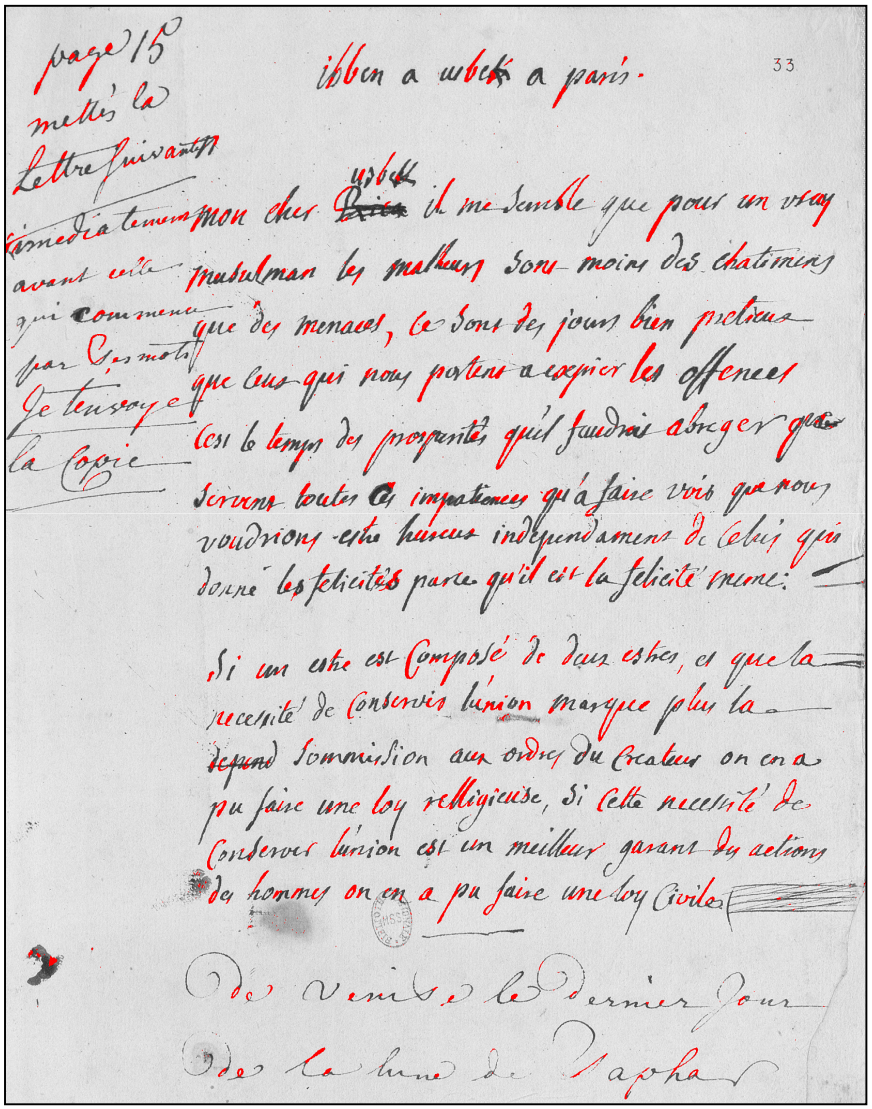
- Vers une caractérisation des écritures par invariants par le calcul des occurrences de formes

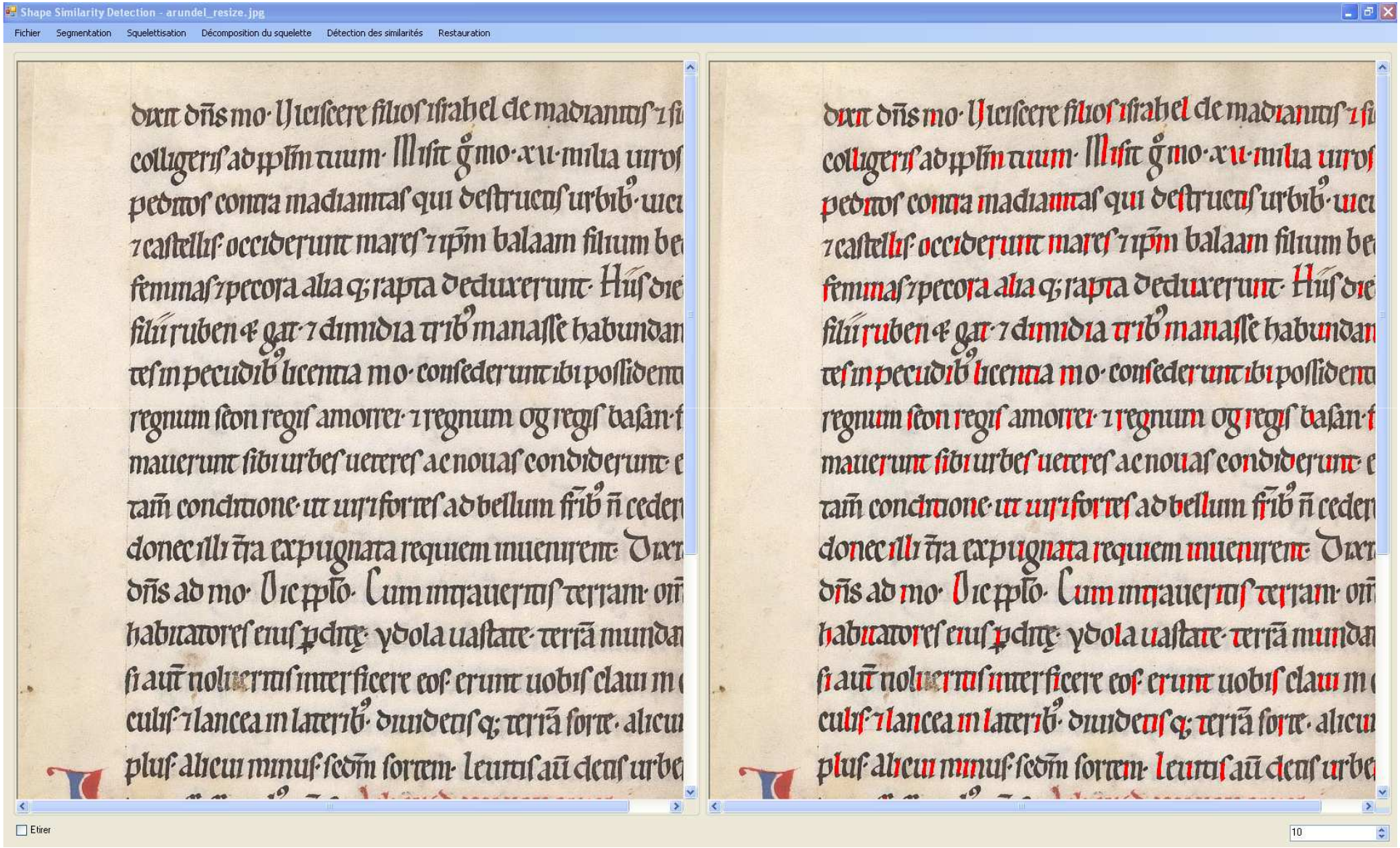
Montesquieu – Esprit des Lois



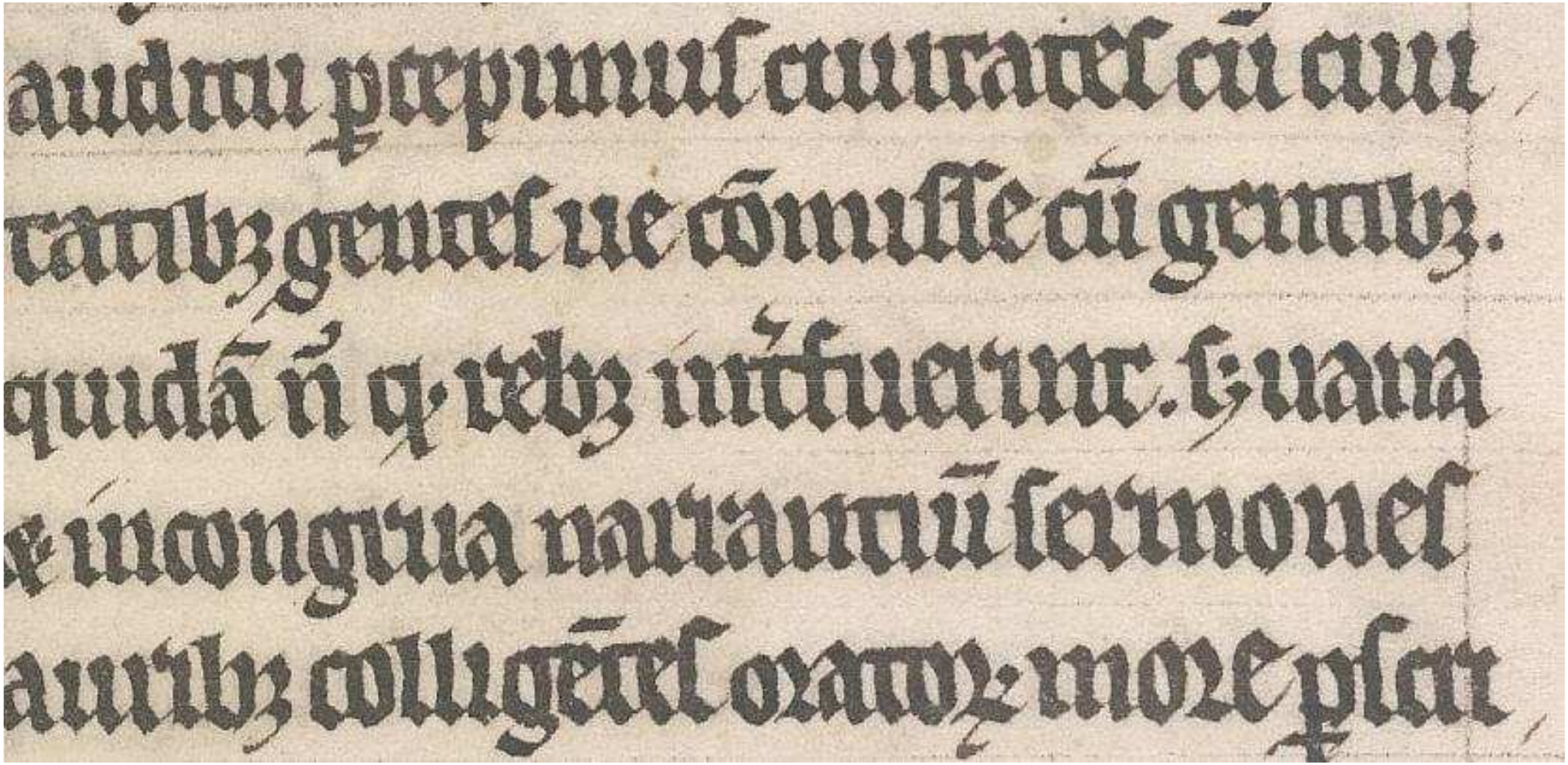


Exemple de tables de similarités à partir d'une décomposition en 31 graphèmes

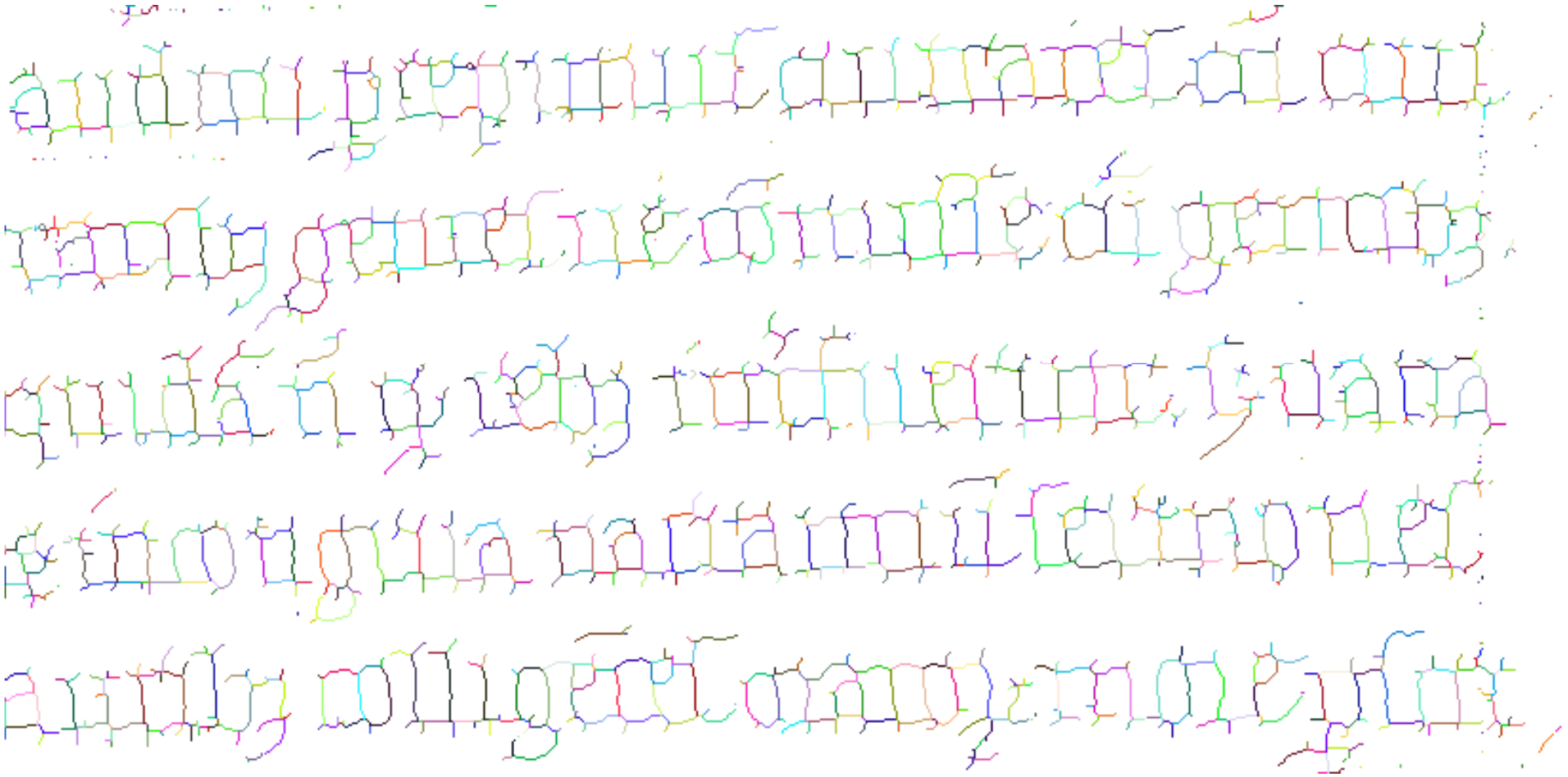


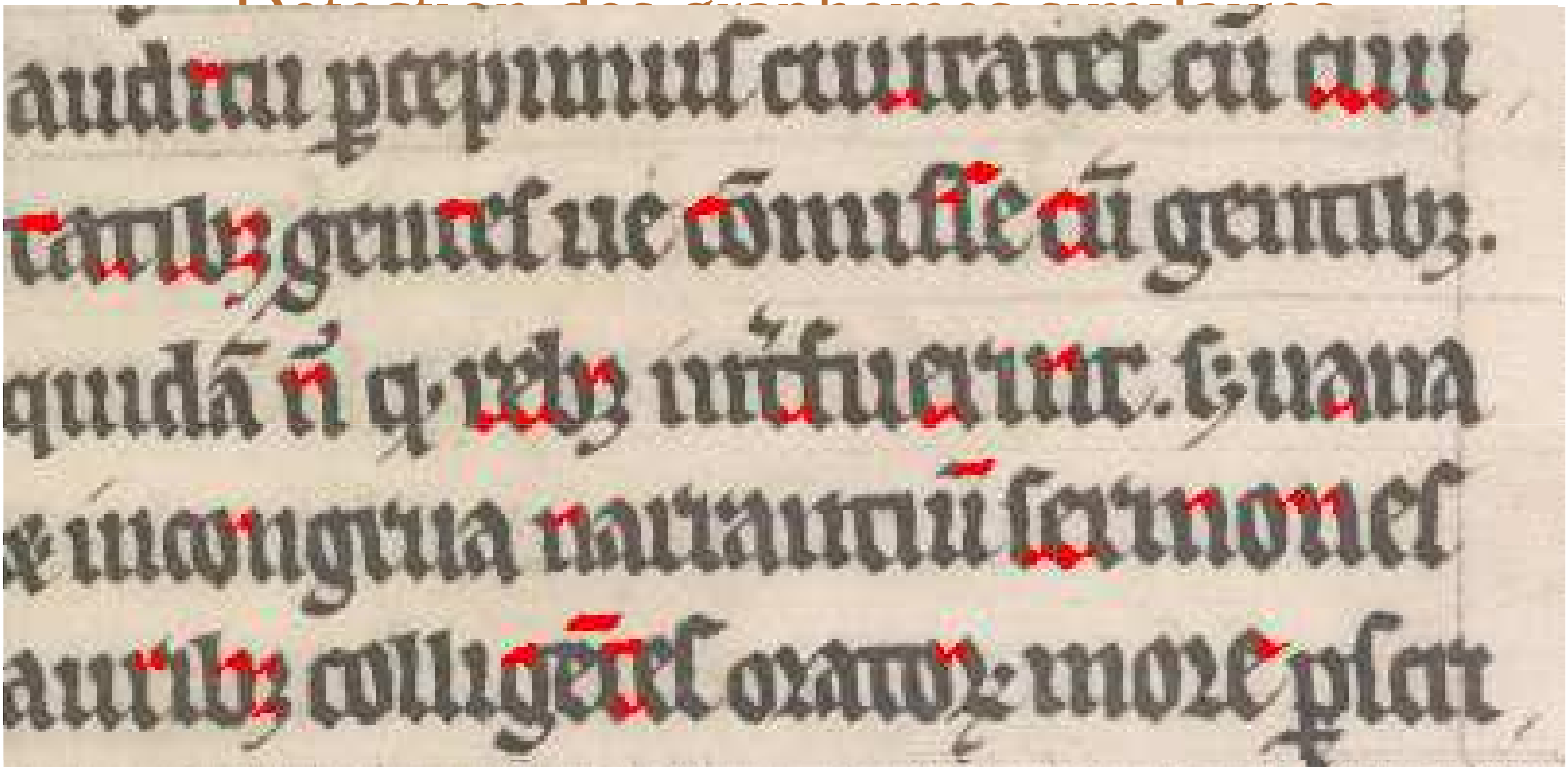

Shape Similarity Detection - arundel_resize.jpg
Fichier
Segmentation
Squelettisation
Décomposition du squelette
Détection des similarités
Restauration
Etirer
10

Détection des graphèmes similaires



Détection des graphèmes similaires

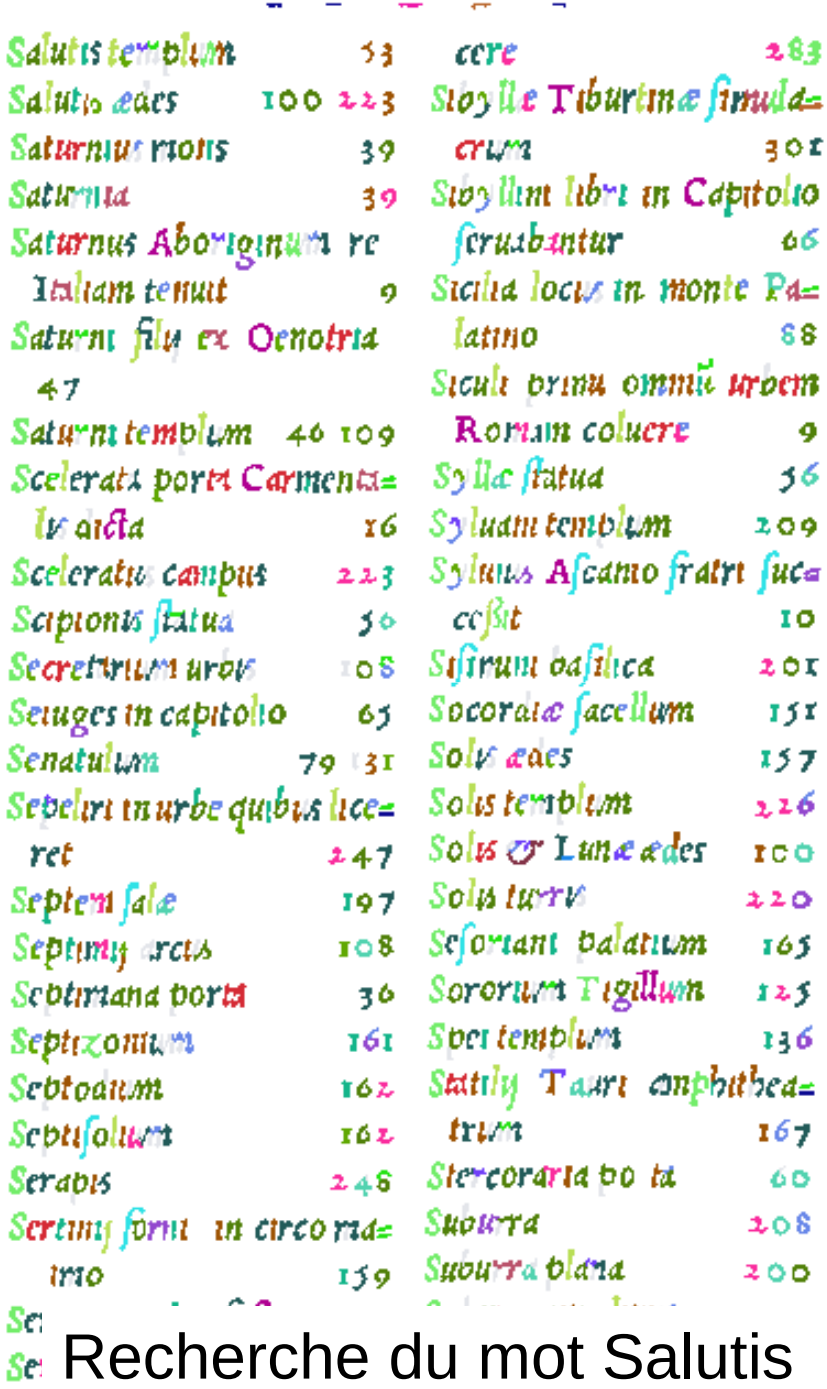




Application à la recherche de mots en mode image

Recherche du mot Salutis

Application à la recherche de mots en mode image

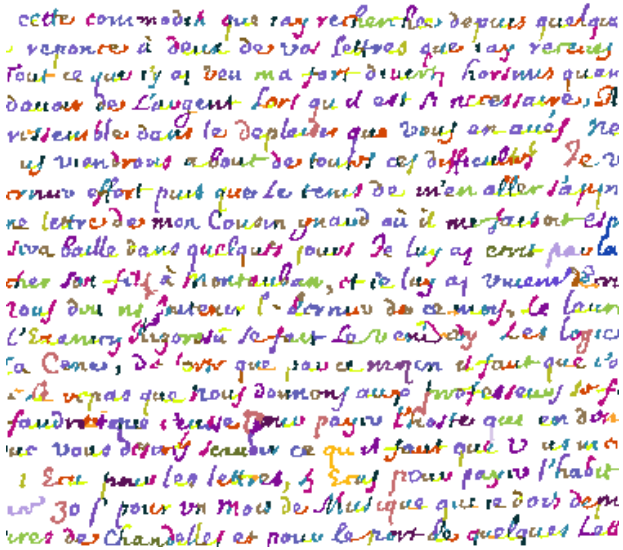


Recherche du mot Salutis

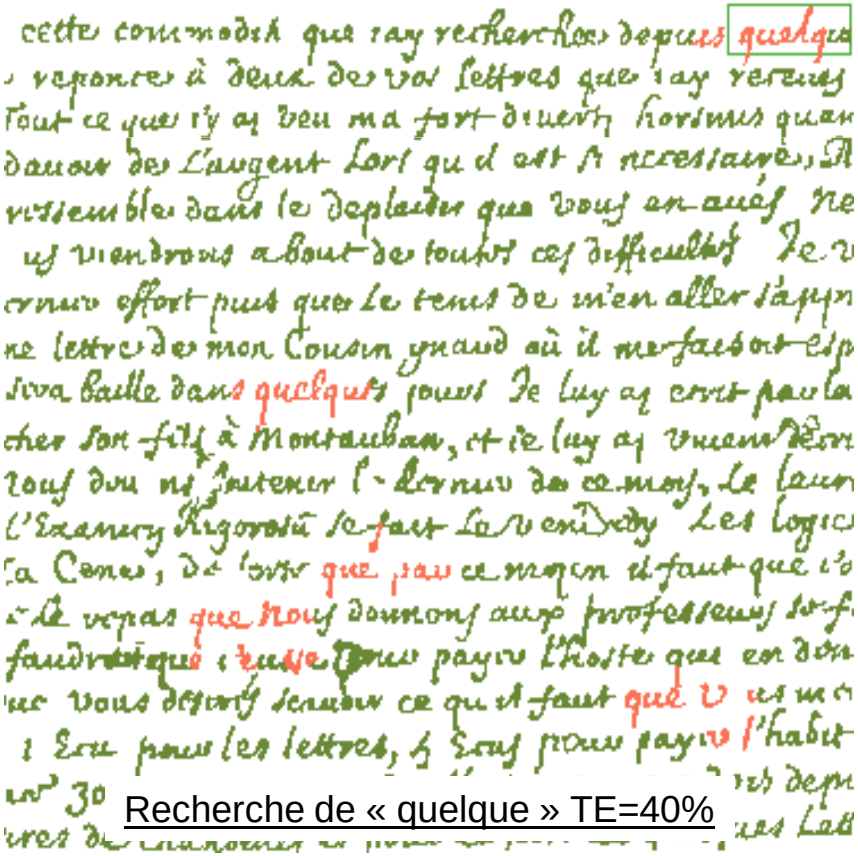
Application à la recherche de mots en mode image

Recherche du mot Salutis

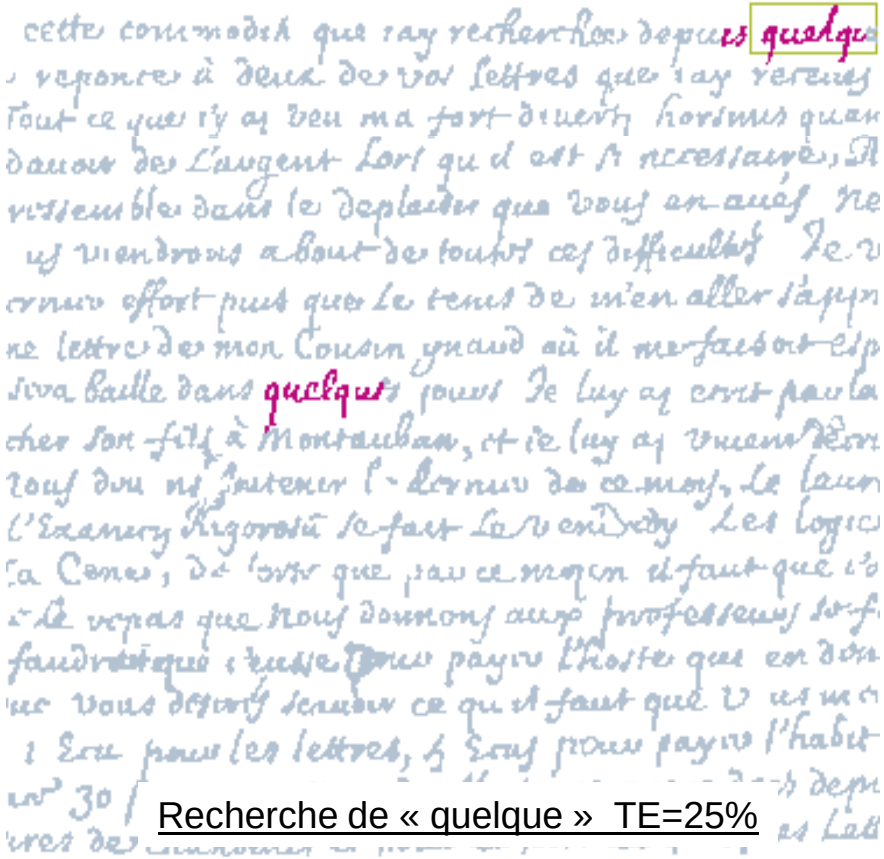
Application à la recherche de mots en mode image



Manuscrits des Lumières, 18è
ENS Lettres, A. McKenna



Recherche de « quelque » TE=40%

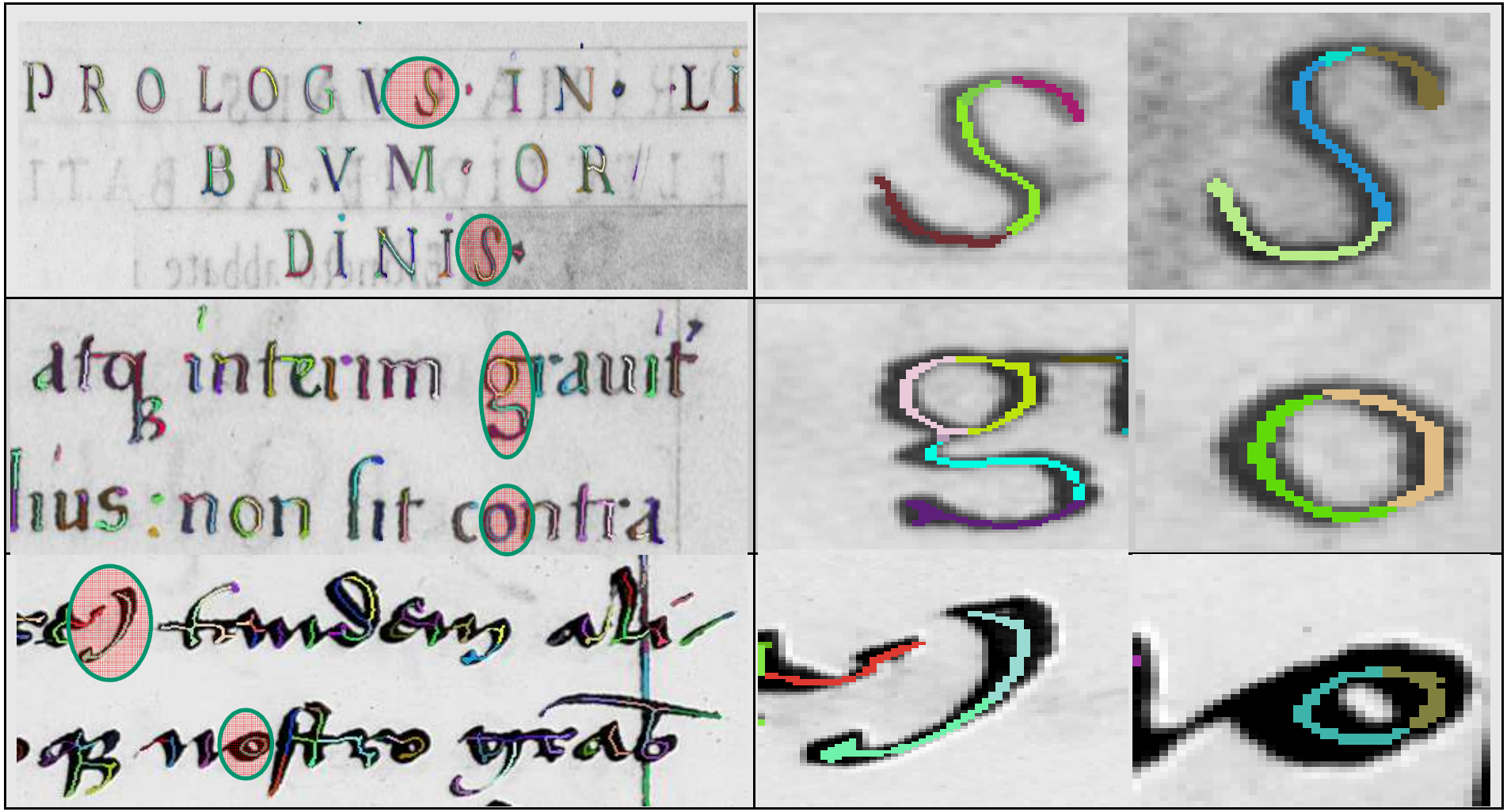


Recherche de « quelque » TE=25%

Cas d'étude avancé:

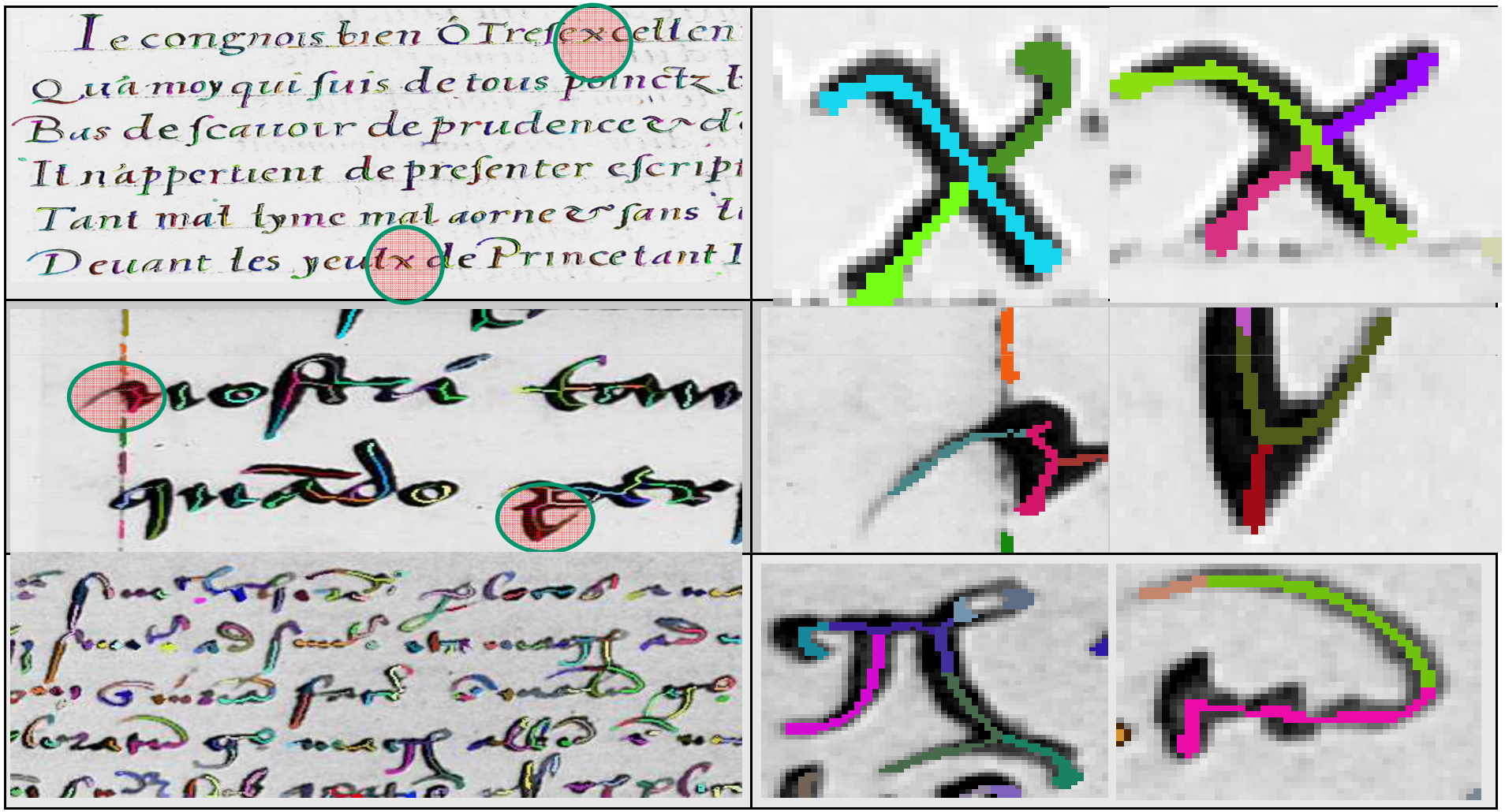
Décomposition selon l'exécution des formes

Et la localisation des points de « posers » et « levers » de plume



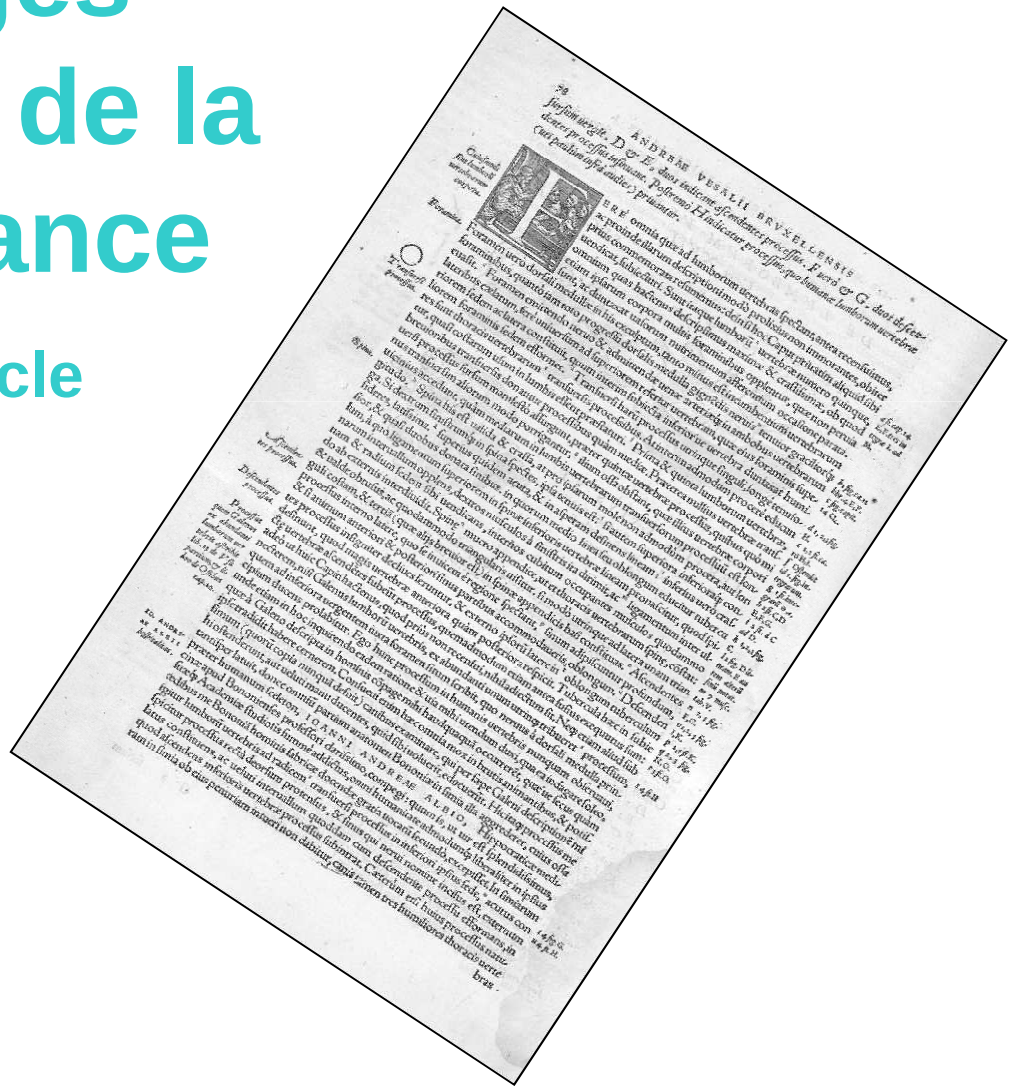
Contribution: Décomposition des manuscrits en graphèmes

Illustration: croisements et cas d'erreur

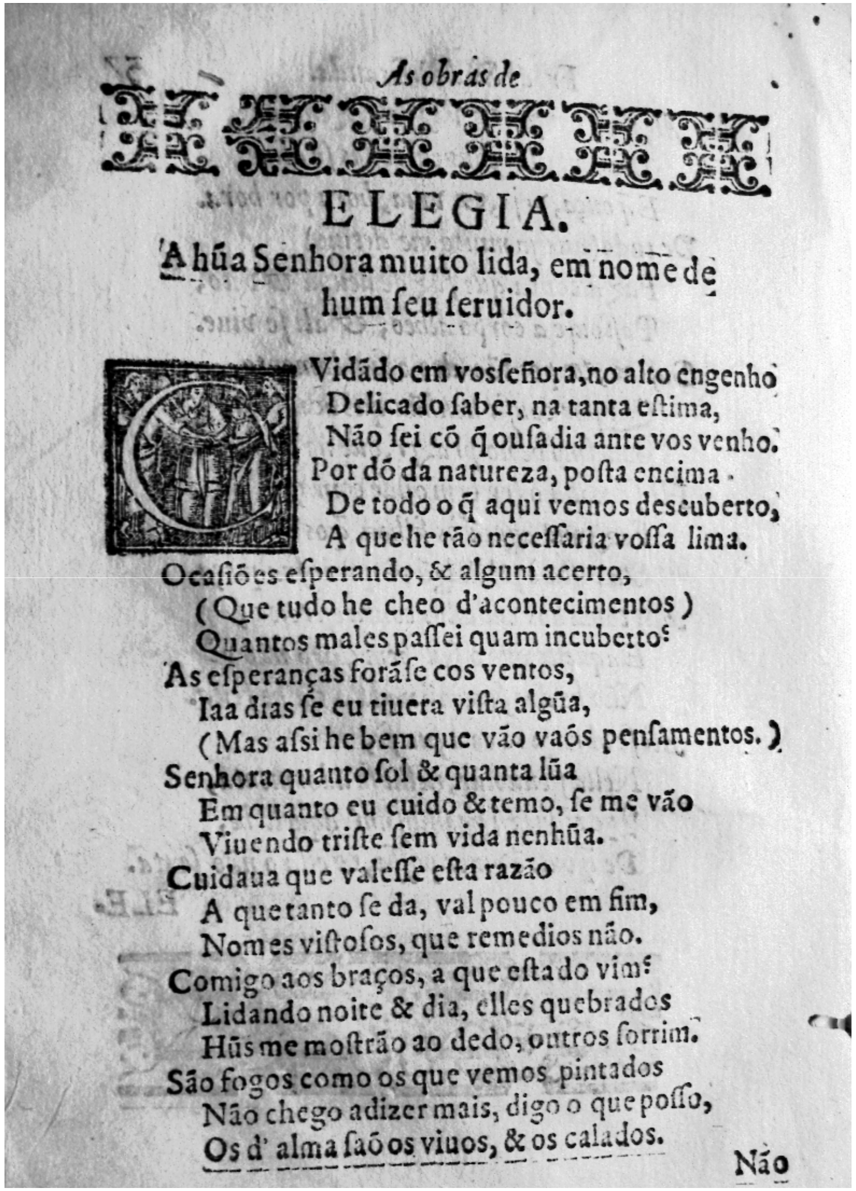


Ouvrages imprimés de la Renaissance

16ème siècle



Imprimés de la Renaissance XVI


As obrás de

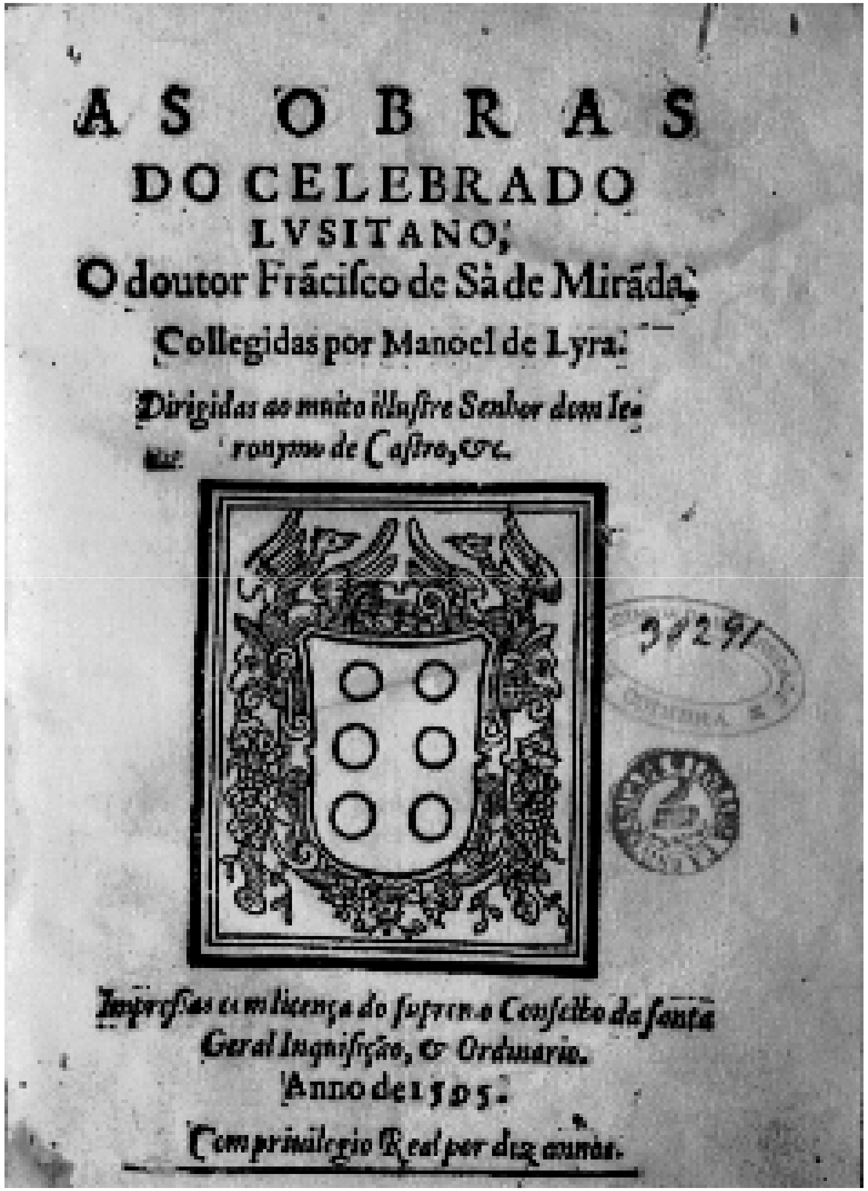
ELEGIA.

A hũa Senhora muito lida, em nome de hum ſeu ſeruidor.

CVidãdo em voſſẽñora, no alto engenho
Delicado ſaber, na tanta eſtima,
Não ſei cõ q̃ ouſadia ante vos venho.
Por dó da natureza, poſta encima
De todo o q̃ aqui vemos deſcuberto,
A que he tão neceſſaria voſſa lima.
Ocaſiões eſperando, & algum acerto,
(Que tudo he cheo d'acontecimentos)
Quantos males paſſei quam incuberto?
As eſperanças forãſe cos ventos,
Iaa dias ſe eu tiuera viſta algũa,
(Mas aſsi he bem que vão vaõs penſamentos.)
Senhora quanto ſol & quanta lũa
Em quanto eu cuido & temo, ſe me vão
Viuendo triſte ſem vida nenhũa.
Cuidaua que valeſſe eſta razão
A que tanto ſe da, val pouco em fim,
Nomes viſtoſos, que remedios não.
Comigo aos braços, a que eſta do vim?
Lidando noite & dia, elles quebrados
Hũs me moſtrão ao dedo, outros ſorrim.
São fogos como os que vemos pintados
Não chego adizer mais, digo o que poſſo,
Os d' alma ſaõ os viuos, & os calados.

Não

Déformations géométriques


AS OBRAS
DO CELEBRADO
LVSITANO,
O doutor Frãciſco de Sà de Mirãda.
Collegidas por Manoel de Lyra.
Dirigidas ao muito illuſtre Senhor dom Ieronymo de Caſtro, &c.

Impreſſas com licença do ſupremo Conſelho da ſanta Geral Inquiſição, & Ordinario.
Anno de 1595.
Com priuilegio Real por dez annos.

Dégradations

Imprimés de la Renaissance XVI

Méta-données recherchées : riches et complexes

Structure du livre *(pages de titres, frontispice, pièces préliminaires, pages de texte, Index, tables, pièces finales..)*

Transcription fidèle *(manuelle, OCR ou TAO)*

Structure physique *(texte principal/annotations, blocs, lignes, mots, position de tous les objets non textuels)*

Identification des objets graphiques *(Lettrines, enluminures,Frontispices, illustrations, bandeaux, colophons, marques de l'imprimeur, culs-de-lampe, miniatures…)*

Informations textuelles *(signatures, titres, typographie, notes, pagination, foliotation)*

Mots-clés comme « incipit »…

…. **liste non exhaustive**

Digital **A**cc**E**ss to **BO**ok of the **R**en**A**issance

- Analyse des usages des utilisateurs d'ouvrages du 16ème siècle (chercheurs, historiens, lecteur éclairé, néophytes…)
- Spécification et développement d'une chaîne de numérisation,
- Etude des coûts de la numérisation.
- Traitement et analyse des images (restoration, segmentation, interprétation, compression)
- Conception de postes de consultation

Développement d'une plate-forme d'aide à l'analyse

DEBORA

▶ Utiliser des méthodes **d'analyse d'images de documents** pour à la fois les **comprimer** et les **indexer**

▶**Extraction automatique de méta-données** par analyse d'image

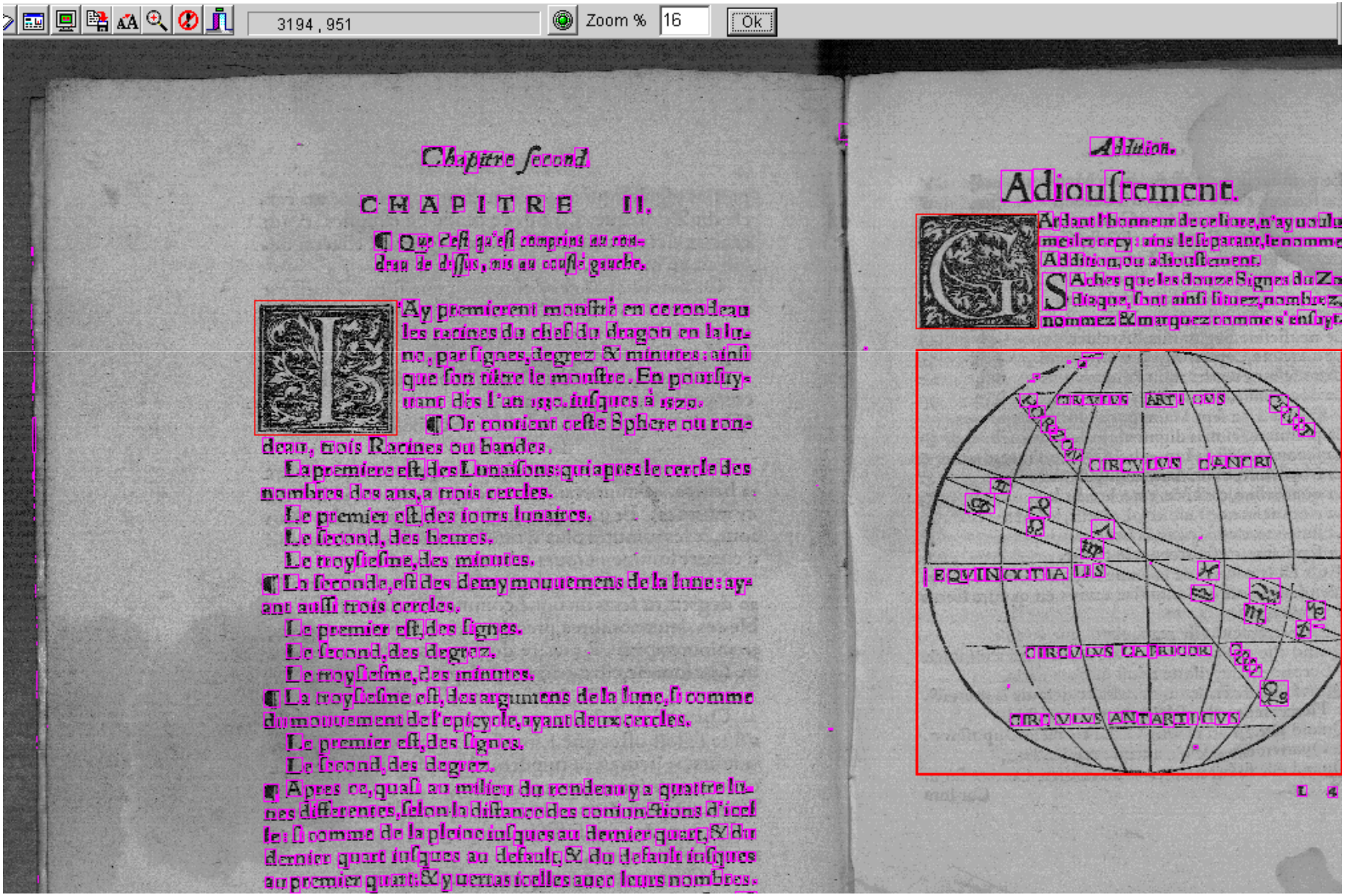
▶**Nouvelle forme de compression des images de documents**

▶**Un nouveau format de données** qui véhicule les métadonnées dans les données images : (1 livre + Métadonnées = 1 fichier)

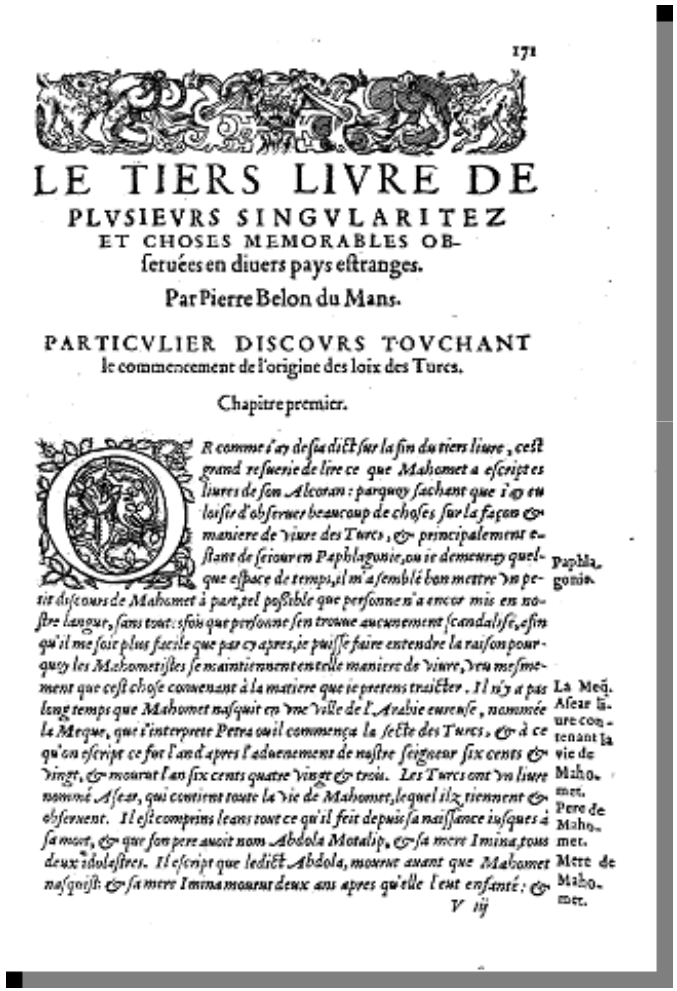
▶**Développement d'outils de consultation, de recherche, d'édition et d'annotation des documents numérisés** pour les enrichir et les structurer par des méta-données

Accès à la structure physique

Extraction de la structure physique et différenciation Texte/graphique



Classification Lettrine/bandeaux



171

LE TIERS LIVRE DE
PLVSIEVRS SINGVLARITEZ
ET CHOSES MEMORABLES OB-
ſeruées en diuers pays eſtranges.

Par Pierre Belon du Mans.

PARTICVLIER DISCOVRS TOVCHANT
le commencement de l'origine des loix des Turcs.

Chapitre premier.

OR comme i'ay deſia dict ſur la fin du tiers liure, c'eſt grand reſuerie de lire ce que Mahomet a eſcript es liures de ſon Alcoran: parquoy ſachant que i'ay eu loiſir d'obſeruer beaucoup de choſes ſur la façon & maniere de viure des Turcs, & principalement eſtant de ſeiour en Paphlagonie, ou ie demeuray quelque eſpace de temps, il m'a ſemblé bon mettre vn petit diſcours de Mahomet à part, tel poſſible que perſonne n'a encor mis en noſtre langue, ſans tout: ſfois que perſonne ſen trouue aucunement ſcandaliſé, afin qu'il me ſoit plus facile que par cy apres, ie puiſſe faire entendre la raiſon pourquoy les Mahometiſtes ſe maintiennent en telle maniere de viure, veu meſmement que c'eſt choſe conuenant à la matiere que ie pretens traicter. Il n'y a pas long temps que Mahomet naſquit en vne ville de l'Arabie eureuſe, nommée la Meque, que l'interprete Petra ouil commença la ſecte des Turcs, & à ce qu'on eſcript ce fut l'an d'apres l'aduenement de noſtre ſeigneur ſix cents & vingt, & mourut l'an ſix cents quatre vingt & trois. Les Turcs ont vn liure nommé Aſcar, qui contient toute la vie de Mahomet, lequel ilz tiennent & obſeruent. Il eſt compris leans tout ce qu'il feit depuis ſa naiſſance iuſques à ſa mort, & que ſon pere auoit nom Abdola Motalip, & ſa mere Imina, tous deux idolaſtres. Il eſcript que ledict Abdola, mourut auant que Mahomet naſquiſt: & ſa mere Imina mourut deux ans apres qu'elle l'eut enfanté: &

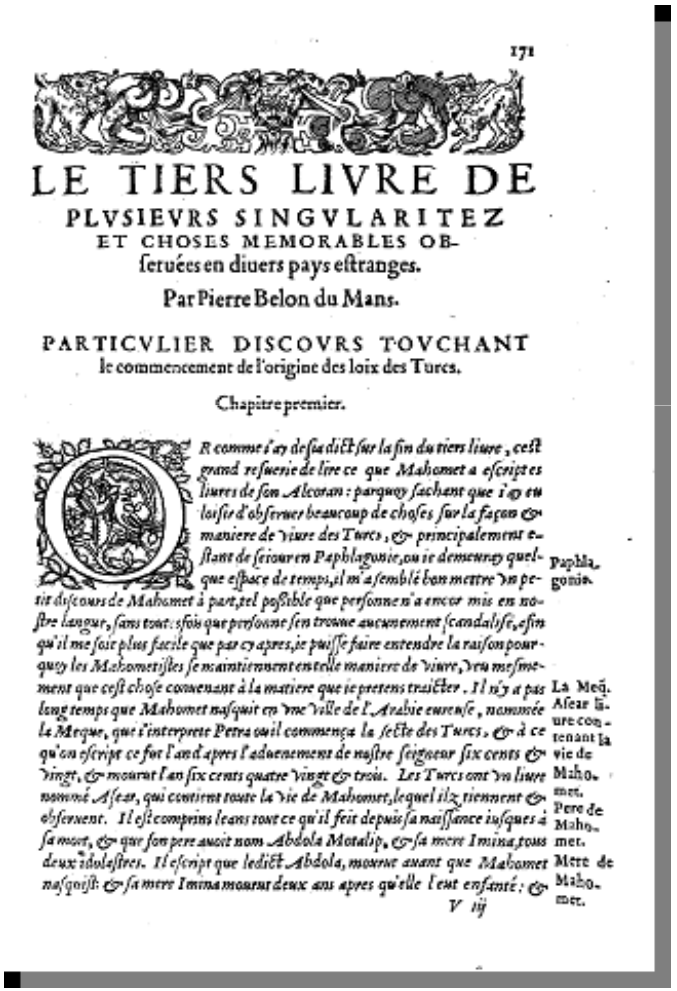
Paphlagonie.

La Meq. Aſcar liure contenant la vie de Mahomet.

Pere de Mahomet.

Mere de Mahomet.

V iij



171

LE TIERS LIVRE DE
PLVSIEVRS SINGVLARITEZ
ET CHOSES MEMORABLES OB-
ſeruées en diuers pays eſtranges.

Par Pierre Belon du Mans.

PARTICVLIER DISCOVRS TOVCHANT
le commencement de l'origine des loix des Turcs.

Chapitre premier.

OR comme i'ay deſia dict ſur la fin du tiers liure, c'eſt grand reſuerie de lire ce que Mahomet a eſcript es liures de ſon Alcoran: parquoy ſachant que i'ay eu loiſir d'obſeruer beaucoup de choſes ſur la façon & maniere de viure des Turcs, & principalement eſtant de ſeiour en Paphlagonie, ou ie demeuray quelque eſpace de temps, il m'a ſemblé bon mettre vn petit diſcours de Mahomet à part, tel poſſible que perſonne n'a encor mis en noſtre langue, ſans tout: ſfois que perſonne ſen trouue aucunement ſcandaliſé, afin qu'il me ſoit plus facile que par cy apres, ie puiſſe faire entendre la raiſon pourquoy les Mahometiſtes ſe maintiennent en telle maniere de viure, veu meſmement que c'eſt choſe conuenant à la matiere que ie pretens traicter. Il n'y a pas long temps que Mahomet naſquit en vne ville de l'Arabie eureuſe, nommée la Meque, que l'interprete Petra ouil commença la ſecte des Turcs, & à ce qu'on eſcript ce fut l'an d'apres l'aduenement de noſtre ſeigneur ſix cents & vingt, & mourut l'an ſix cents quatre vingt & trois. Les Turcs ont vn liure nommé Aſcar, qui contient toute la vie de Mahomet, lequel ilz tiennent & obſeruent. Il eſt compris leans tout ce qu'il feit depuis ſa naiſſance iuſques à ſa mort, & que ſon pere auoit nom Abdola Motalip, & ſa mere Imina, tous deux idolaſtres. Il eſcript que ledict Abdola, mourut auant que Mahomet naſquiſt: & ſa mere Imina mourut deux ans apres qu'elle l'eut enfanté: &

Paphlagonie.

La Meq. Aſcar liure contenant la vie de Mahomet.

Pere de Mahomet.

Mere de Mahomet.

V iij

Classification Lettrine/bandeaux

Classification illustrations

qui tient ſa teſte droicte & eleuée. Son bec eſt large & canelé, pointu & recroché par le bout. Il porte des plumes ſur ſa teſte par le derriere, qui luy font qui ſi vne creſte comme à vn Vanneau, & quand il volle, va battant des alles comme vn Cigne. Il ſe paiſt auſſi bien ſur l'eau ſalée, qu'en l'eau doulce. Ie prouueray en autre mien œuure, ou i'ay mis le portraict des oiſeaux, que ceſtuy eſt le Pelican, dont me tais pour ceſte heure à cauſe de Briefueté. Entre les choſes ſingulieres de ceſte iſle, ay veu le ſerpent nommé Iaculus, moucheté de petites taches deſſus le dos, reſſemblantes à des petis yeux, tout ainſi que ſont les taches de deſſus le dos d'vn Tremble, nommé en Latin Torpedo. Ie le trouuay deſſoubs vn Caprier eſpineux hors la ville, celle part ou le Turc auoit planté ſon artillerie quand il aſſiega Rhodes. Les Grecs le nomment maintenant en leur vulgaire Saetta, c'eſt à dire Sagitta, & les Turcs Ochilanne, les anciens Acontius. Il a trois paulmes de longueur, & n'eſt plus gros que le petit doigt. Sa couleur eſt cendrée tirant ſur la couleur de laict, & eſt totalement blanc deſſoubs le ventre, ayant des eſcailles deſſus le dos, & tablettes deſſoubs le ventre à la maniere des autres. Il eſt noir deſſus le col, & taché de deux lignes blanches, qui commencent dés la teſte, & ſuiuent tout le long du dos iuſques à la queue. Les taches dont il eſt moucheté, ne ſont plus larges qu'eſt vne Lentille. Mais eſtant ſon dos cendré, les taches noires ſont rondes, entournées d'vn cercle blanc. Ie parleray de ſon anatomie ailleurs plus à plain en deſcriuant tous ſerpents par le menu. Toutesfois ayant eu ſon naif portraict, ie l'ay mis en ce lieu.

Pelican.

Tremble. Torpedo

Saetta, Acontius.

Le portraict du Iaculus, autrement dit Acontius.

Ie vei auſſi deſcharger vn brigantin deſſus la riue du port, plein d'vne drogue propre en medecine, appellée Storax rouge. Les Grecs la nomment maintenant Maurocapno. Et m'a lon dit qu'il croiſt en l'iſle. Mais pource que ceux qui font voyages par mer, ne peuuent s'abſenter loing de leur vaiſſeau, ie n'ay eu loiſir de m'eſcarter pour aller veoir ſon arbre: car quand les mariniers ont le temps à propos, ils ne retarderoient pour homme viuant. Ie vueil inferer par cela

Storax rouge. Maurocapno.

lité rendra la terre d'Egypte. Et pource qu'il n'a pas accouſtumé croiſtre tant vne année que l'autre, ilz ont diuers ſignes pour ſçauoir à peu pres ce que le pays rendra l'année à venir. On trouue par eſcript que le reuenu d'Egypte eſtoit moult grand du temps que les Rommains en eſtoient ſeigneurs, lequel a beaucoup diminué depuis: mais il fault entendre que pour lors les Rommains n'eſpargnoient rien à y faire deſpenſe pour le rendre fertile. I'ay prins grande merueille d'auoir veu ſi grande quantité de Caſſiers es iardins du Caire, & par Egypte, & toutesfois les autheurs anciens n'en ont faict aucune mention: car meſmement Theophraſte qui a quaſi parlé de toutes autres plantes d'Egypte, n'en faict mention. Mais il fault dire de Theophraſte parlant des plantes, tout ainſi comme d'Ariſtote des animaulx. Car comme diuerſes nations obeiſſants aux commandements d'Alexandre apportoient diuerſes eſpeces d'Animaulx à Ariſtote, lors qu'il en eſcripuoit l'hiſtoire, auſſi eſtoit il neceſſaire que par meſme moyen diuerſes nations feiſſent rapport des plantes à Theophraſte quād il les deſcripuoit. Et appert à ſon hiſtoire qu'il ne l'a faicte ſans grande deſpenſe, & d'hommes qui ont eſté expreſſement enuoyez en diuers endroicts du monde, pour les obſeruer.

Caſſiers.

Portraict du Caſſier.

Liberalité d'Alexandre. enuers Theophraſte, & Ariſtote.

Classification illustrations

Fonctionnalités du poste client

DEBORA

▶ **Enrichissement des documents numérisés** *(annotations de n'importe quel élément avec des annotations de n'importe quelle nature…)*

▶ **Recomposition de n'importe quel élément physique** *(image, caractère, dessin..)* **ou logique** *(métadata,annotation)*, **pour une vue personnalisée.**

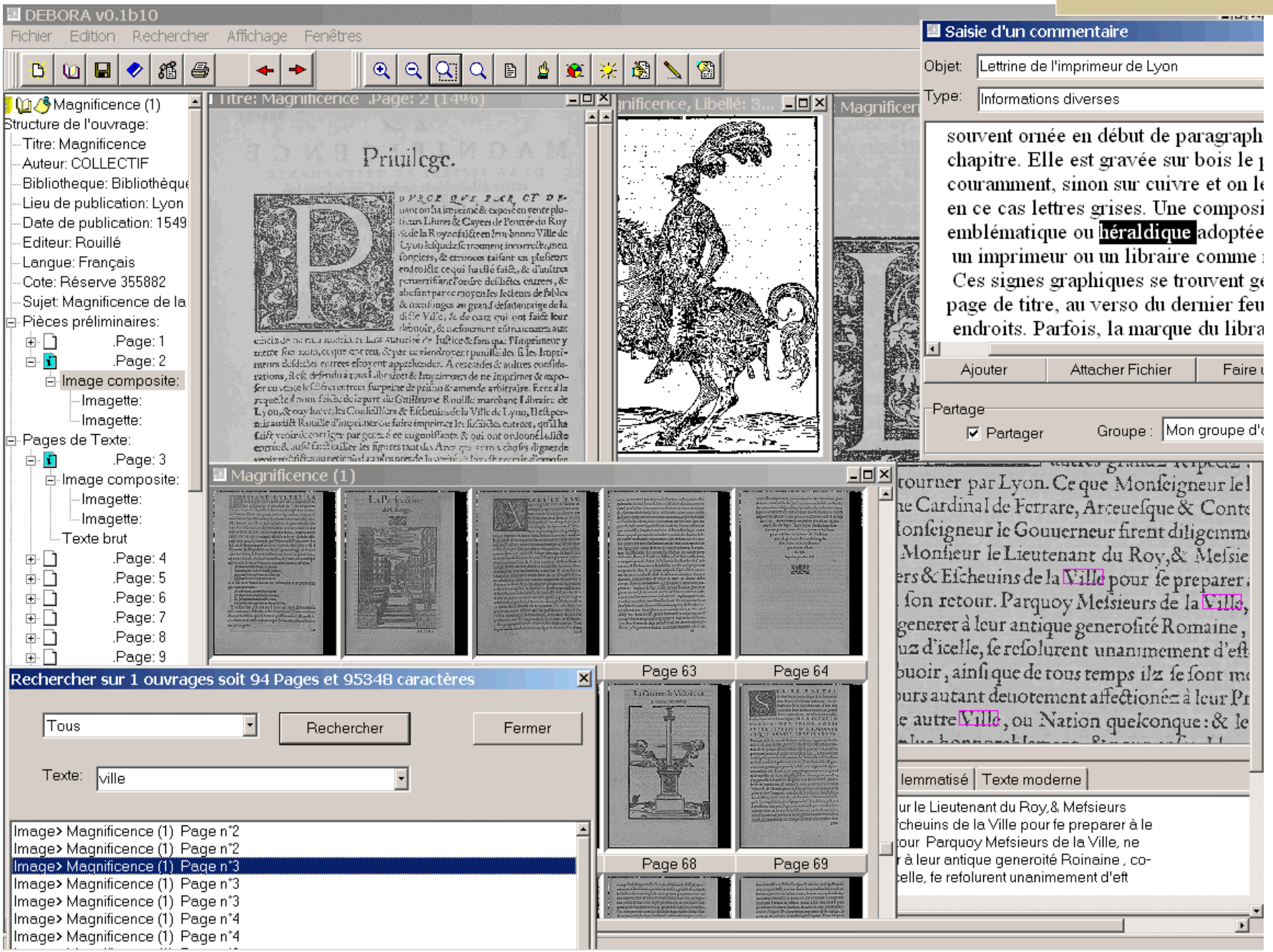
▶ **Éditer, suivre les annotations d'un travail public ou privé**

▶ **Partager les travaux** *(par l'intermédiaire d'un serveur ou directement)*

▶ **Possibilité de télécharger totalement ou partiellement les ouvrages et de travailler localement**

▶ **Chercher un objet physique** *(caractère, mot, lettrine…),* **une méta-donnée** (fiche bibliographique, élément de la structure logique, annotations..)

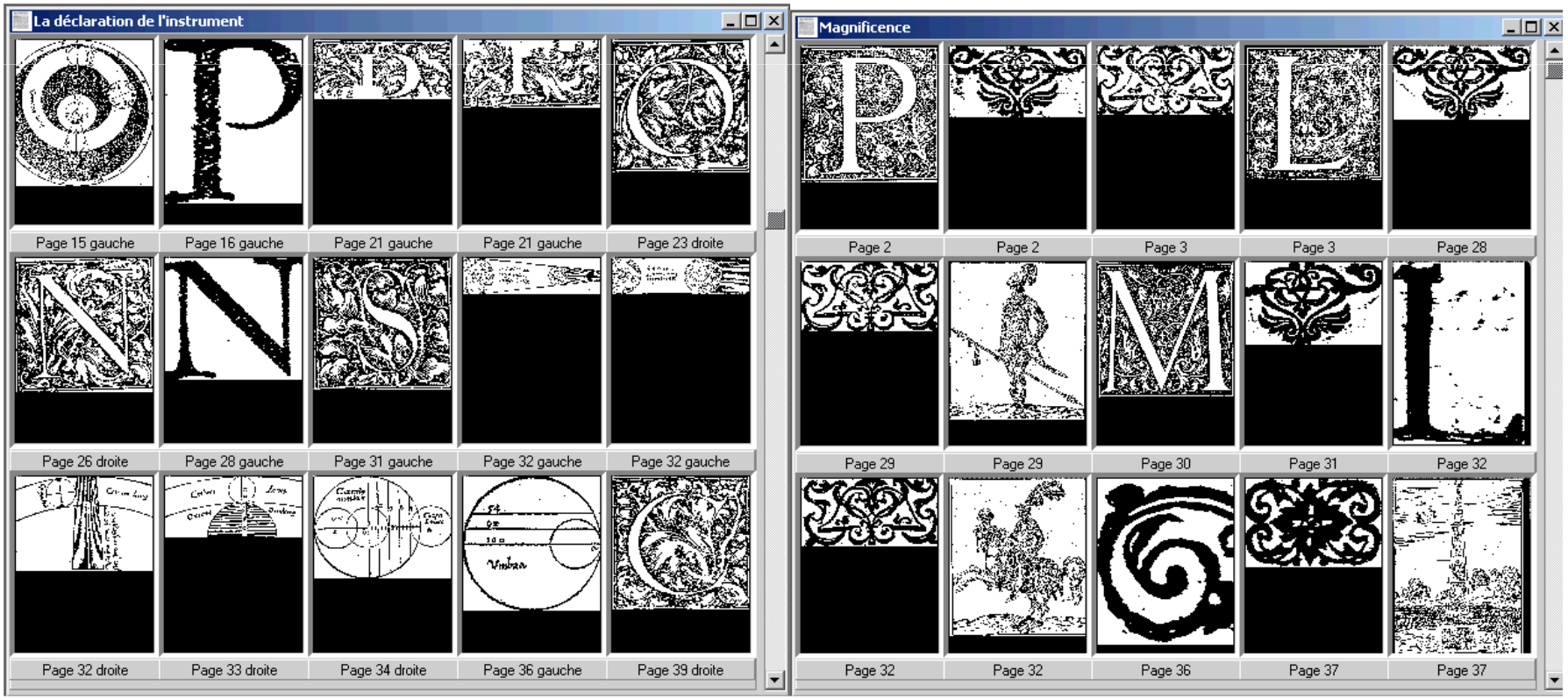
Poste de consultation DEBORA



Fonctionnalités de recherche étendue

Chercher une donnée, une méta-donnée, un élément d'une structure :

- Une annotation (par auteur, date, contenu, sujet...),
- Un élément particulier (mots dans les textes, dans les libellés des structures, des illustrations…)

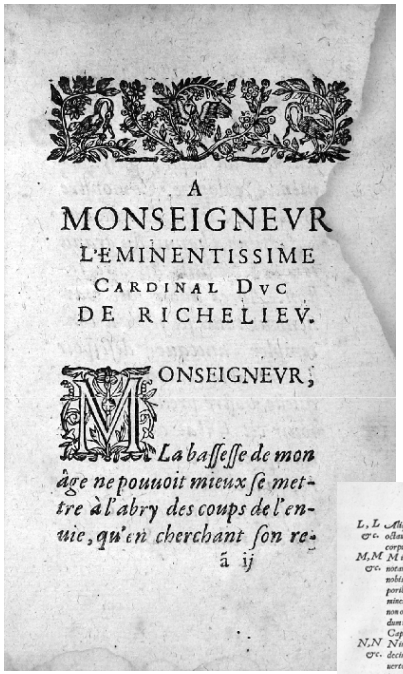


Indexation d'images de documents de la renaissance

MADONNE

Que peut-on indexer ?

- structure de l'ouvrage
- Séparation texte/image
- Construction d'index d'images


A
MONSEIGNEVR
L'EMINENTISSIME
CARDINAL DVC
DE RICHELIEV.

MONSEIGNEVR,

La baſſeſſe de mon âge ne pouuoit mieux ſe mettre à l'abry des coups de l'enuie, qu'en cherchant ſon re-

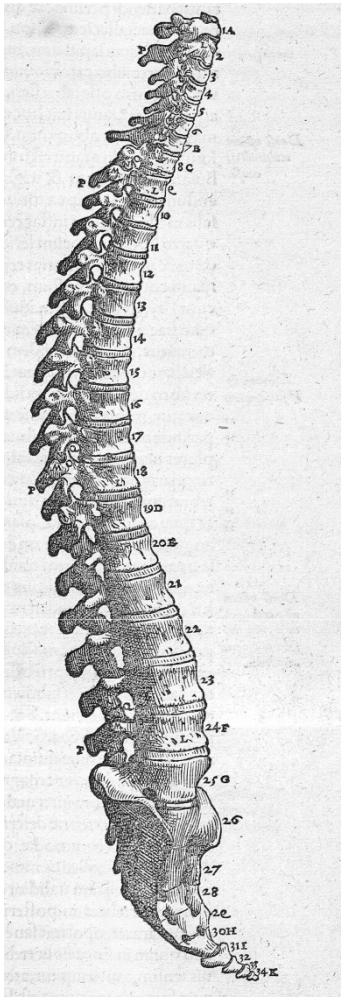

DE HVMANI CORPORIS FABRICA LIBER I.

Complexité:

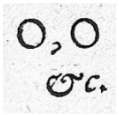
- Forte variabilité de la typographie, de la mise en page, des dessins...
- Nécessité d'inclure un expert dans le processus d'indexation
- Grosse quantité de données

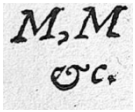
MADONNE

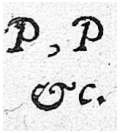
Gravure

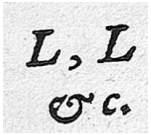


Index de lignes

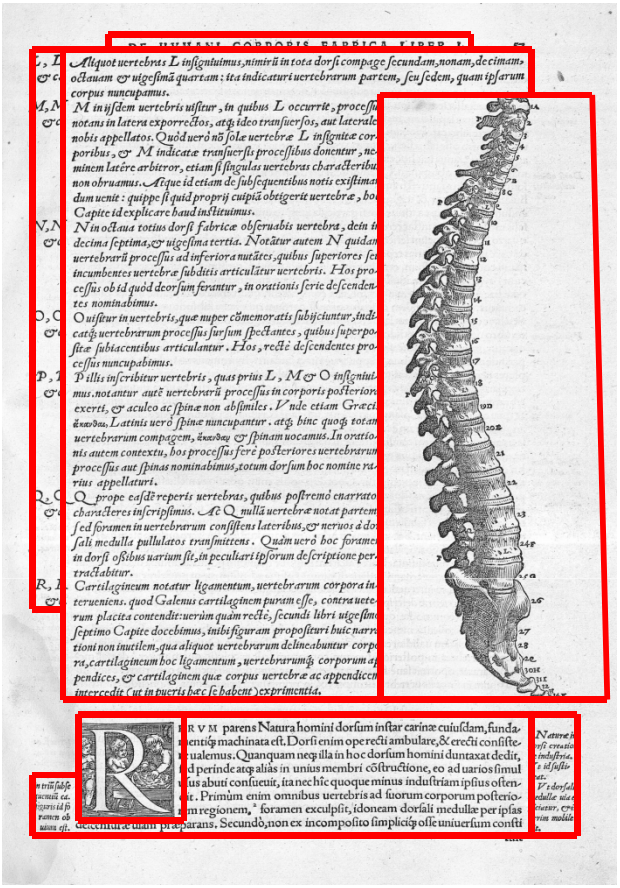




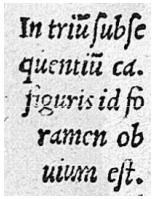


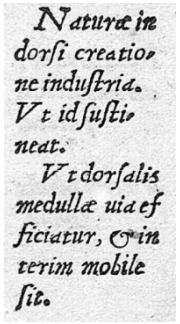


N° de page

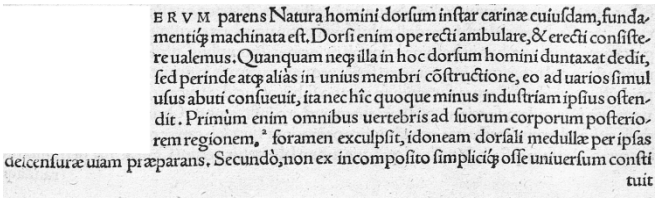


Notes





Paragraphe fonte normal



ERVM parens Natura homini dorſum inſtar carinæ cuiuſdam, fundamentiq; machinata eſt. Dorſi enim ope recti ambulare, & erecti conſiſtere ualemus. Quanquam neq; illa in hoc dorſum homini duntaxat dedit, ſed perinde atq; aliàs in unius membri cõſtructione, eo ad uarios ſimul uſus abuti conſueuit, ita nec hîc quoque minus induſtriam ipſius oſtendit. Primùm enim omnibus uertebris ad ſuorum corporum poſterioremregionem,[a] foramen exculpſit, idoneam dorſali medullæ per ipſas deſcenſuræ uiam præparans. Secundò, non ex incompoſito ſimpliciq; oſſe uniuerſum conſtituit

Titre

Lettrine

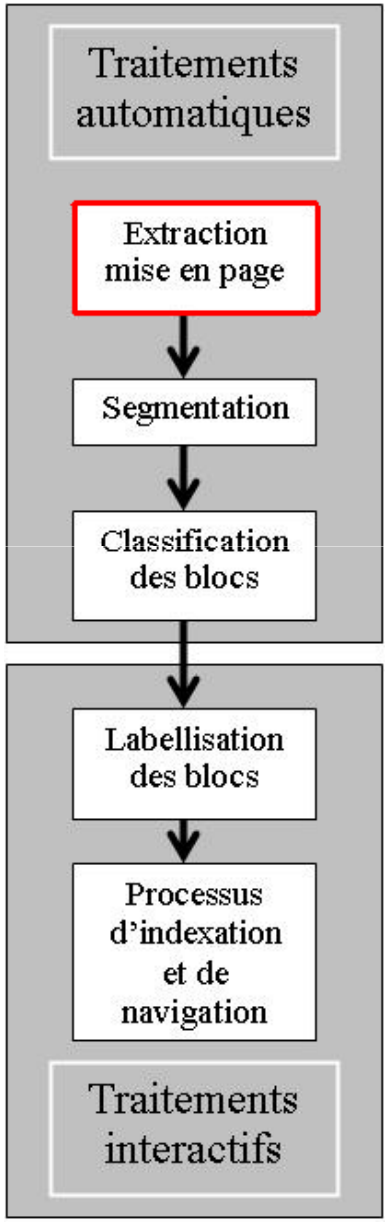
Un problème déjà résolu?

MADONNE

- **Nos hypothèses**
  - Pas de connaissances sur le modèle
    - Hypothèses réduites d'anisotropie du texte, pas de connaissances sur les caractéristiques physiques
    - Aucun modèle sur l'organisation des entités
  - Données accessibles?
    - Pas d'ontologie
    - Utilisateur non spécialiste du traitement des images et de l'ARD

Première étape: Extraction de la mise en page

MADONNE



Pourquoi?

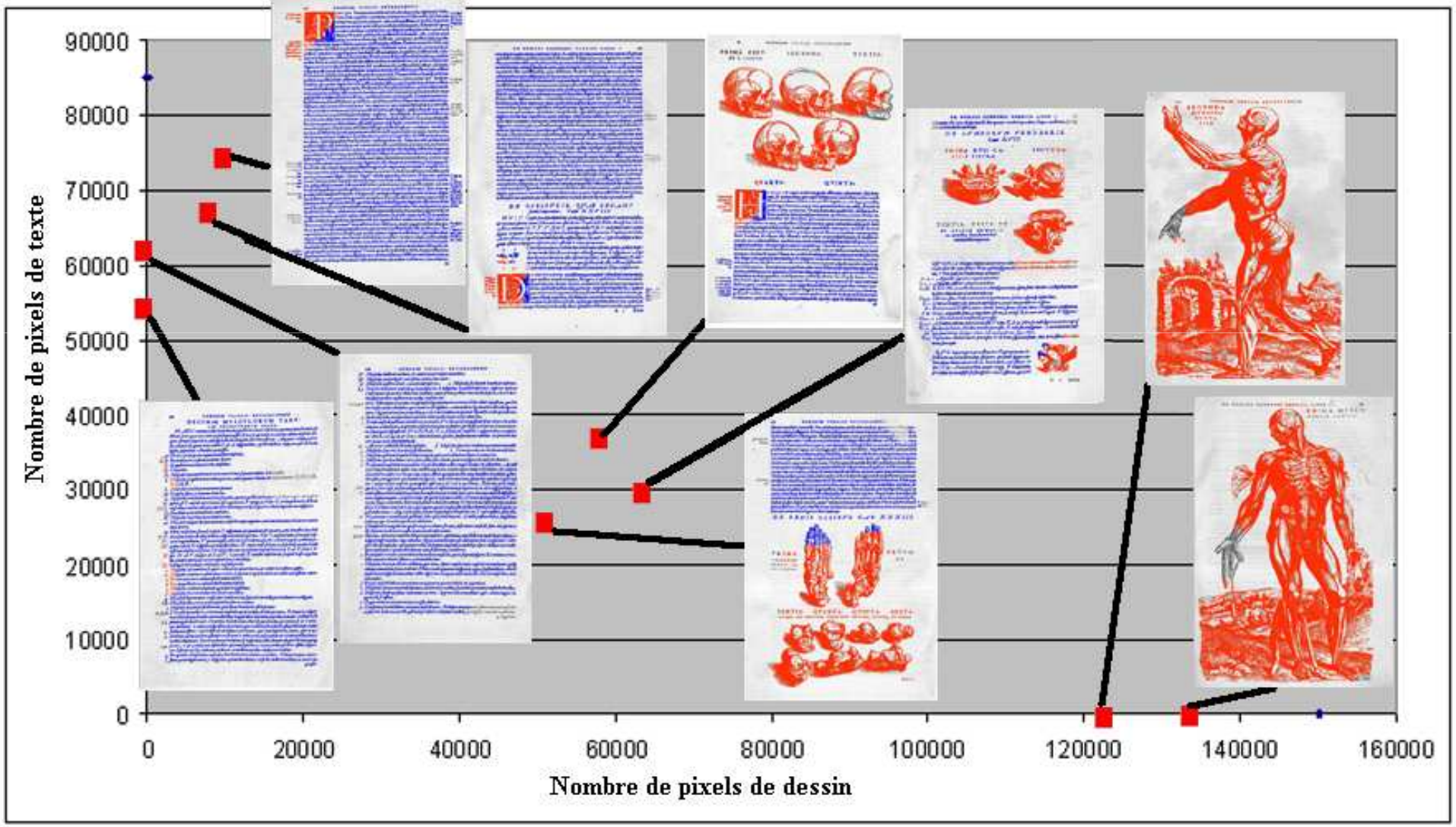
- Localiser l'information (marquage générique)
- Avoir une première idée du modèle (première exploitation par indexation rapide)
- Simplifier la phase de segmentation

Comment?

- Séparation texte/image (pré - étiquetage)
- Sans connaissance a priori des dispositions ni des contenus (indépendamment des typographies, fontes, bruits et résolutions) : peu de paramètres
- Adaptabilité à tout type d'images de documents hétérogènes
- Utilisation de faibles résolutions (anisotropie du texte)

MADONNE

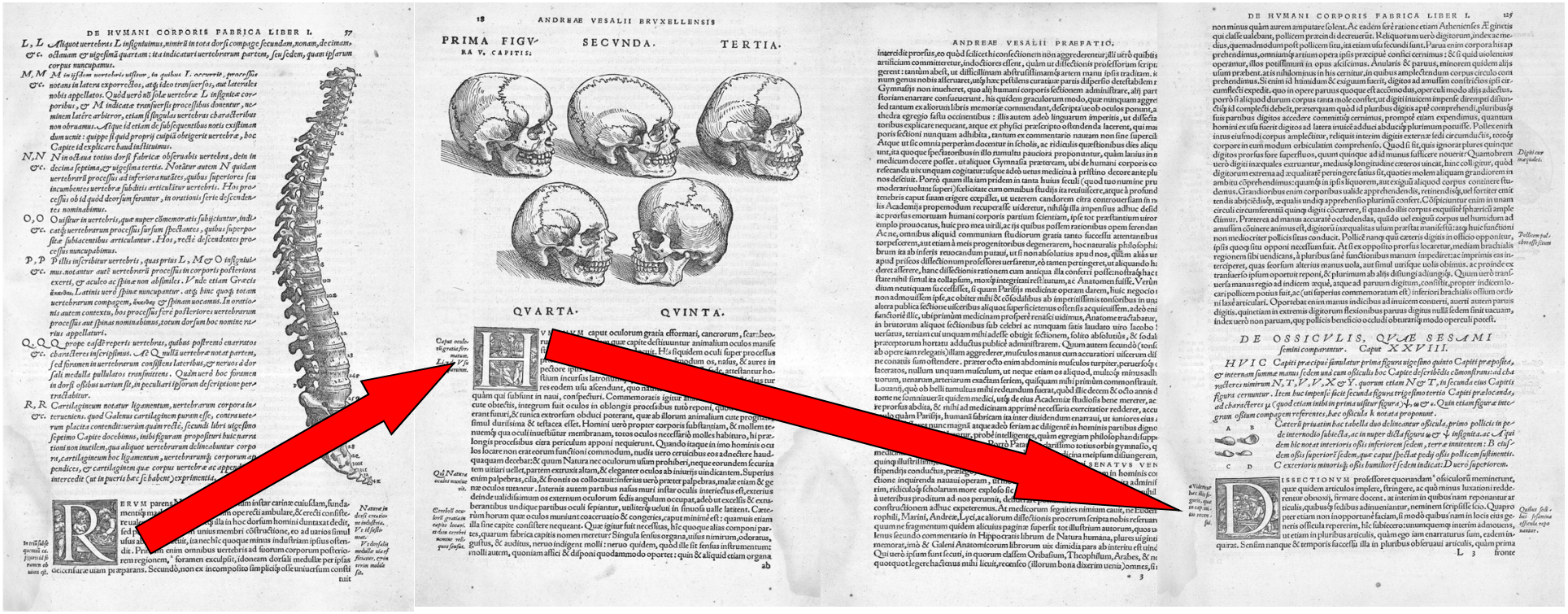
Résultats



Indexation/aide à la navigation

MADONNE

- Module de navigation: indexation du contenu
- A partir de la connaissance même partielle de la mise en page (marquage ≠ segmentation)



MADONNE

Indexation/aide à la navigation

- Module de recherche :
  - indexation de la structure et de la mise en page
  - Notion de “similarité” de mise en page

Image requête

Images similaires