

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 :

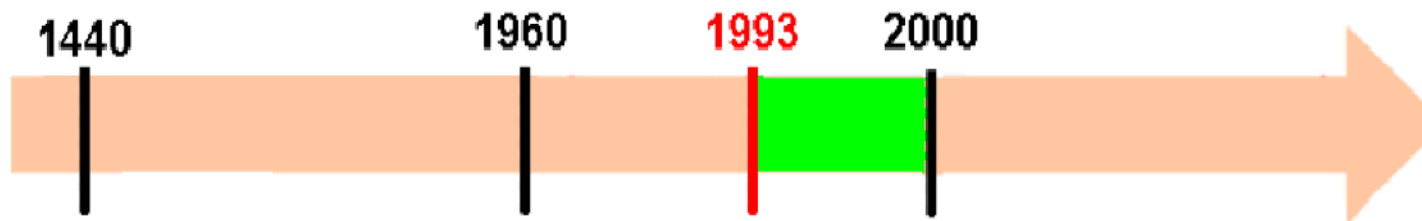
- *La période de l'histoire*
- *Le lieu où le document original a été produit*
- *Les intentions d'usage, les besoins*

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...*)

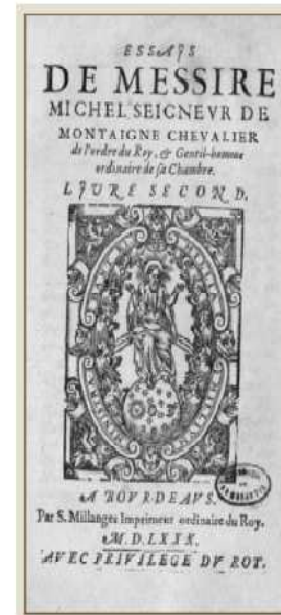
Mais quelles solutions...?

Quelles solutions?

Exploiter les documents en mode texte...



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

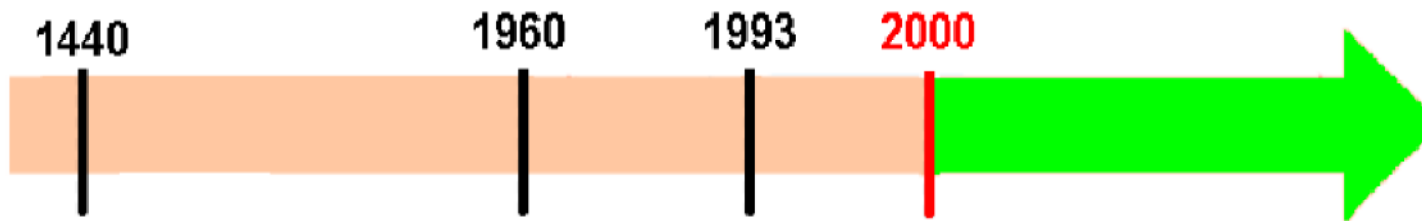


Auteur: Montaigne, Michel de
Titre: Essais de Messire Michel seigneur de Montaigne, chevalier de l'ordre du Roy, & Gentil-homme ordinaire de sa Chambre. Livre Second
Edition: A bourdeaus par S. Millanges
Format: 8°
Collation: [2]f., 650, [3] p.
Langue: Français
Matière: Littérature
Numérisation: CESR - Digibook - 2006

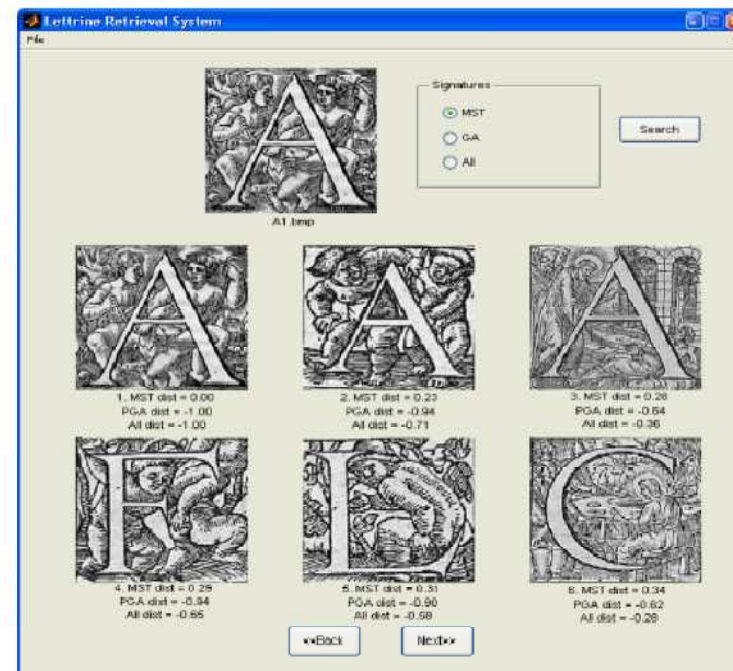
Bibliothèque numérique du CESR

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

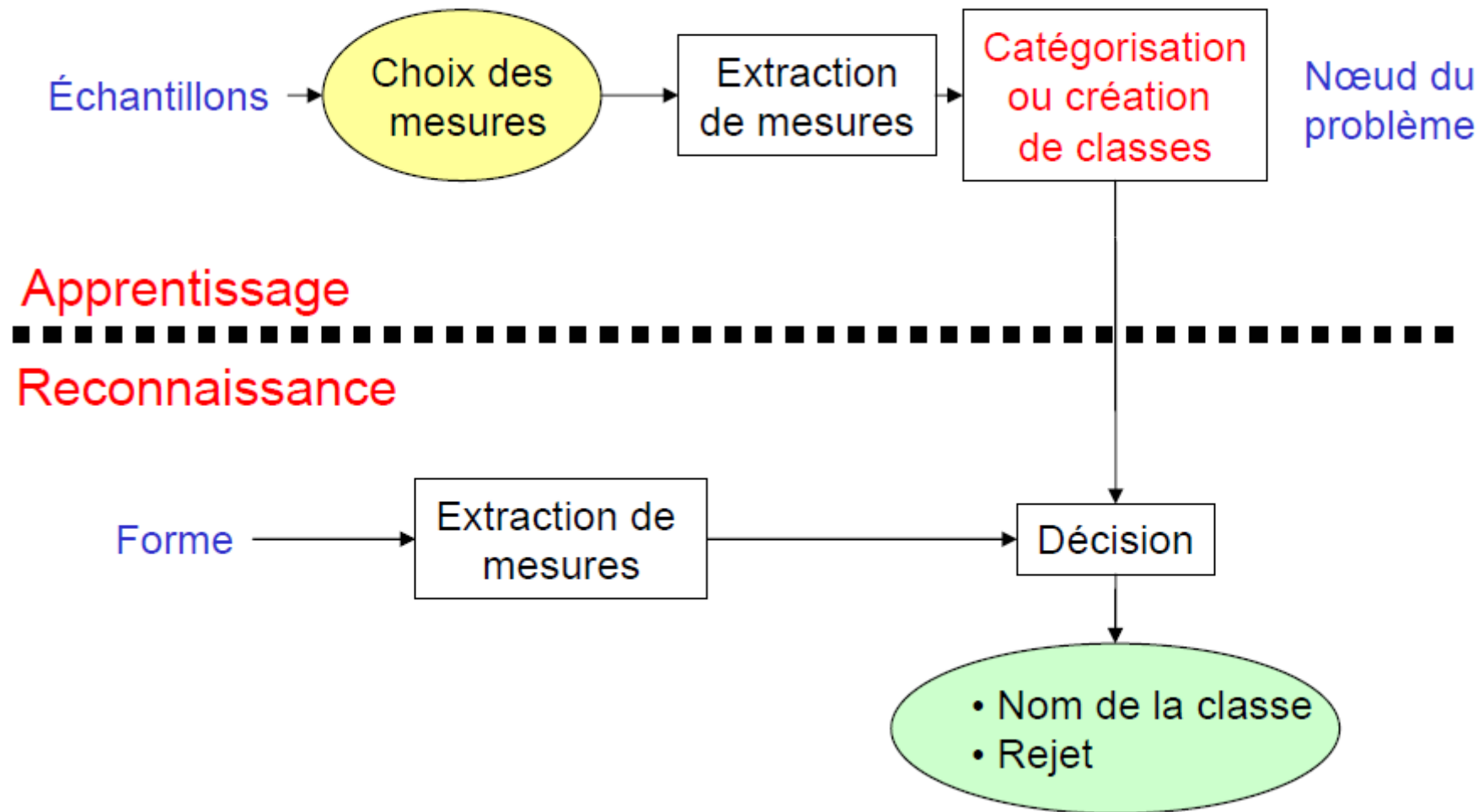
Quelques domaines d'applications existants:

- ▶ 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

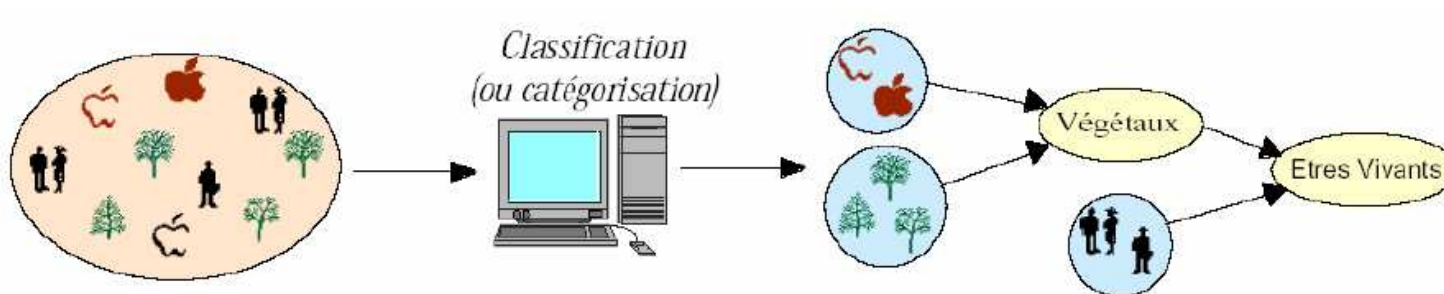


Analyse automatique des images

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

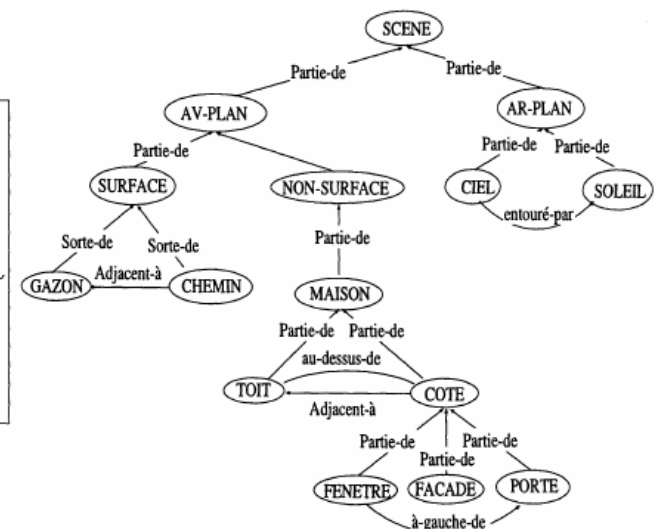
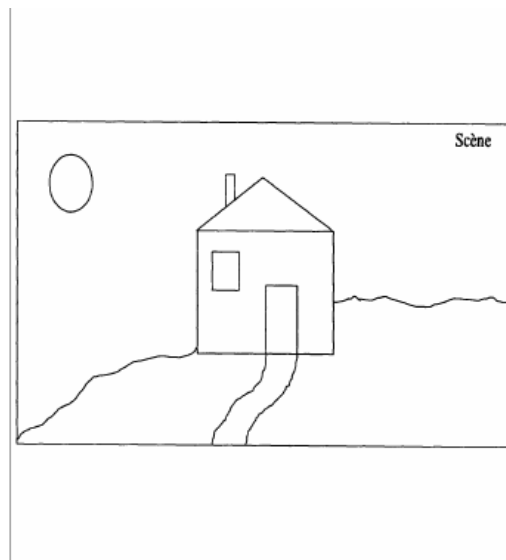
La catégorisation (reconnaissance) structurelle

■ 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

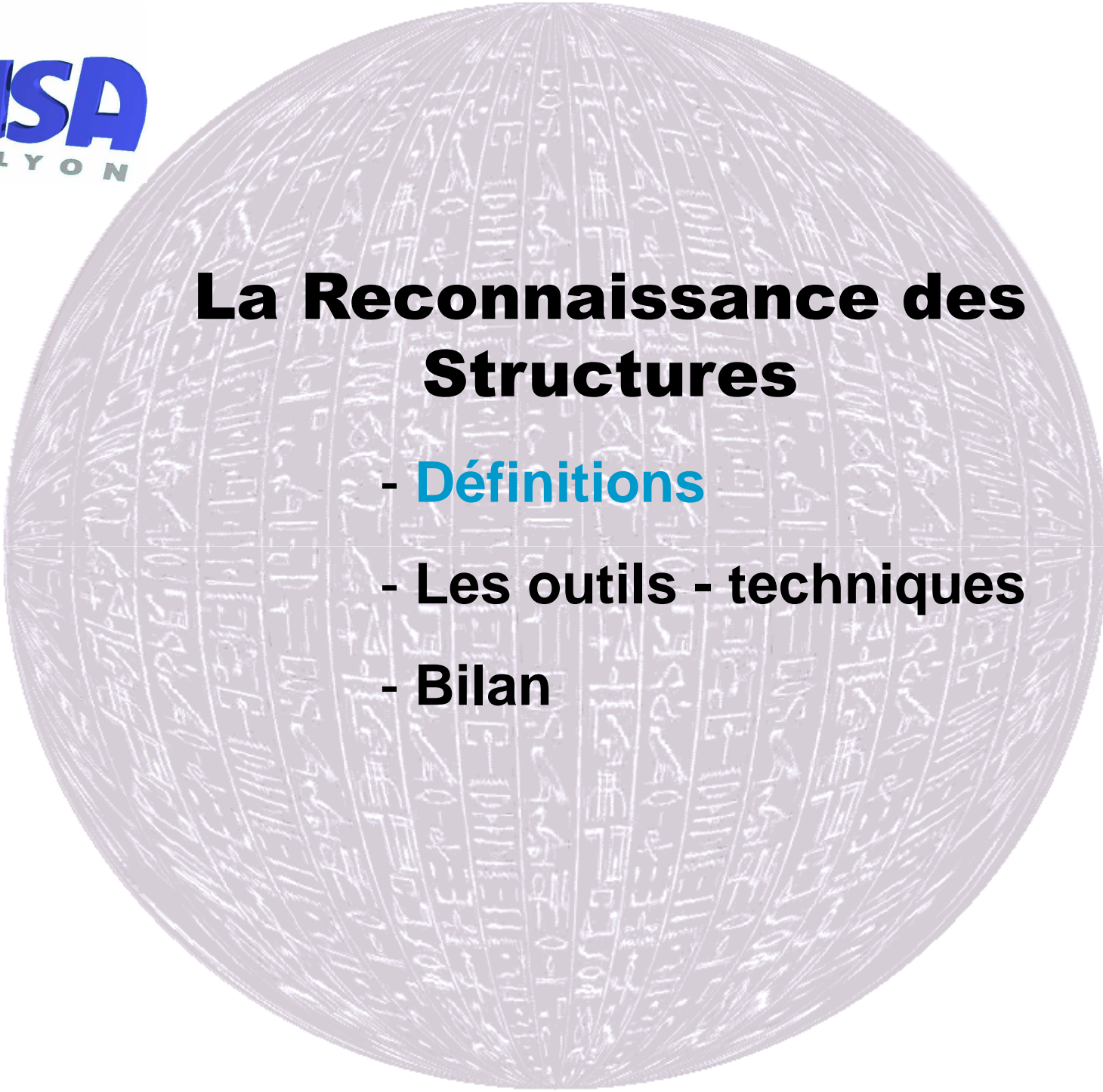
Les experts doivent fournir des critères sûrs et non ambigus des objets qu'ils veulent reconnaître

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...

A large, semi-transparent sphere occupies the center of the slide. The sphere's surface is covered with faint, white mathematical formulas and equations, including trigonometric functions, algebraic expressions, and differential equations, arranged in a grid-like pattern across its surface.

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

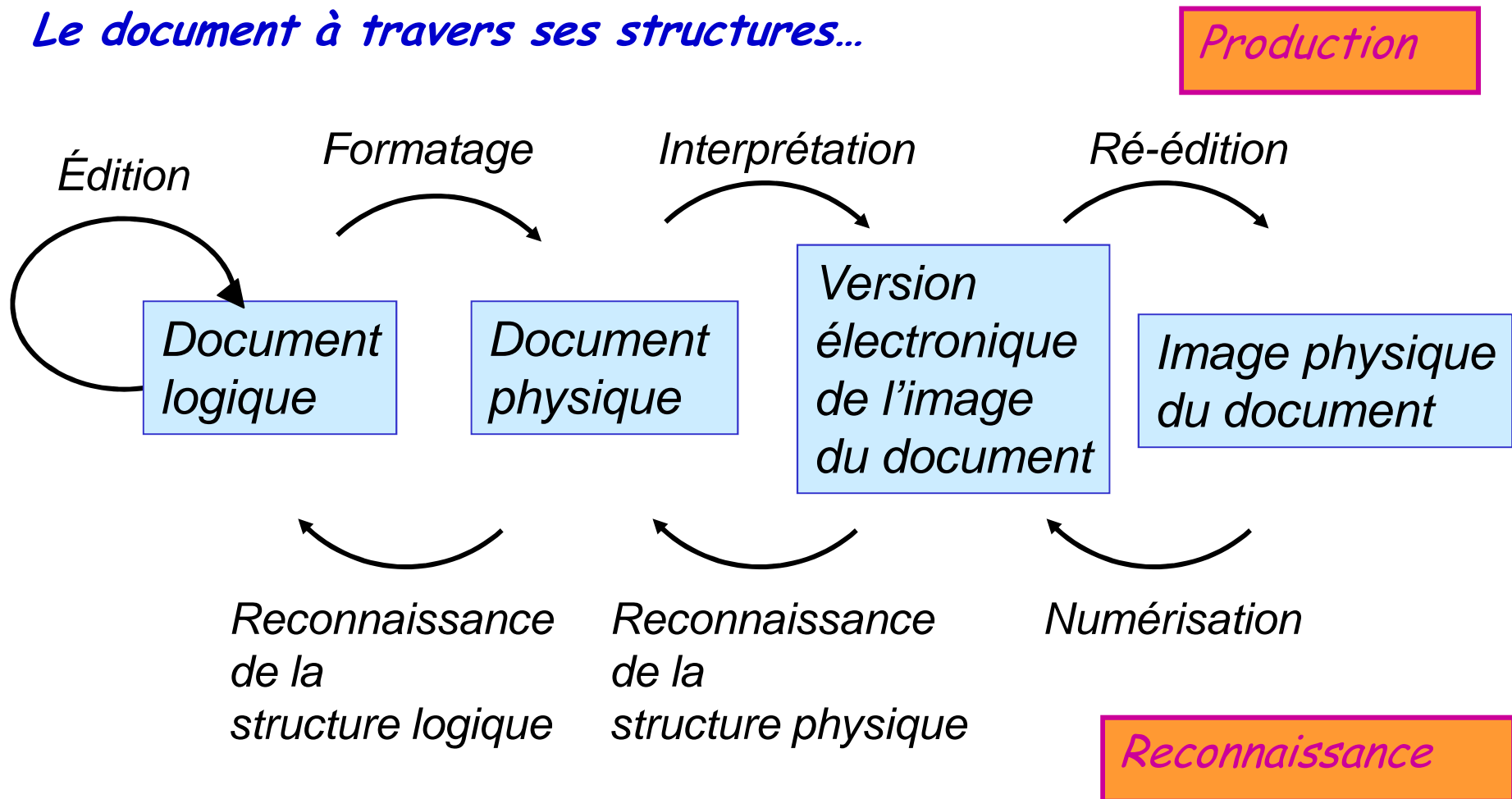
			
<i>Manuscrit médiéval (source IRHT²) XIII</i>	<i>Livre Imprimé de la renaissance XVI^{ème}</i>	<i>Imprimé du XVII^{ème}</i>	<i>Journal du XX^{ème}</i>

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

Rolf Brügger
 Information Institute of the University of Fribourg
 Chemin du Musée 2
 CH-1700 Fribourg (Switzerland)
 E-mail: rolf.bruegger@unifr.ch

1 Introduction

The importance of structured documents has been widely accepted in the past few years. In this context, the reconstruction or recognition of the structure of a document from a given document image has become an important research topic. In other words, to recognize the structure means to translate a printed document to a structured electronic format like SGML or similar. Existing systems for this task have different advantages and drawbacks: the most of them have in common, that only with difficulty they can be adapted to new document structures and that they are not error tolerant enough. In our Project (STRS) we are facing these problems by trying to construct a highly interactive and adaptive system.

In the following paragraphs we present a novel approach to model document structures, a learning method and an adapted recognition algorithm. Finally, some test results are presented and discussed.

2 Modeling with Generalized n-Grams

In order to be able to recognize the structure of documents, they have to be represented by a model. Conventional systems are using grammar or knowledge based approaches or decision trees [1, 2, 3]. Most of these systems do not provide automatic learning of the model. Moreover, it would be difficult to integrate it subsequently. We therefore propose a statistical approach, which includes a considerably simplified construction of the model as well as other advantages like inherent error tolerance

in the recognition phase.

It is commonly accepted to represent the logical structure of specific documents by means of a tree. A set of such tree structures belonging to the same document type can be represented in a generic way by reducing them to probabilities of local neighborhood relations of tree nodes. Our document model contains all probabilities of ancestor-child and sibling-sibling configurations. This approach is inspired by a 'gram model' which can be used to represent sequences of symbols. We therefore call it 'generalized n-gram model for tree structures' [4].

The main advantage of such a representation lies in the fact, that the n-gram probabilities can easily be estimated. The system simply has to analyze statistically a set of documents whose logical structure is known. Likewise it is easy to integrate additional documents later in order to refine an existing model. This incremental learning can be done by adapting the conditional probabilities.

3 Recognition of Scanned Documents

The recognition of a document is split up into several phases. First, the document is being scanned. Then it is segmented and the characters and their font is recognized. The last phase then consists of building the logical structure on top of the previously obtained layout structure. This can be attained a step by step by iteratively aggregating tree nodes to the logical tree structure. The thus created new partial tree

```

graph TD
    ARTICLE[ARTICLE] --> HEADER[HEADER]
    ARTICLE --> BODY[BODY]
    ARTICLE --> BIBLIOGRAPHY[BIBLIOGRAPHY]
    HEADER --> TITLE1[TITLE]
    HEADER --> AUTHOR[AUTHOR]
    HEADER --> AFFILIATION[AFFILIATION]
    AFFILIATION --> ADDRESS[ADDRESS]
    AFFILIATION --> EMAIL[EMAIL]
    BODY --> SECTION1[SECTION]
    BODY --> SECTION2[SECTION]
    SECTION1 --> TITLE2[TITLE]
    SECTION1 --> PARAGRAPH[PARAGRAPH]
    TITLE2 --> INTRODUCTION[Introduction]
    PARAGRAPH --> PARAGRAPH_CONTENT["The importance of structured documents ..."]
    BIBLIOGRAPHY --> REFER1[REFER.]
    BIBLIOGRAPHY --> REFER2[REFER.]
    
```


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 de 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 recourir 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 "explorer, 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 "démantique" 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 :

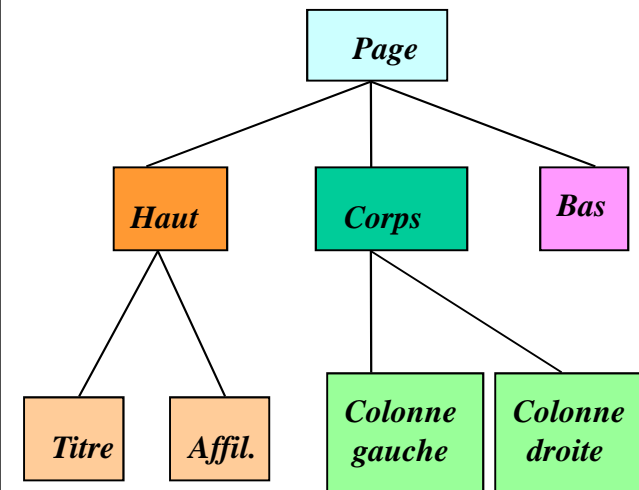
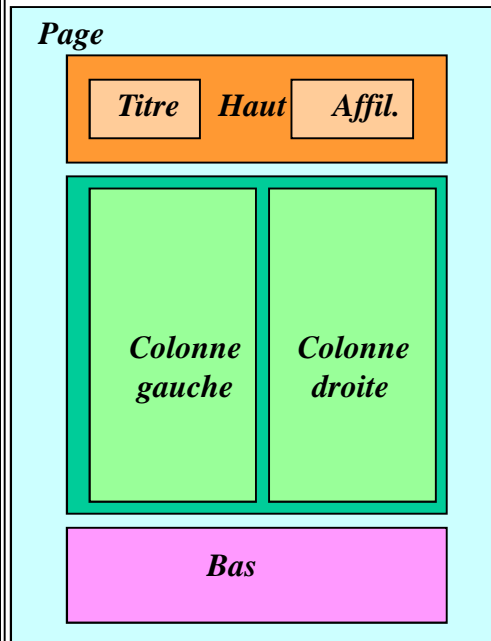
- D'une part les documents ayant une forte structuration logique et un style typographique riche, qui peuvent, après une étude poussée, être reconnus et traités. (cf. les travaux de [INGOLD 1989], [DERRIEN-PELDEN 1990], [JAKINDELE 1993] ou [MOHAID 1994]).
- D'autre part une seconde catégorie de documents n'obéissant, le plus souvent, à aucune règle précise et reconnus 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	
VEP	
TITRE	
PUBCO	
TICOLD	
NOTE	
	LIBR

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



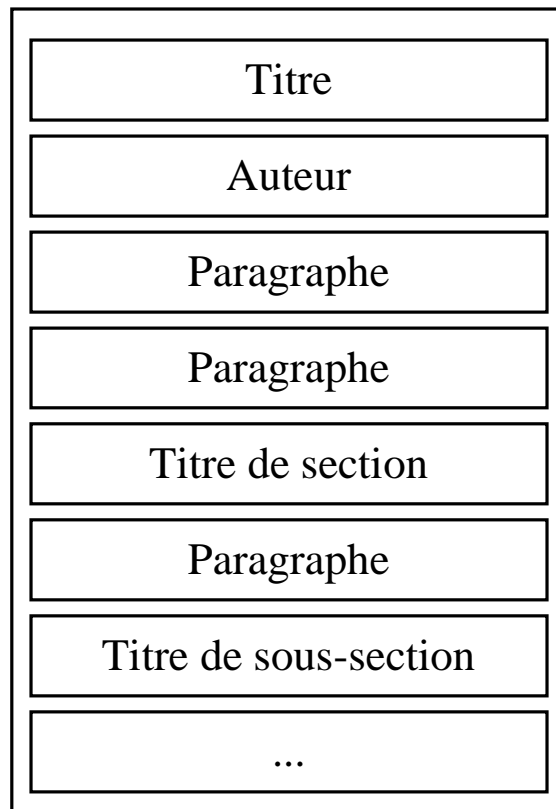
Structure PHYSIQUE d'un document (2)

<p>Juridiction et administration épiscopales. Marriages.</p> <p>G. Maurienne. 61. — Liasse. — 6 pièces parchemin, 147 pièces papier.</p> <p>1640-1791. — <i>Juridiction et administration épiscopales : mariages.</i> — 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.</p> <p>G. Maurienne. 62. — Liasse. — 1 cahier, 8 pièces papier.</p> <p>1710-1761. — <i>Juridiction et administration épiscopales.</i> — 1. Excommunication lancée par le vicaire général contre un habitant d'Albiez-le Jeune</p>	<p>Juridiction et administration épiscopales. Marriages.</p> <p>G. Maurienne. 61. — Liasse. — 6 pièces parchemin, 147 pièces papier.</p> <p>1640-1791. — <i>Juridiction et administration épiscopales : mariages.</i> — 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.</p> <p>G. Maurienne. 62. — Liasse. — 1 cahier, 8 pièces papier.</p> <p>1710-1761. — <i>Juridiction et administration épiscopales.</i> — 1. Excommunication lancée par le vicaire général contre un habitant d'Albiez-le Jeune</p>
--	--

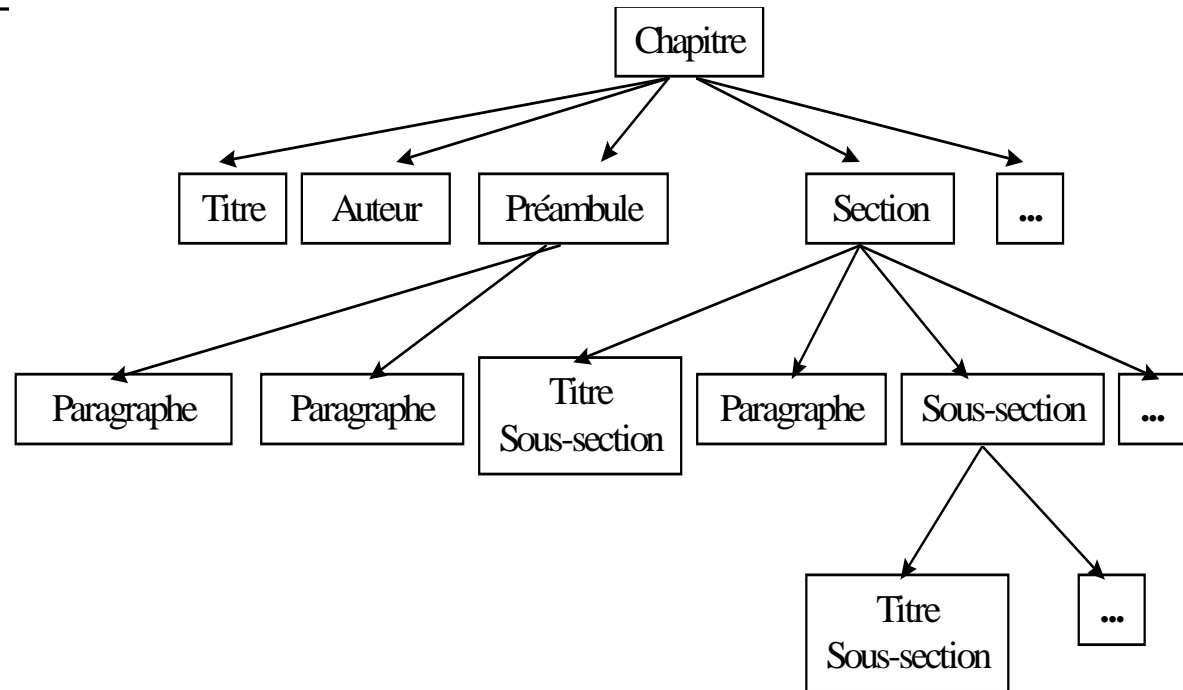
Pour une **localisation** des informations dans la page

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

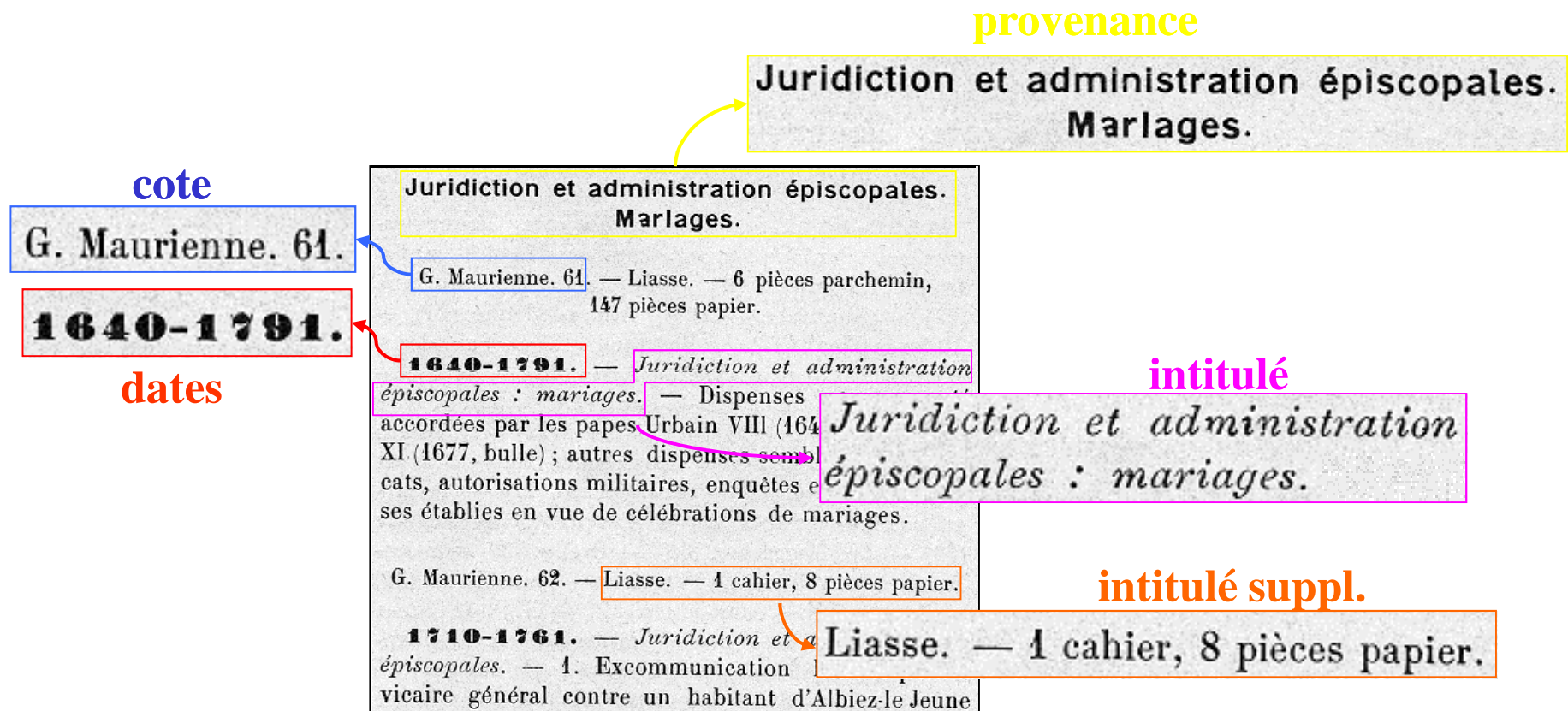


Exemple de structure logique de liste



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; sur la tête laurée de Napoléon 1^{er}). 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.

Article

NOM (prénom)

Bibliographie

Profession

Date de naissance

Lieu de naissance

Date de décès

Fonctions date,

Réalisations

Nom, lieu, date

Publications

Titre, éditeur, page

Titres honorifiques,
vie associative

Titre & médailles

Membre de ..

Informations diverses

Références

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

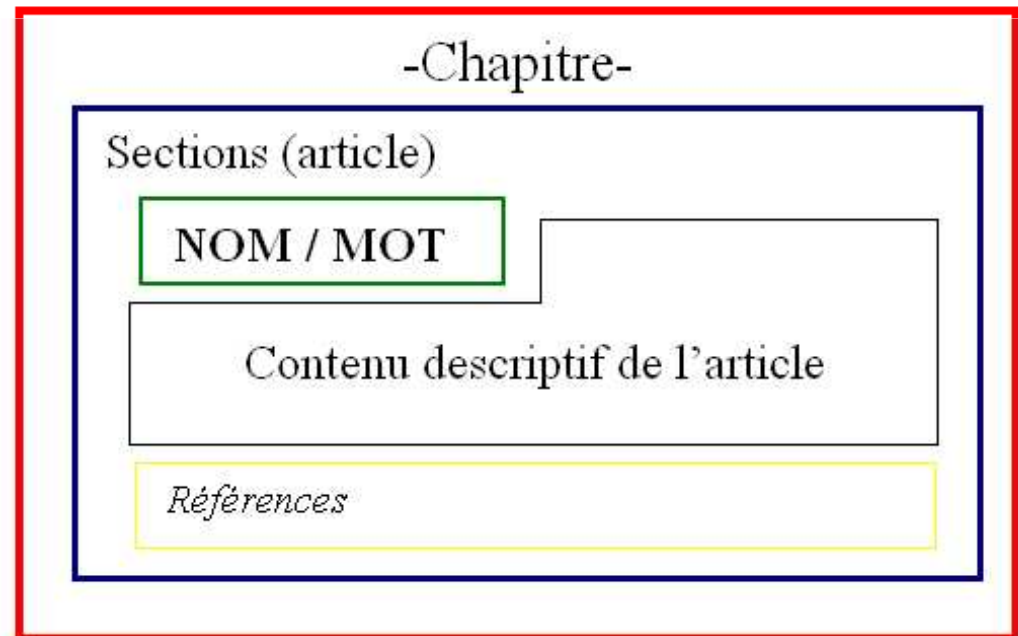
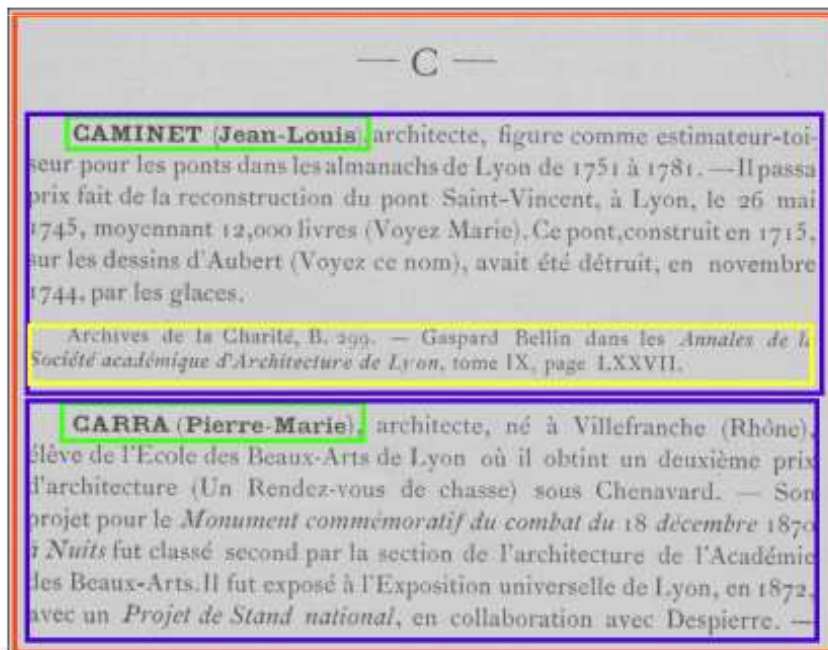
Vers une structure de l'ouvrage...

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

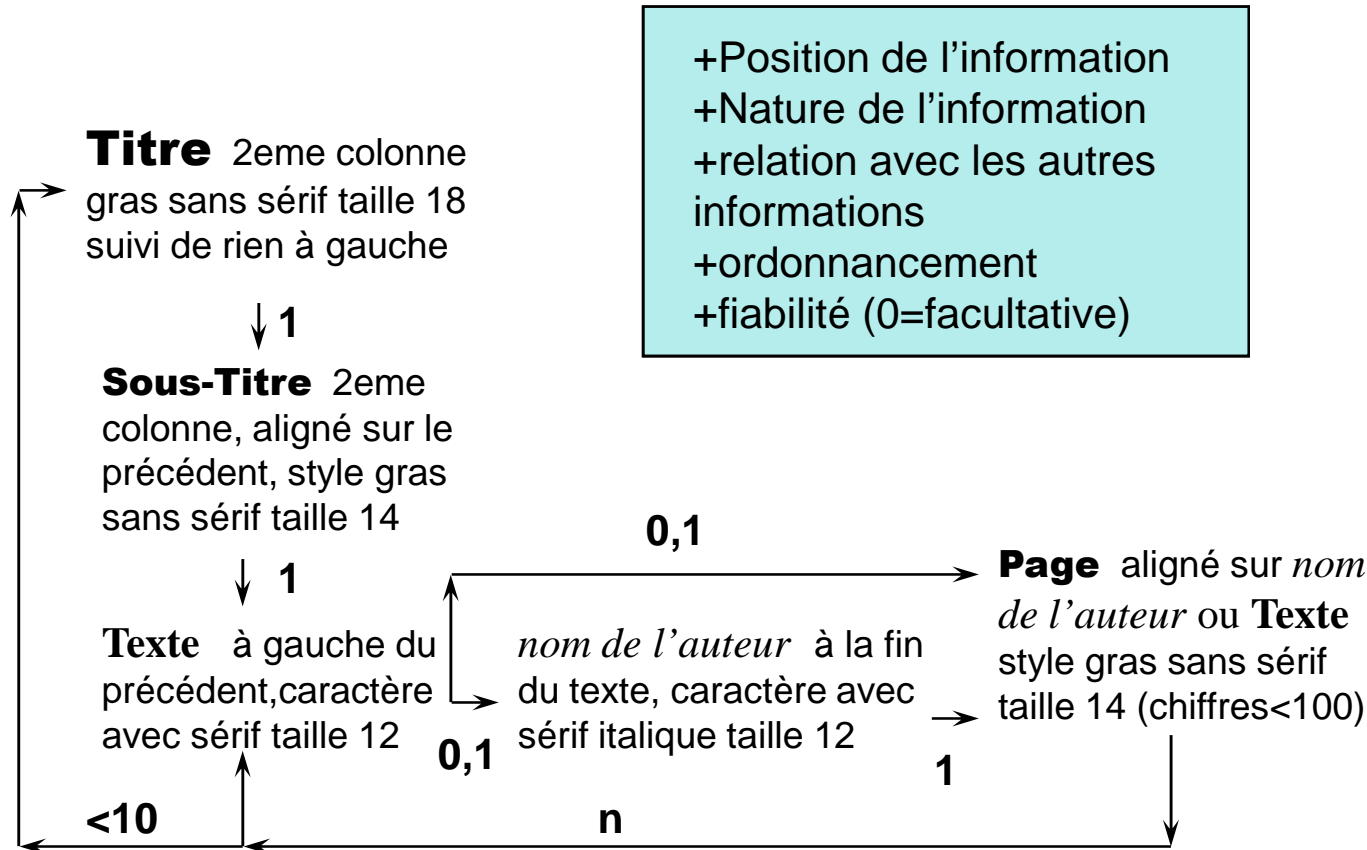


Les structures: pour quels usages?

- 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)

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	Newsmakers 31
Cyberscope	4	Transition 31
Perspectives	9	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

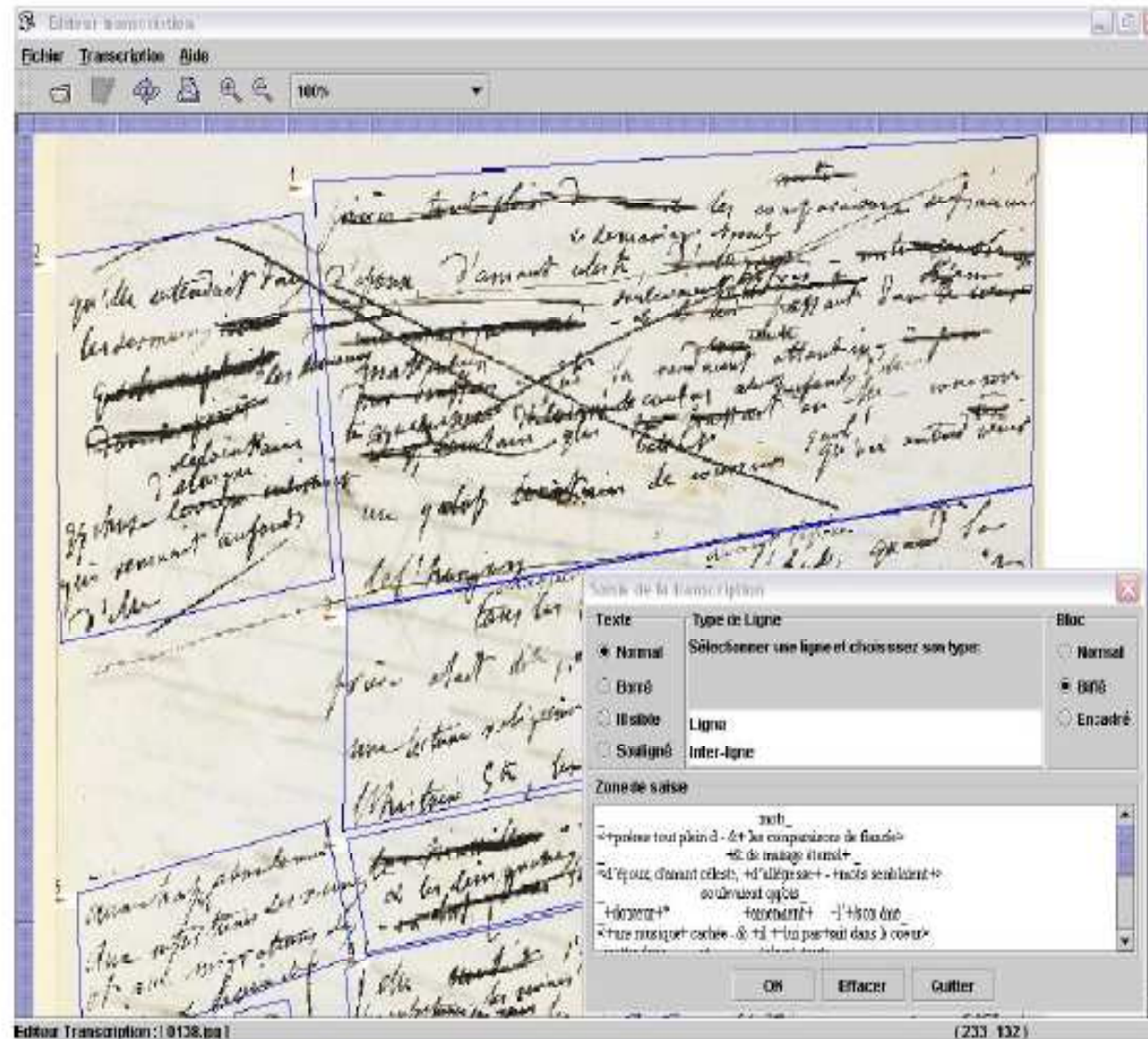
Toute segmentation est le produit d'un choix d'une méthode et de ses paramètres !

Plusieurs méthodes de segmentation :

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

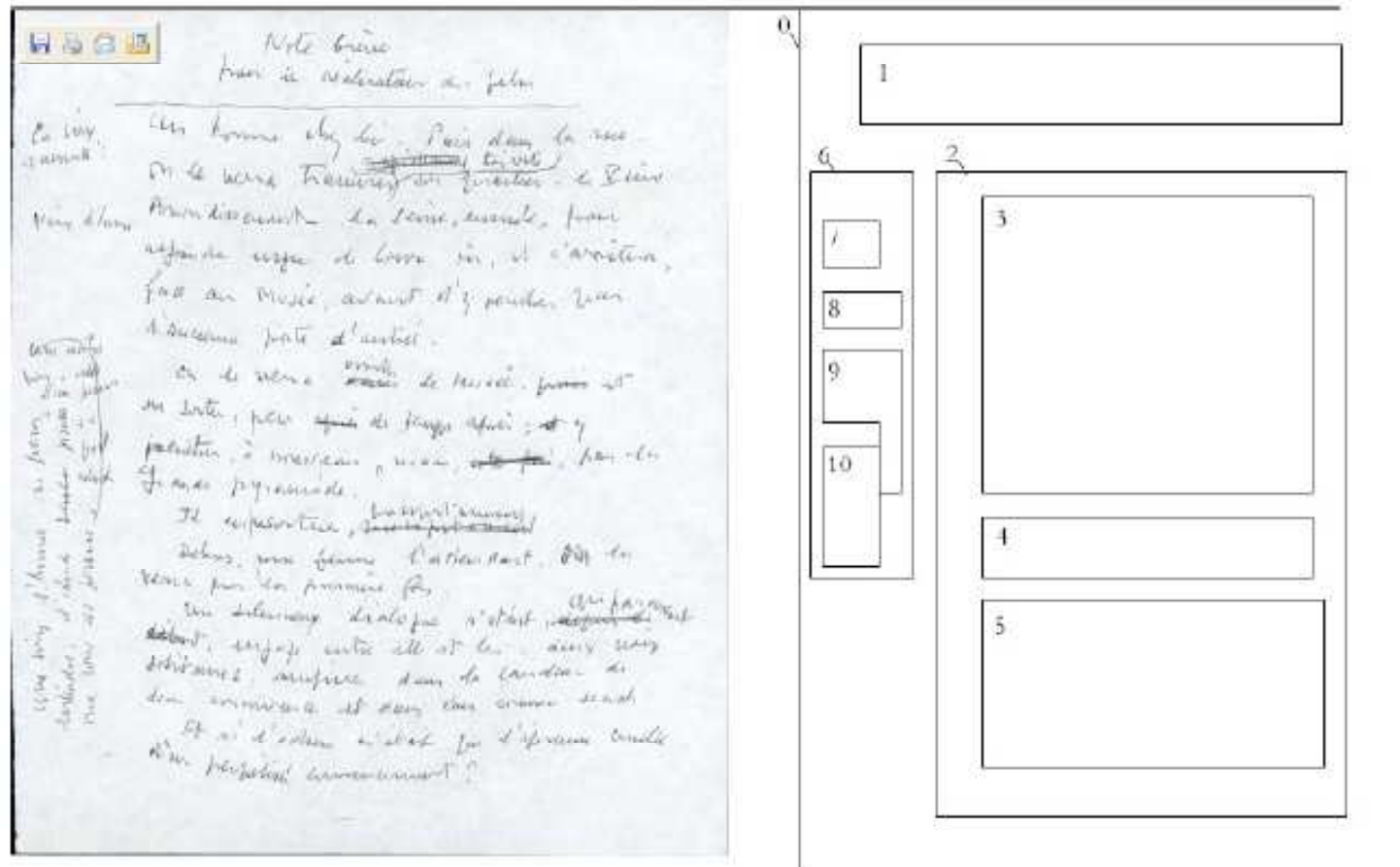
Cette opération n'est pas toujours réalisable, dépend de la qualité initiale de l'image

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

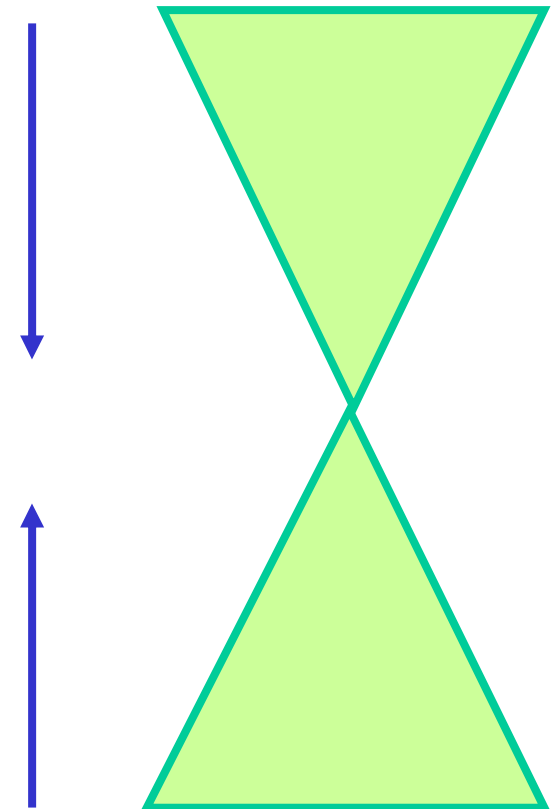
- Cas des documents réguliers composites
- Cas des courriers d'entreprises
- Cas des documents du patrimoine et des manuscrits irréguliers

La structure physique du document

La structure physique : les méthodes d'analyse

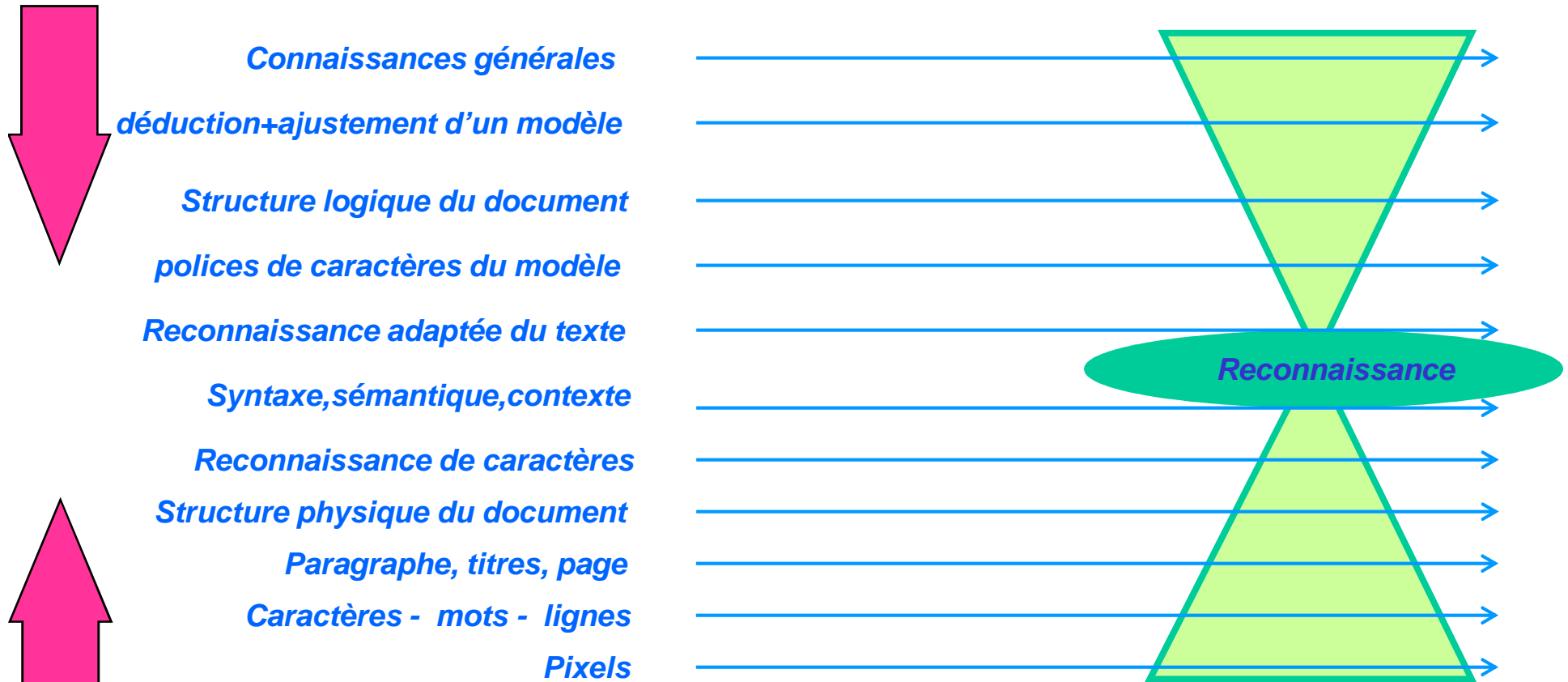
- *Les méthodes ne sont pas concurrentes*
- *Elles sont complémentaires*
- *On distingue plusieurs classes*

- *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		<p>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.</p> $F(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \cdot e^{-j2\pi u x} \cdot e^{-j2\pi v y} \, dx \, dy$ <p>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</p>	
<p>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 réduisant les écarts présentant un « intérêt » moindre.</p>		<p>Pour cela, nous utilisons des registres de sélection directement dérivées de certains constats psychologiques de capture de l'information visuelle. En particulier, des mesures expérimentales ont 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</p> <p>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 fixation successives que la description de la structure du document évoluera et aboutira à une représentation segmentée marquée par des résolutions différentes.</p> $G * H(i, j) = \sum_{m=0}^M \sum_{n=0}^N G(m, n) * H(i - m, j - n)$ <p>Les techniques permettant de sélectionner ces zones informative seront évoquées dans la suite. Elles son</p>	
1. Etat de l'art sur la segmentation des documents	1.2 Introduction de la multirésolution	2. La simulation des processus perceptifs	
1.1 Approches traditionnelles	1.2.1 La perception	<p>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.</p>	
<p>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 <i>pixel</i> 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 <i>physique</i> (correspondant aux objets physiques présents dans le document) et une information <i>logique</i> (relevant</p>	<p>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 <i>fenêtrage</i>), 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 reste du document (dans la zone périphérique). La résolution, de ce fait, n'est plus <i>uniforme</i> sur toute la surface du document, et les techniques utilisées pour</p>	<p>Un modèle de simulation du regard pour la structuration de documents composites</p> <p>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...), sont blocs écrits avec des polices différentes)</p>	
	1.2.2 Techniques d'extraction des données	<p>De manière générale, pour obtenir la segmentation physique du document, on a recours à des contourness.</p>	
	<p>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</p>	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 multi-résolution. 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 résommant 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 focale) 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

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 de documents en blocs homogènes (photographies, graphiques, textes, ... sous blocs écrits avec des polices différentes)

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?

Voilà 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

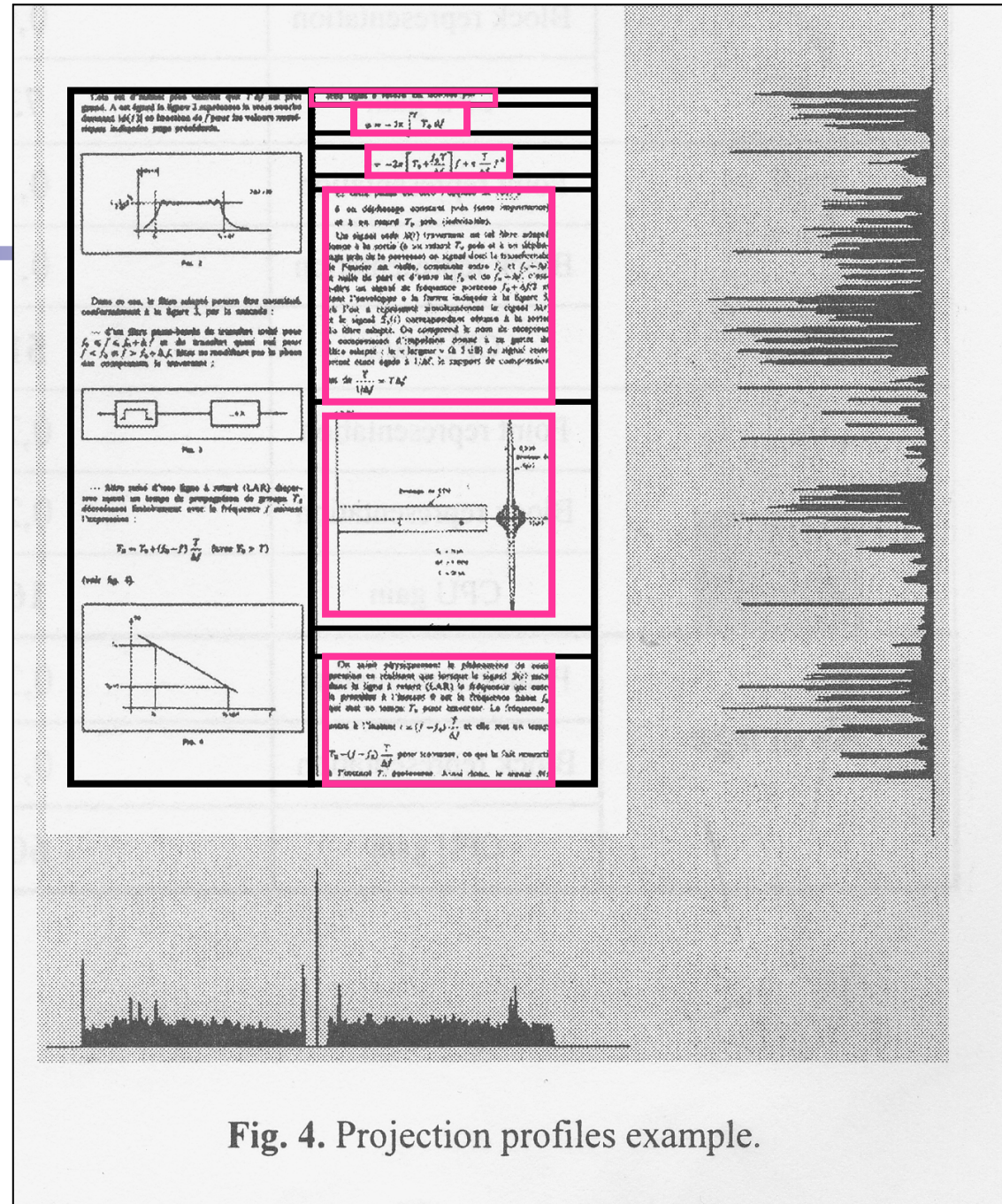


Fig. 4. Projection profiles example.

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 documents écrits : 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 réduisant les zones qui ne sont pas d'intérêt.

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

2. La simulation des processus perceptifs

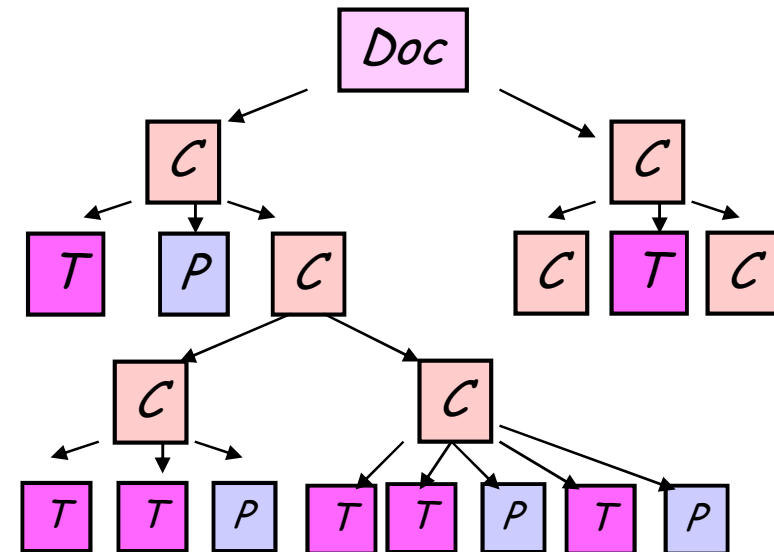
Dans notre cas, analyser l'image à différentes résolutions ne signifie pas simplement considérer des portions de documents limitant des zones entières (comme dans les techniques de fenêtrage), mais au contraire, définir un mode d'observation en zones, conservant une forte résolution au point de focalisation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

2.2 Techniques d'extraction des données

Cet appareil donc utile, peut-être même indispensable pour certaines applications, de reconstituer l'image à des résolutions variables, de façon à privilégier l'information

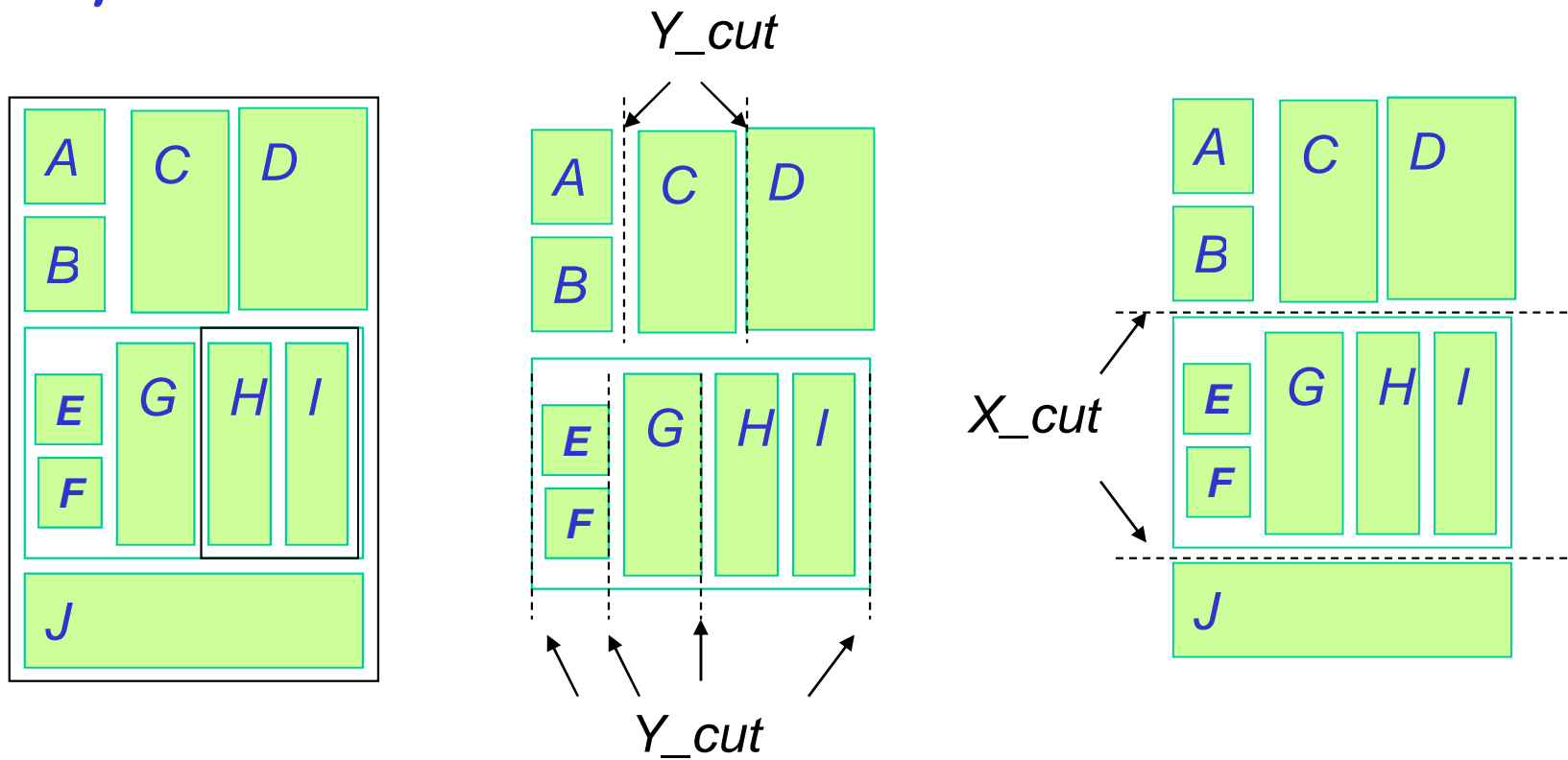
Un modèle de simulation du regard pour la structuration de documents composites

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



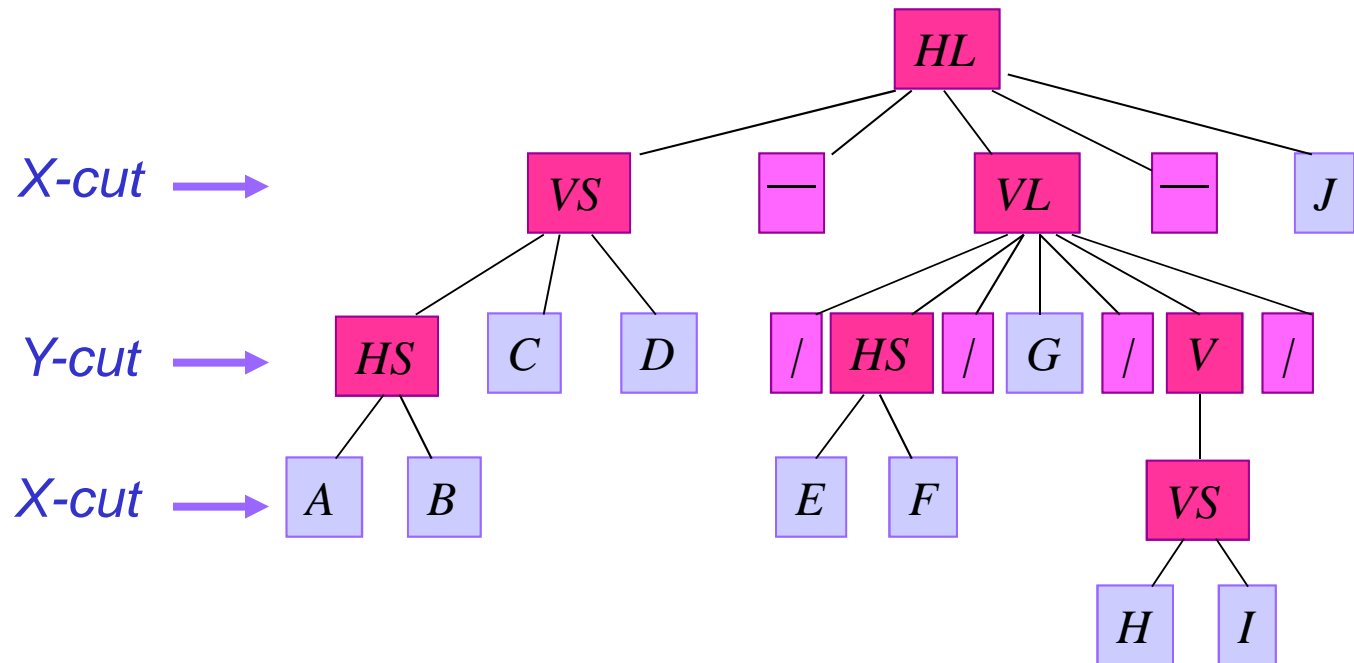
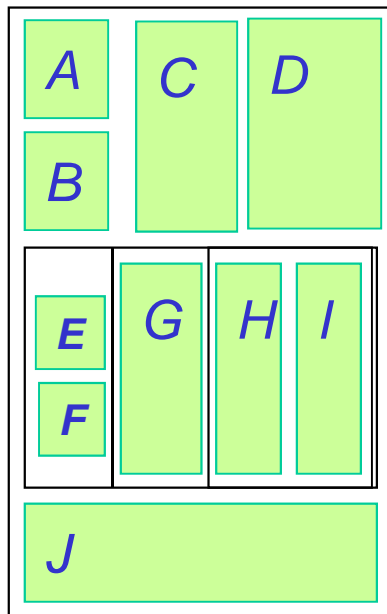
Approches descendantes (top-down)

La représentation X-Y Tree



Approches descendantes (top-down)

La représentation X-Y Tree



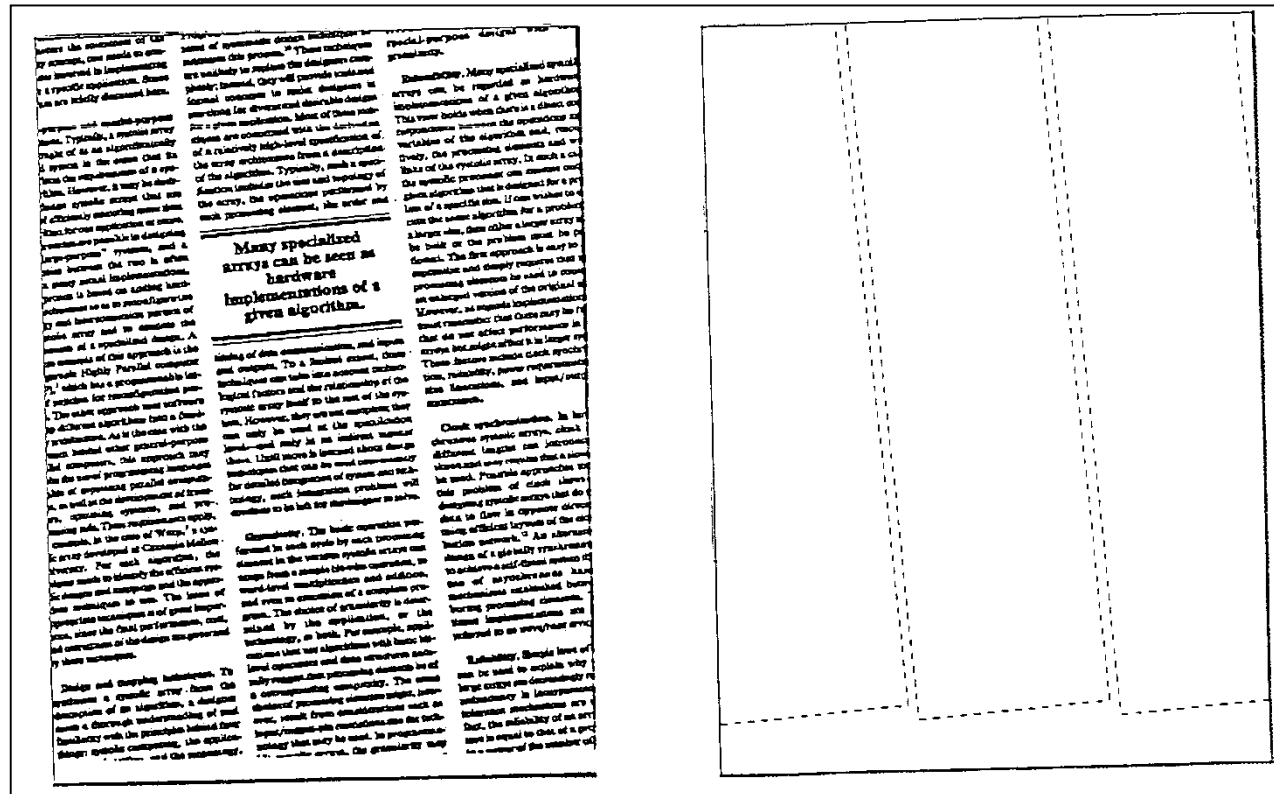
HL VL : horizontal and vertical lines

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]



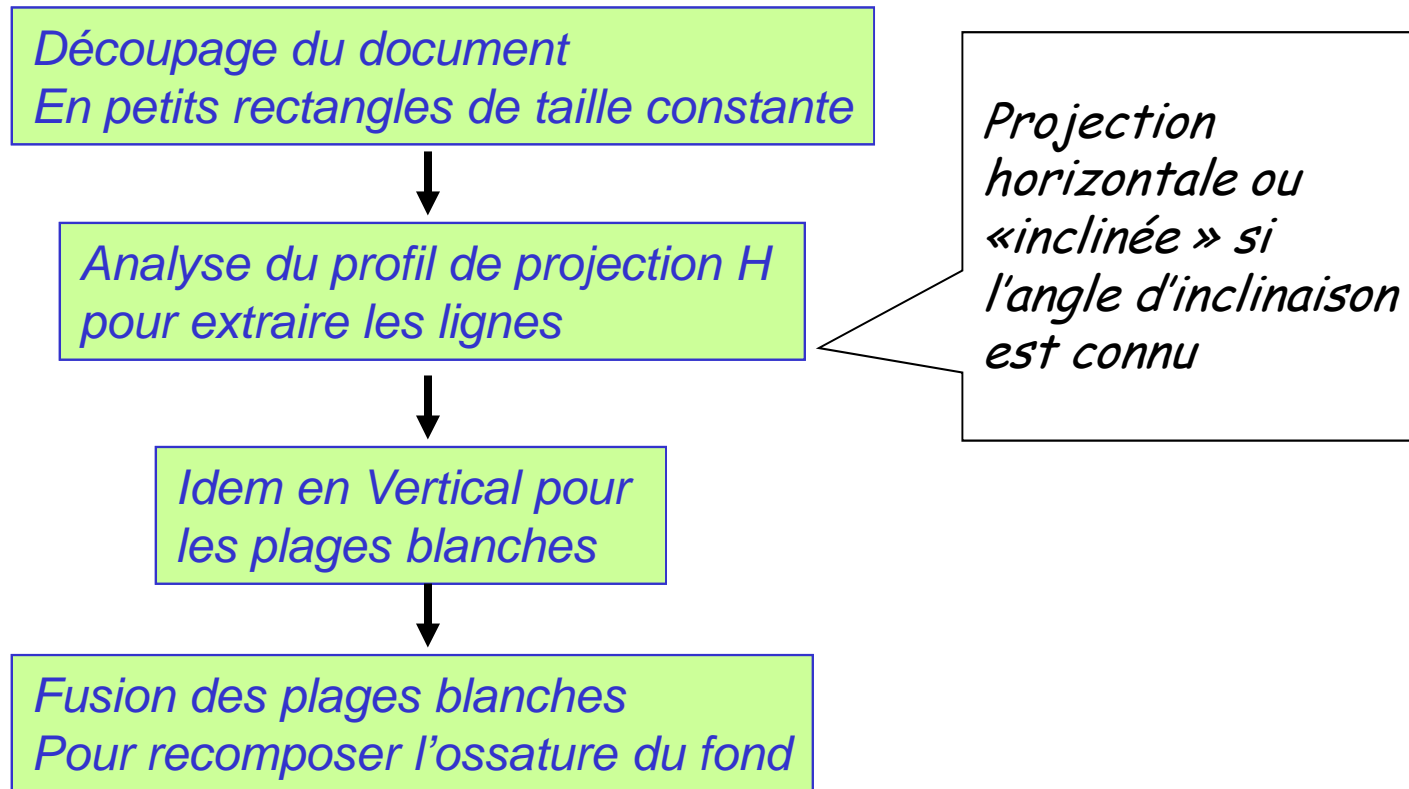
Approches descendantes (top-down)

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 multirésolution

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

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 multirésolution

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

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 multirésolution

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 :

- ⇒ diminution des détails
- ⇒ fusion des composantes connexes en mots
- ⇒ des mots en lignes
- ⇒ des lignes en blocs

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'analyse de contours à plusieurs résolutions. Cette approche 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 multirésolution

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

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'analyse de contours à plusieurs résolutions. Cette approche 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 multirésolution

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

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'analyse de contours à plusieurs résolutions. Cette approche 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 multirésolution

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)

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.*

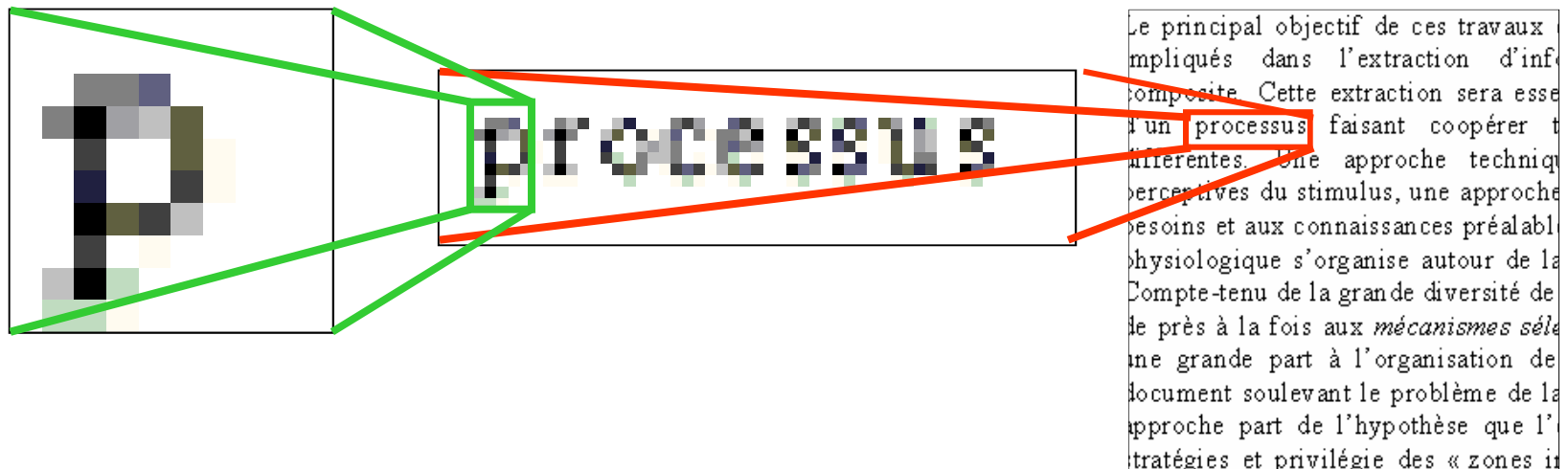
*Les approches
ascendantes (bottom-up)*

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

The image displays a three-stage process of image segmentation and character recognition. On the left is the original document page, which includes a table of contents and several columns of text. The middle section shows the same page with individual objects (text blocks, images, and graphics) segmented and outlined. The right section shows these segmented objects grouped into characters, with each character represented by a small icon and a corresponding label.

Image originale

Segmentation des objets

Regroupement des objets en caractères

Des méthodes pour une segmentation physique

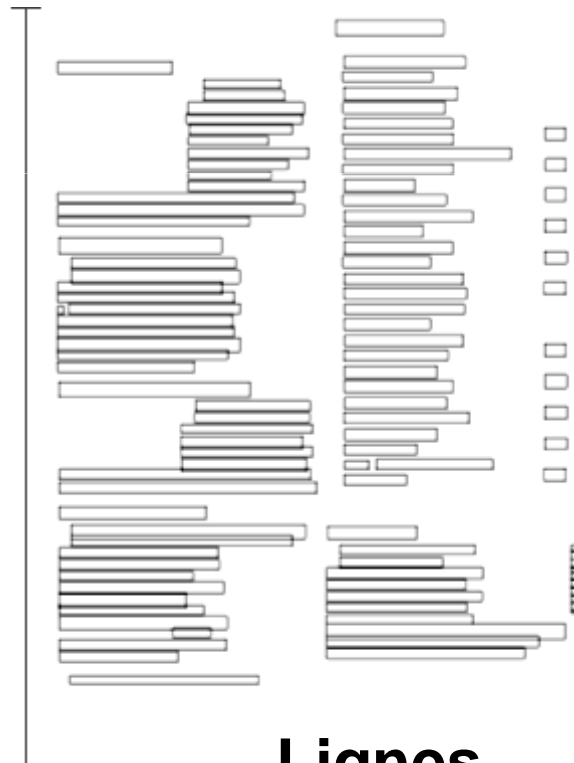
Mots = Regroupements de caractères faiblement espacés

Lignes = Regroupement de mots adjacents faiblement espacés

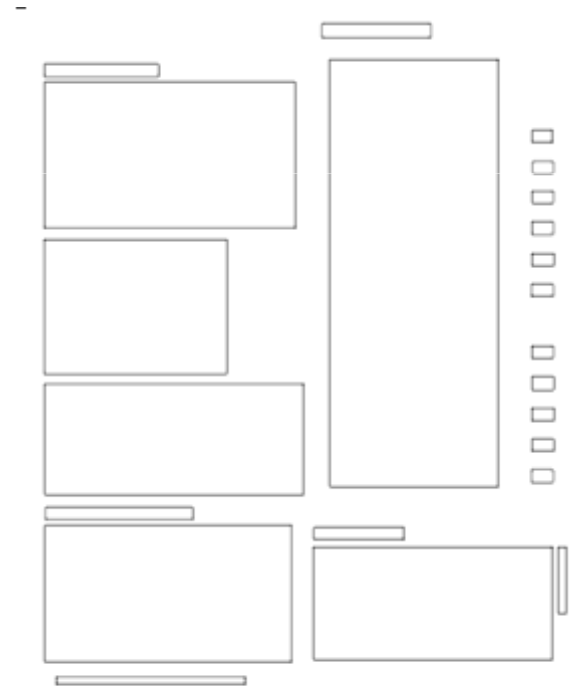
Paragraphes = Regroupement de lignes de faible interligne



Mots



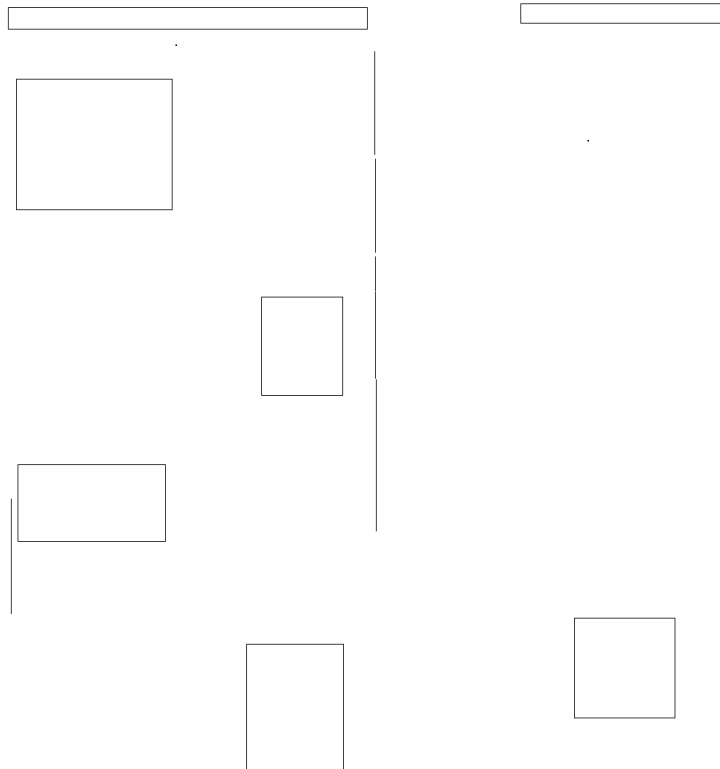
Lignes



Paragraphes

Des méthodes pour une segmentation physique

Illustrations = Objets non classés comme du texte



Illustrations

SOMMAIRE

Fini et sans limites

H eriter du terme grec *ἄπειρον* (de *peira* qui signifie limite), le mot latin *infinitum* et son dérivé français « in-fini » ont longtemps été considérés comme le synonyme de « illimité ». On peut cependant envisager, sans contradiction aucune, des espaces finis ne possédant aucune limite. C'est notamment le cas de la surface fluide d'une sphère ou d'un tore (espace à deux dimensions) qui ne comporte aucune frontière. (*La clôture du monde. Page 19.*)

Le passage du fini à l'infini
E n 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 allons 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 76.*)

A la conquête de l'espace infini

P our 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
L on ne 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'un 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 dodécèdres déformés par des illusions d'optique. (*Notre univers est-il chiffonné? Page 12.*)

Les infinis en question
par Jean Seidenhart

Les figures de l'infini
par Jean-Louis Ferrier

La clôture du monde
par Jean-Pierre Luminet **10**

Notre univers est-il chiffonné ?
par Jean-Pierre Luminet **12**

L'œil intégral
par Jean-Louis Ferrier **18**

Les angoisses de l'infini
par Jean Coltraux **20**

Répétition indéfinie
par Claude Frontisi **24**

L'infiniment complexe
par Gilles Cohen-Tannoudji **26**

Combinaisons sans fin
par Claude Frontisi

L'infini en perspective
par Jean-Pierre Le Goff **32**

L'espace défiguré
par Jean-Pierre Luminet **38**

L'infini en pratique
par Jean-Marc Lévy-Leblond **40**

Jardin labyrinthe
par Sophie Rigat **44**

Dieu, l'infini et les nombres
par Tony Lévy **46**

4 **HORS-SÉRIE SCIENCES ET Avenir - MARS 1996**

Superposition avec l'image

Structure physique hiérarchique du texte

(Colonnes>paragaphes>Lignes>mots>caractères>accents et ponctuation...)

Future Shock by John Barry
Which Countries Have the Best
Who's Sorry Now? by Bill Power
No Apologies by Kenneth Auchi

Business

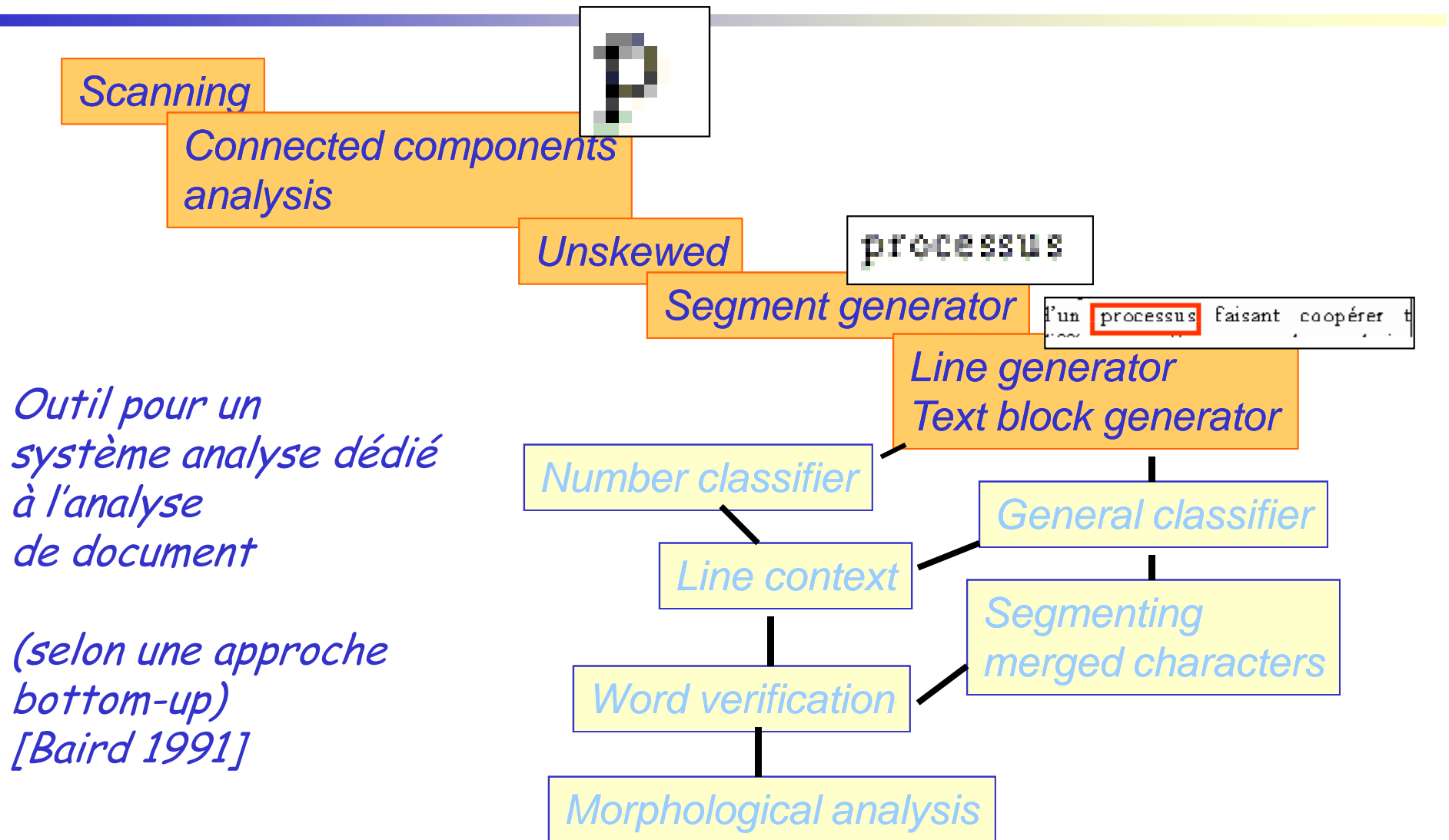
Chile: No Tequila Hangover

by Michael Hirsh

Stocks: Lusting After Wall Street

Résultat de la Segmentation de la structure physique

Approches ascendantes (bottom-up)

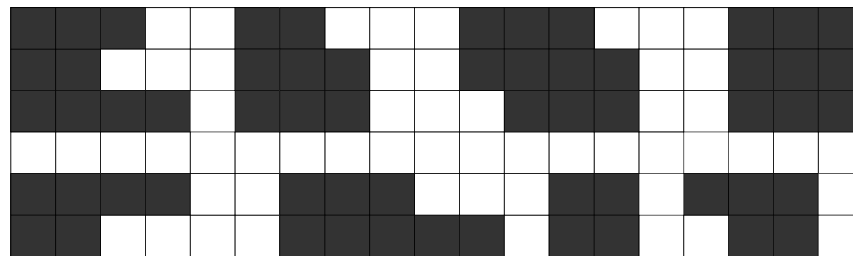


Approches ascendantes (bottom-up)

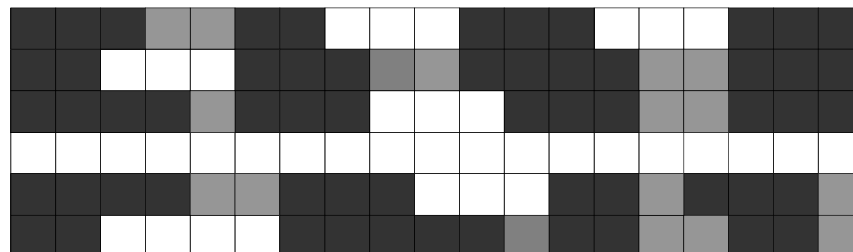
La technique de lissage directionnel




La technique du RLSA

(Run-Length Smoothing Algorithm, Wong 82)



Seuil = 3



-  *Pixel objet*
-  *Pixel bouché*
-  *Pixel blanc*

Approches ascendantes (bottom-up)

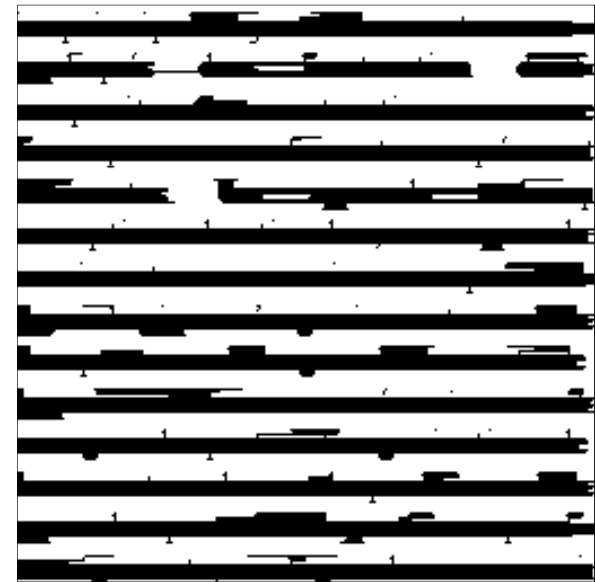
La technique du RLSA

Le principal objectif de ces travaux est d'être impliqués dans l'extraction d'informations d'un document composite. Cette extraction sera essentiellement un processus faisant coopérer techniques différentes. Une approche technique basée sur les besoins et aux connaissances préalables physiologiques s'organise autour de la lecture. Compte-tenu de la grande diversité de documents, on se propose de près à la fois aux *mécanismes sélectifs* de la lecture, une grande part à l'organisation de l'information dans le document soulevant le problème de la lecture. Cette approche part de l'hypothèse que l'information est traitée par des stratégies et privilégie des « zones in-

Image

Le principal objectif de ces travaux est d'être impliqués dans l'extraction d'informations d'un document composite. Cette extraction sera essentiellement un processus faisant coopérer techniques différentes. Une approche technique basée sur les besoins et aux connaissances préalables physiologiques s'organise autour de la lecture. Compte-tenu de la grande diversité de documents, on se propose de près à la fois aux *mécanismes sélectifs* de la lecture, une grande part à l'organisation de l'information dans le document soulevant le problème de la lecture. Cette approche part de l'hypothèse que l'information est traitée par des stratégies et privilégie des « zones in-

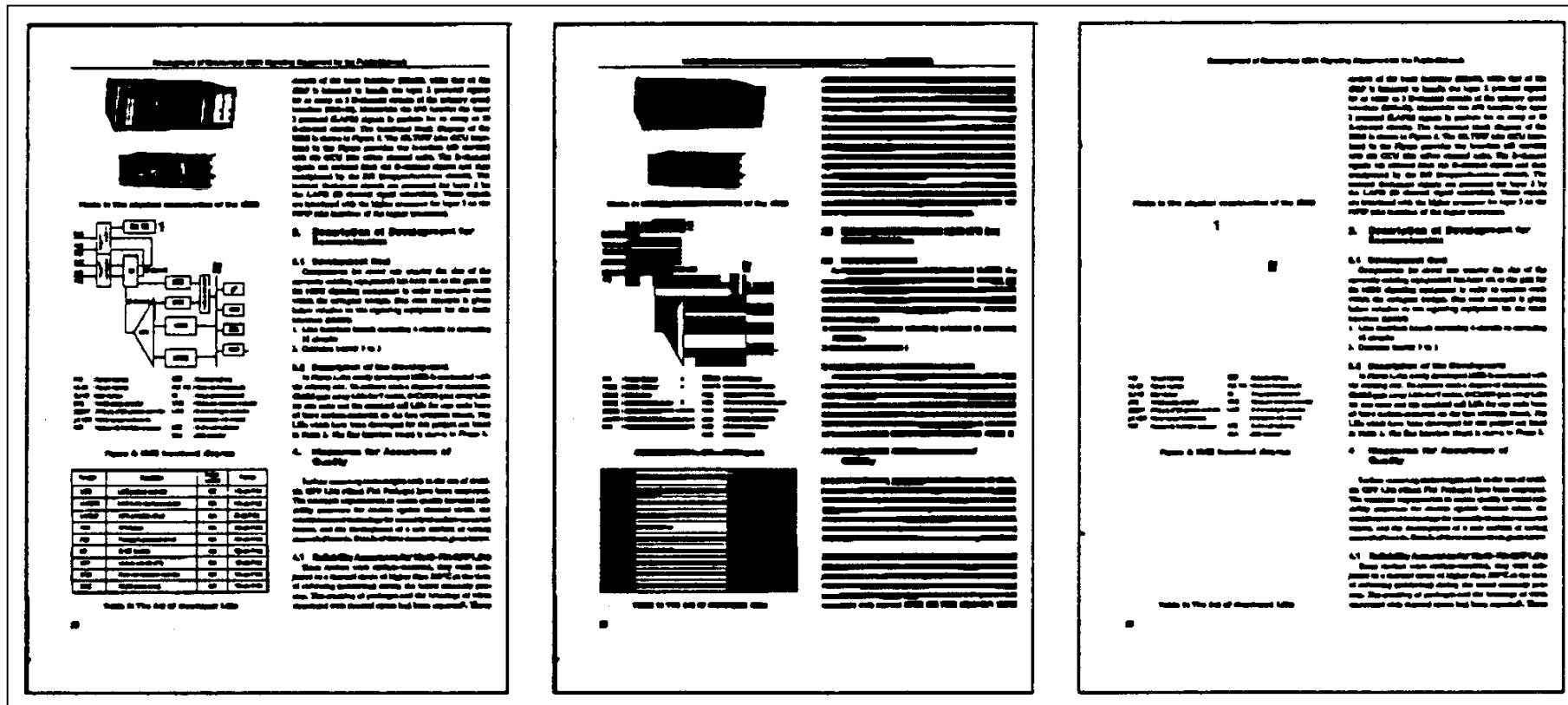
Seuil =3



Seuil =20

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

The image shows two pages from a technical document, labeled 4 and 5. Page 4 is titled "4. Profils de projection" and page 5 is titled "5. Résultats". Both pages contain dense text, diagrams, and tables.

Page 4: Profils de projection

This page contains several sections of text, some of which are partially obscured by black redaction bars. There are also some diagrams and tables, though they are mostly illegible due to the redaction.

Page 5: Résultats

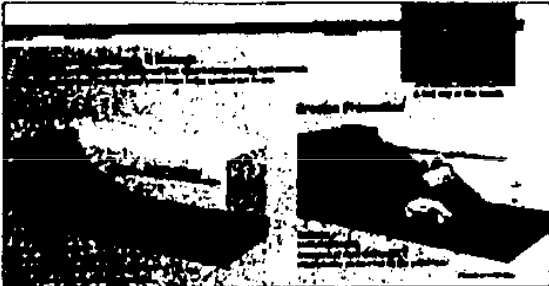
This page contains a large table with multiple columns and rows, likely representing data from an experiment or simulation. The table is partially obscured by black redaction bars. There are also some diagrams and text blocks on this page.

4. Profils de projection

5. Résultats

Approches ascendantes (bottom-up)

Suivi de segments pour l'extraction de blocs polygonaux



Drawing the Line on Erosion

Coastal engineers are scrambling to keep beachfront properties from washing away.

As the ocean's reach extends further inland, coastal engineers are scrambling to keep beachfront properties from washing away. The problem is not just the loss of sand, but the loss of the very structure of the beach itself. In some areas, the ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.



The erosion is not uniform. In some places, the ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.



Coastal Erosion

The ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.



The erosion is not uniform. In some places, the ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.



Coastal Erosion

The ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.



The erosion is not uniform. In some places, the ocean is eating away at the land so fast that homes and businesses are in jeopardy. Engineers are trying to find ways to hold the line, but it's a constant battle against the forces of nature.

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) :

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

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

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

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

On plane formed by $\hat{u}_{i,j}$ and the focal point of the camera must include $\hat{v}_{i,j}$. On this plane the projection of $\hat{v}_{i,j}$ is constant $c_{i,j}$.

$$c_{i,j} = \hat{u}_{i,j} \times \hat{u}_{i,j+1} \quad (1)$$

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

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

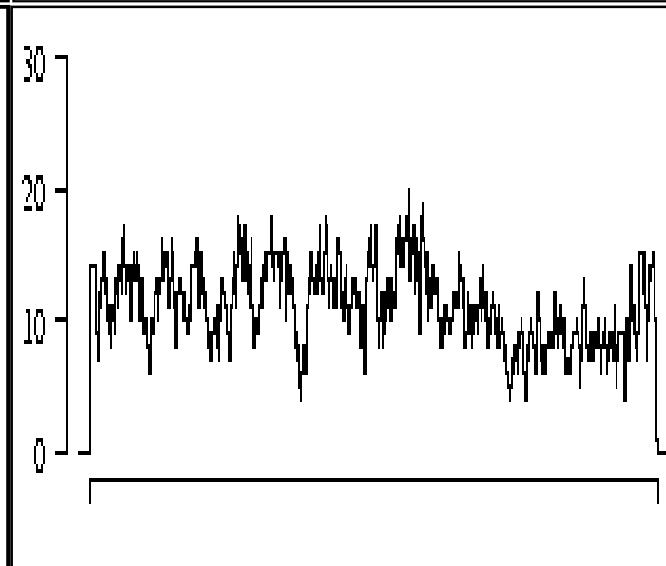
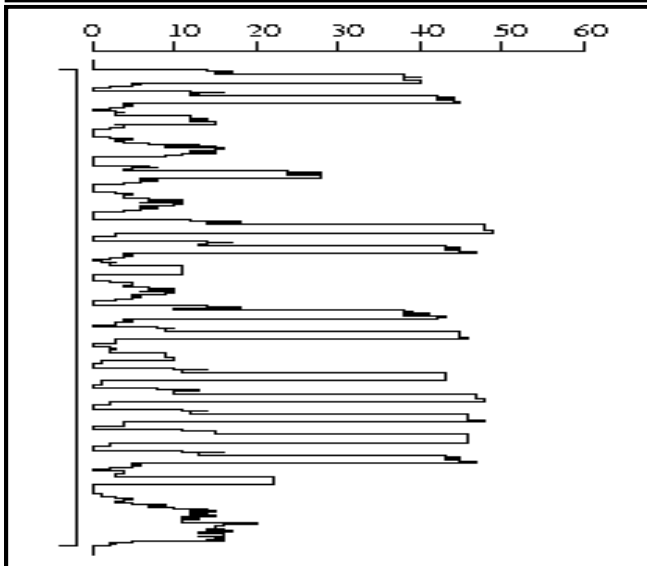
In the case of purely translational motion, the direction of $\hat{v}_{i,j}$ is constant for all i . Therefore, Equation 2 can be rewritten as

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

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 $c_{i,j}$ from Equation 1, the error measure is defined as

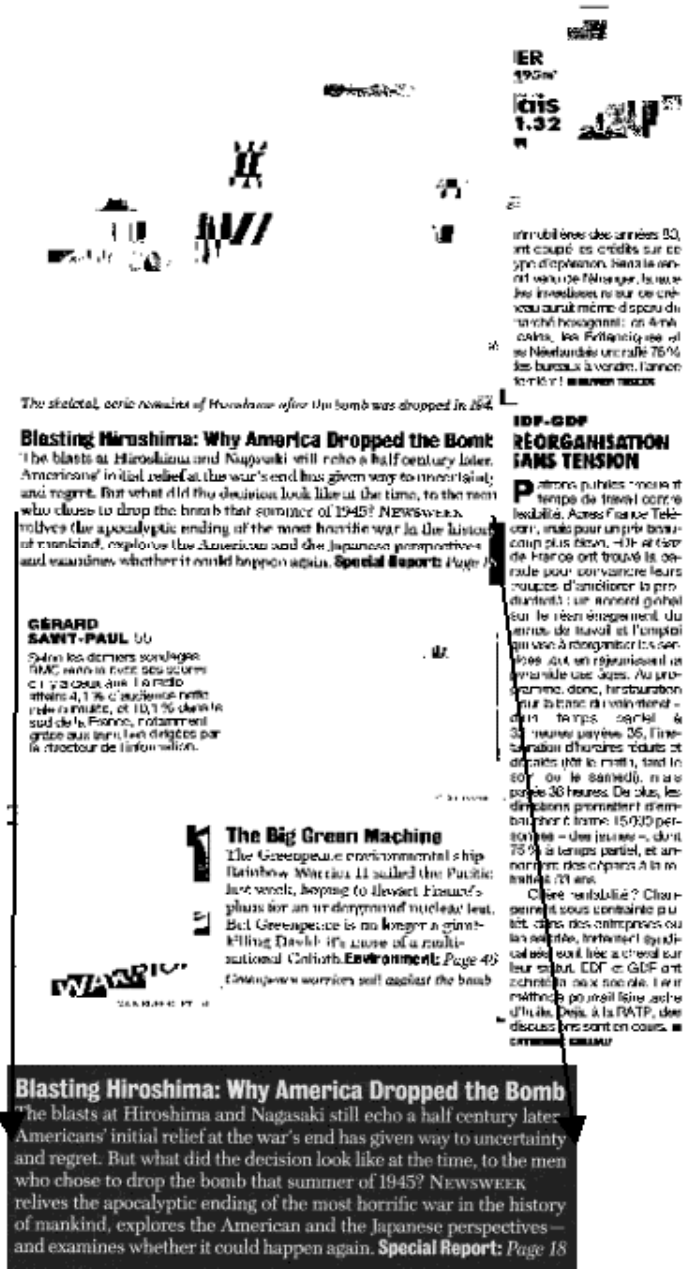
$$\frac{1}{m} \sum_{i=1}^m \left| \sin^{-1} \left(\frac{c_{i,j} \cdot \hat{v}_j}{\|c_{i,j}\| \|\hat{v}_j\|} \right) \right| \quad (4)$$



2. Segmentation par BINARISATION

Binarisation adaptative dans chaque ligne

Image binaire obtenue



ICIET
LVMM

Le numéro en minuscule du...
L'année dernière, le chiffre d'affaires de la vente d'un tiers de sa participation dans Guinness. Une cagnotte appréciée au moment où LVMM possédait le réseau de l'américain Unity Auto Shopper (UAS). L'année des pièces perforées de Guinness, un allié qui avait été à portée de canon de LVMM.

Bernard Arnault a cédé la semaine passée 7% du capital du bureau de retombée. « Pro formes de rétrocession, les performances unitaires de LVMM et de Guinness sont moindres que celles d'autres », explique-t-il à l'analyse. Mais la France - qui termine deuxième dans le classement des entreprises françaises - a été devancée par les États-Unis.

LE NOUVEAU PEL

Les Français ne s'y sont pas trompés. Pour continuer à profiter des plans d'épargne-logement à 5,25%, ils se sont rendus massivement dans leurs agences bancaires. Le taux de placement est, dans l'ensemble, de 900 milliards de francs, soit en effet d'être abaissé à 4,25%. Pour assurer cette mesure, le taux des prêts immobiliers obligés grâce à ces nouveaux PEL a été abaissé de 5,54% à 4,50%. Une compensation qui n'est certes pas seulement au lieu des souscripteurs de PEL, mais aussi des emprunteurs de PEL, car les banques ont pu profiter en plein du taux de ce placement.

The skeletal, ceramic remains of Hiroshima after the bomb was dropped in 1945.

Blasting Hiroshima: Why America Dropped the Bomb

The blasts at Hiroshima and Nagasaki still echo a half-century later. Americans' initial relief at the war's end has given way to uncertainty and regret. But what did the decision look like at the time, to the men who chose to drop the bomb that summer of 1945? NEWSWEEK relives the apocalyptic ending of the most horrific war in the history of mankind, explores the American and the Japanese perspectives—and examines whether it could happen again. **Special Report: Page 18**

GÉRARD SAINT-PAUL 60

Quels les derniers sondages RMC ont vu ses scores chuter à 1,3% et à 1,1% d'adhésion nette (de 1,1% à 1,0%) dans le sud de la France, notamment grâce aux services dirigés par le directeur de l'information.

The Big Green Machine

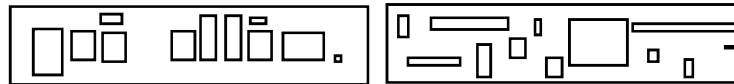
The Greenpeace environmental ship Rainbow Warrior II sailed the Pacific last week, hoping to dissuade Russia's plans for an underground nuclear test. But Greenpeace is no longer a grinning David: it's a cause of a multinational Colossus. **Environment: Page 40**

www.warrior.org will assist the ship

Blasting Hiroshima: Why America Dropped the Bomb
The blasts at Hiroshima and Nagasaki still echo a half-century later. Americans' initial relief at the war's end has given way to uncertainty and regret. But what did the decision look like at the time, to the men who chose to drop the bomb that summer of 1945? NEWSWEEK relives the apocalyptic ending of the most horrific war in the history of mankind, explores the American and the Japanese perspectives—and examines whether it could happen again. **Special Report: Page 18**

2. Segmentation par BINARISATION

Séparation texte/graphique : cohérence des alignements

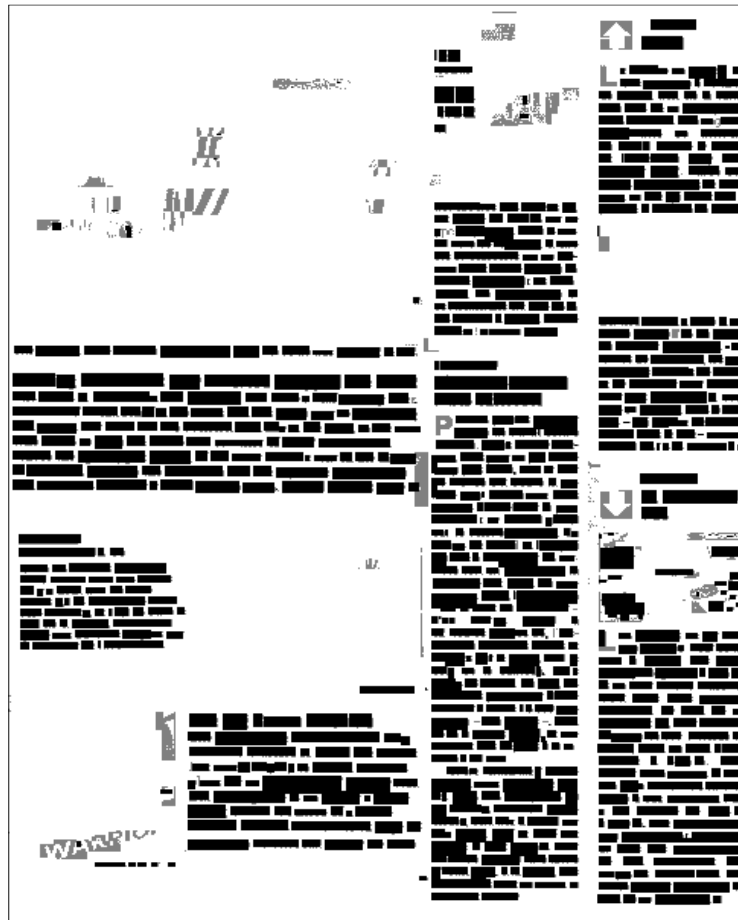


Texte

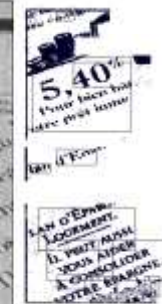
graphique

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

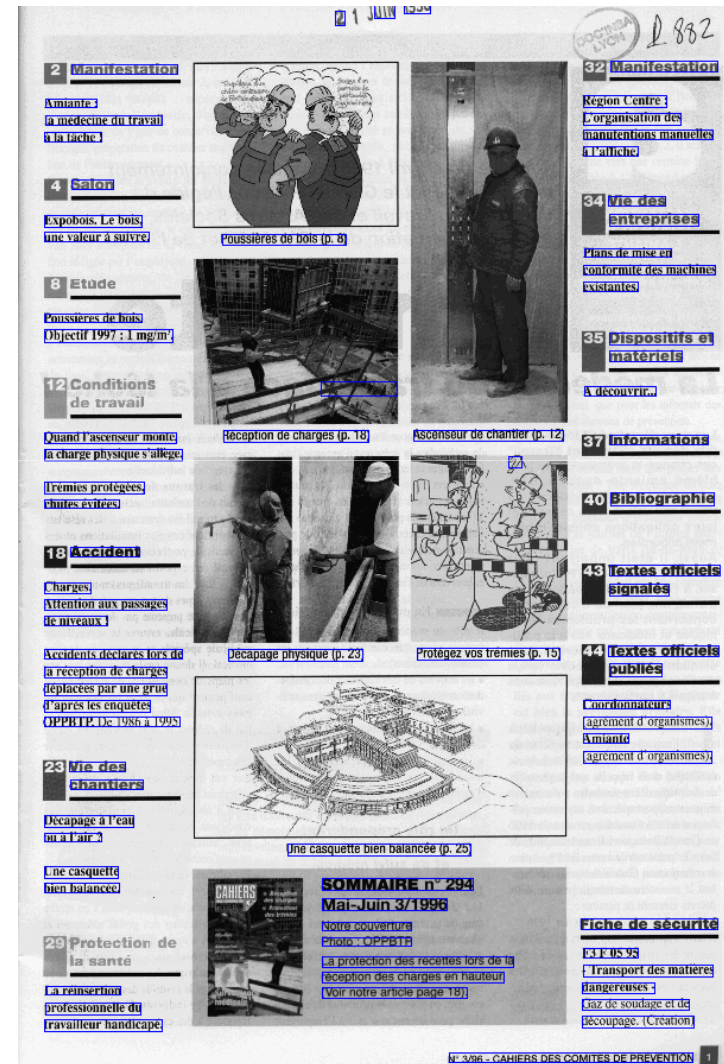
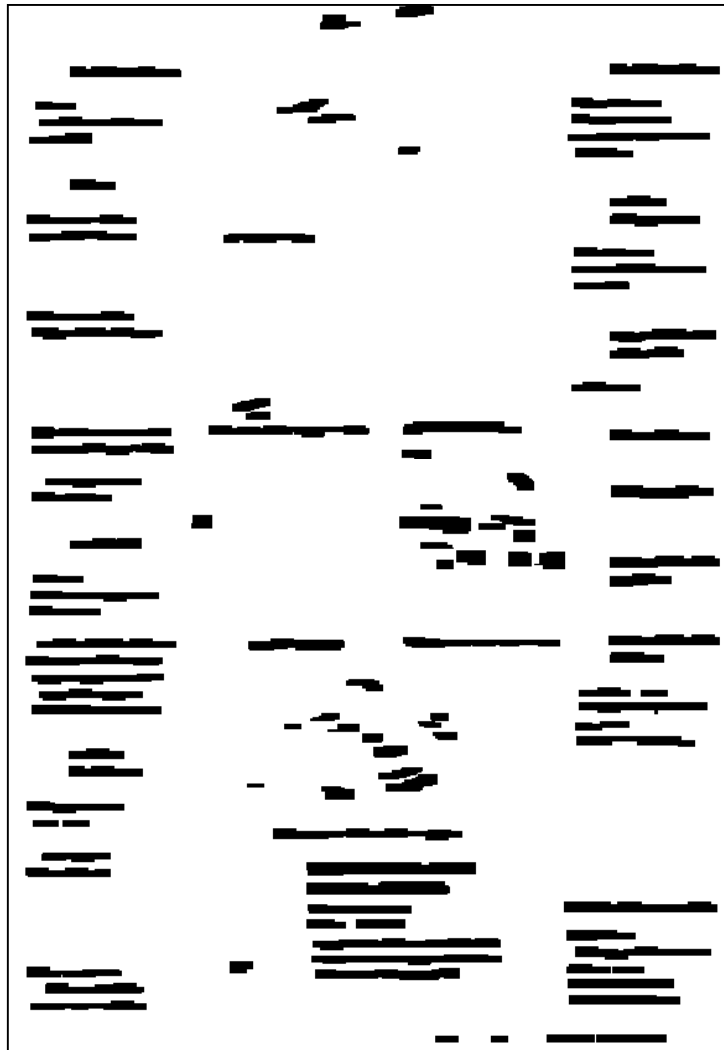


Image binaire



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

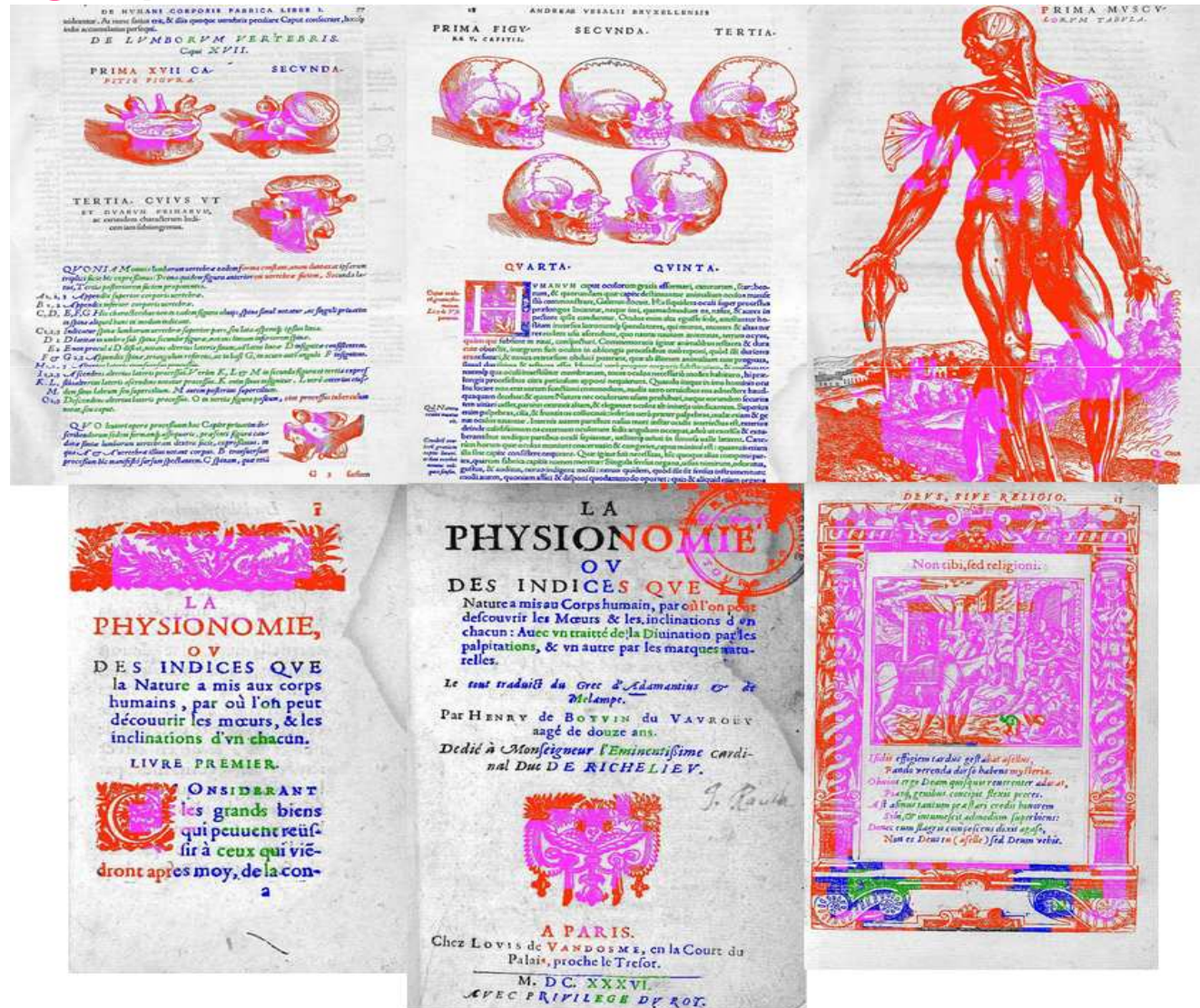
Par filtrage des variations d'intensité lumineuse de l'image (nuances de gris)

- ▶ 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 un seuil global.

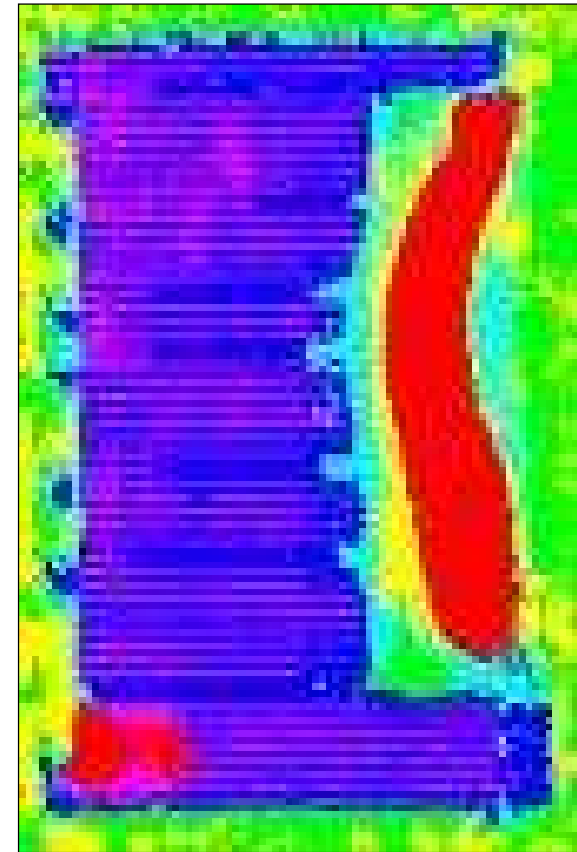
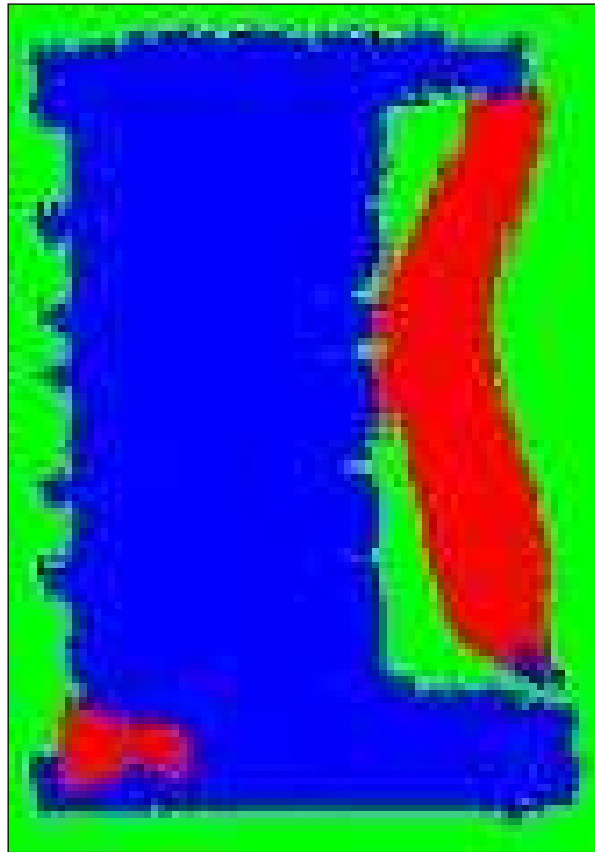
3. Séparation texte/graphique par texture

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

Détection et interprétation des orientations
(thèse Nicholas Journet – La Rochelle 2006)

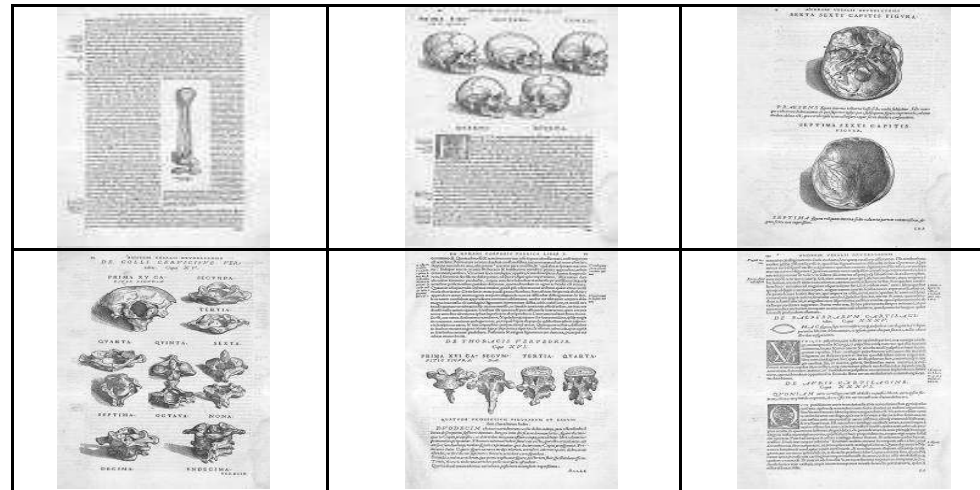


3. Séparation texte/graphique par texture

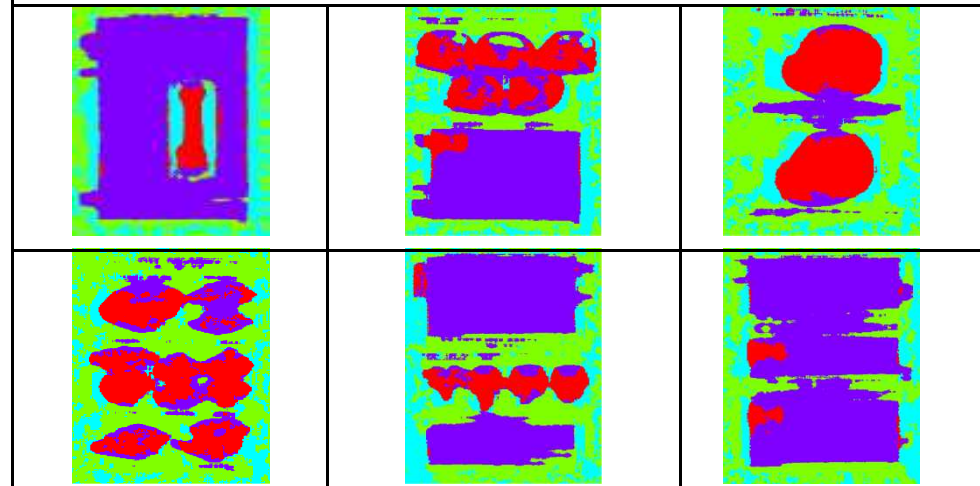


Détection et interprétation de critères de texture par classification: orientations, densité, complexité - Nicholas Journet – La Rochelle 2006

3. Séparation texte/graphique par texture



Tested pages (from the same book)



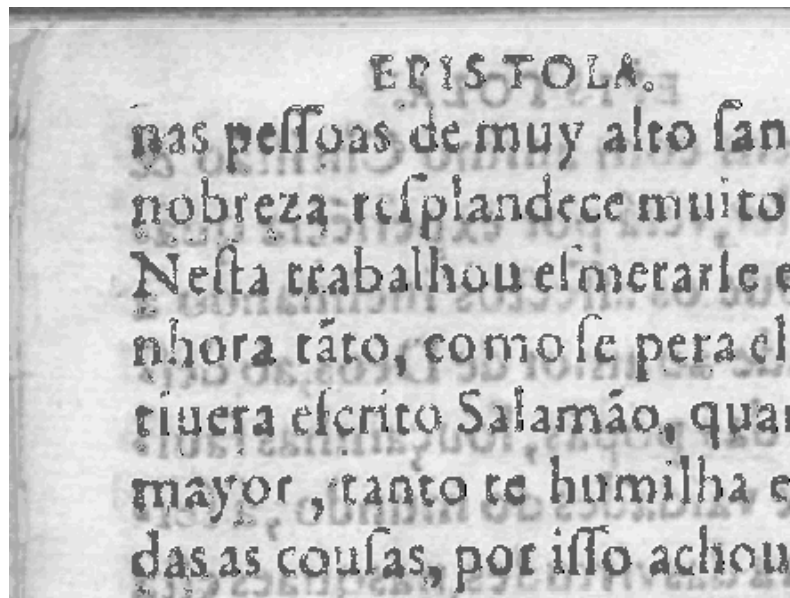
Results of a 4 classes classification
(simulation of text/drawings separation)

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.

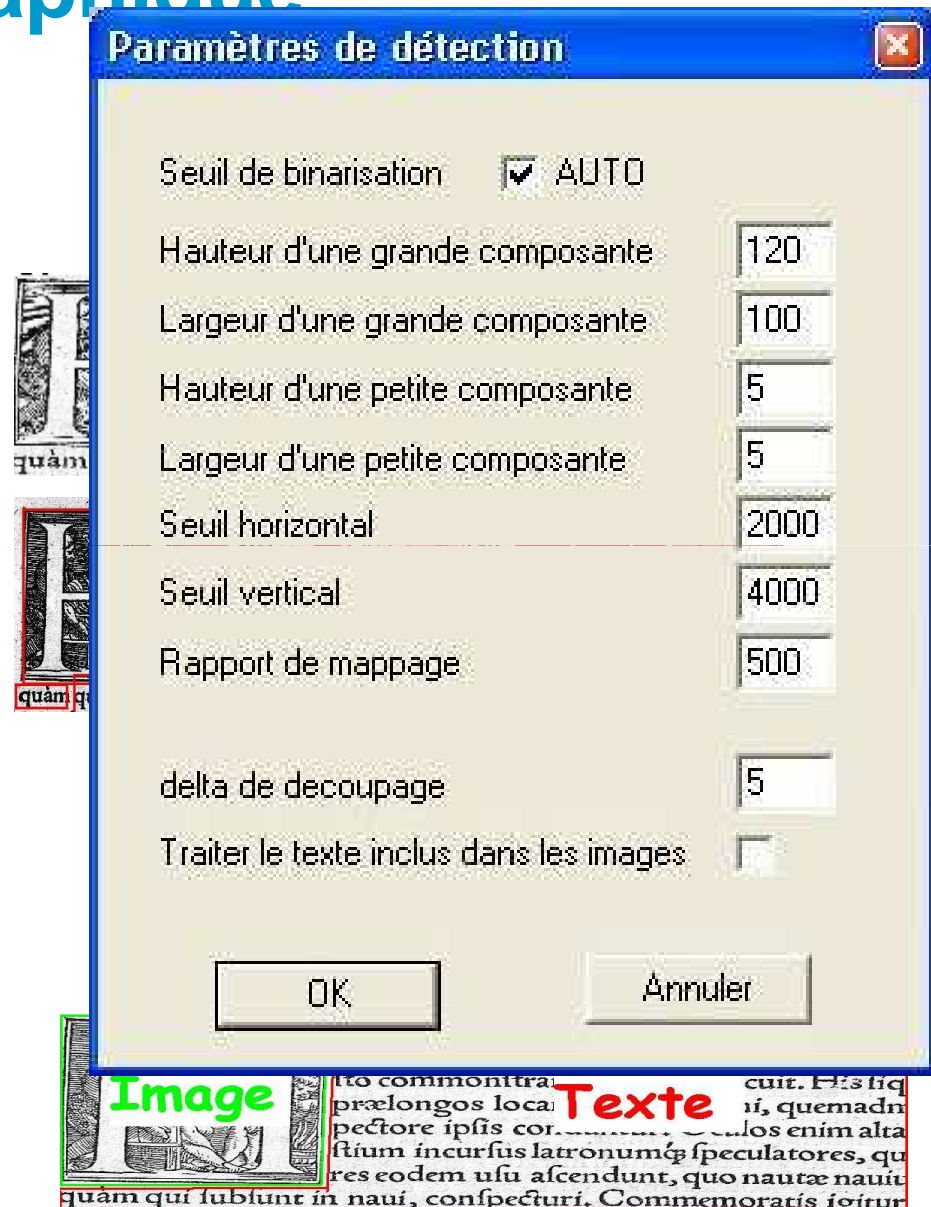
Une segmentation optimale correspond à un taux maximal de redondance des formes de caractères !



EPIS TOLA.
nas pessoas de muy alto san
nobreza resplandece muito
Nesta trabalhou elmerarle e
nhora tão, como se pera el
tiuera escrito Salamão, qua
mayor, tanto te humilha e
das as cousas, por isso achou

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 généralité (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'utilisateur) 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

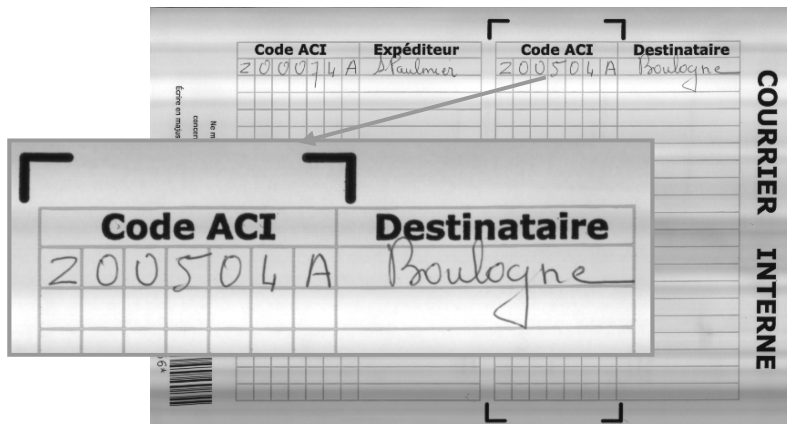
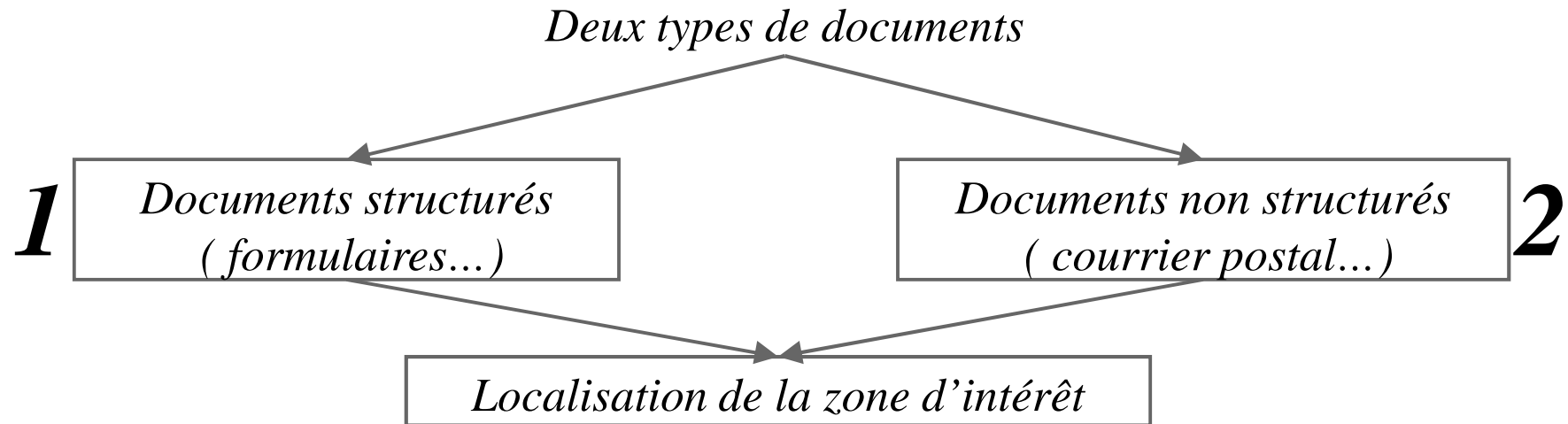
Le tri automatique des documents : problématique et besoins récents

Familles de documents et de courriers d'entreprises

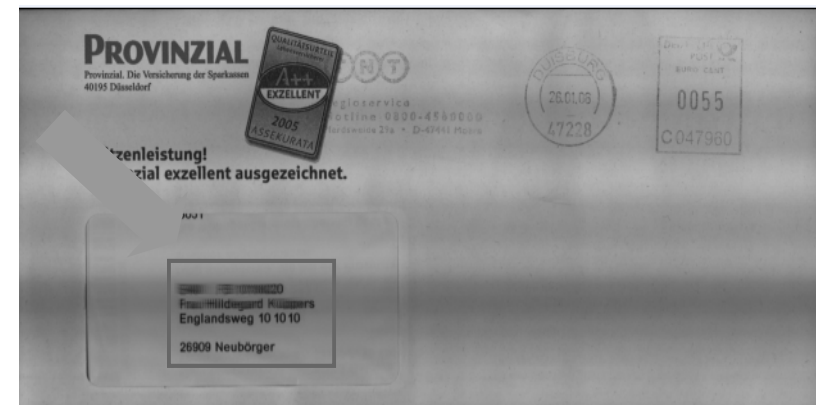
- Courrier postal
- Courrier interne manuscrit
- Courrier interne dactylographique
- Formulaire
- Planus
- Carte bleue
- Listing A3
- Listing A4
- NPAI
- Chèque circulant
- Pochettes DVD, etc.



Le tri automatique des documents : problématique et besoins récents



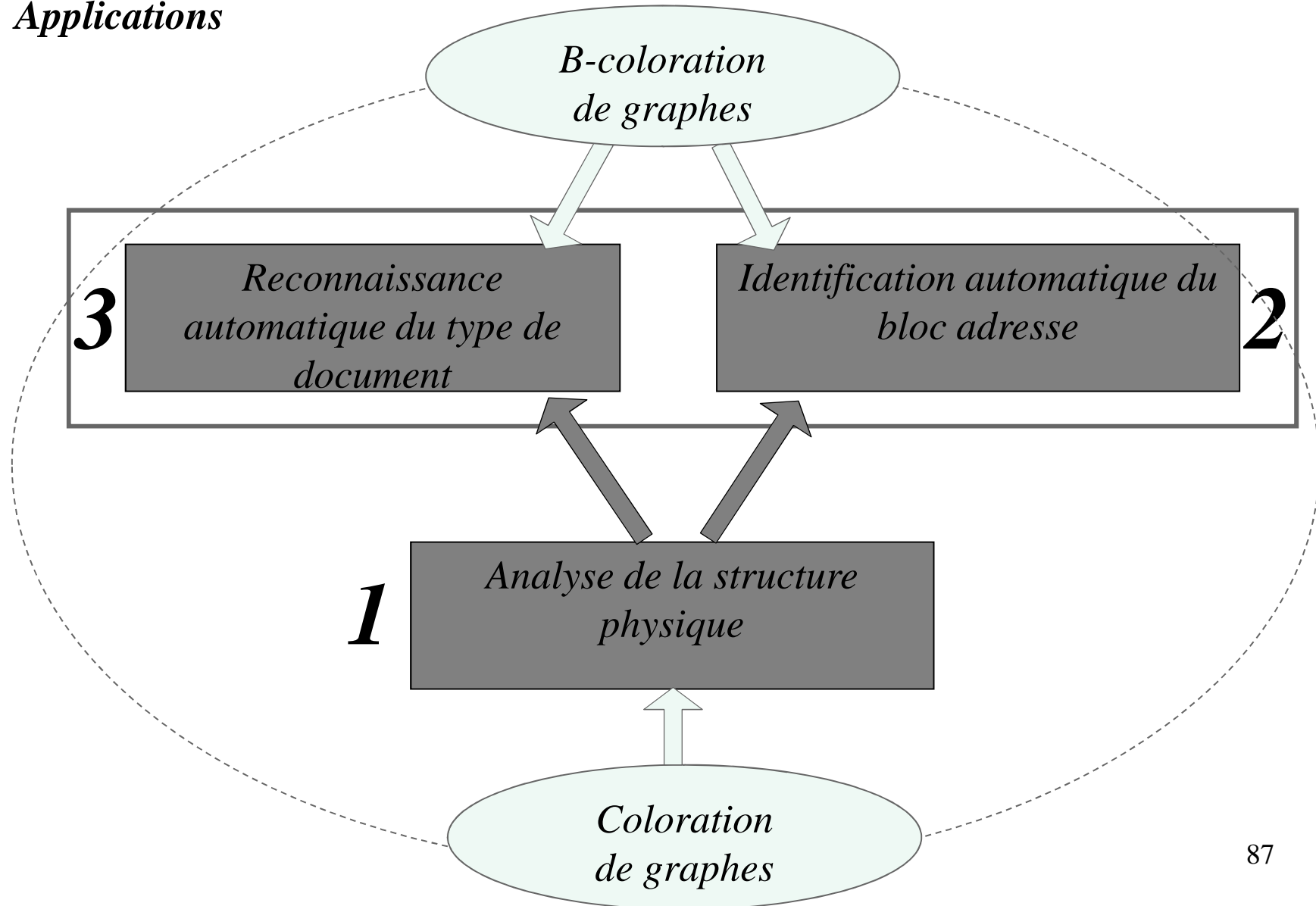
Lecture de code ACI manuscrit



Lecture de l'adresse de destination

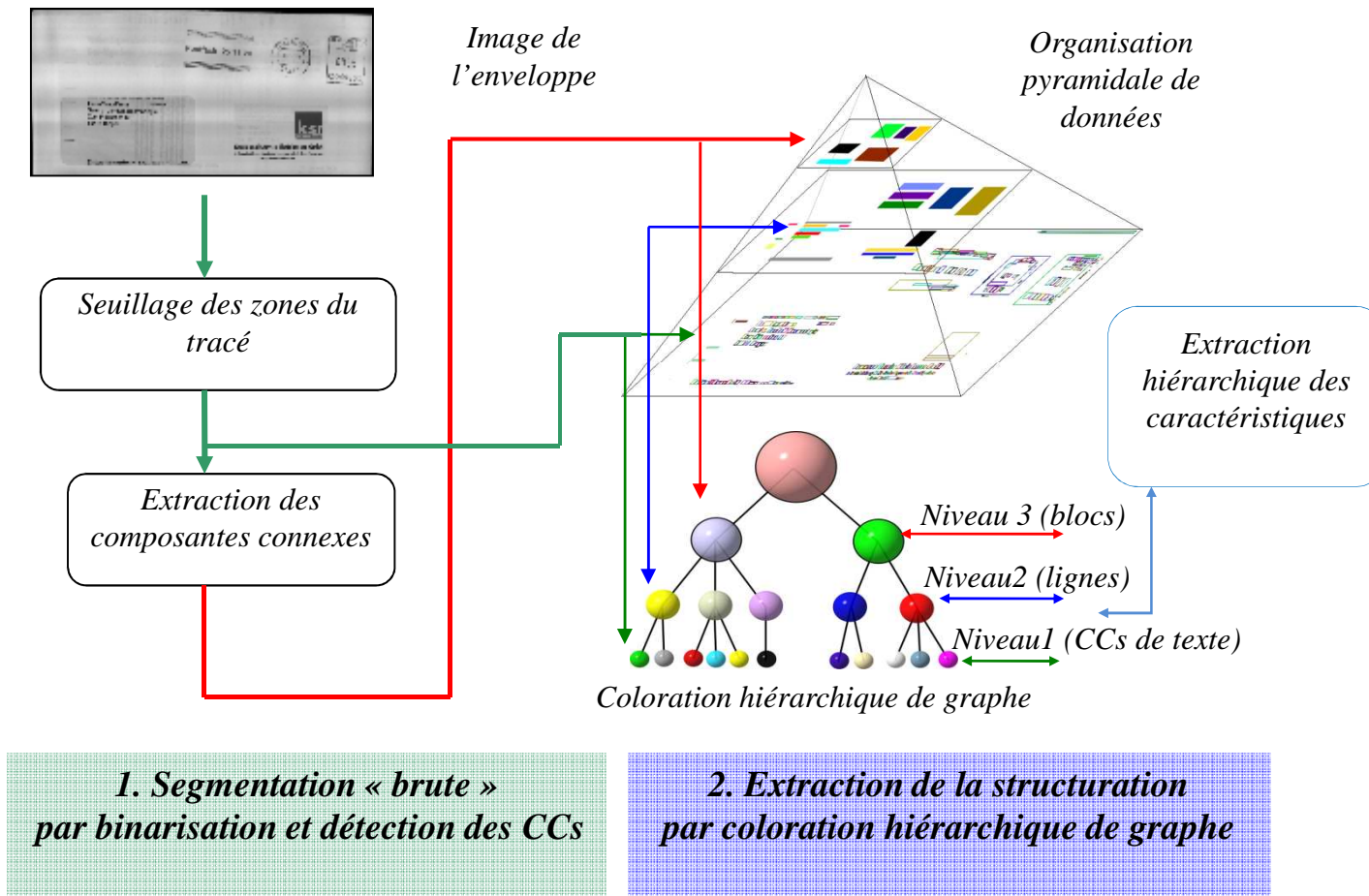
Des approches conventionnelles aux solutions industrielles

Applications



Partie 1 → Extraction de la structure physique par coloration hiérarchique

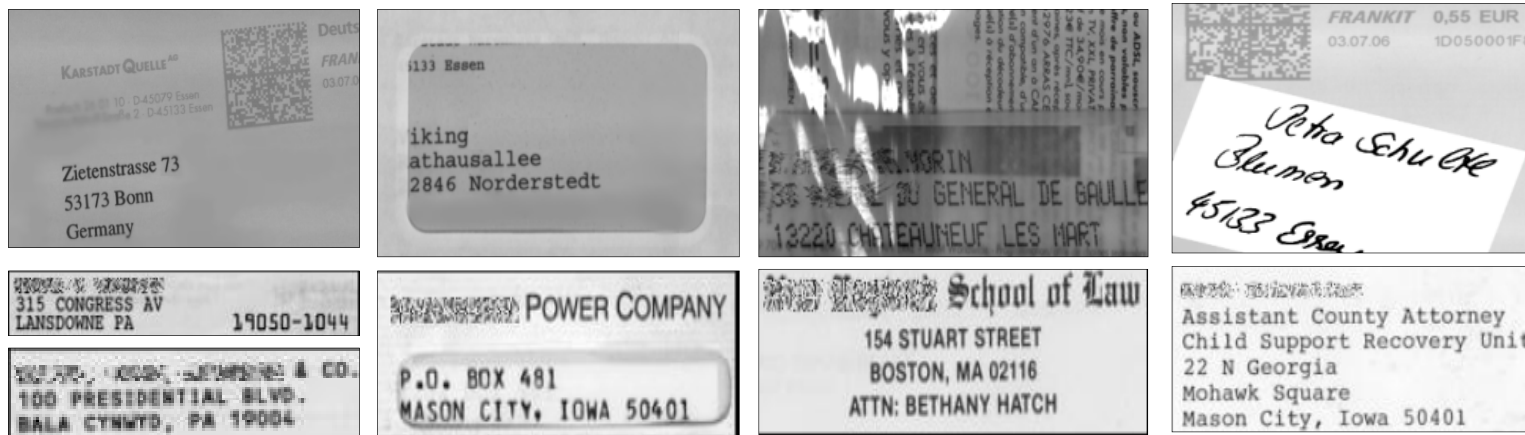
□ Diagramme fonctionnel de notre méthode



Partie 2 → Localisation automatique du bloc adresse (LBA)

Grande variabilité de la zone d'adresse

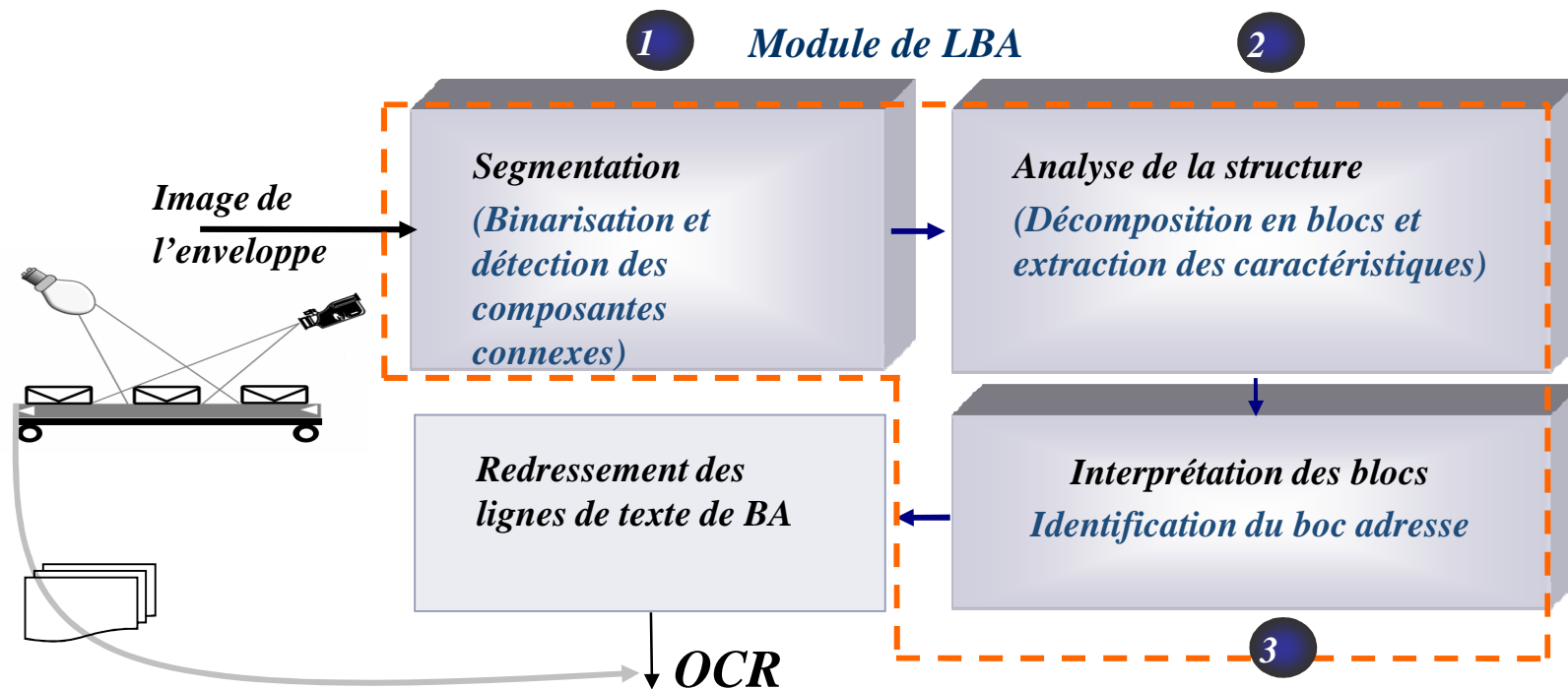
- ❑ Adresse de taille variable.
- ❑ Adresse manuscrite de styles variables ou imprimée de différentes polices de mises en forme variable (taille, espacement...).
- ❑ Technologies d'impression différentes : machines à écrire, imprimante matricielle, laser, à jet d'encre...etc.
- ❑ Adresse dégradée (faible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



Partie 2 → Localisation automatique du bloc adresse

□ 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

SOMMAIRE

**Climat :
demain
l'apocalypse**
p. 30



Les consensus d'un rapport d'été 2009 l'ajout de 1,5°C sont formels : à l'avenir, certains d'accroissent à 2°C. Plus l'effet de serre, le 2009 sera considéré une année clé.

**Génétique
Les sauveurs
du troisième
millénaire**
p. 33



À l'occasion de son anniversaire, permet de voir pourquoi l'impact d'une nouvelle technologie, celle de l'ADN-médecine. Le défi des chercheurs : soigner 4000 maladies génétiques.

**Comment
vieillir
en bonne
santé**
p. 40

Présentées comme les « pures de la jeunesse », la méditerranée et la Océan sont l'objet des recherches les plus récentes. Quelle sera l'espérance de vie dans 50 ans ? Où ira-t-on en bonne santé ? Enquête auprès des scientifiques, en France et aux États-Unis.



Les grandes invasions
p. 54

En introduisant des espèces animales exotiques au Mexique, les chercheurs ont découvert que les espèces natives, comme les araignées, souffrent de maladies mortelles causées par une introduction d'un parasite du système.



**Pays de l'Est :
situation
explosive** p. 62



Défenses, sources et eau, les tensions de l'Est font passer le monde à un nouveau Tchernobyl.

Reconnaissance de la structure des sommaires de revues

CAHIERS
PREVENTION

SOMMAIRE
JULIEN 1971 N° 223

- 1. **COMPTES RENDUS DE LA REUNION**
- 2. **LES BREVETS**
- 3. **LES BREVETS**
- 4. **LES BREVETS**
- 5. **LES BREVETS**
- 6. **LES BREVETS**
- 7. **LES BREVETS**
- 8. **LES BREVETS**
- 9. **LES BREVETS**
- 10. **LES BREVETS**
- 11. **LES BREVETS**
- 12. **LES BREVETS**
- 13. **LES BREVETS**
- 14. **LES BREVETS**
- 15. **LES BREVETS**
- 16. **LES BREVETS**
- 17. **LES BREVETS**
- 18. **LES BREVETS**
- 19. **LES BREVETS**
- 20. **LES BREVETS**
- 21. **LES BREVETS**
- 22. **LES BREVETS**
- 23. **LES BREVETS**
- 24. **LES BREVETS**
- 25. **LES BREVETS**
- 26. **LES BREVETS**
- 27. **LES BREVETS**
- 28. **LES BREVETS**
- 29. **LES BREVETS**
- 30. **LES BREVETS**
- 31. **LES BREVETS**
- 32. **LES BREVETS**
- 33. **LES BREVETS**
- 34. **LES BREVETS**
- 35. **LES BREVETS**
- 36. **LES BREVETS**
- 37. **LES BREVETS**
- 38. **LES BREVETS**
- 39. **LES BREVETS**
- 40. **LES BREVETS**
- 41. **LES BREVETS**
- 42. **LES BREVETS**
- 43. **LES BREVETS**
- 44. **LES BREVETS**
- 45. **LES BREVETS**
- 46. **LES BREVETS**
- 47. **LES BREVETS**
- 48. **LES BREVETS**
- 49. **LES BREVETS**
- 50. **LES BREVETS**

SOMMAIRE

- 1. **Chinot : domine l'apocalypse** p. 30
- 2. **Comment vieillir en bonne santé** p. 40
- 3. **Généraliste Les sauveurs du troisième millénaire** p. 33
- 4. **Les grandes invasions** p. 64
- 5. **Pays de l'Est : situation explosive** p. 62
- 6. **Les trafiquants du jurassique** p. 100

SOMMAIRE

- 1. **Chinot : domine l'apocalypse** p. 30
- 2. **Comment vieillir en bonne santé** p. 40
- 3. **Généraliste Les sauveurs du troisième millénaire** p. 33
- 4. **Les grandes invasions** p. 64
- 5. **Pays de l'Est : situation explosive** p. 62
- 6. **Les trafiquants du jurassique** p. 100

SOMMAIRE

- 1. **Nouveautés pour le plaisir** p. 70
- 2. **Actualités**
- 3. **AGIR**
- 4. **DOSSIER**
- 5. **VIVRE**
- 6. **TECHNO**
- 7. **COMPRENDRE**
- 8. **AUJOURD'HUI**
- 9. **S'ÉVALUER**

Sommaire
17 MAI 1997

- 1. **ENTREPRISE**
- 2. **FORMATION**
- 3. **GRAND ANGLE**
- 4. **AGENDA**
- 5. **SOCIÉTÉ**
- 6. **AGENDA**
- 7. **SOCIÉTÉ**
- 8. **AGENDA**
- 9. **SOCIÉTÉ**
- 10. **AGENDA**
- 11. **SOCIÉTÉ**
- 12. **AGENDA**
- 13. **SOCIÉTÉ**
- 14. **AGENDA**
- 15. **SOCIÉTÉ**
- 16. **AGENDA**
- 17. **SOCIÉTÉ**
- 18. **AGENDA**
- 19. **SOCIÉTÉ**
- 20. **AGENDA**
- 21. **SOCIÉTÉ**
- 22. **AGENDA**
- 23. **SOCIÉTÉ**
- 24. **AGENDA**
- 25. **SOCIÉTÉ**
- 26. **AGENDA**
- 27. **SOCIÉTÉ**
- 28. **AGENDA**
- 29. **SOCIÉTÉ**
- 30. **AGENDA**
- 31. **SOCIÉTÉ**
- 32. **AGENDA**
- 33. **SOCIÉTÉ**
- 34. **AGENDA**
- 35. **SOCIÉTÉ**
- 36. **AGENDA**
- 37. **SOCIÉTÉ**
- 38. **AGENDA**
- 39. **SOCIÉTÉ**
- 40. **AGENDA**
- 41. **SOCIÉTÉ**
- 42. **AGENDA**
- 43. **SOCIÉTÉ**
- 44. **AGENDA**
- 45. **SOCIÉTÉ**
- 46. **AGENDA**
- 47. **SOCIÉTÉ**
- 48. **AGENDA**
- 49. **SOCIÉTÉ**
- 50. **AGENDA**

Sommaire
N° 316 SEPTEMBRE 1996

- 1. **Galileo : cartes postales de Jupiter**
- 2. **La galaxie qui défie le big bang**
- 3. **Mars, le retour en force**
- 4. **Mission Cluster : le prix du savoir**
- 5. **LES MILLS ET UNE RUSSIE DES MICROBES INTRACELLULAIRES**
- 6. **DE LA MÉMOIRE HUMAINE À LA MÉMOIRE ARTIFICIELLE**

SOMMAIRE

- 1. **RECHERCHE**
- 2. **LES MILLS ET UNE RUSSIE DES MICROBES INTRACELLULAIRES**
- 3. **DE LA MÉMOIRE HUMAINE À LA MÉMOIRE ARTIFICIELLE**

SOMMAIRE N° 272 FÉVRIER 1995

- 1. **LA SOCIÉTÉ FRANÇAISE À L'ÉPREUVE DU SIDA**
- 2. **L'IMPÉRIUM MÉDICAL DANS LE CASSE**
- 3. **LES ÉLÉMENTS PRATIQUES**
- 4. **LES ÉLÉMENTS PRATIQUES**
- 5. **LES ÉLÉMENTS PRATIQUES**
- 6. **LES ÉLÉMENTS PRATIQUES**
- 7. **LES ÉLÉMENTS PRATIQUES**
- 8. **LES ÉLÉMENTS PRATIQUES**
- 9. **LES ÉLÉMENTS PRATIQUES**
- 10. **LES ÉLÉMENTS PRATIQUES**
- 11. **LES ÉLÉMENTS PRATIQUES**
- 12. **LES ÉLÉMENTS PRATIQUES**
- 13. **LES ÉLÉMENTS PRATIQUES**
- 14. **LES ÉLÉMENTS PRATIQUES**
- 15. **LES ÉLÉMENTS PRATIQUES**
- 16. **LES ÉLÉMENTS PRATIQUES**
- 17. **LES ÉLÉMENTS PRATIQUES**
- 18. **LES ÉLÉMENTS PRATIQUES**
- 19. **LES ÉLÉMENTS PRATIQUES**
- 20. **LES ÉLÉMENTS PRATIQUES**
- 21. **LES ÉLÉMENTS PRATIQUES**
- 22. **LES ÉLÉMENTS PRATIQUES**
- 23. **LES ÉLÉMENTS PRATIQUES**
- 24. **LES ÉLÉMENTS PRATIQUES**
- 25. **LES ÉLÉMENTS PRATIQUES**
- 26. **LES ÉLÉMENTS PRATIQUES**
- 27. **LES ÉLÉMENTS PRATIQUES**
- 28. **LES ÉLÉMENTS PRATIQUES**
- 29. **LES ÉLÉMENTS PRATIQUES**
- 30. **LES ÉLÉMENTS PRATIQUES**
- 31. **LES ÉLÉMENTS PRATIQUES**
- 32. **LES ÉLÉMENTS PRATIQUES**
- 33. **LES ÉLÉMENTS PRATIQUES**
- 34. **LES ÉLÉMENTS PRATIQUES**
- 35. **LES ÉLÉMENTS PRATIQUES**
- 36. **LES ÉLÉMENTS PRATIQUES**
- 37. **LES ÉLÉMENTS PRATIQUES**
- 38. **LES ÉLÉMENTS PRATIQUES**
- 39. **LES ÉLÉMENTS PRATIQUES**
- 40. **LES ÉLÉMENTS PRATIQUES**
- 41. **LES ÉLÉMENTS PRATIQUES**
- 42. **LES ÉLÉMENTS PRATIQUES**
- 43. **LES ÉLÉMENTS PRATIQUES**
- 44. **LES ÉLÉMENTS PRATIQUES**
- 45. **LES ÉLÉMENTS PRATIQUES**
- 46. **LES ÉLÉMENTS PRATIQUES**
- 47. **LES ÉLÉMENTS PRATIQUES**
- 48. **LES ÉLÉMENTS PRATIQUES**
- 49. **LES ÉLÉMENTS PRATIQUES**
- 50. **LES ÉLÉMENTS PRATIQUES**

➔ Variabilité des types de sommaires

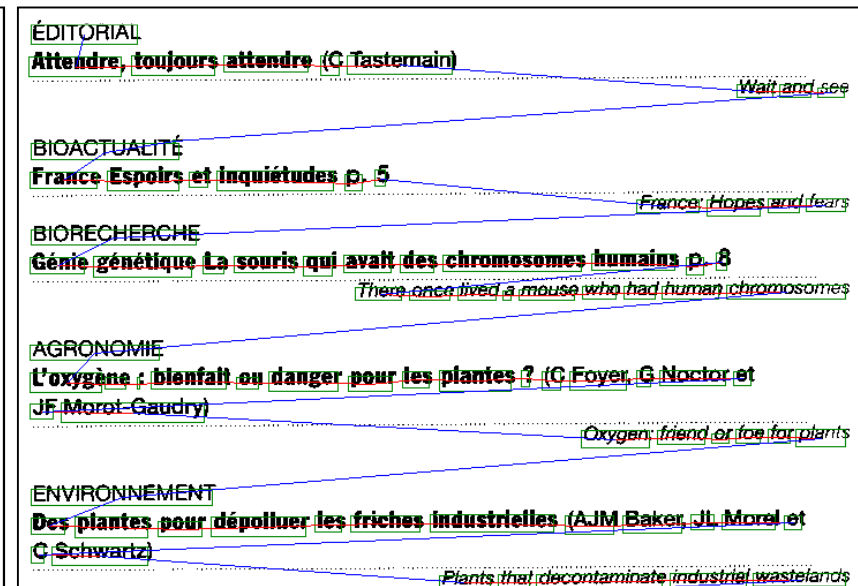
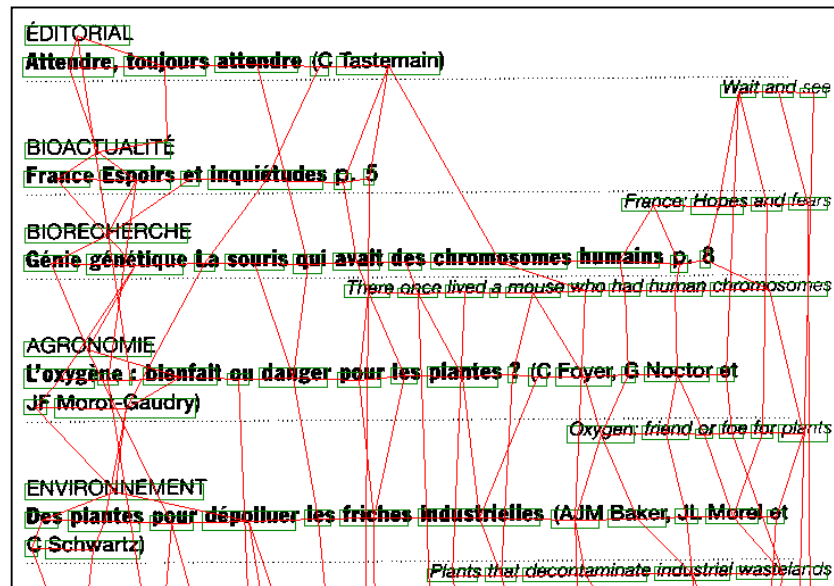
Reconnaissance des structures par apprentissage

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 :

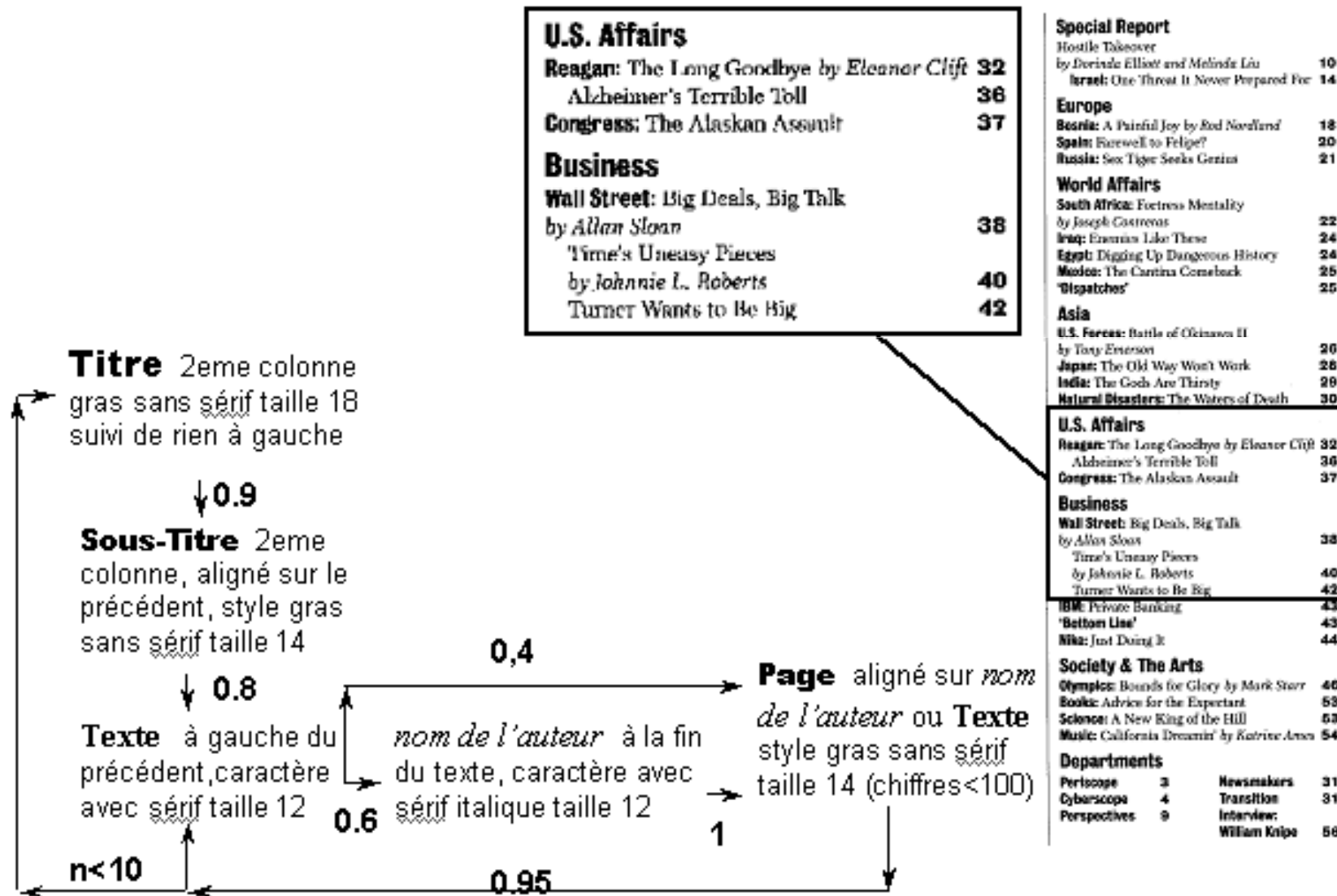
➔ *apprentissage des relations spatiales entre les blocs de texte*



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

2819, 1882 Zoom % 16 Ok

Revue : pcdirect2 Page : Typographie No Etiquette : SelectEtq

Sommaire

B I O F U T U R N ° 1 6 7 M A I 1 9 9 7



3 EDITORIAL
Entre sécheresse et qualité de l'eau (C Tastemain)
From drought to water quality

4 BIOACTUALITÉ
Allemagne La biotechnologie à la recherche de modèles p. 5
German biotechnology looks for role models

8 BIORECHERCHE
Biologie Les premiers chromosomes artificiels p. 8
The first artificial chromosomes

29 SANTÉ
Les nouveaux inhibiteurs de métalloprotéinases à zinc (V Dive, M Kaczorek, C Roussel et F Roux)
New inhibitors of zinc metalloproteases

34 AGRICULTURE
Les traitements biologiques du lin (A Jauneau, F Bert, C Rihouey et C Morvan)
Enzymatic treatments for flax

38 SOCIÉTÉ
Le stress des jeunes chercheurs américains (P Stephan et V Mangematin)
Young American scientists under stress

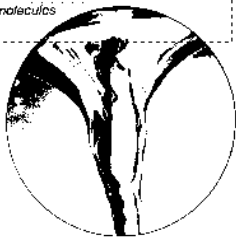
41 VIE DES SOCIÉTÉS
Atlantic Pharmaceuticals : nouvelles molécules antisens p. 41
Atlantic Pharmaceuticals : new antisense molecules

45 BREVETS
49 AGENDA
50 NOUVEAUX PRODUITS
51 PETITES ANNONCES
52 BULLETIN D'ABONNEMENT

OSIIFR

cess à l'eau potable
Biotechnology for drinking water

urces en eau
HILIPON
bactéries plus sélectives
ARANE



Technoscope


Apprentissage manuel des zones

3238, 332 Zoom % 16 Ok

Revue : pcdirect2 Page : Typographie i Etiquette : SelectEtq

Sommaire

B I O F U T U R N ° 1 6 7 M A I 1 9 9 7



3 EDITORIAL
Entre sécheresse et qualité de l'eau (C Tastemain)
From drought to water quality

4 BIOACTUALITÉ
Allemagne La biotechnologie à la recherche de modèles p. 5
German biotechnology looks for role models

8 BIORECHERCHE
Biologie Les premiers chromosomes artificiels p. 8
The first artificial chromosomes

29 SANTÉ
Les nouveaux inhibiteurs de métalloprotéinases à zinc (V Dive, M Kaczorek, C Roussel et F Roux)
New inhibitors of zinc metalloproteases

34 AGRICULTURE
Les traitements biologiques du lin (A Jauneau, F Bert, C Rihouey et C Morvan)
Enzymatic treatments for flax

38 SOCIÉTÉ
Le stress des jeunes chercheurs américains (P Stephan et V Mangematin)
Young American scientists under stress


41 VIE DES SOCIÉTÉS
Atlantic Pharmaceuticals : nouvelles molécules antisens p. 41
Atlantic Pharmaceuticals : new antisense molecules

45 BREVETS
49 AGENDA
50 NOUVEAUX PRODUITS
51 PETITES ANNONCES
52 BULLETIN D'ABONNEMENT

OSIIFR

cess à l'eau potable
Biotechnology for drinking water

urces en eau
HILIPON
bactéries plus sélectives
ARANE



Technoscope

Apprentissage des titres

Résultats de la reconnaissance

WORLD AFFAIRS	
Mexico: Who Can We Trust Anymore? by Martha Brani	40
Opinion: It's Time for the Big Bang by Jorge G. Castañeda	43
Indonesia: Grabbing for Gold	44
Humanitarians: We Leave. You Die.	45

Recognition of the four Watson–Crick base pairs in the DNA minor groove by synthetic ligands 468
S White, J W Szewczyk, J M Turner, E E Baird
& P B Dervan **N&V**

A new perspective on the dynamical link between the stratosphere and troposphere 471
D F Hartley, J T Villarin, R X Black & C A Davis

Trypanosoma brucei
W Bitter, H Gerrits, R I
& P Borst

Structure of the V δ 1 T-cell antigen receptor
H Li, M I Lebedeva, A S
M B Brenner & R A M

PRODUITS LAITIERS	
Le lait dans tous ses états (S Carantino)	Dairy products
AQUACULTURE	
Le saumon, favori de l'aquaculture / Biotechnologie, coquillages et crustacés (P Ceccaldi)	Salmon: the favourite with fish farming / Biotechnology, shellfish and seafood
PRODUCTION VÉGÉTALE	
Priorité à la conservation (S Carantino)	Priority for preservation

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

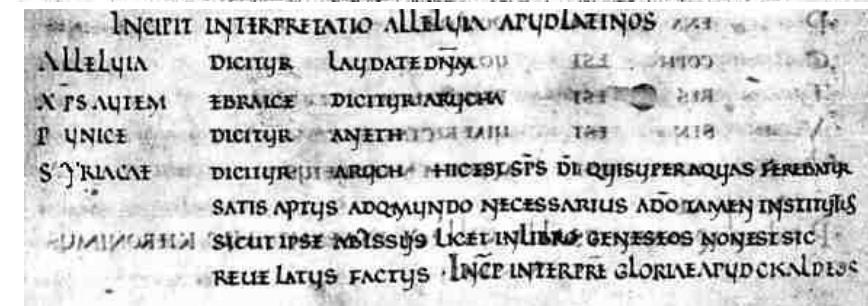
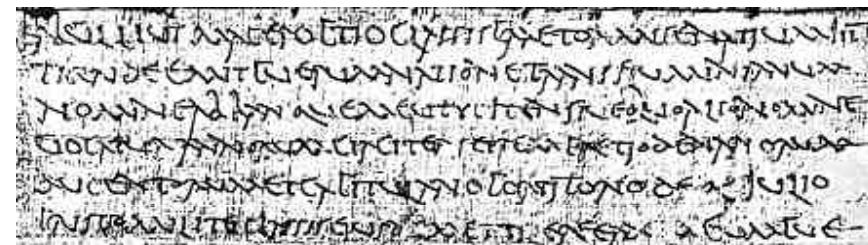
Caroline

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

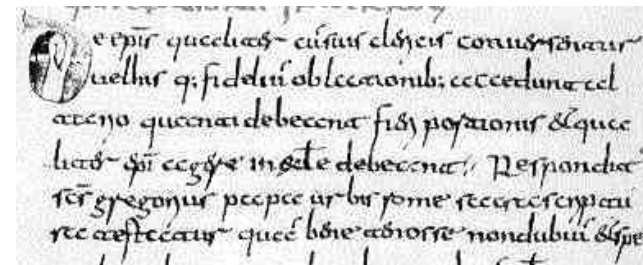
- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

Humanistic scripts

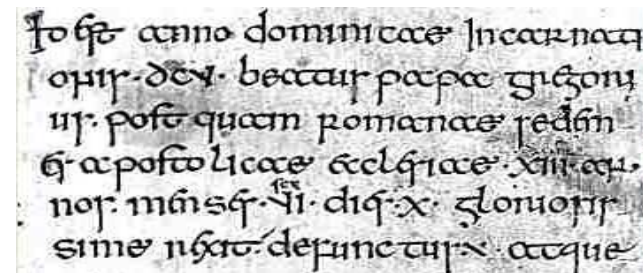
Contemporary scripts



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

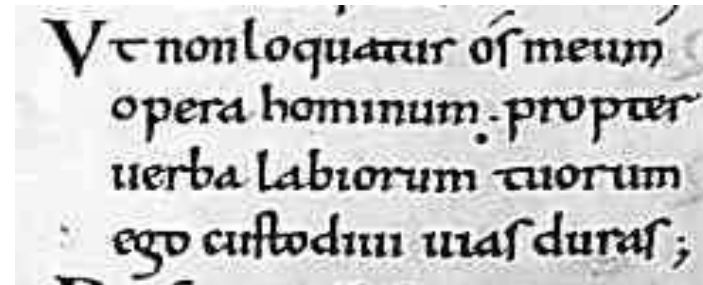
Caroline

Gothic

- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

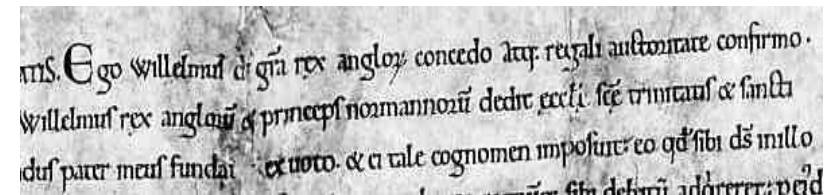
Humanistic scripts

Contemporary scripts



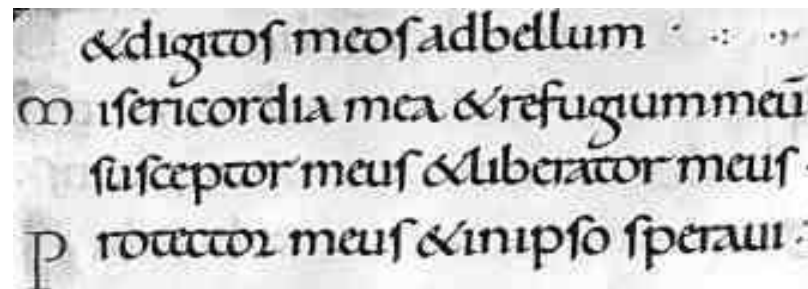
Vt non loquatur os meum
opera hominum. propter
uerba labiorum tuorum
ego custodiu uias duras;

10th



ms. Ego Willhelmus di grā rex angloy concedo atq. regali auctoritate confirmo.
Willhelmus rex angloy & princeps normannoꝝ dedit ecclie scē trinitatis & sancta
dus pater meus fundat. ex uoto. & ei tale cognomen imposuit: eo qđ sibi dñs in illo

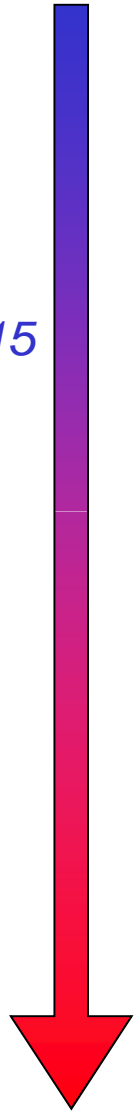
11th



& digitos meos ad bellum
in misericordia mea & refugium meū
susceptor meus & liberator meus
protector meus & in ipso speraui.

11th

Les documents à travers les périodes



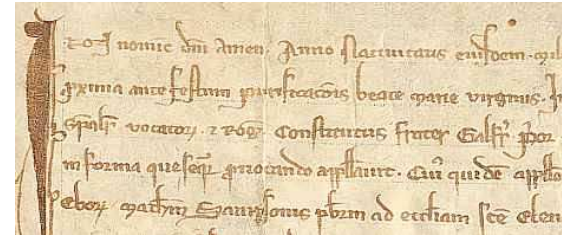
12-15

Old / New Roman scripts
Pre-carolingian scripts
Caroline

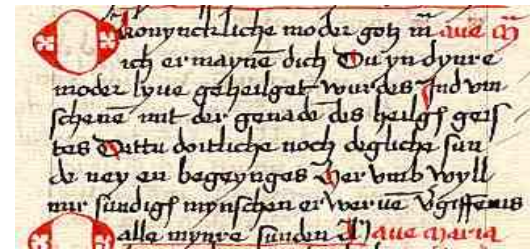
Gothic

- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

Humanistic scripts
Contemporary scripts



13th



15th

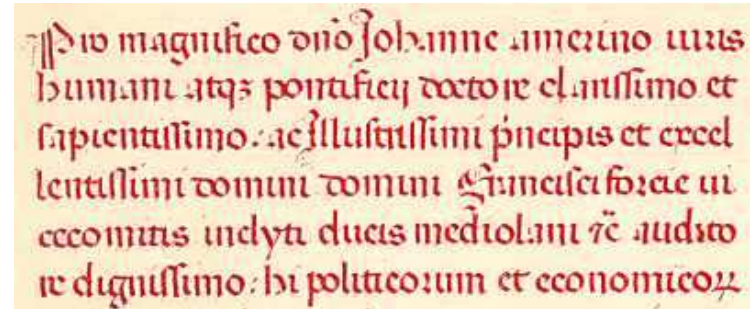


1. Calligraphic (Private coll.)
2. Cursive Gothic (Private coll.)
3. Textura (NDL)
4. Bastarda (Private coll.)

Les documents à travers les périodes

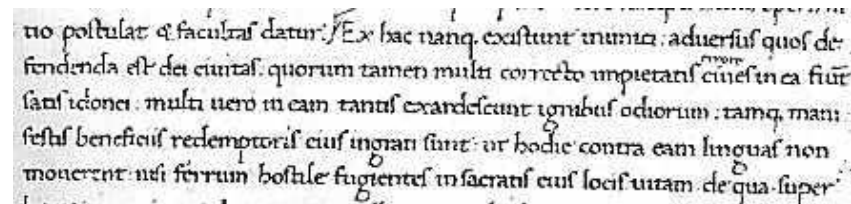
Old / New Roman scripts
Pre-carolingian scripts
Caroline
Gothic

- Protogothic
- Textura
- Rotunda
- Cursiva
- Bastarda

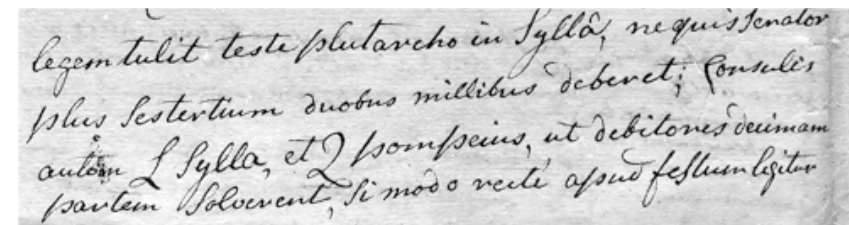


¶ Ad magnifico dno Johanne Amerino iuris
humani atq; pontificij doctore clarissimo et
sapientissimo. ac Illustrissimi pncipis et excel
lentissimi domini domini Guineisei forae ui
cecomitis inelyti ducis mediolani ꝛc audito
re dignissimo. hi politico:um et economicoꝝ

15th



no postulat et facultas datur. Ex hac namq; existunt inimici: aduersus quos de
fendenda est dei ciuitas. quorum tamen multi correcto impietatis eius in ea sunt
satis idonei. multi uero in eam tantis exardescunt ignibus ochorum: tamq; mani
festis beneficis redemptoris eius ingratii sunt: ut hodie contra eam linguas non
mouerent nisi ferrum hostale fugientes in sacris eius locis uitam de qua super



legem tulit teste Plutarcho in Sylla, ne quis senator
plus sestertium duobus millibus deberet; Consules
autem L Sylla, et Q Pompeius, ut debitores decimam
partem soluerent, si modo uelit apud festum legitor

18th

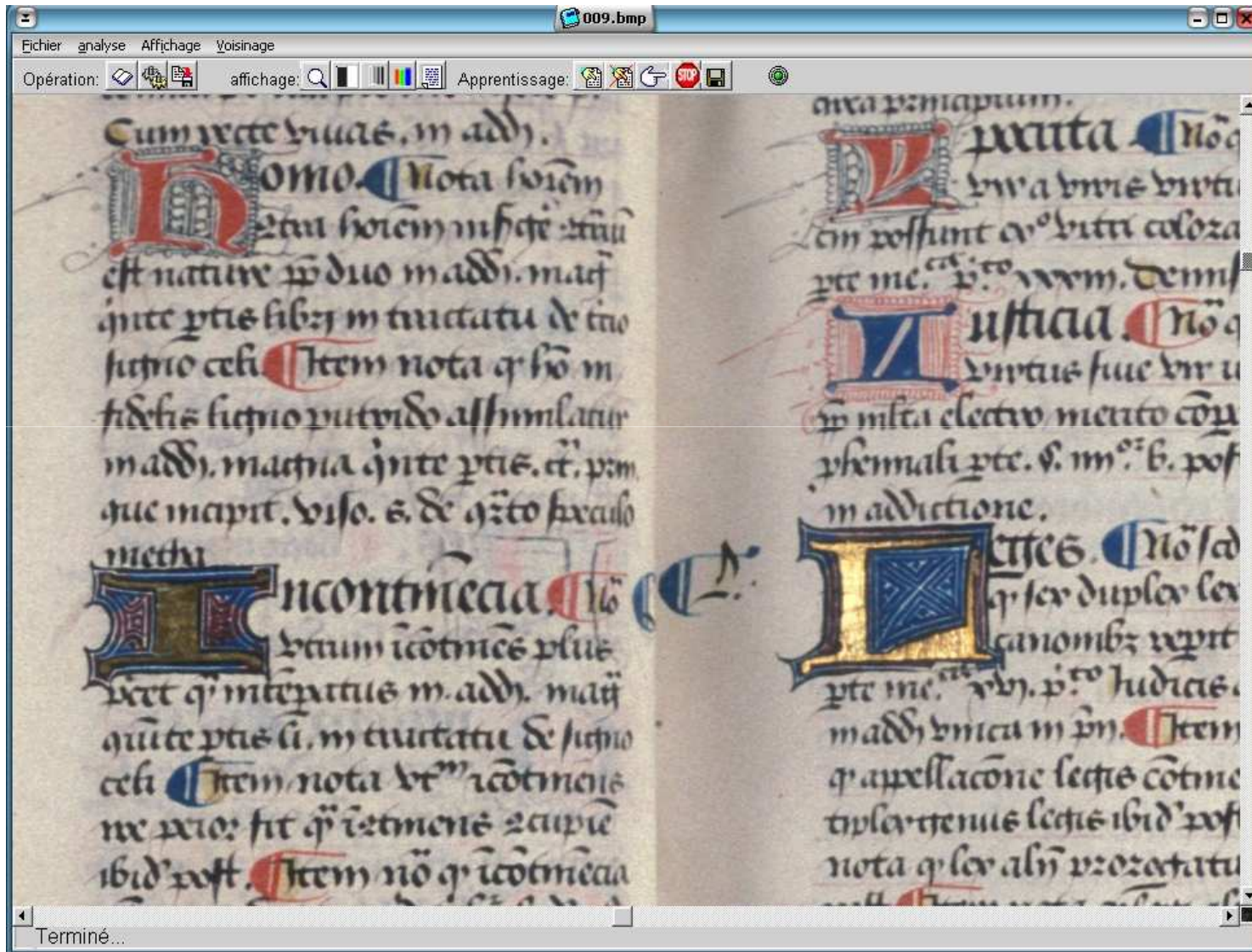
15-19

Humanistic scripts
Author's scripts

1&2. Humanistic (BL)

3. Montesquieu's collection (BNF)

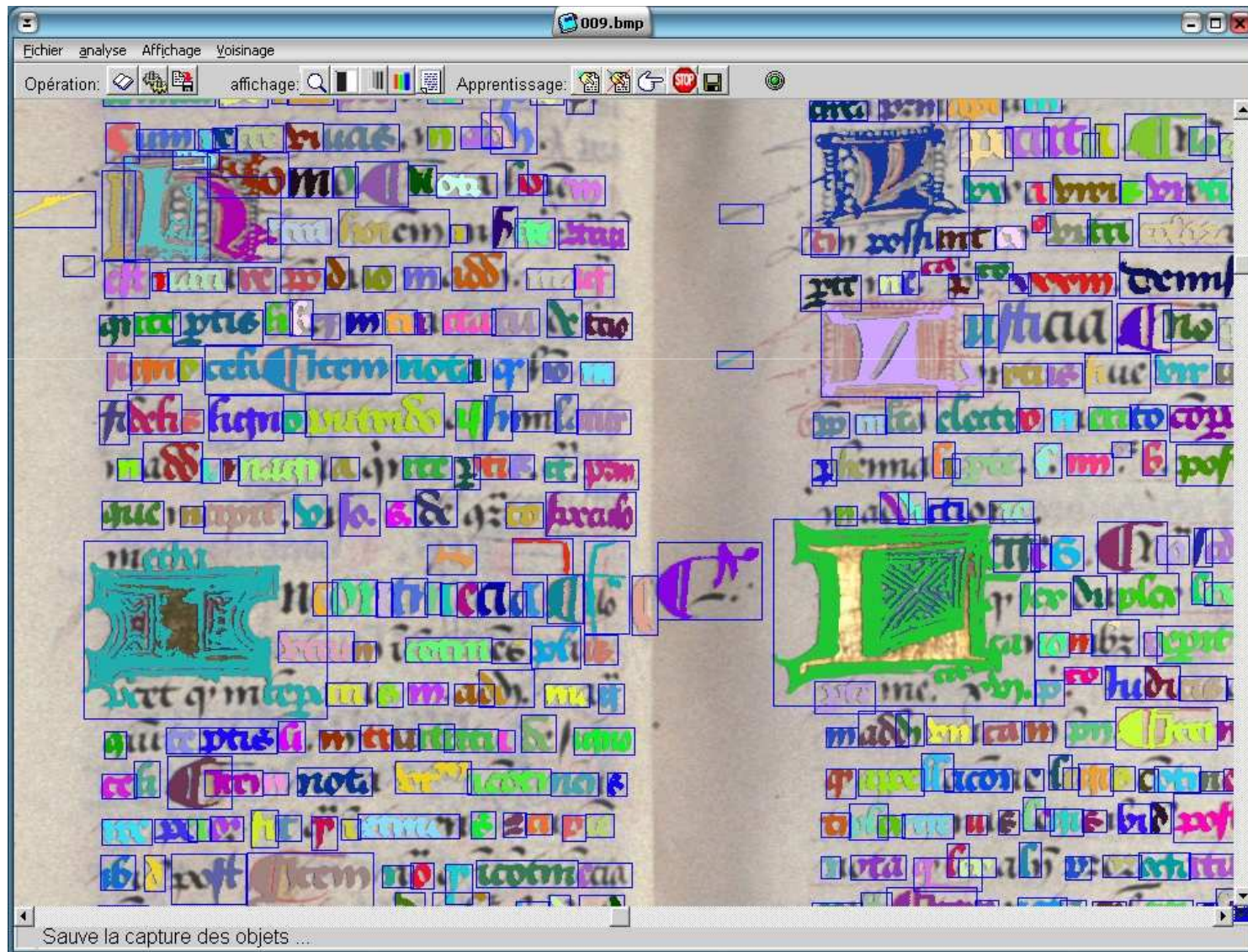
Segmentation et capture des objets



Segmentation et capture des objets

Mesures et caractérisation des objets

Caractéristiques de Formes, géométrie, et couleur...

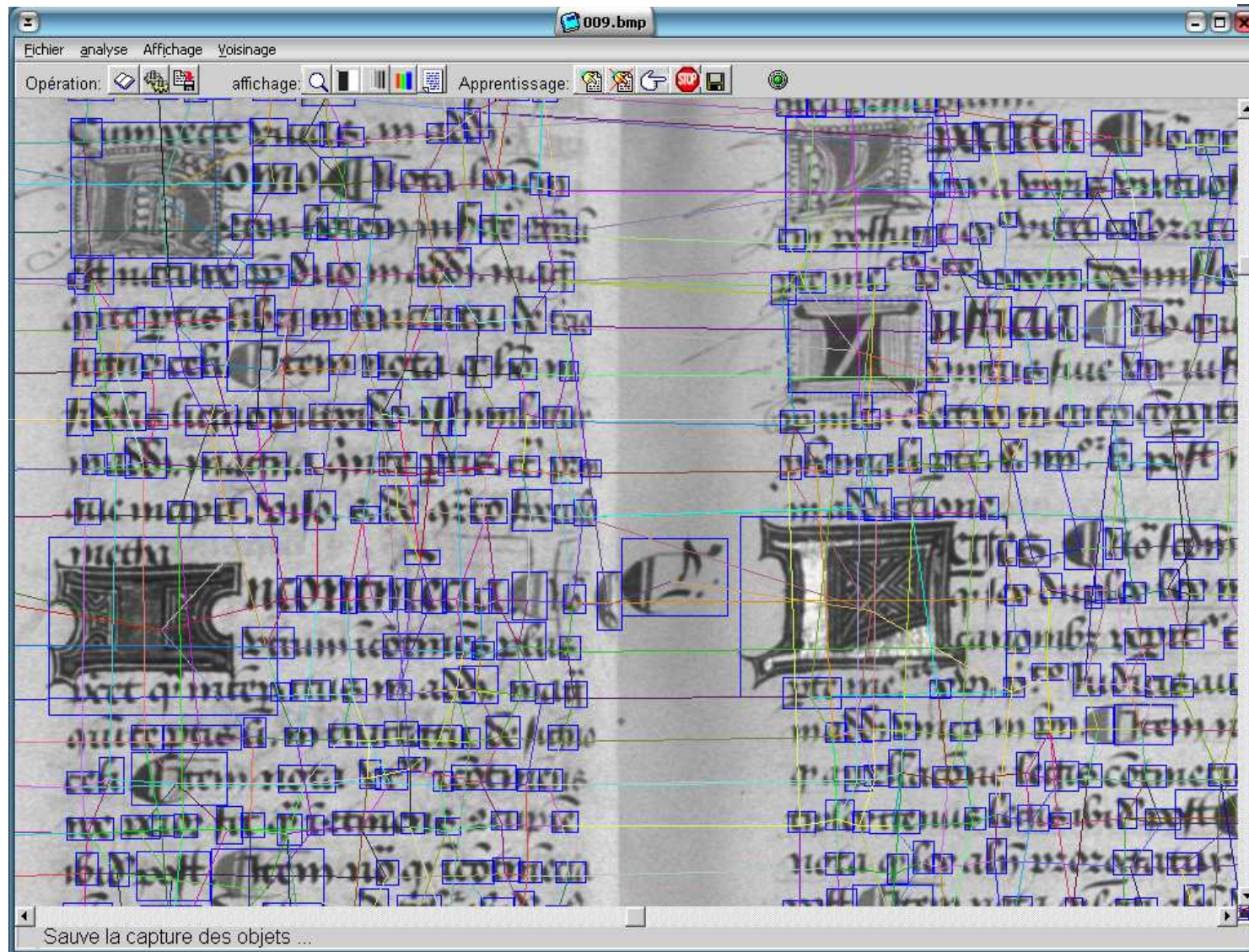


Extraction des informations sur les relations spatiales entre les objets

Exemple...

- Centre de gravité C_x et C_y .
- 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 :
$$Ff = \frac{4 \pi S}{P^2} \quad \text{Droite : } Ff = 0$$
- ...
$$\text{Cercle : } Ff = 1$$

Extraction des informations sur les relations spatiales entre les objets



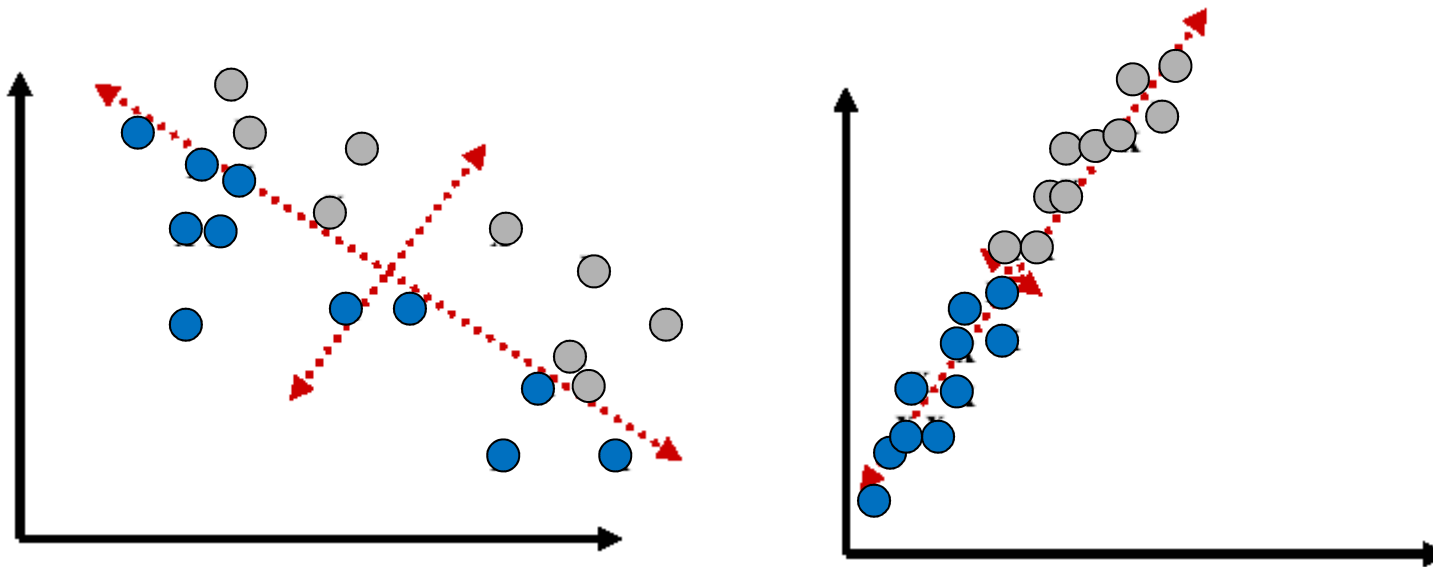
Extraction des informations sur les relations spatiales entre les objets

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.
 - *Élimine 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.

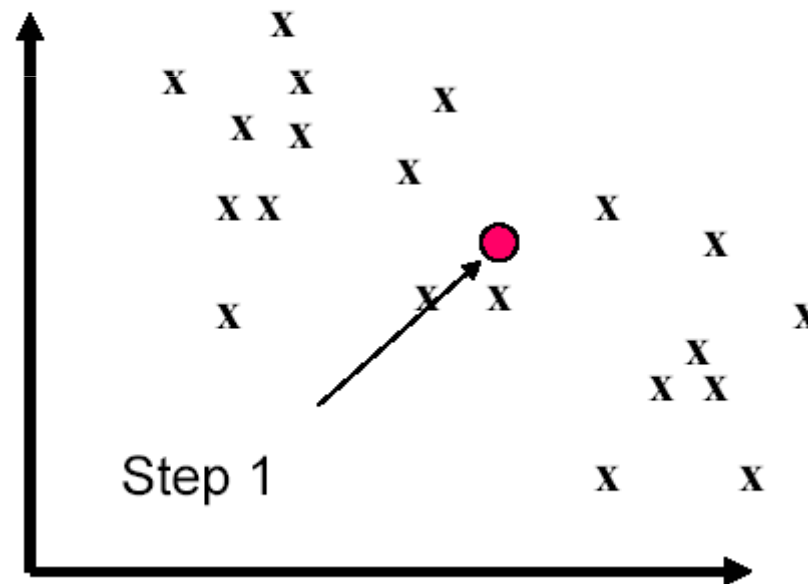
- **Etape 6**

Projeter les données dans ce nouvel espace propre.

Analyse en Composantes Principales

- **Etape 1**

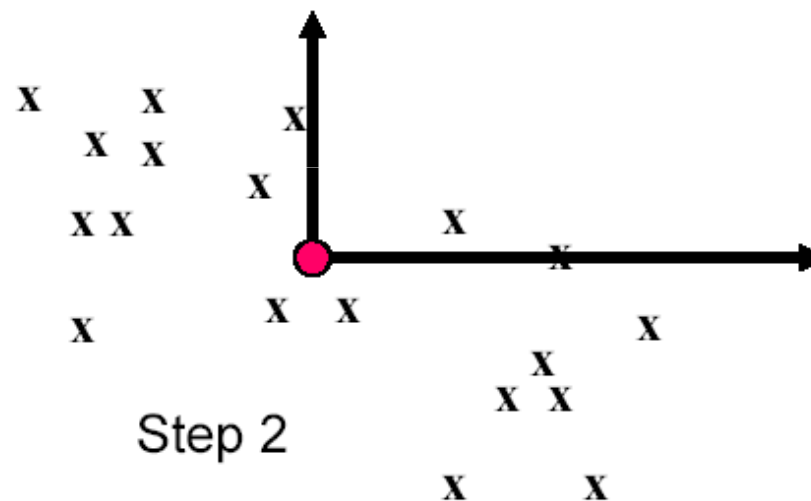
Calculer la moyenne de chaque vecteur de caractéristiques.



Analyse en Composantes Principales

- **Etape 2**

Soustraire la moyenne de chaque vecteur de caractéristiques.



- **Etape 3**

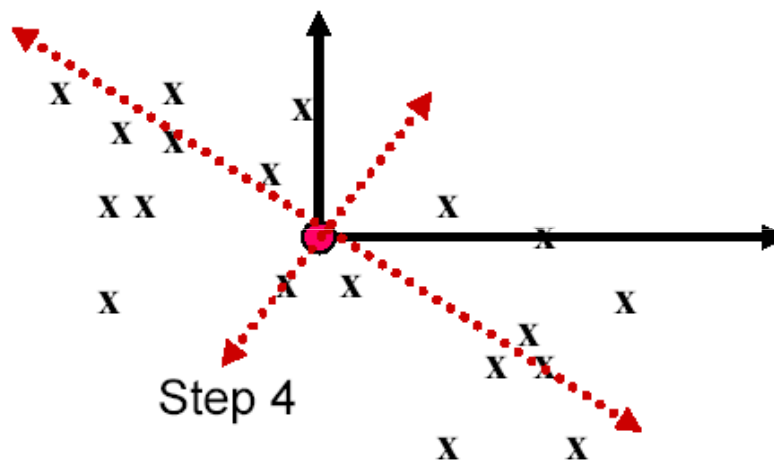
Calculer la matrice des covariances.

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

Analyse en Composantes Principales

- **Etape 4**

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

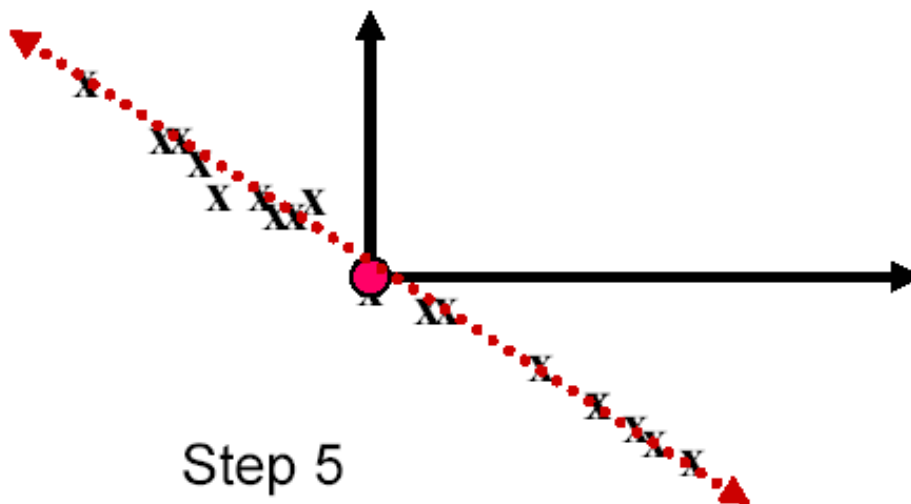


- Les vecteurs propres de la matrice de covariance représentent une base orthonormée d'axes principaux (significatifs) de l'ensemble des données.
- Chaque valeur propre exprime l'importance du vecteur propre associé : plus la valeur propre est grande, plus le vecteur propre associé est significatif

Analyse en Composantes Principales

- **Etape 5**

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



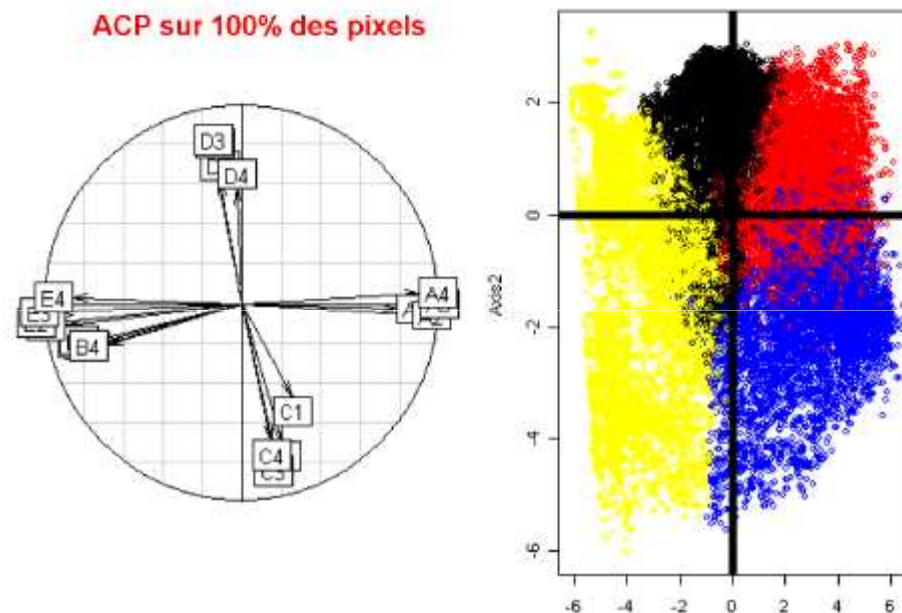
- **Etape 6**

Projeter 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.

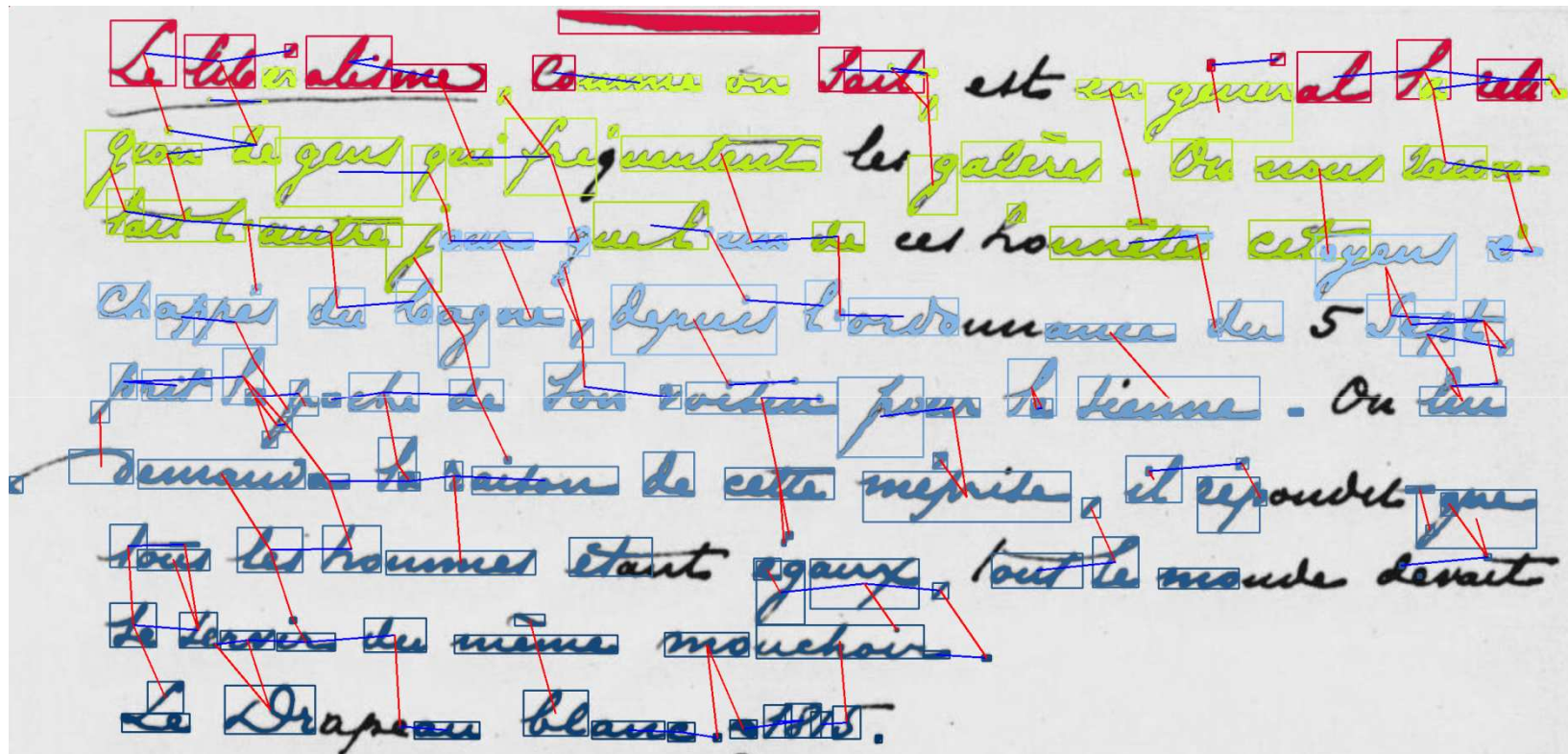
Exemple : classification des objets d'un document



Etude des corrélations entre dimensions
Caractéristiques (cercle des corrélations)

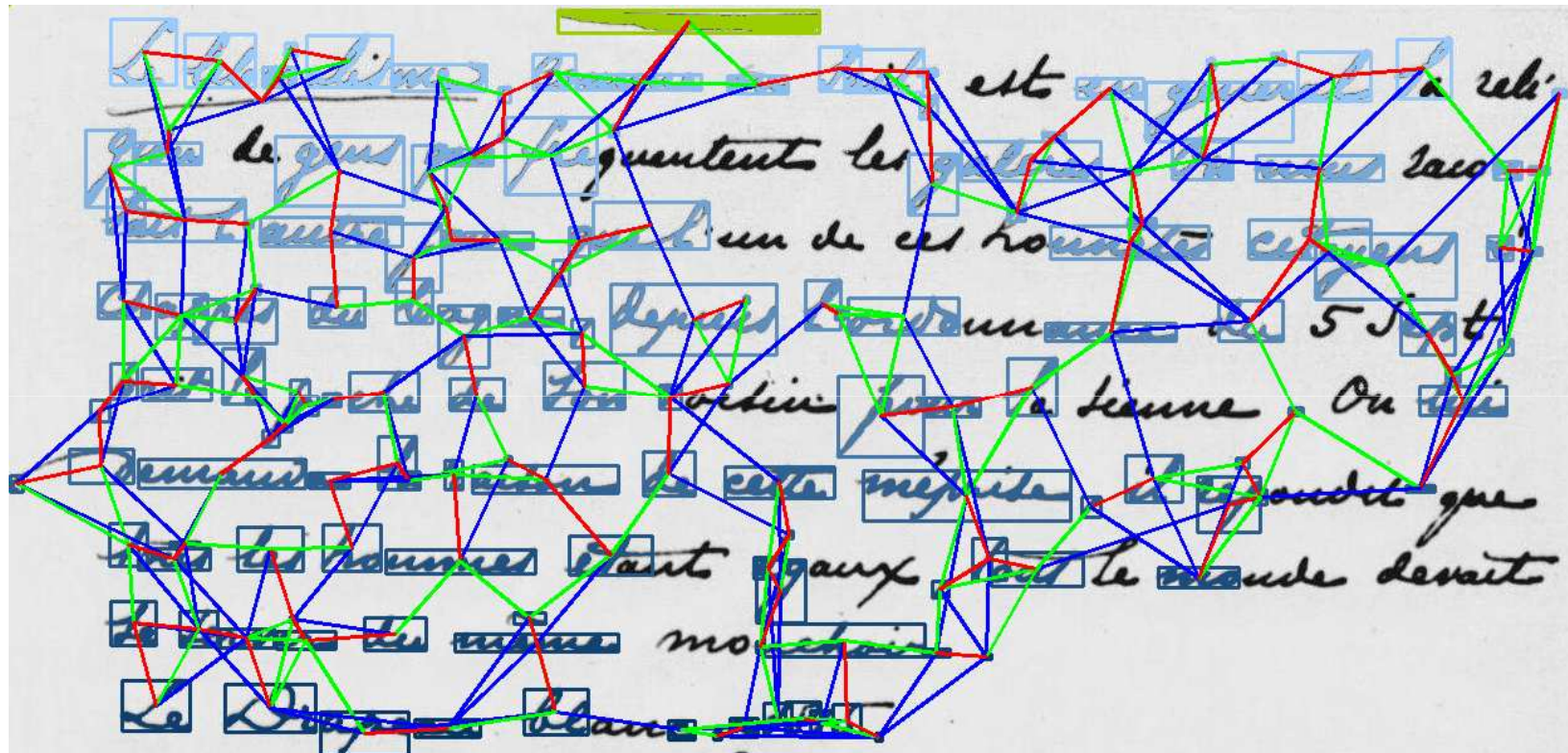
- contribution
- inter-dépendances

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

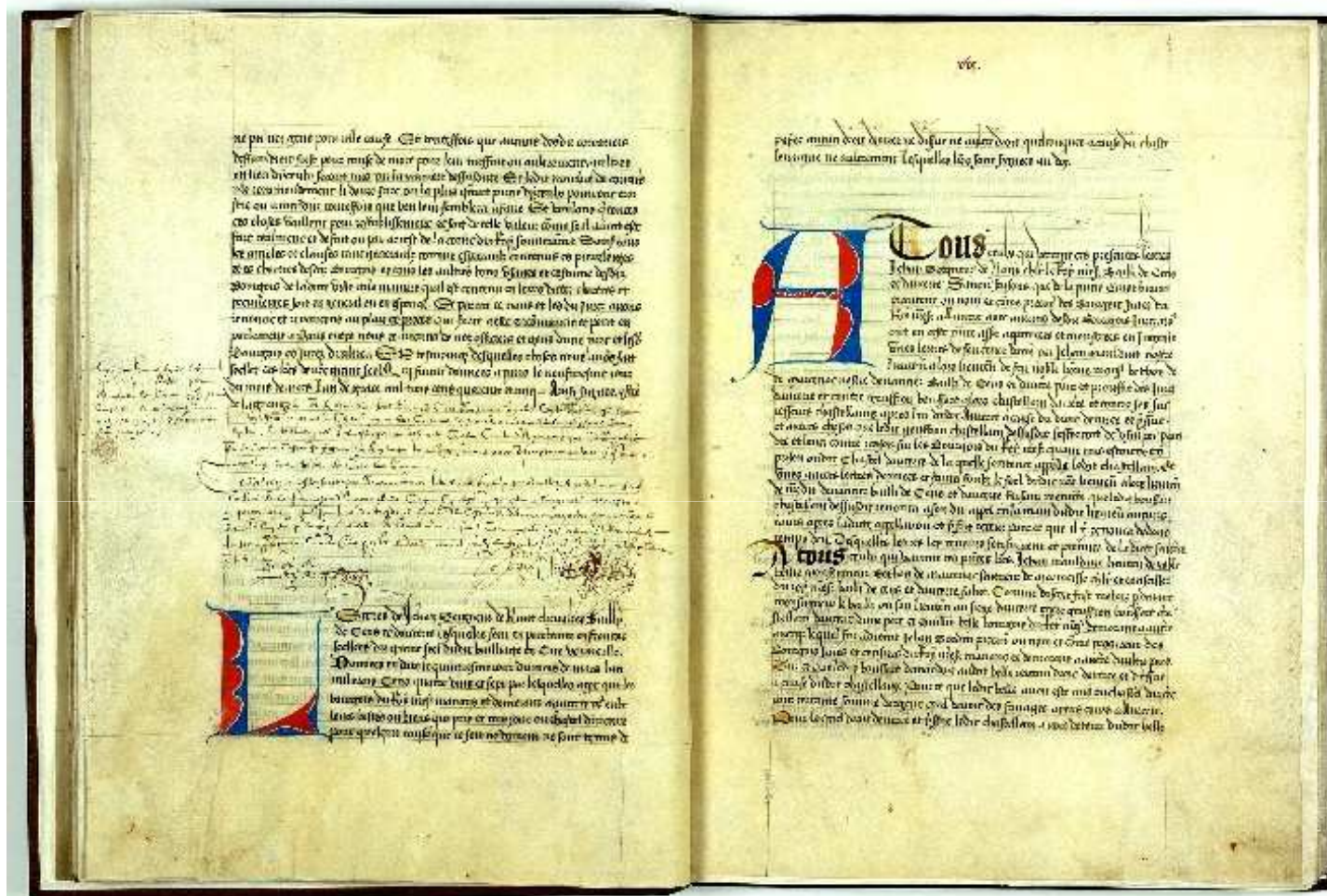
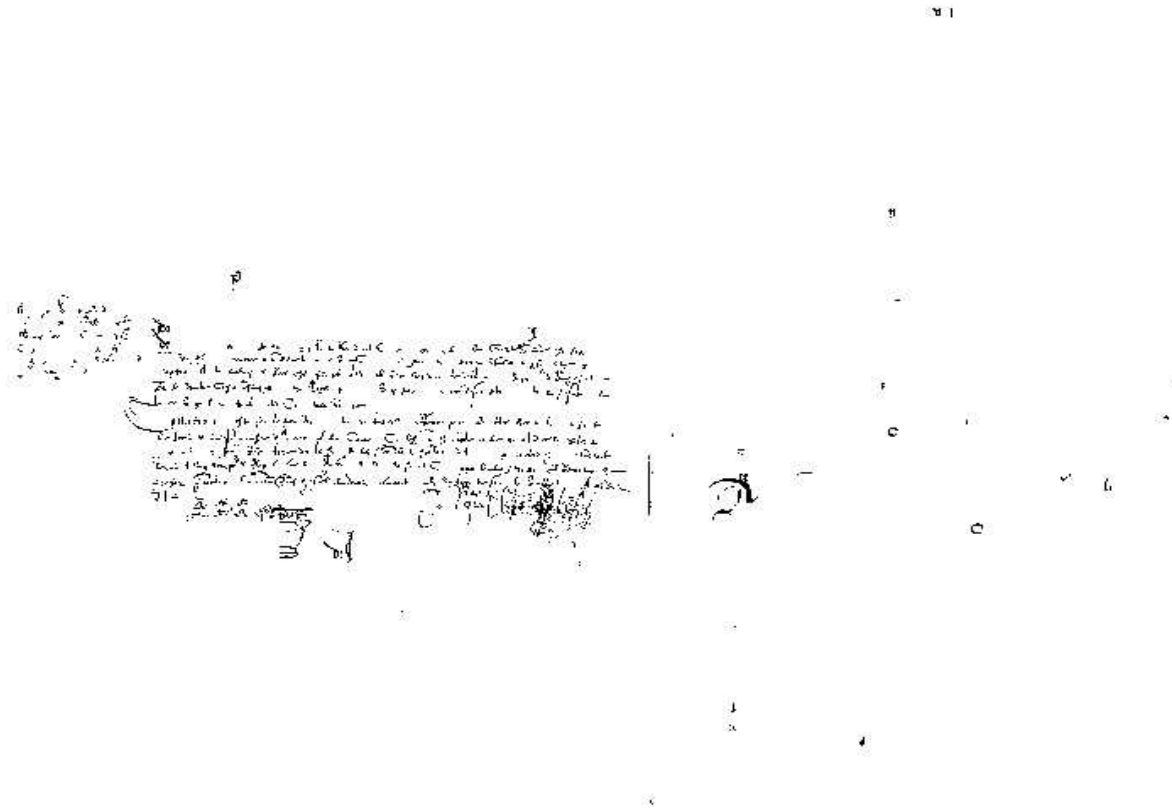


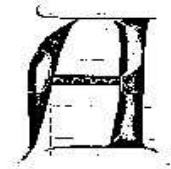
Image originale

Segmentation en type d'écriture



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

Extraction automatique des lettrines

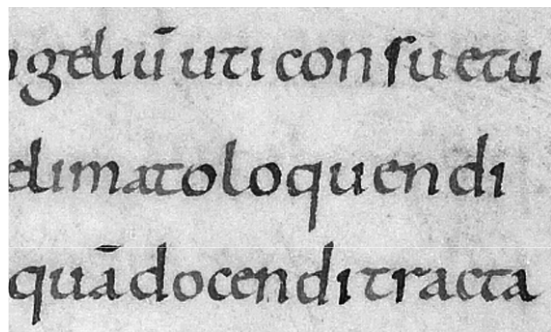
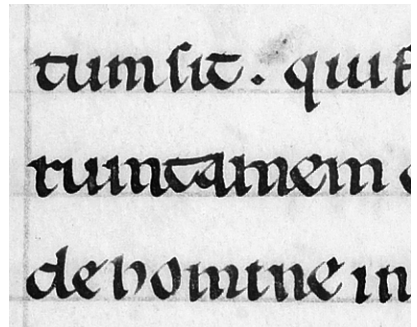


Classe n°3 : Lettrines de couleur

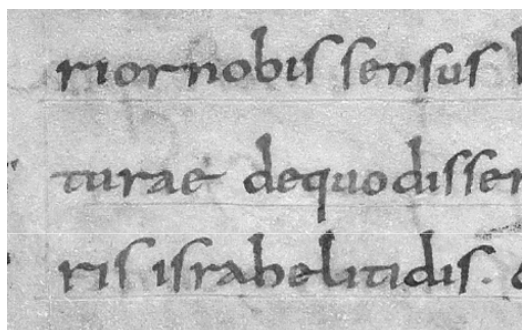
Paléographie sur manuscrits réguliers

Requête : Trouver le style d'écriture la plus proche

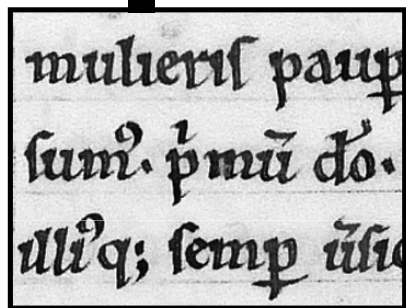
Hypothèse de date : XII



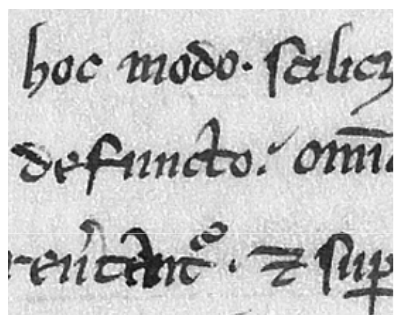
Corbie 850



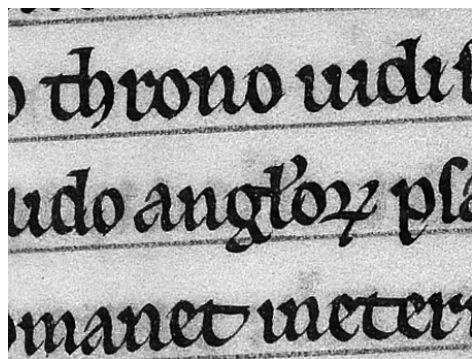
IX



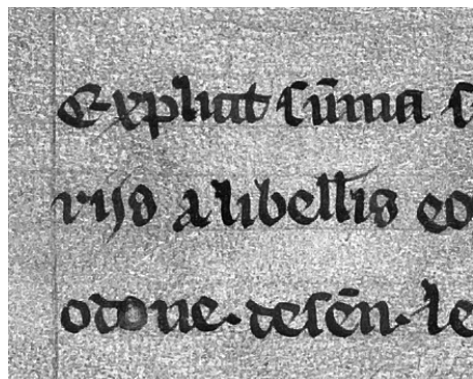
Vauclaire 1150



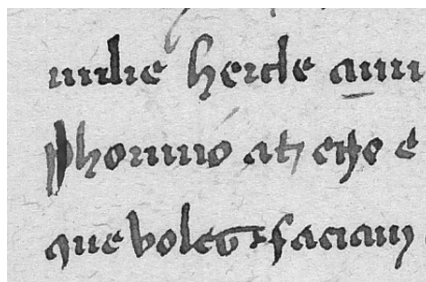
Laon 1262



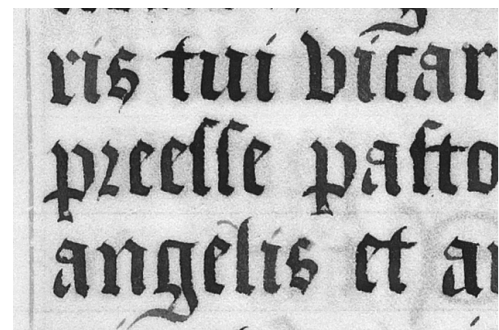
1222 Cistercien



1325

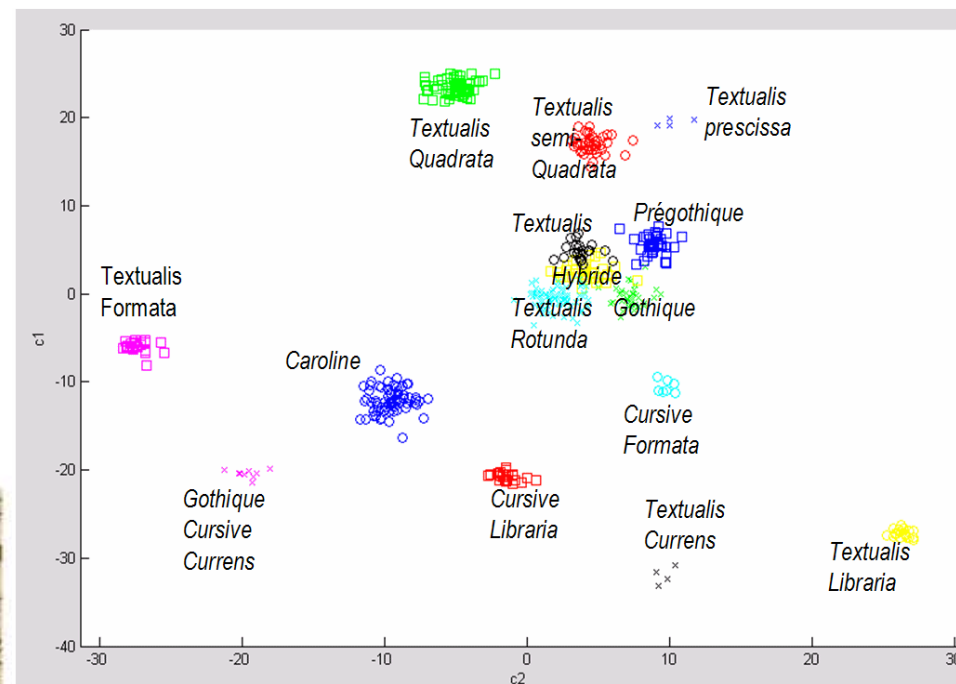
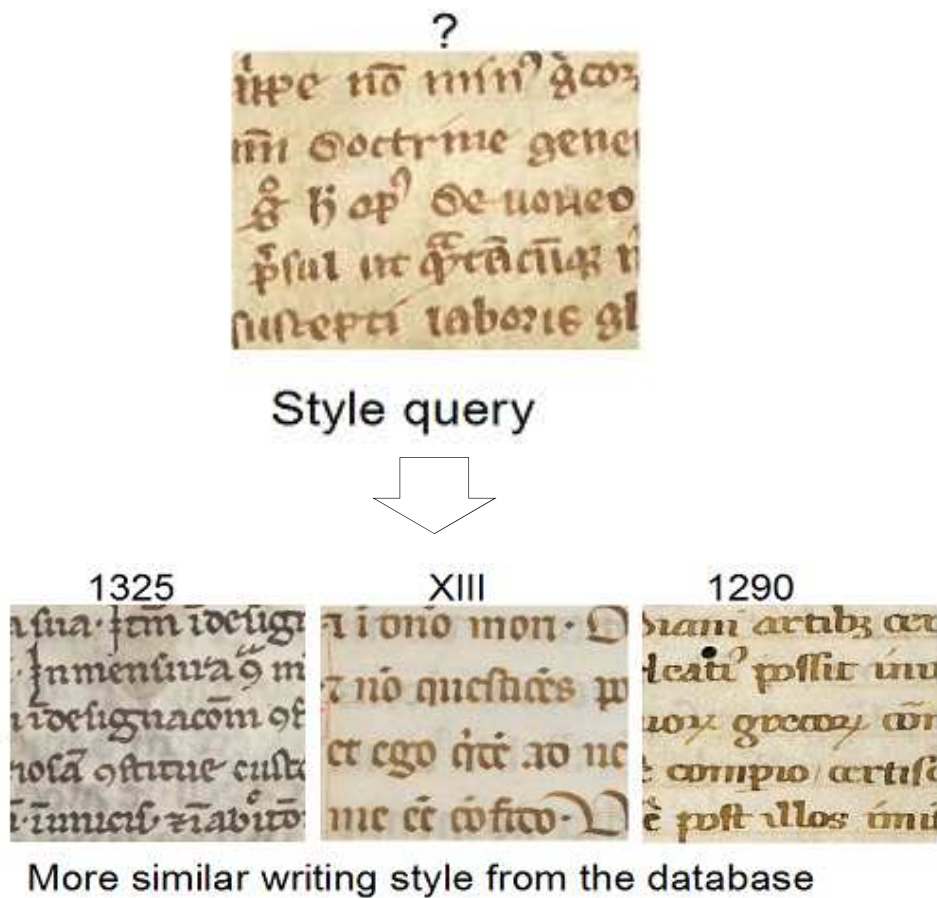


1401-1450



1535

Paléographie sur manuscrits réguliers



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

Paléographie sur manuscrits réguliers

Image traitée :   011ARU000000014U00002000.bmp

 <p>99.4523 c5487-08.bmp</p>	 <p>99.3757 c1333-06.bmp</p>	 <p>99.3204 c1329-05.bmp</p>	 <p>99.2169 c1570-07.bmp</p>	 <p>99.1468 c5720-03.bmp</p>
 <p>99.0737 c1334-07.bmp</p>	 <p>98.9809 c5460-02.bmp</p>	 <p>98.9659 c1333-08.bmp</p>	 <p>98.9515 011ARU000000016U00002000.bmp</p>	 <p>98.9485 c5488-01.bmp</p>

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

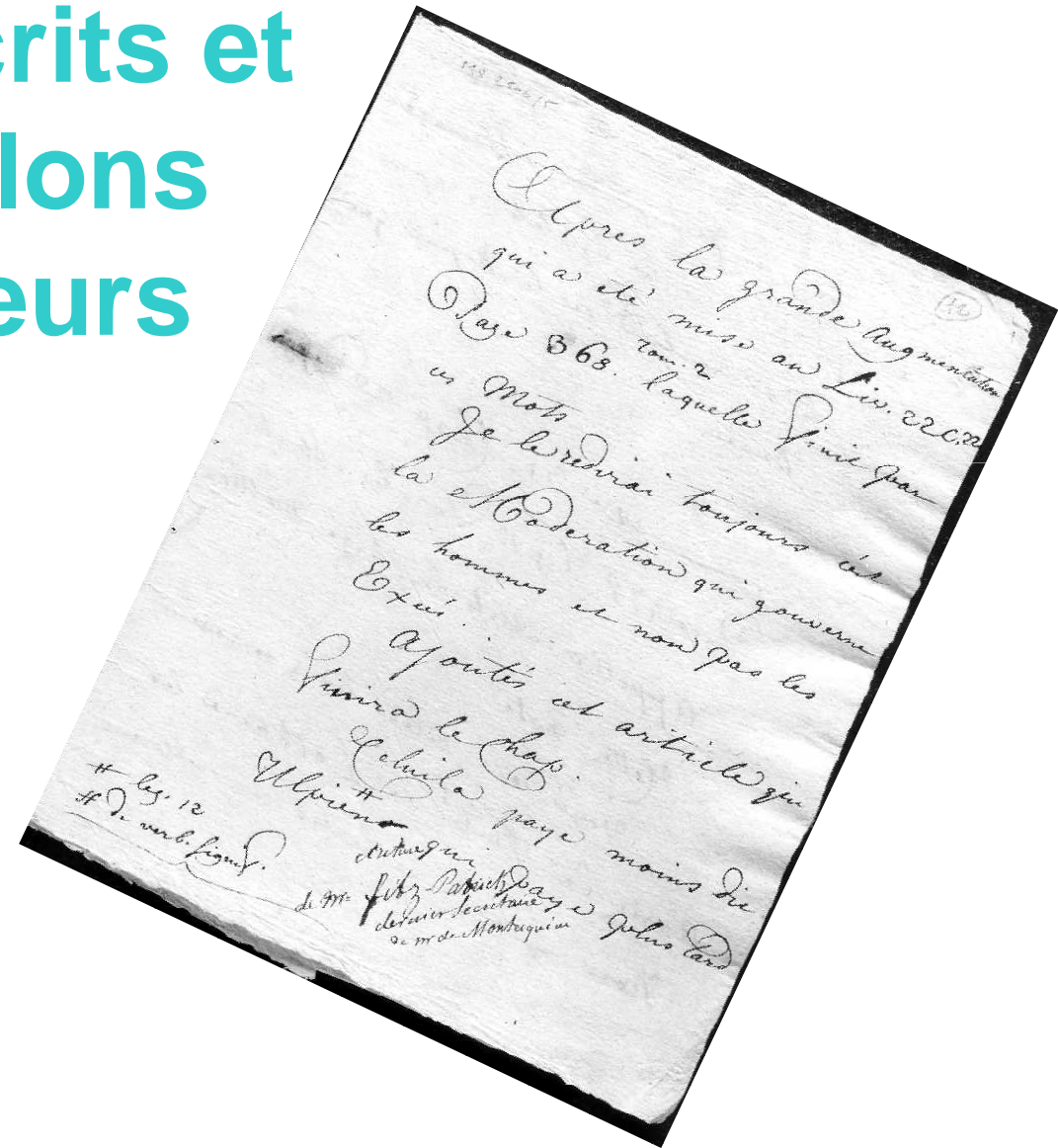
Paléographie sur manuscrits réguliers

Image traitée :   Flaubert/FLAUBERT_MB10.bmp

 98.7534 FLAUBERT_MB2.bmp	 98.5632 FLAUBERT_MB12.bmp	 98.5038 FLAUBERT_MB14.bmp	 97.8494 FLAUBERT_MB3.bmp	 97.1213 FLAUBERT_MB4.bmp
 97.085 FLAUBERT_MB7.bmp	 97.025 FLAUBERT_MB22.bmp	 96.428 FLAUBERT_MB13.bmp	 95.7674 MONTES-PHOTO2.bmp	 95.3565 FLAUBERT_MB17.bmp

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

Manuscrits et brouillons d'auteurs



Analyse des manuscrits contemporains

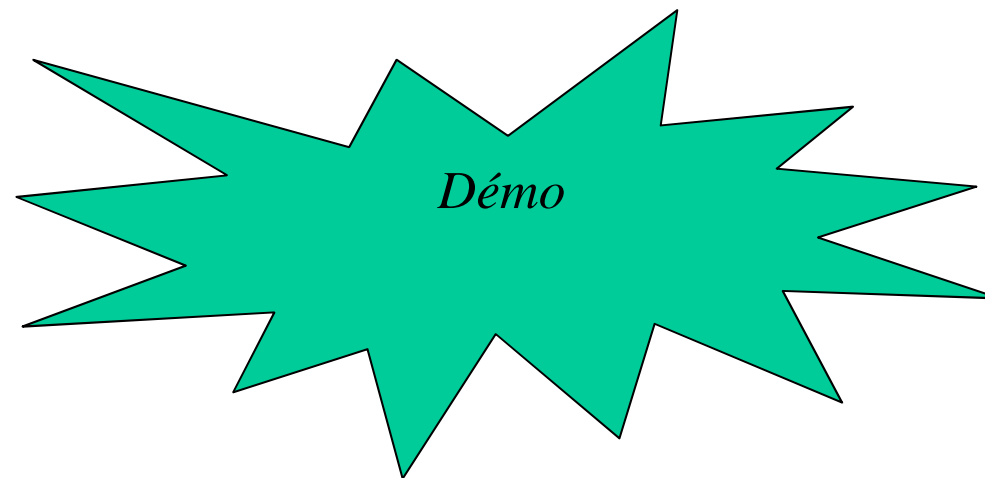
- **Exploitation scientifique des fonds manuscrits d'auteurs**
 - Rendre accessible aux chercheurs des textes de grande fragilité
 - Proposer à l'étude des textes dont l'authenticité reste à prouver
 - Préserver et valoriser les contenus
- **La problématique du brouillon d'auteur (Montesquieu- 1750)**
 - 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

^{o pour le liv. 27}
27: mettre la citation de *regulae de A. de Comma*
~~de deo, d'alicunus et Honor~~
28
mal effai
revoir tout
les chronologies
sur romes
J'ay mis au livre 22 chapitre 22 ala page 368
a la note que le pretur Sempronius fut tue
par les evaniers Ban de rome 663. J'ay
trouve en Sigonius Ban 664 voir qui le trompe
de Sigonius ou de moy ou de l'editeur
J'ajoute que Sigonius page 107 ajoute *ffine postea*
anno (cest a dire 665) *ps Sulpicius in tribunitia*
legem tulit teste Plutarcho in Sylla, ne quis senator
plus sestertium duobus milibus deberet; Consul
autem L. Sylla, et P. Pompeius, ut debitorum pecuniam
paucam solverent, si modo recte opus festum legitur

Quelques exemples...

Histoire Vraisemblable
Livre Cinquieme 4^e

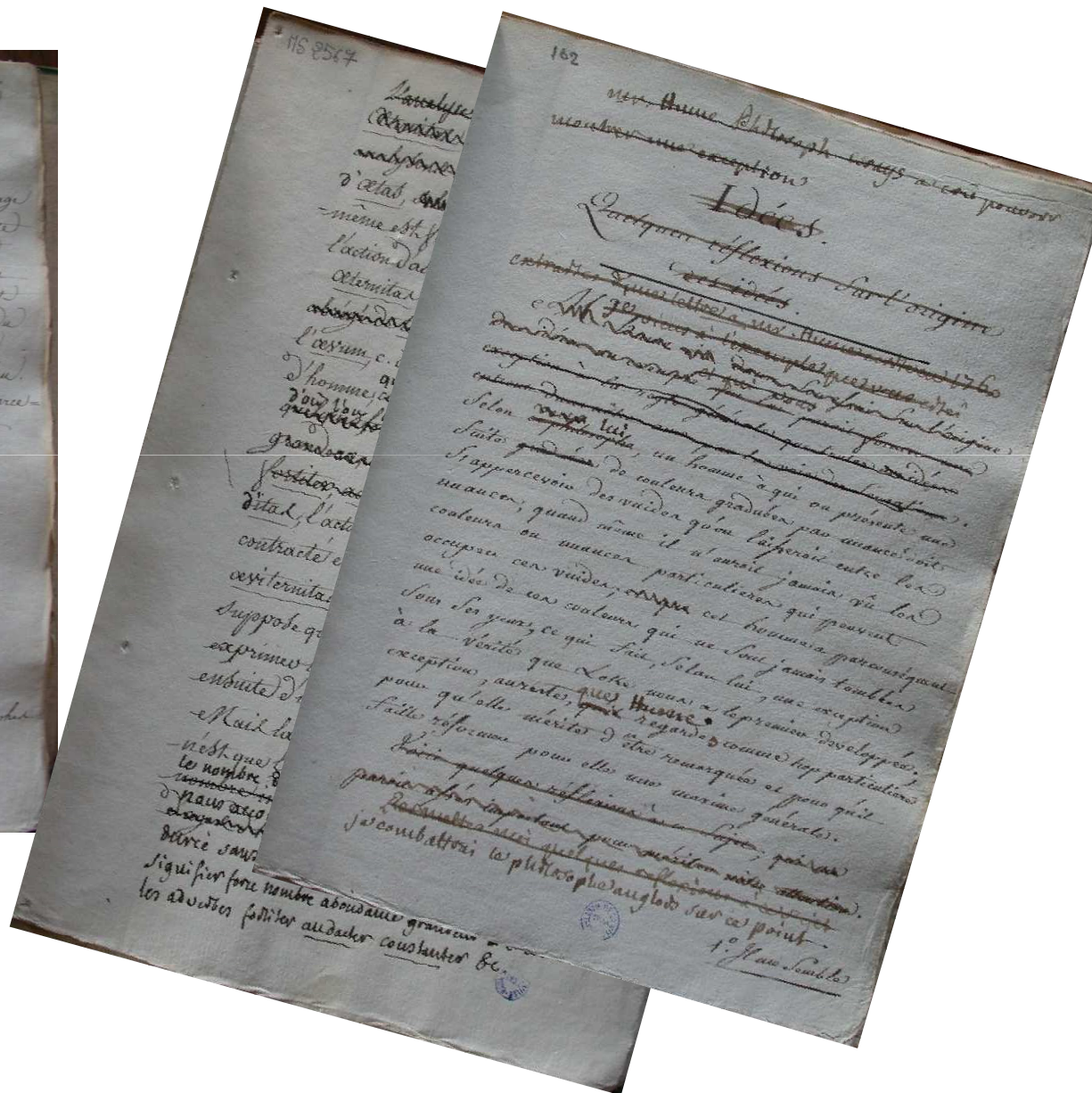
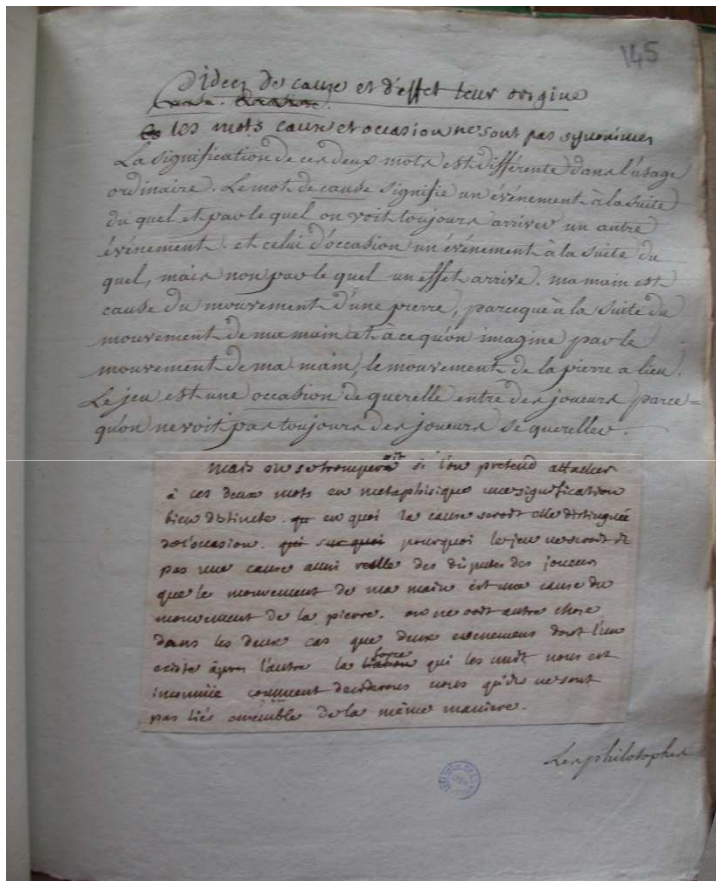
Dans cette lettre je me plains tant et si
et que mon genie pendant qu'on me
et il y a longtemps que tu
tu que selon le pouvoir que j'en
je te metamorphose tout
que c'est un homme? c'est, selon
dirige toute l'eternité. Oh bien
recommande le Roi de son
genie, il est si desiré que
doux mort a Vienne, dont le

²⁹
Dont il vint les autres
~~quand il vint les autres~~
mais celle des lois qui est mille fois plus enragée
devoir faire: J'allois mais c'est a regret: des que j'en
fait ~~ce~~ offrir le Roi ~~ce~~ vent flatter, mais j'étais si
indigné contre lui que j'ai lui donnai un coup de
bâton et la fille a dix pas de là.

Tout de bon les courtisans m'embarrassent
si j'en mis mille darts, tous contre moi, j'allois
~~seul~~
mort: J'indai chacun briser les armes, plusieurs même
virent me Carrier et un instant après tout le
monde disparut.

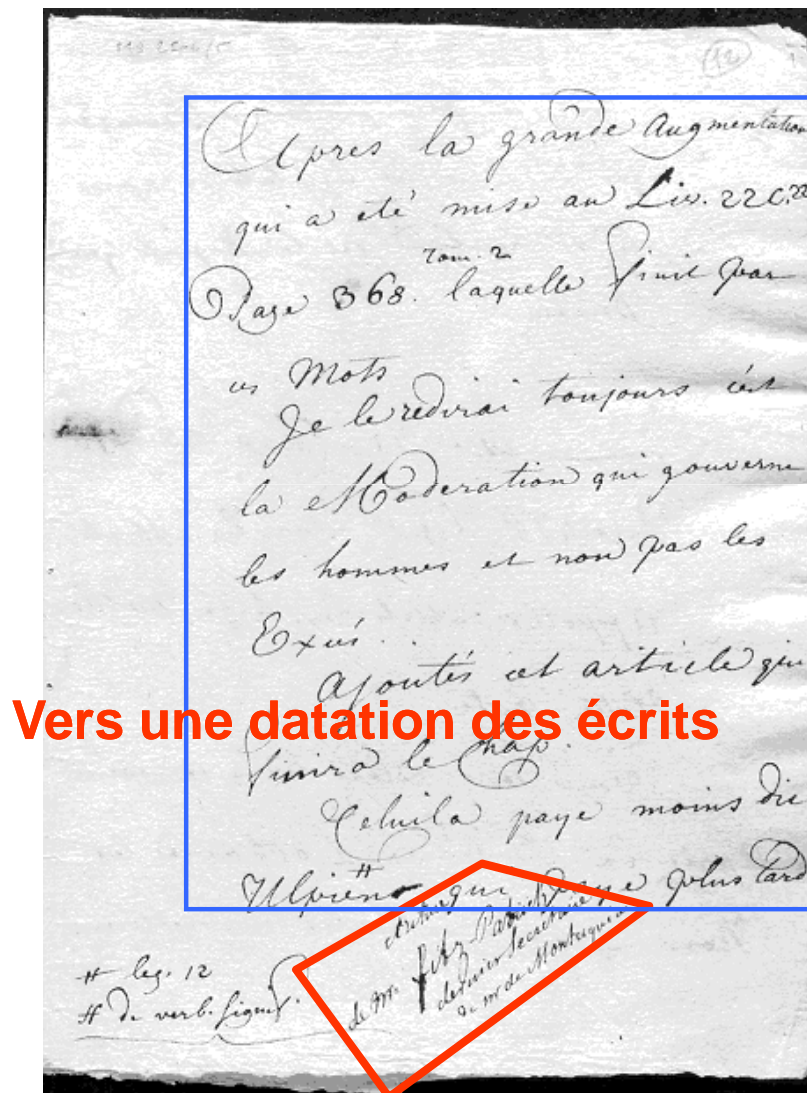
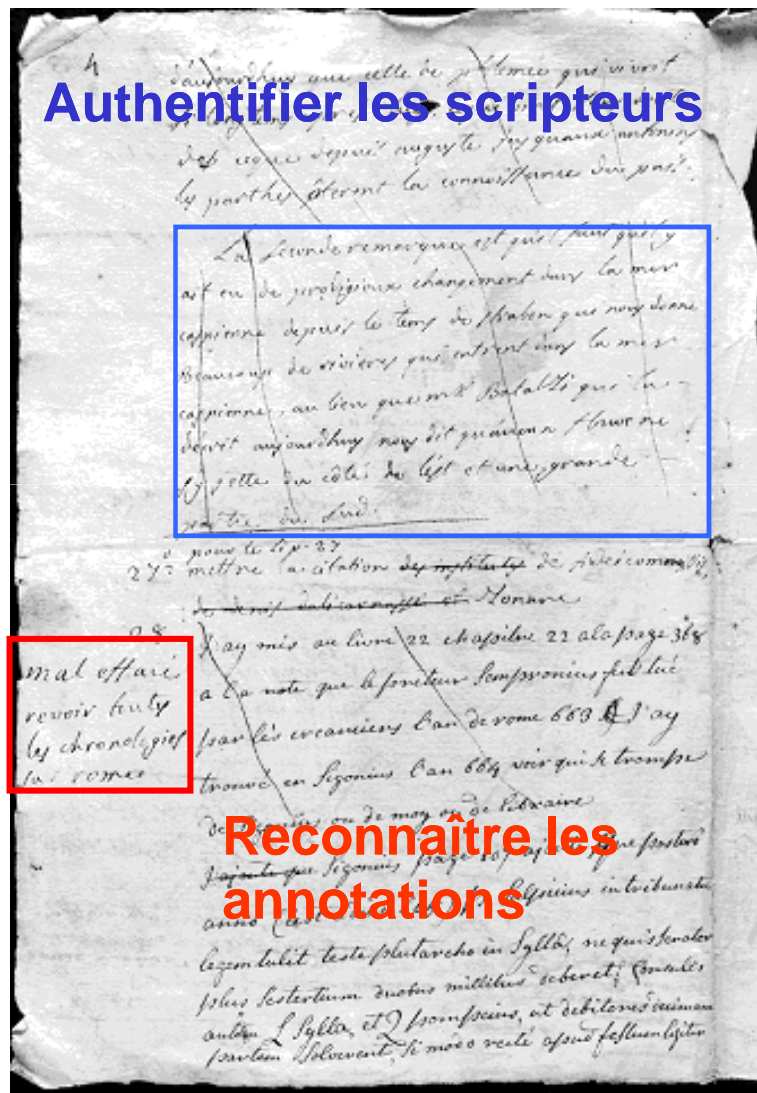
Analyse des manuscrits contemporains

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



Analyse des manuscrits contemporains

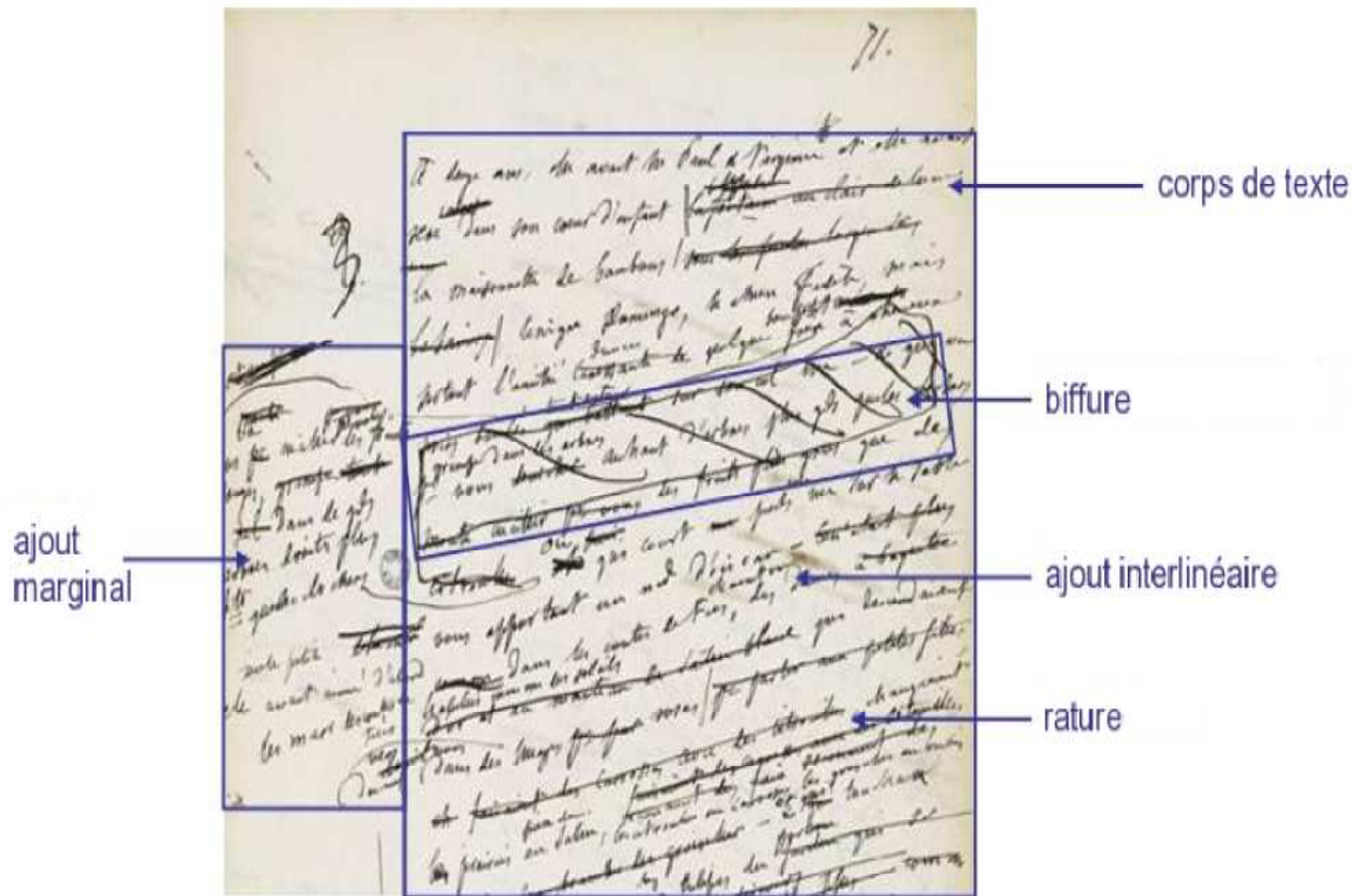
Découpage idéal



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

Analyse des manuscrits contemporains

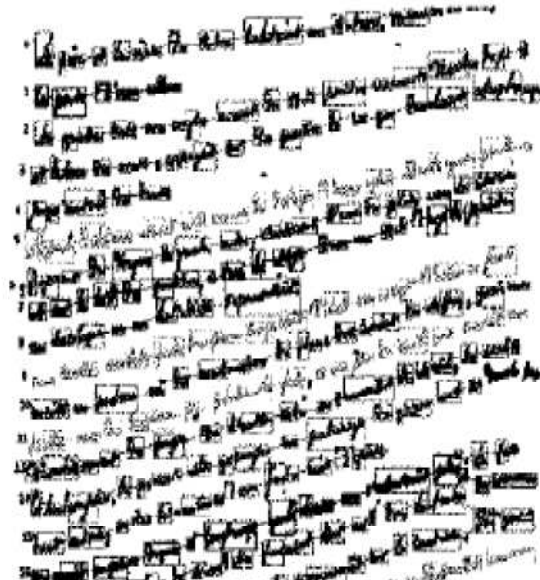
Découpage idéal



G. Flaubert – Madame Bovary

Analyse des manuscrits contemporains

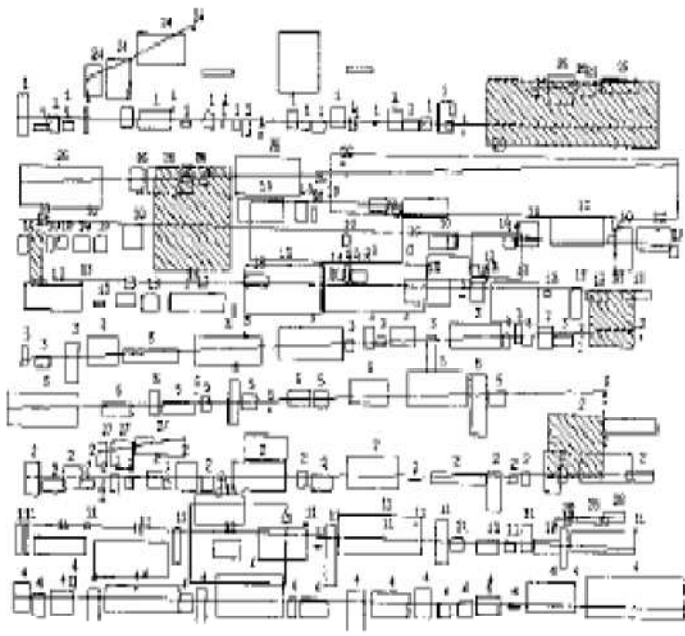
de nombreux aspects appartenant
 de manière non binaire de leur. Il y a une certaine ambiguïté
 de la... de possession...
 de son fait... de son fait... de son fait...
 de son fait... de son fait... de son fait...
 de son fait... de son fait... de son fait...



Au niveau des lignes et du bloc de texte

Séparation des lignes.
 Likforman, 2000

^{de gobierno} - I -
 Anti - Dios, Contra - Cristo, ^{Anti - humanos}
 Pueblo Anti - pueblo, ~~...~~
 servicio el ~~...~~ ^{en un caso anti - país,}
 anti - ma suon, ~~...~~ ^{anti - tiempo por un}
 su gobierno ~~...~~ ^{anti - todo y controlados,}
 no tiene fin, ~~...~~ ^{anti - fin.}
 Anti - se doler, ^{con} sus anti - amigos (a nadie
 Homo amigo) ^{anti - familia} (a nadie) ^{con}
 de su parente) ~~...~~ ^{anti - sus partidarios - anti -}



^{de gobierno} - I - ^{Anti - humanos}
 Anti - Dios, Contra - Cristo, ~~...~~
 Pueblo Anti - pueblo, ~~...~~
 servicio el ~~...~~ ^{en un caso anti - país,}
 anti - ma suon, ~~...~~ ^{anti - tiempo por un}
 su gobierno ~~...~~ ^{anti - todo y controlados,}
 no tiene fin, ~~...~~ ^{anti - fin.}
 Anti - se doler, ^{con} sus anti - amigos (a nadie
 Homo amigo) ^{anti - familia} (a nadie) ^{con}
 de su parente) ~~...~~ ^{anti - sus partidarios - anti -}
 todos sus anti - ~~...~~ ^{anti - todo lo que era y representaba el}
~~...~~ ^{anti - tiempo por un}
 anti - humanos, ~~...~~ ^{anti - injusticia,}

Analyse des manuscrits contemporains

Hough + Détection de cellules

cerfa N° 10164

Compte de résultat et A du 1^{er} au 31^{er} de l'exercice

4

COMPTÉ DE RÉSULTAT

Désignation de l'entreprise : JANTIA

		Exercice N	Exercice N-1
Produits exceptionnels	Produits exceptionnels sur opérations de gestion	III	
	Produits exceptionnels sur opérations en capital	III	219 800
	Répères sur provisions et transferts de charges	III	219 800
	Total des produits exceptionnels (I + III)	III	5 415
Charges exceptionnelles	Charges exceptionnelles sur opérations de gestion	III	125 057
	Charges exceptionnelles sur opérations en capital	III	5 200
	Charges exceptionnelles sur provisions et transferts de charges	III	176 400
	Total des charges exceptionnelles (I + III)	III	83 048
X - RESULTAT EXCEPTIONNEL (II - III)		III	
Prévisions des résultats sur opérations de gestion		III	
Impôts sur les bénéfices		III	5 350 905
TOTAL DES PRODUITS (I + II + X + Y)		III	7 275 680
TOTAL DES CHARGES (I + III + VI + VII + IX + X)		III	65 240
Z - BÉNÉFICE OU PÉRIE (total des produits - total des charges)		III	

100

100

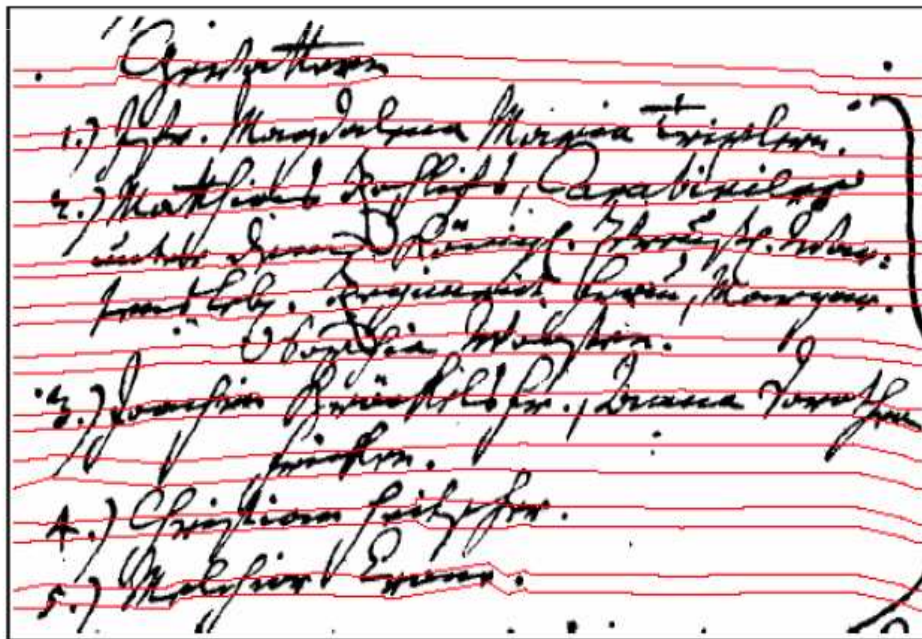
Analyse des manuscrits contemporains

Au niveau des lignes et du bloc de texte

trop choquer ce sujet ce
pourrait éclater en mo
r plutôt un ^{trou}seule : e
as de reposer trop brua
à un peu de champ e

Erons sont utiles et o
le monde est petit
choquer ce sujet contre
éclater en morceaux
un ^{trou}seule : le ^{trou}isol
poser trop brutalement

Il ne sort
liquide d'un
dans un plus p
dont on pourra
Mais il est

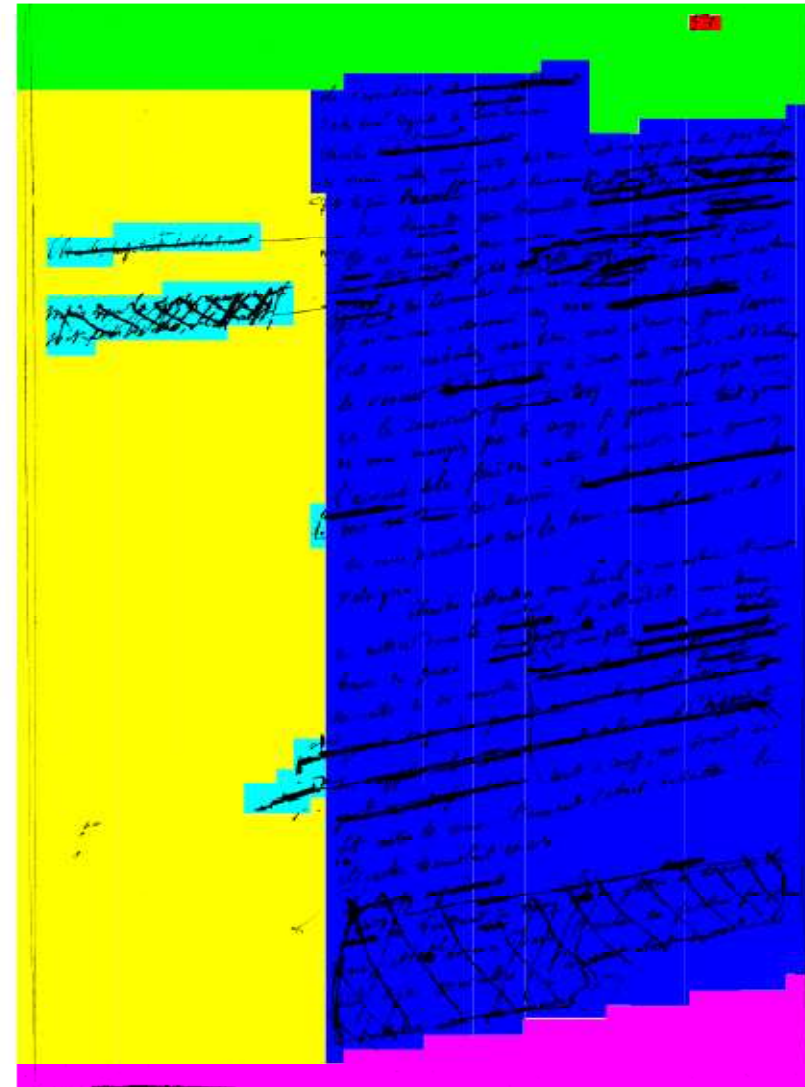
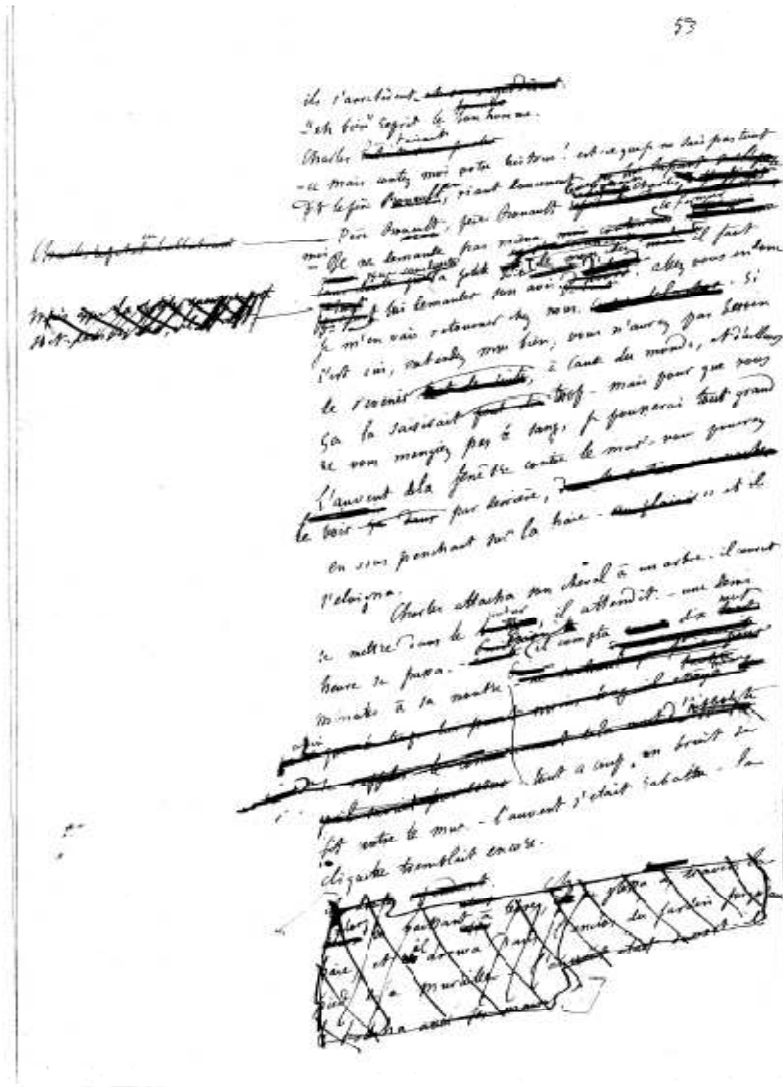


Laisse ça. Comme dirait: "Glorie", n'est-ce pas? Ce
peut-être en sa folie, c'est qu'elle se "protège", du beau jour,
des séductions de "Pélie", : jamais je ne me suis vu l'en-
propriétaire d'un "talent" : la seule affaire est de me seu-
rien dans les mains, rien dans les poches - par le travail et les-
sances. D'après ma pure option ce n'est pas en-dessus de personnel:
équipement, sans outillage pour sur tous tout entier à l'œuvre
pour me savoir tout entier. Si prange impossible Salut au sage-
des accessoires, qu'est-t-il? Tout un homme, fait de tous les
mes et qui le veut tous et qui veut à l'imp. qui s'écrit

par tous les sens et au sein
de prunedans, s'allorment

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

profondeurs d'impressions, comme le visage d'un
ovale à quatre balanciers ~~en l'air~~, elle commençait
par trois balanciers enroulés, sa sur le fond du bras enroulé
de prunelle, l'alternant séparé par une ~~une~~ ^{grande rouge}
des lozanges de couleur et de point de la figure
... de fait que le terrain est par
... en l'air

(a)

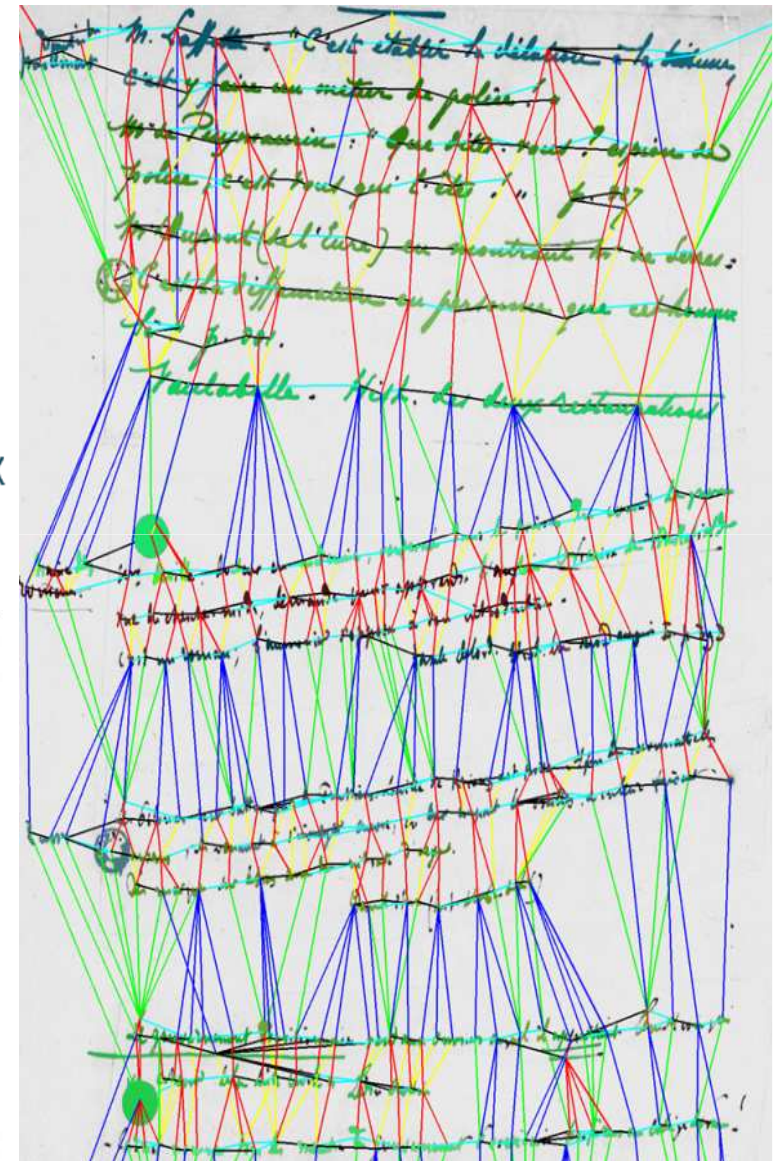
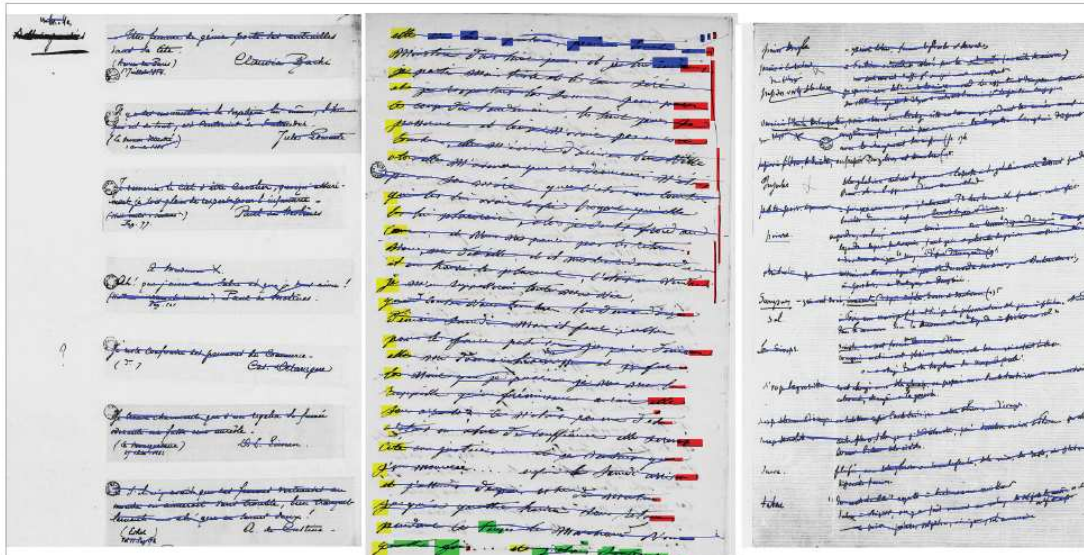
profondeurs d'impressions, comme le visage d'un
ovale à quatre balanciers ~~en l'air~~, elle commençait
par trois balanciers enroulés, sa sur le fond du bras enroulé
de prunelle, l'alternant séparé par une ~~une~~ ^{grande rouge}
des lozanges de couleur et de point de la figure
... de fait que le terrain est par
... en l'air

(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 entre composantes
- Segmentation du graphe en n sous-graphes correspondant aux lignes



Application à l'analyse des manuscrits de Flaubert



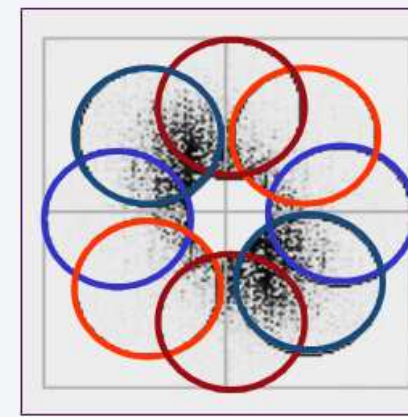
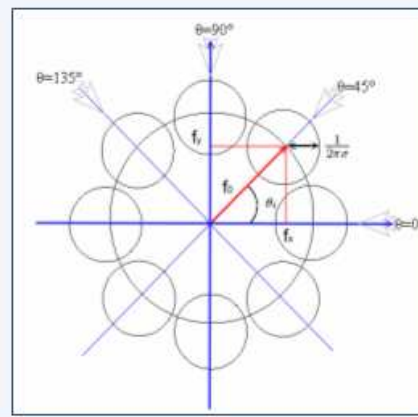
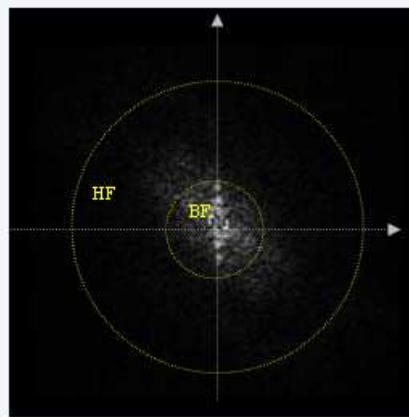
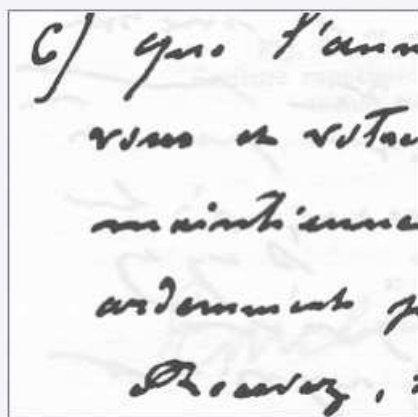
Jeune homme.
 Et toujours farceur, serait même incou-
 rable s'il n'en était pas. Comment, dit
 un jeune homme.
 Tout ce qu'il doit faire : chanter, danser,
 avoir des belles, pas trop cependant.

Jeunesse.
 Il faut toujours citer ces vers italiens,
 même sans les comprendre.
 "Gioventù! primavera bella vita."
 "Primavera! gioventù bel tempo."
 "Ah! c'est beau la jeunesse."

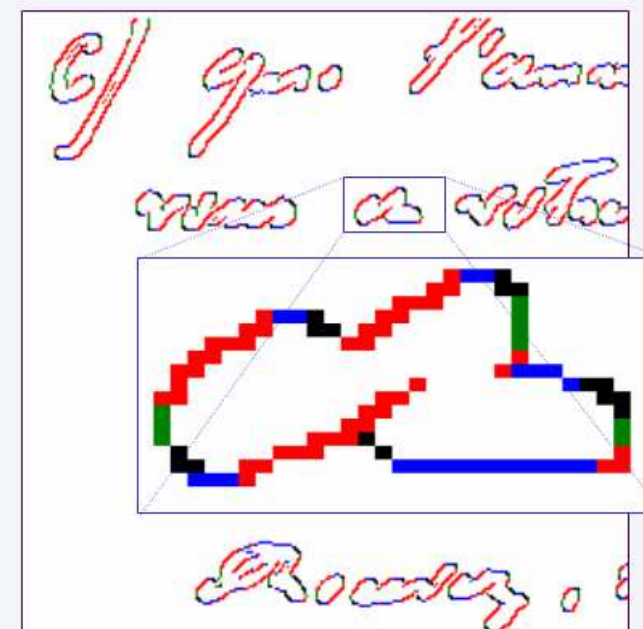
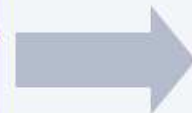
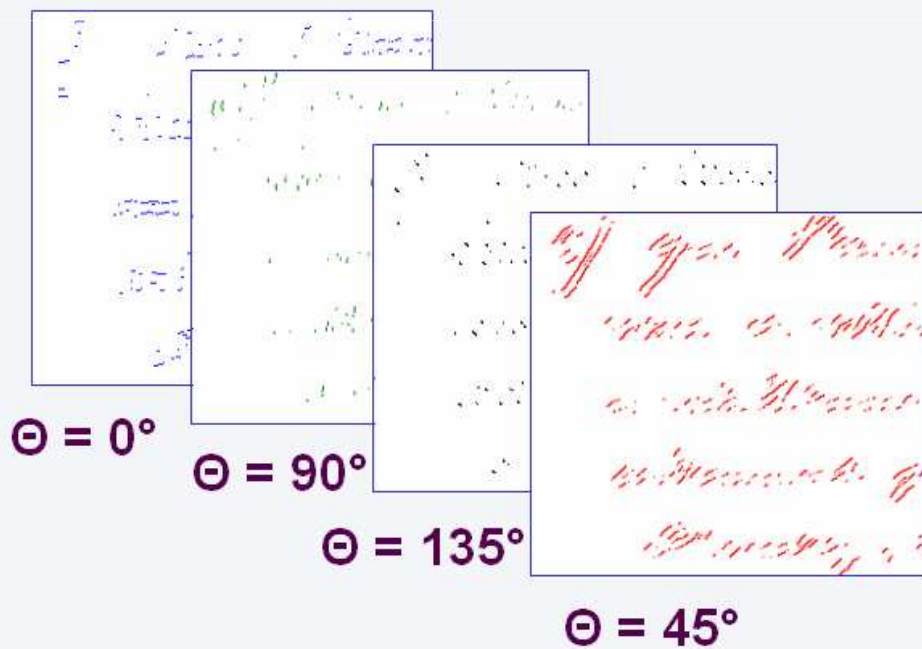
John Bull.
 Quand on ne sait pas le nom d'une capitale
 on l'appelle : John Bull.

Voilà.
 On mène des jeux de bel air ; on ne
 sait pas parler de ses filles.





Gabor filtering – outlines marking and directional maps decomposition

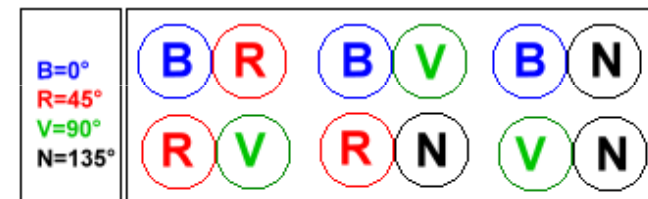


Je vous envoie les quatre premiers liers
 de ma Tragedie et je vous envoie le
 cinquiesme. Mais que je l'envoye
 vous supplie, avec respect. Pour le
 la police de l'écriture, et de marquer les fautes
 que je puis avoir faites contre la langue. C'est
 vous aller en de vos plus excellentes manières.
 Et vous en envoie quelques fautes. Mais c'est
 nature de vous par. Mais le bon. Mais les
 envoie vous. Mais. Je vous prie encore
 de faire part de cette lecture au R. Pour luy
 Il veut bien en donner quelques exemplaires.
 Je les envoie tres bon et tres bellement
 de luy. Mais.

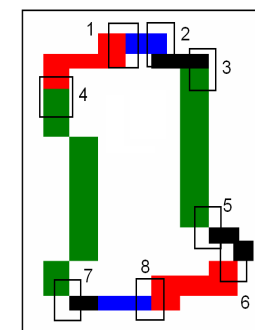
Forme recherchée

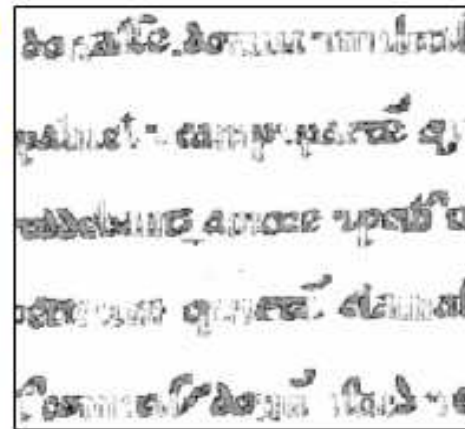
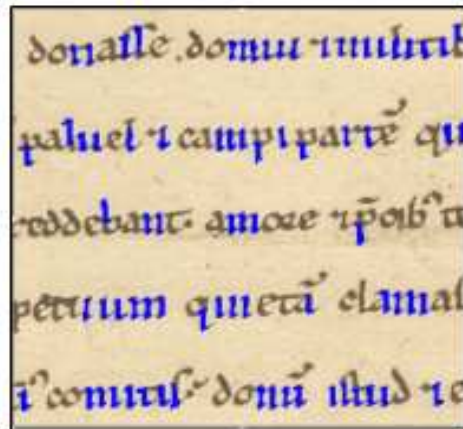
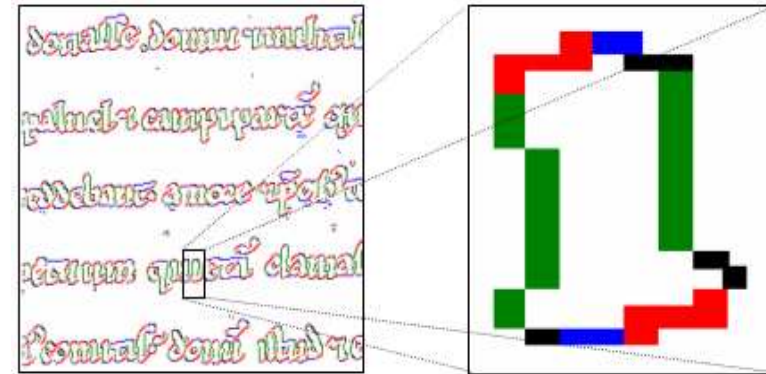
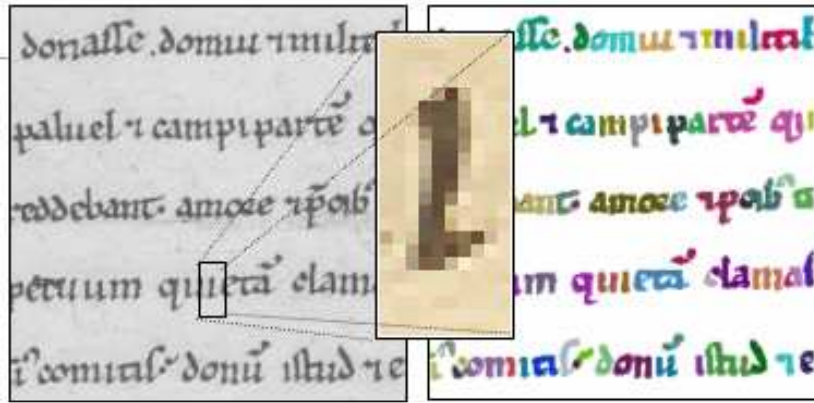


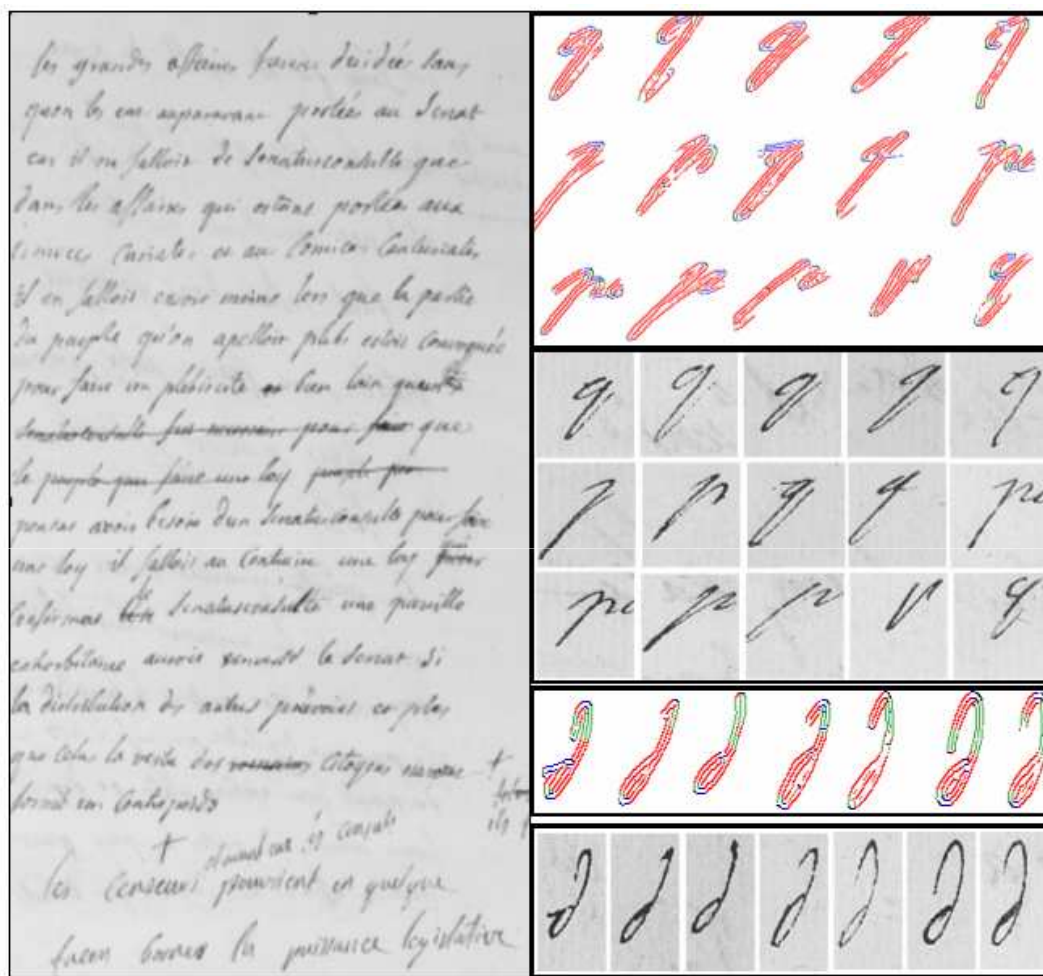
A partir de sa version codée



Junctions coding







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

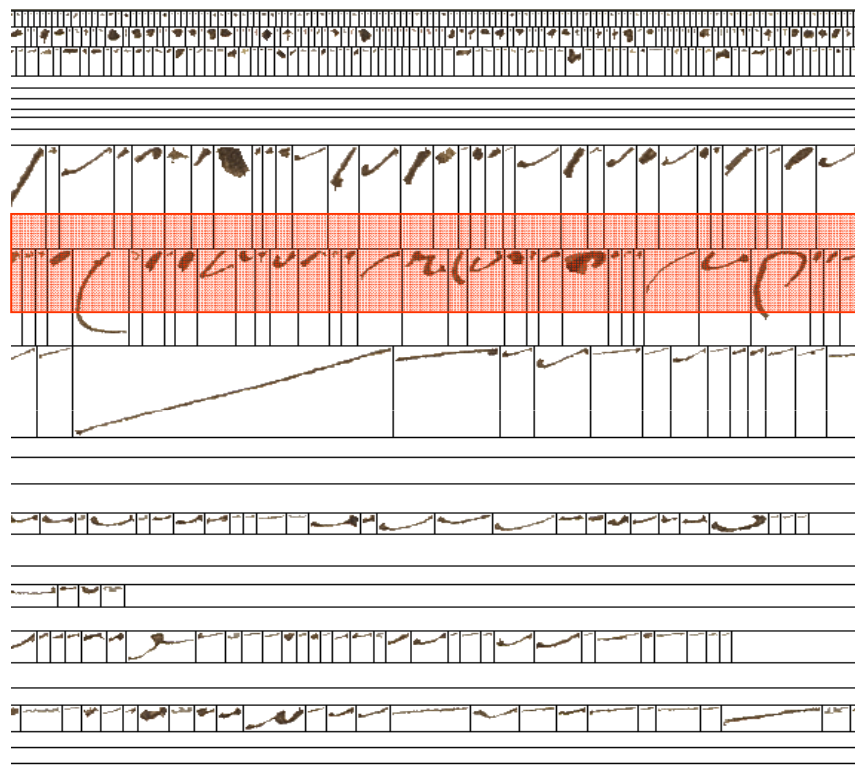
page 15
mettes la
lettre suivante

ibben a ubek a paris.

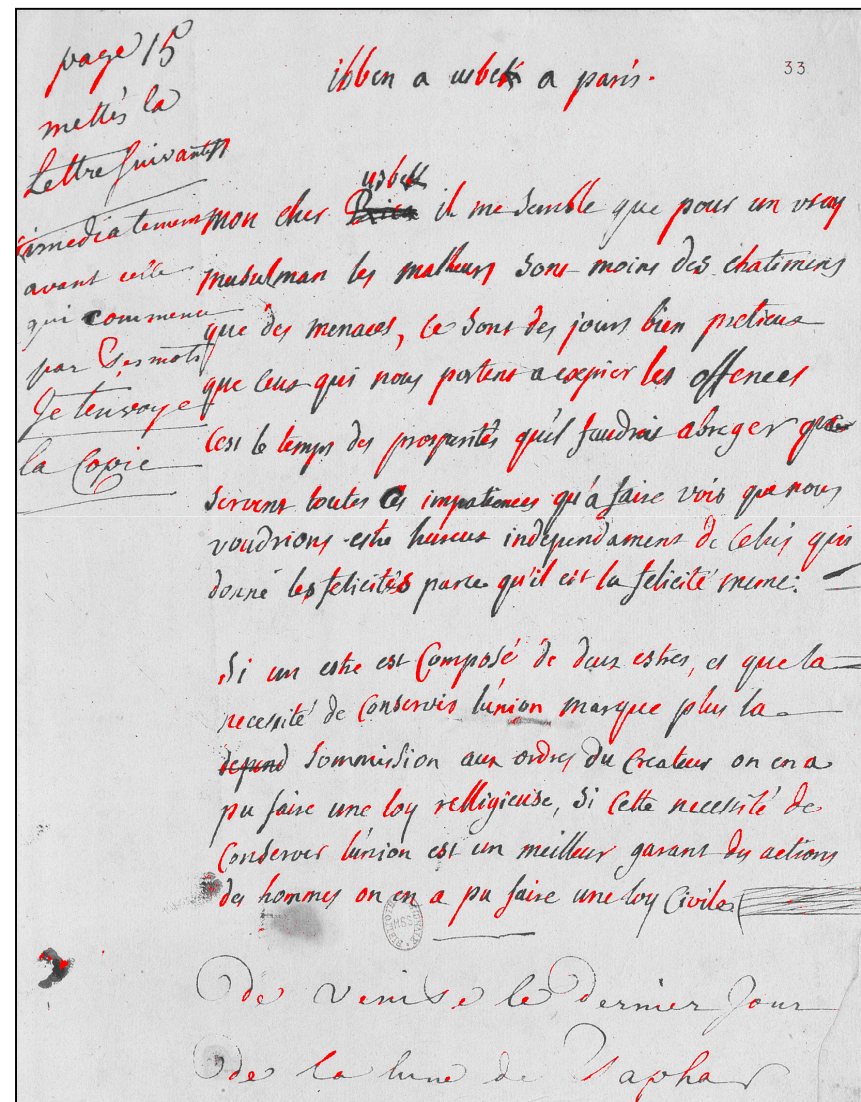
si media tenu mon cher ^{ubek} il me semble que pour un usage
avant cela musulman les malheur sous-main des châtiments
qui commencent que des menaces, ce sont des jours bien précieux
par ces motifs que ceux qui nous portent a expier les offenses
Je t'embrasse
la copie — c'est le temps des propitiations qu'il faudroit abréger que
servent toutes ces impatences qu'à faire voir que nous
voudrions être heureux indépendamment de celui qui
donne les félicités parce qu'il est la félicité même.

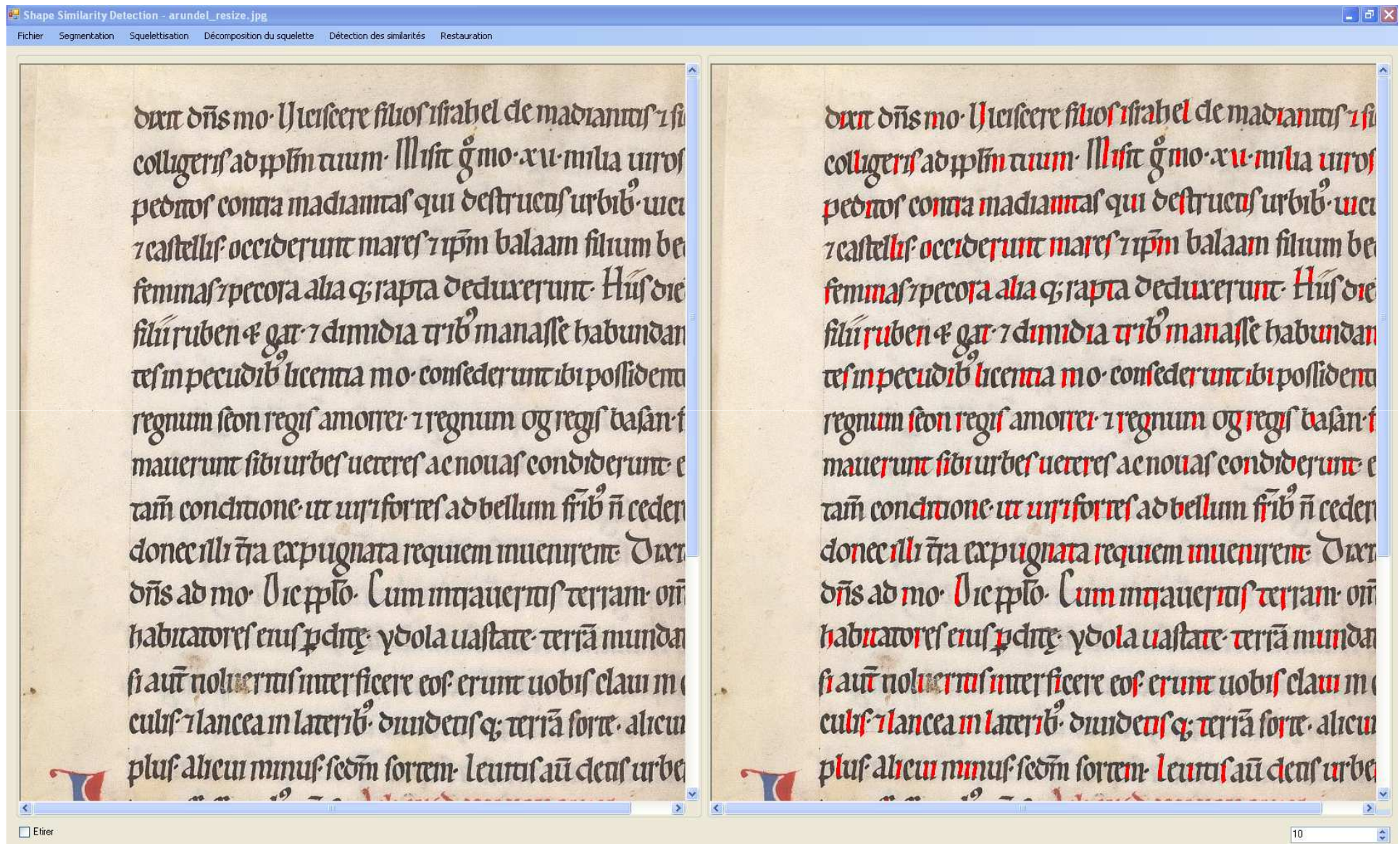
si un acte est composé de deux actes, et que la
nécessité de conserver l'union marque plus la
dépendance soumission aux ordres de l'acteur on en a
pu faire une loi religieuse, si cette nécessité de
conserver l'union est un meilleur garant des actions
des hommes on en a pu faire une loi civile.

De Venise le dernier jour
De la lune de Vapha



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

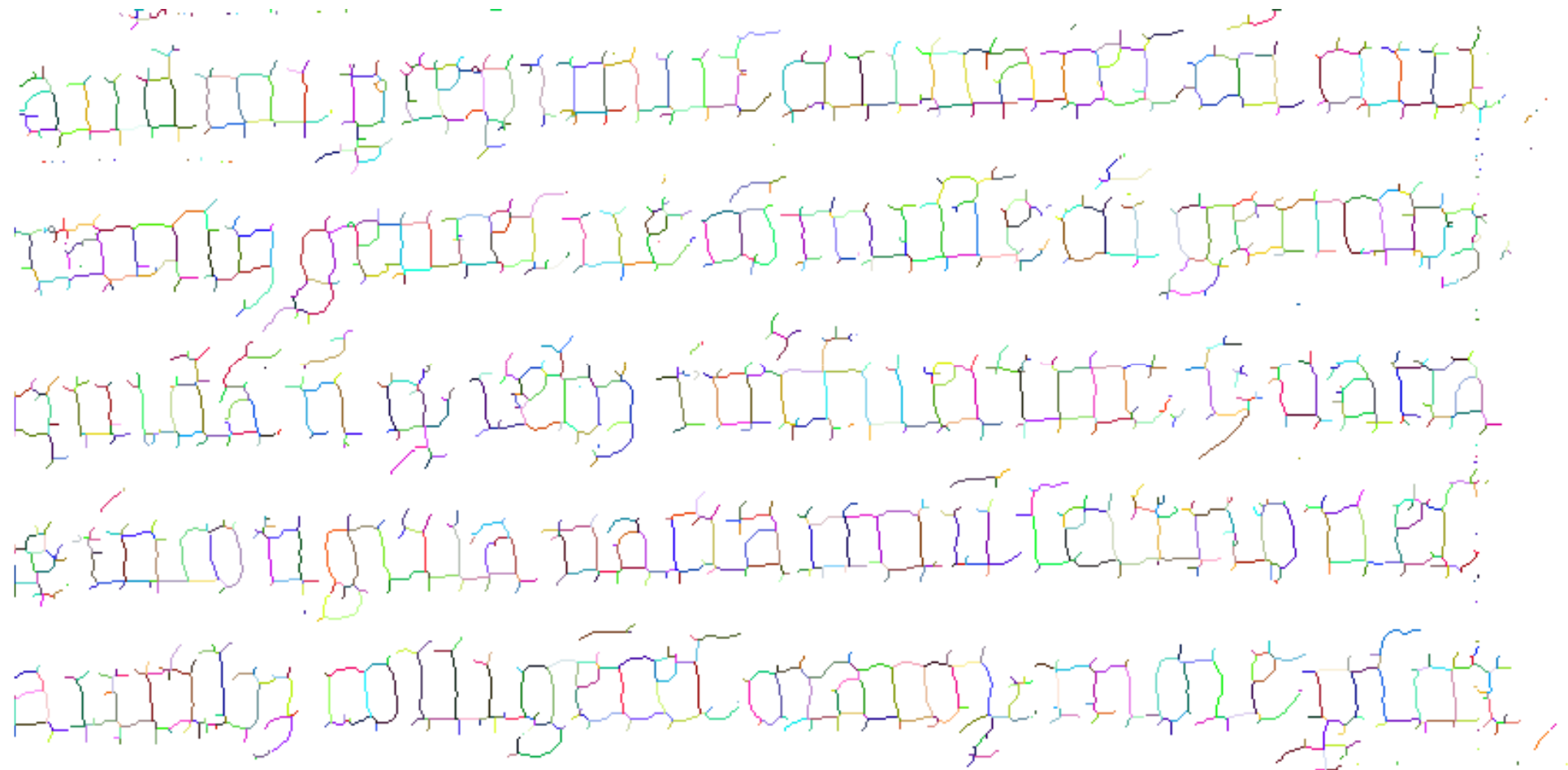




Détection des graphèmes similaires

auditi p̄cepimus civitates cū civi-
tatiḃz gentes ne cōmisse cū gentiḃz.
quidā n̄ q̄ rebz interfuerunt. Guana
e incongrua narratiū sermones
auriḃz colligētēs oratorz more p̄fer

Détection des graphèmes similaires



audita precepimus civitates cum civi-
 tatis gentes ne commiserant gentibus.
 quidam tamen quibus interfuerunt. Guana
 et incongrua narratum sermones
 auribus colligentes oratores more plerumque

Application à la recherche de mots en mode image

Salutis templum	53	cere	283
Salutis aedes	100.223	Sibyllæ Tiburtinæ simulacrum	301
Saturnius mons	39	Sibyllini libri in Capitolio seruabantur	66
Saturnia	39	Sicilia locus in monte Palatino	88
Saturnus Aboriginum rex Italiam tenuit	9	Siculi primi omnium urbem Romanam coluere	9
Saturni filij ex Oenotria	47	Syllæ statua	56
Saturni templum	46.109	Syluani templum	209
Scelerata porta Carmentalis dicta	16	Syluius Ascanio fratri successit	10
Sceleratus campus	223	Sisimuni basilica	201
Scipionis statua	56	Socordiae sacellum	151
Secretarium urbis	108	Solis aedes	157
Seiuges in capitolio	65	Solis templum	226
Senatulum	79.131	Solis & Lunæ aedes	100
Sepeliri in urbe quibus liceret	247	Solis turris	220
Septem salæ	197	Sesoriani palatium	165
Septimij arcus	108	Sororium Tigillum	125
Septimana porta	36	Spei templum	136
Septizonium	161	Statilij Tauri amphitheatrum	167
Septodium	162	Stercoraria porta	60
Septifolium	162	Suburra	208
Serapis	248	Suburra plana	200
Sertinij fornix in circumaximo	159	Suburrani cliuius	204
Seruorum dies festus	183		

Recherche du mot Salutis

Application à la recherche de mots en mode image

Salutis templum	53	cere	283
Salutis aedes	100 223	Sibyllae Tiburtinae simulacrum	301
Saturnus mons	39	Sibyllum libri in Capitolio	66
Saturnia	39	seruabantur	66
Saturnus Aboriginum re Italiam tenuit	9	Sicilia locus in monte Palatino	88
Saturni filii ex Oenotria	47	Siculi primum omnium urbem Romam coluere	9
Saturni templum	46 109	Syllae statua	36
Scelerata porta Carmentalis	16	Sylvanum templum	209
Sceleratus campus	223	Sylvius Ascanio fratri successit	10
Scipionis statua	36	Sifirum basilica	201
Secretarium urbis	108	Socoratae sacellum	151
Seiuges in capitolio	63	Solis aedes	157
Senatum	79 131	Solis templum	226
Sebeliri in urbe quibus liceret	247	Solis & Lunae aedes	100
Septem sale	197	Solis turris	220
Septimij arcus	108	Sesoriani palatium	163
Septimana porta	36	Sororum Tigillum	123
Septizonium	161	Sper templum	136
Septoaium	162	Saturni Tauri amphitheatrum	167
Septifolium	162	Stercoraria porta	60
Serapis	248	Suaurra	208
Seruij fons in circo riuus	139	Suaurra plana	200

Se: Recherche du mot Salutis

Application à la recherche de mots en mode image

Salutis templum	3	cere	283
Salutis aedes	100 23	Sibyllæ Tiburtinæ simula-	
Saturnius mons	39	crum	301
Saturnia	39	Sibyllin libri in Capitolio	
Saturnus Aboriginum re-		seruabantu	66
Italia n tenuit	9	Sicilia locus in monte Pa-	
Saturni filij e Oenotria		latino	88
47		Siculi p inu o nniū urbem	
Saturni templum	46 109	Roman colucre	9
Scelerata porta Carmenta		Sylla statua	36
Is dicta	16	Sylvanū templum	209
Sceleratus campus	223	Sylvius Ascamo frat i succo-	
Scipionis statua	36	cessit	10
Secretarium urbis	108	Sisinnū basilica	01
Seiuges in capitolio	63	Socordiae facellum	151
Senatulum	79 131	Solis aedes	157
Sepeliri in urbe quibus lice-		Solis templum	226
et	47	Solis & Lunæ aedes	100
Septem sala	197	Solis turris	220
Septimij arcus	108	Seiso iani palatium	163
Septimana porta	36	Sororium Tigillum n	123
Septizonium	161	Spei templum	136
Septodium	62	Statilij Tauri amphithea-	
Septisolum	16	trum	167
Serapis	2 8	Stercoraria po ta	60
Sertimij formi in circo ma-		Suburra	208
imo	139	Suburra plana	200

S
S Recherche du mot Salutis 3

Application à la recherche de mots en mode image

cette commodité que j'ay recherchée depuis quelque
reponces à deux de vos lettres que j'ay receues
tout ce que j'y ay veu ma fort diuenty horsmis quan
dauoir de L'argent Lors qu'il est si necessaire, Il
ressemble dans le deplaidr que vous en aués ne
us viendrons a bout de toutes ces difficultés Je v
crainv effort puis que le tems de mien aller s'aym
ne lettre de mon Cousin ynaud où il me faisoit esp
sua baille dans quelques jours Je luy ay escrit par la
cher son filz à Montauban, et ie luy ay vniens d'ou
vous deu ne s'entener l' - l'ornu de ce moy, Le laun
l'Examery Rigorosa se fait la v en d'ou Les logic
la Cene, de tout que par ce moyen il faut que i'o
c le veras que nous donnons aux professeurs s'of
faudroit que i'usse pour payer l'hoste que en don
ne vous de tout s'enouir ce qu'il faut que U us m'c
i Eau pour les lettres, s' Eau pour payer l'habit
m' 30 f pour un mois de Musique que ie dois de p
vres de Chandelles et pour la port de quelques Let

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

cette commodité que j'ay recherchée depuis **quelque**
reponces à deux de vos lettres que j'ay receues
tout ce que j'y ay veu ma fort diuenty horsmis quan
dauoir de L'argent Lors qu'il est si necessaire, Il
ressemble dans le deplaidr que vous en aués ne
us viendrons a bout de toutes ces difficultés Je v
crainv effort puis que le tems de mien aller s'aym
ne lettre de mon Cousin ynaud où il me faisoit esp
sua baille dans **quelques** jours Je luy ay escrit par la
cher son filz à Montauban, et ie luy ay vniens d'ou
vous deu ne s'entener l' - l'ornu de ce moy, Le laun
l'Examery Rigorosa se fait la v en d'ou Les logic
la Cene, de tout que par ce moyen il faut que i'o
c le veras que nous donnons aux professeurs s'of
faudroit que i'usse pour payer l'hoste que en don
ne vous de tout s'enouir ce qu'il faut que U us m'c
i Eau pour les lettres, s' Eau pour payer l'habit
m' 30 f pour un mois de Musique que ie dois de p
vres de Chandelles et pour la port de quelques Let

Recherche de « quelque » TE=40%

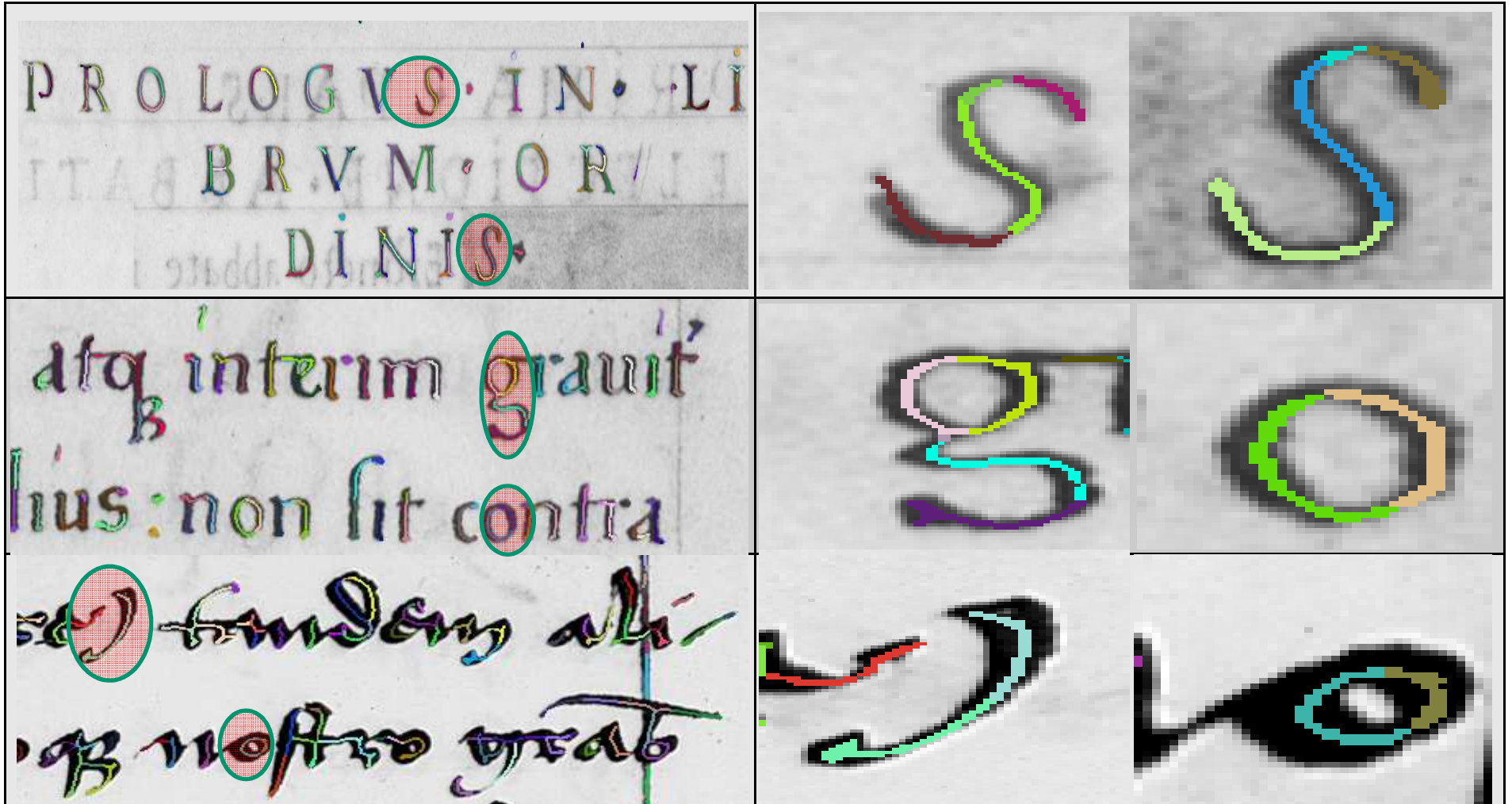
cette commodité que j'ay recherchée depuis **quelque**
reponces à deux de vos lettres que j'ay receues
tout ce que j'y ay veu ma fort diuenty horsmis quan
dauoir de L'argent Lors qu'il est si necessaire, Il
ressemble dans le deplaidr que vous en aués ne
us viendrons a bout de toutes ces difficultés Je v
crainv effort puis que le tems de mien aller s'aym
ne lettre de mon Cousin ynaud où il me faisoit esp
sua baille dans **quelques** jours Je luy ay escrit par la
cher son filz à Montauban, et ie luy ay vniens d'ou
vous deu ne s'entener l' - l'ornu de ce moy, Le laun
l'Examery Rigorosa se fait la v en d'ou Les logic
la Cene, de tout que par ce moyen il faut que i'o
c le veras que nous donnons aux professeurs s'of
faudroit que i'usse pour payer l'hoste que en don
ne vous de tout s'enouir ce qu'il faut que U us m'c
i Eau pour les lettres, s' Eau pour payer l'habit
m' 30 f pour un mois de Musique que ie dois de p
vres de Chandelles et pour la port de quelques Let

Recherche de « quelque » TE=25%

Cas d'étude avancé:

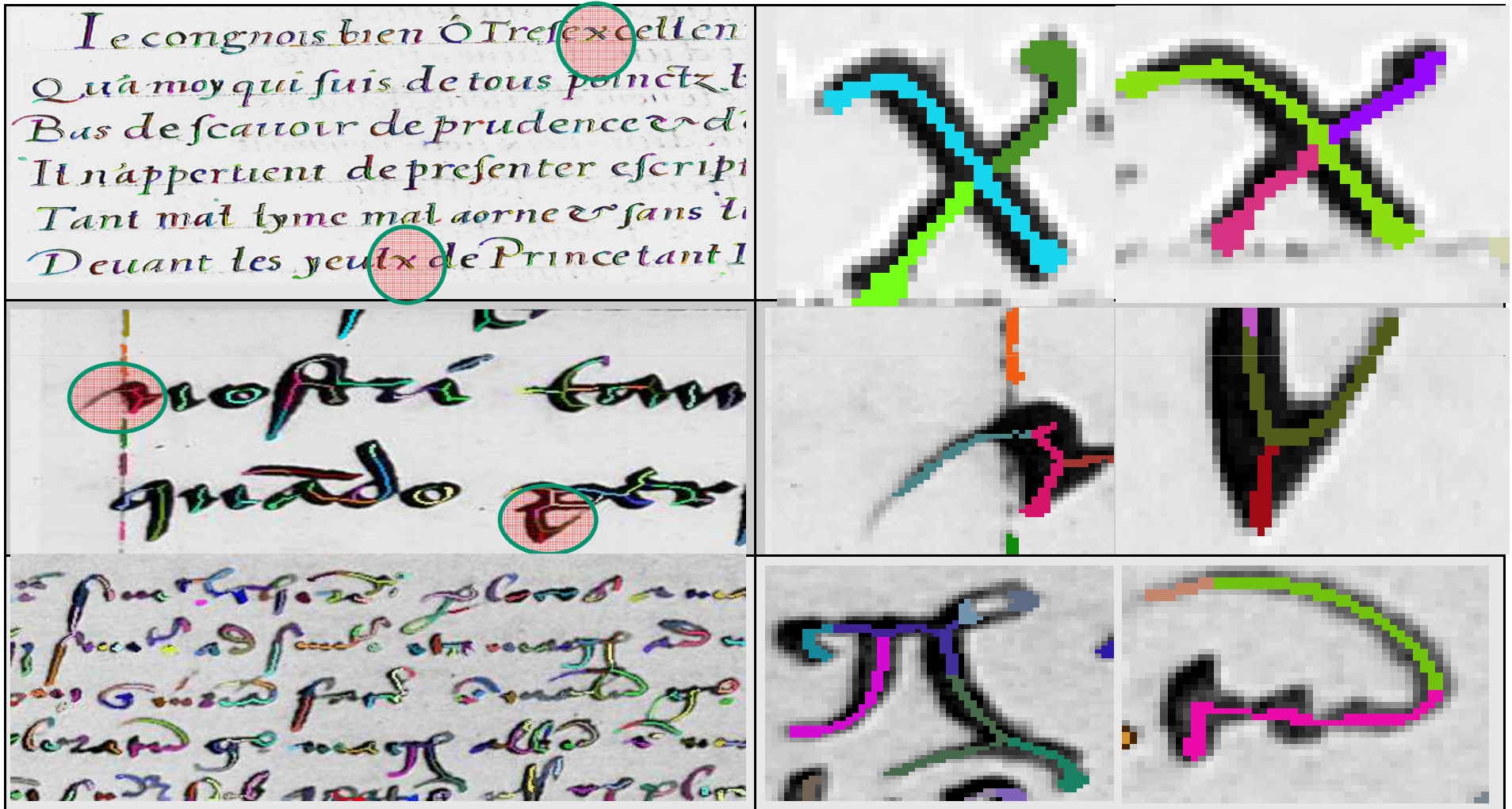
Décomposition selon l'exécution des formes

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



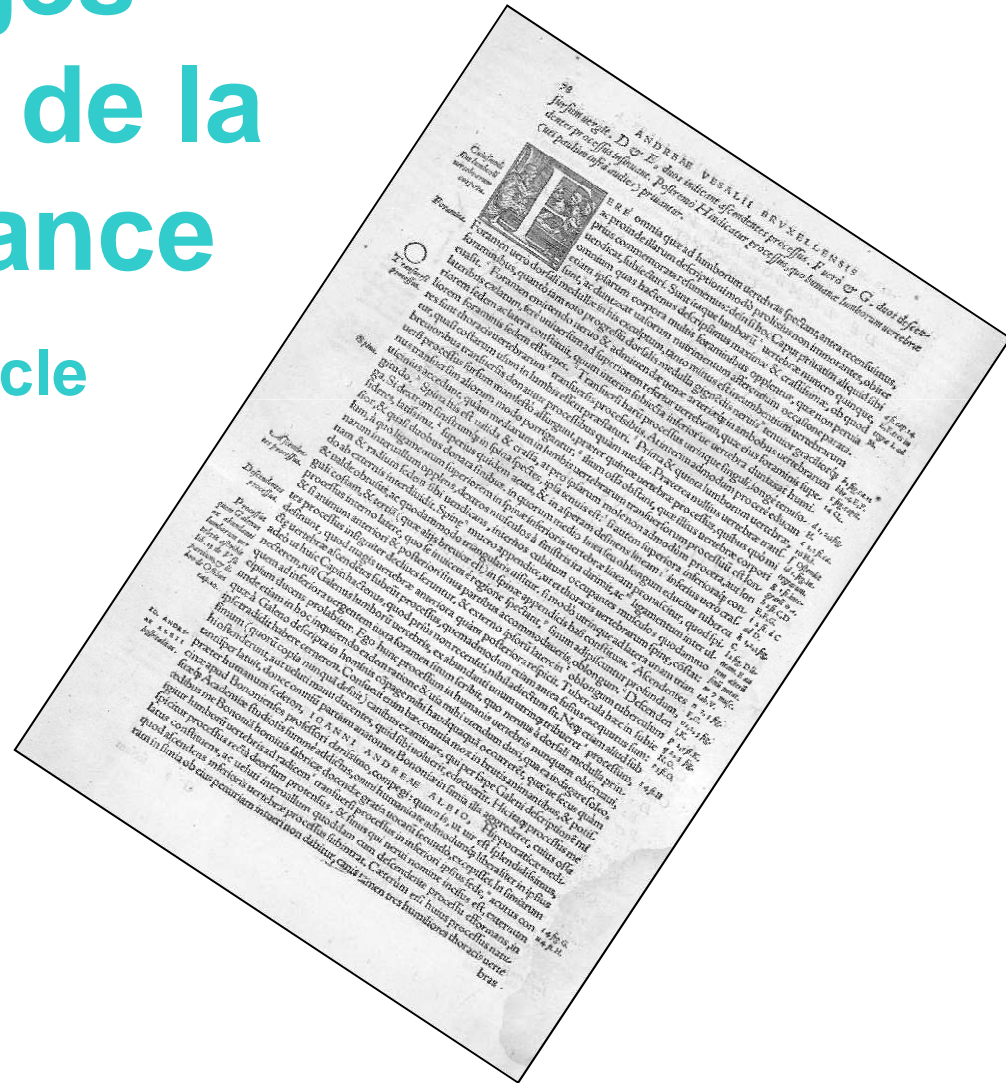
Contribution: Décomposition des manuscrits en graphèmes

Illustration: croisements et cas d'erreur

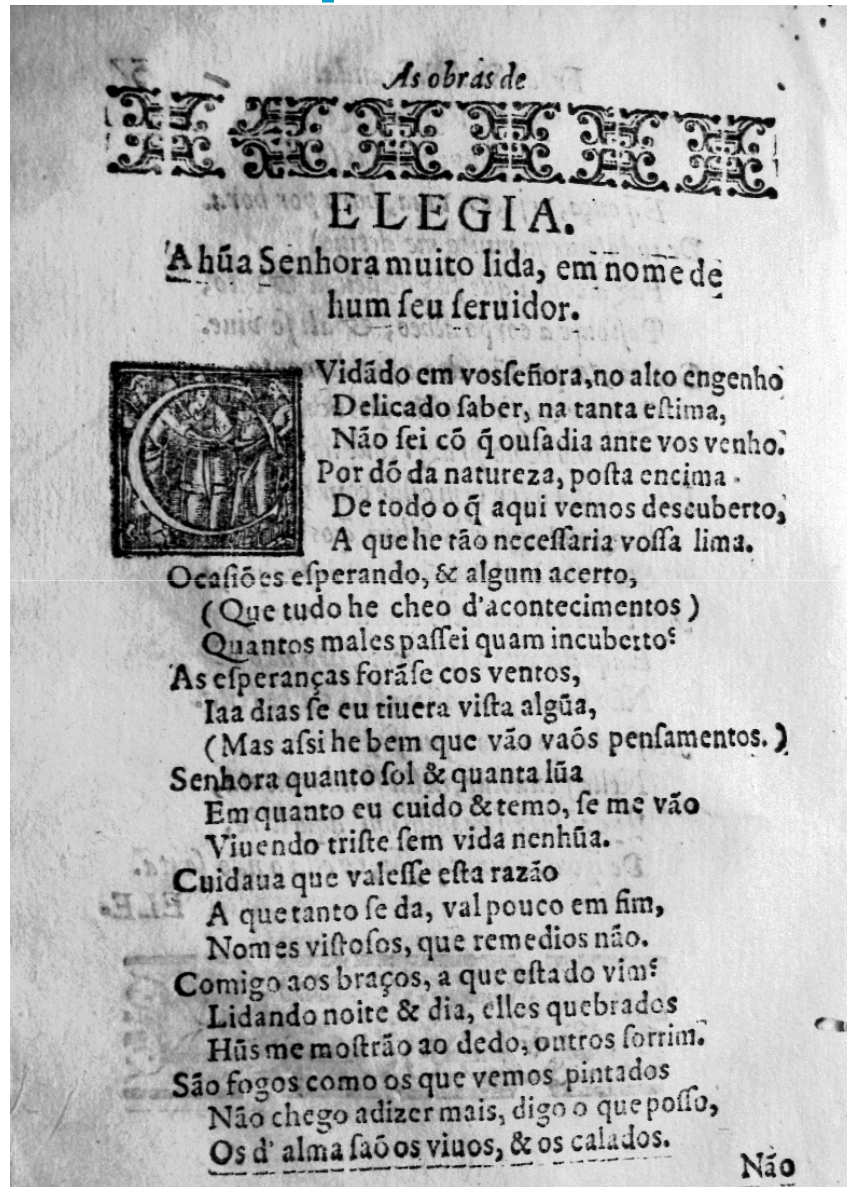


Ouvrages imprimés de la Renaissance

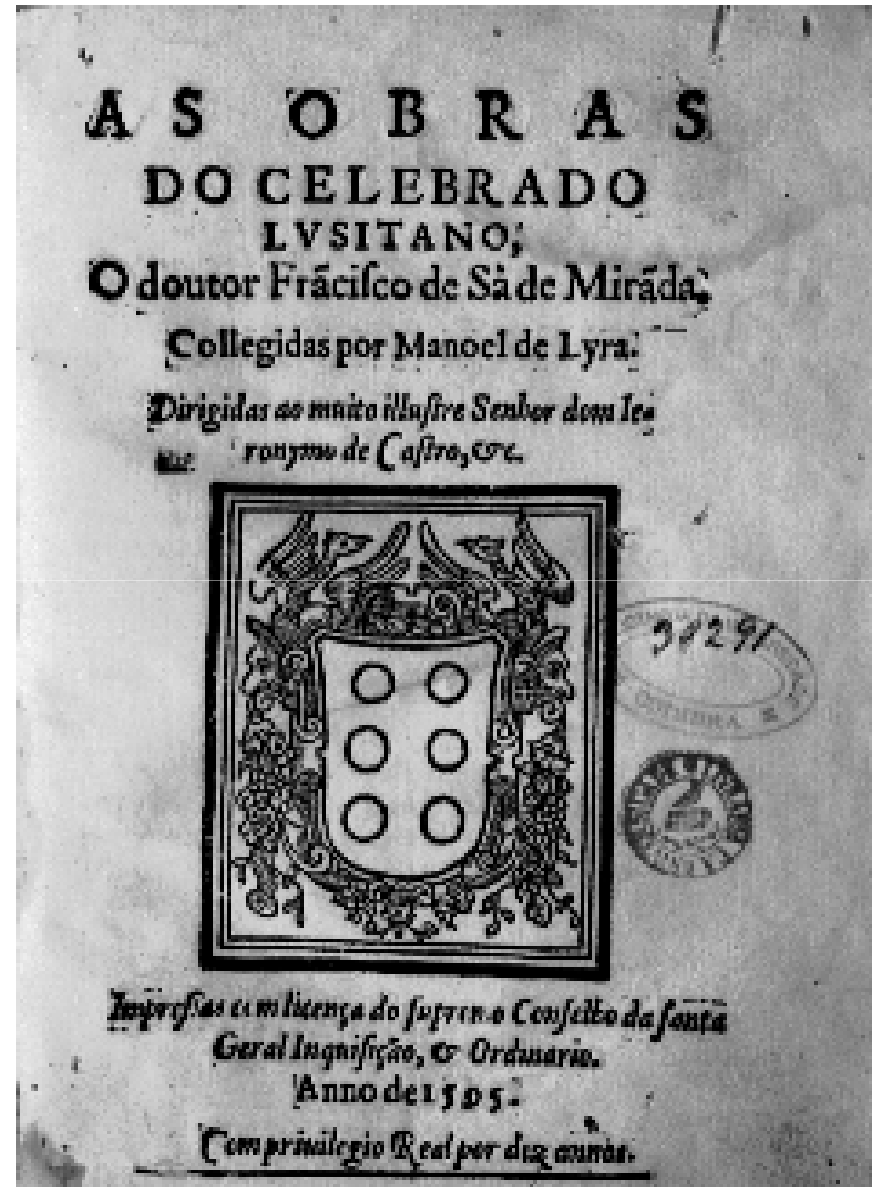
16ème siècle



Imprimés de la Renaissance XVI



Déformations géométriques



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 AccEss to BOok of the RenAissance

- Analyse des usages des utilisateurs d'ouvrages du 16^{ème} 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 (restauration, segmentation, interprétation, compression)
- Conception de postes de consultation

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



- ▶ Utiliser des méthodes **d'analyse d'images de documents** pour à la fois les **compresser** 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

DEBORA

Extraction de la structure physique et différenciation Texte/graphique

3194, 951 Zoom % 16 Ok

Chapitre second.
CHAPITRE II.

¶ Que cest qu'est compris au rondeau de dessus, mis au costé gauche.

Ay premierent monstré en ce rondeau les racines du chief du dragon en la ligne, par signes, degrez & minutes ainsi que son titre le moustre. En poursuivant des l'an 1530. jusques à 1570.

¶ Or contient ceste Sphère ou rondeau, trois Racines ou bandes.

- La premiere est, des Lunaisons qui apres le cercle des nombres des ans, a trois cercles.
 - Le premier est, des jours lunaires.
 - Le second, des heures.
 - Le troysiesme, des minutes.
- La seconde, est des demy-mouvements de la lune: ayant aussi trois cercles.
 - Le premier est, des signes.
 - Le second, des degrez.
 - Le troysiesme, des minutes.
- La troysiesme est, des argumens de la lune, si comme du mouvement de l'epicycle, ayant deux cercles.
 - Le premier est, des signes.
 - Le second, des degrez.

¶ Apres ce, quasi au milieu du rondeau y a quatre lunes differentes, selon la distance des conjunctions d'iceles: si comme de la pleine jusques au dernier quart, & du dernier quart jusques au defaut, & du defaut jusques au premier quart: & y vertas icelles avec leurs nombres.

Adition.
Adioustement.

Autant l'honneur de ce liure, n'ay voulu mesler icy: ains le separant, le nomme Adioustement, ou adioustement.


Saches que les douze Signes du Zodiaque, sont ainsi lineez, nommez, nommez & marquez comme s'ensuyt.

I 4

Classification Lettrine/bandeaux

DEBORA

171



LE TIERS LIVRE DE
PLVSIEVRS SINGVLARITEZ
ET CHOSIS 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.

O R comme l'ay de sa dict sur la fin du tiers liure, cest grand resuerie de lire ce que Mahomet a escriptes liures de son Alcoran: parquoy sachant que i'ay eu loisir d'observer beaucoup de choses sur la façon & maniere de viure des Turcs, & principalement estant de sejourner en Paphlagonie, ou ie demeury quelque espace de temps, il m'a semblé bon mettre en petit discours de Mahomet à part, tel possible que personne n'a encor mis en nostre langue, sans tout: sçou que personne s'en 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, & en mesme-ment que cest chose conuenant à la matiere que ie pretens traiter. Il n'y a pas long temps que Mahomet nasquit en vne ville de l'Arabie eurense, nommée la Meque, que l'interprete Petra ou il commença la secte des Turcs. & à ce qu'on s'escrive ce fut l'and apres l'aduenement de nostre seigneur six cents & vingt, & mourut l'an six cents quatre vingt & trois. Les Turcs ont vn liure nommé A'fez, qui contient toute la vie de Mahomet, lequel ilz tiennent & observent. 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 sous deux idoles. Il est escript que ledict Abdola, mourut auant que Mahomet nasquist: & sa mere Imina mourut deux ans apres qu'elle l'eut enfanté: &

Papla-
gonia.

La Meq.
A'fez li-
ure con-
tenant la
vie de
Maho-
met.
Pere de
Maho-
met.
Mère de
Maho-
met.

V iij

171



LE TIERS LIVRE DE
PLVSIEVRS SINGVLARITEZ
ET CHOSIS 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.

O R comme l'ay de sa dict sur la fin du tiers liure, cest grand resuerie de lire ce que Mahomet a escriptes liures de son Alcoran: parquoy sachant que i'ay eu loisir d'observer beaucoup de choses sur la façon & maniere de viure des Turcs, & principalement estant de sejourner en Paphlagonie, ou ie demeury quelque espace de temps, il m'a semblé bon mettre en petit discours de Mahomet à part, tel possible que personne n'a encor mis en nostre langue, sans tout: sçou que personne s'en 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, & en mesme-ment que cest chose conuenant à la matiere que ie pretens traiter. Il n'y a pas long temps que Mahomet nasquit en vne ville de l'Arabie eurense, nommée la Meque, que l'interprete Petra ou il commença la secte des Turcs. & à ce qu'on s'escrive ce fut l'and apres l'aduenement de nostre seigneur six cents & vingt, & mourut l'an six cents quatre vingt & trois. Les Turcs ont vn liure nommé A'fez, qui contient toute la vie de Mahomet, lequel ilz tiennent & observent. 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 sous deux idoles. Il est escript que ledict Abdola, mourut auant que Mahomet nasquist: & sa mere Imina mourut deux ans apres qu'elle l'eut enfanté: &

Papla-
gonia.

La Meq.
A'fez li-
ure con-
tenant la
vie de
Maho-
met.
Pere de
Maho-
met.
Mère de
Maho-
met.

V iij

Classification Lettrine/bandeaux



Classification illustrations

PREMIER LIVRE DES SINGVLA.

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 quasi vne creste comme à vn Vanneau, & quand il volle, va battant des elles comme vn Cigne. Il se paist aussi bien sur l'eau salée, qu'en l'eau douce. Je prouueray en autre mien ouuert, ou i ay mis le portraict des oiseaux, que cestuy est le Pelican, dont me tuis pour ceste heure à cause de Briefueté. Entre les choses singulieres de ceste isle, ay veu le serpent nommè Iaculus, moucheté de petites taches dessus le dos, ressemblantes à des petis yeux, tout ainsi que sont les taches de dessus le dos d'un Tremble, nommè en Latin Torpedo. Je le trouuy des sous vn Caprier espineux hors la ville, celle part ou le Turc auoit planté son artillerie quand il assiégea Rhodes. Les Grecs le nomment maintenant en leur vulgaire Saetta, c'est à dire Sagitta, & les Turcs Ochilanne, les anciens Acontias. Il a trois paulmes de longueur, & n'est plus gros que le petit doigt. Sa couleur est cendrée tirant sur la couleur de lait, & est totalement blanc dessous le ventre, ayant des escailles dessus le dos, & tablettes dessous le ventre à la maniere des autres. Il est noir dessus le col, & taché de deux lignes blanches, qui commencent des la teste, & suruent 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, entourées d'un cercle blanc. Je 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.

Le portraict du Iaculus, autrement dit Acontius.

Je vei aussi descharger vn brigantin dessus la rine du port, plein d'une drogue propre en medecine, appelée Storax rouge. Les Grecs la nomment maintenant Mastrocappo. Et m'a lon dit qu'il croist en l'isle. Mais pource que ceux qui font voyages par mer, ne peunent 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. Je vueil inserer par

Storax
rouge.
Mastrocappo.
pno.

OBSERVEES PAR P. BELON.

108

l'île 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 Romains en estoient seigneurs, lequel a beaucoup diminué depuis: mais il fault entendre que pour lors les Romains n'espargnoient rien à faire despense pour le rendre fertile. I'ay prins grande merueille d'auoir veu si grande quantité de Cassiers es iardins du Caire, & par

Portraict du Cassier.

Egypte, & toutesfois les auteurs anciens n'en ont fait aucune mention: car mesmemēt Theophraste qui a quasi parlé de toutes autres plantes d'Egypte, n'en fait mention. Mais il fault dire de Theophraste parlant des plantes, tout ainsi comme d'Aristote des animaux. Car comme diuerses nations obeissent aux commandemens d'Alexandre apportent diuerses especes d'Animaux à Aristote, lors qu'il en escripuoit l'histoire, aus si estoit il necessaire que par mesme moyen diuerses nations feissent rapport des plantes à Theophraste quand il les descripuoit. Et apert à son histoire qu'il ne l'a fait sans grande despense, & d'hommes qui ont esté expressément enuoyez en diuers endroits du monde, pour les observer.

Liberalité d'Alexandre. Theophraste & Aristote.

Classification illustrations

DEBORA



Fonctionnalités du poste client



- ▶ **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** (*métadatas, 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

DEBORA

DEBORA v0.1b10
Fichier Edition Rechercher Affichage Fenêtres

Structure de l'ouvrage:
 Titre: Magnificence
 Auteur: COLLECTIF
 Bibliothèque: Bibliothèque
 Lieu de publication: Lyon
 Date de publication: 1549
 Editeur: Rouillé
 Langue: Français
 Cote: Réserve 355882
 Sujet: Magnificence de la
 Pièces préliminaires:
 Page: 1
 Page: 2
 Image composite:
 Imagette:
 Imagette:
 Pages de Texte:
 Page: 3
 Image composite:
 Imagette:
 Imagette:
 Texte brut
 Page: 4
 Page: 5
 Page: 6
 Page: 7
 Page: 8
 Page: 9

Titre: Magnificence Page: 2 (14%)

Privilege.

PROVCE QUE PAR CT DE
 quant on ha imprimé & exposé en vstre plu
 rieurs Liures & Cayens de l'entrée du Roy
 de la Royne née en leur ditte Ville de
 Lyon, & qu'il y eut mesme incornu de meu
 sioniers, & etrangers, tant en plusieurs
 endroits ce qui ha fait, & d'autres
 pour servir l'ordre desdictes entrees, &
 abusant par ce moyen les letens de sables
 & d'oranges au grand defaut de la
 dite Ville, & de ce qui ont fait leur
 despit, & auindement estrangeres, sur
 ches de ne rra mot, & au Indice & sans que l'imprimeur y
 entre les deux, ce qui ont en, & par ce moyen rruille de les imprime
 ments de sables entrees estoient appendez. A cesades & autres confu
 sions, il est defendu à tous Libraires & Imprimeurs de ne imprimer & expo
 ser en ce lieu de sables, & de pres, & de amende arbitraire. Fait à la
 requeste & sous la faict de l'aport du Guillemme Rouille, marchand Libraire de
 Lyon, & seoy luy en les Concllillers & Echevins de la Ville de Lyon, & le per
 mis aussi à Rouille, Imprimeur ou faire imprimer les sables entrees, qu'il ha
 fait voir & corriges par gens de ce signiffians, & qui ont ordonné la dite
 entree, & ont fait valloir les sables entrees. En ces choses il gade de
 www.deborasur.com

Saisie d'un commentaire
 Objet: Lettrine de l'imprimeur de Lyon
 Type: Informations diverses

souvent ornée en début de paragraphe chapitre. Elle est gravée sur bois le plus souvent, sinon sur cuivre et on la grave en ce cas lettres grises. Une composition emblématique ou **héraldique** adoptée par un imprimeur ou un libraire comme : Ces signes graphiques se trouvent généralement sur la page de titre, au verso du dernier feuillet de l'ouvrage. Parfois, la marque du libraire se trouve sur la page de titre.

Rechercher sur 1 ouvrages soit 94 Pages et 95348 caractères
 Tous Rechercher Fermer
 Texte: ville

Image> Magnificence (1) Page n°2
 Image> Magnificence (1) Page n°2
Image> Magnificence (1) Page n°3
 Image> Magnificence (1) Page n°3
 Image> Magnificence (1) Page n°3
 Image> Magnificence (1) Page n°4
 Image> Magnificence (1) Page n°4

Page 63 Page 64
 Page 68 Page 69

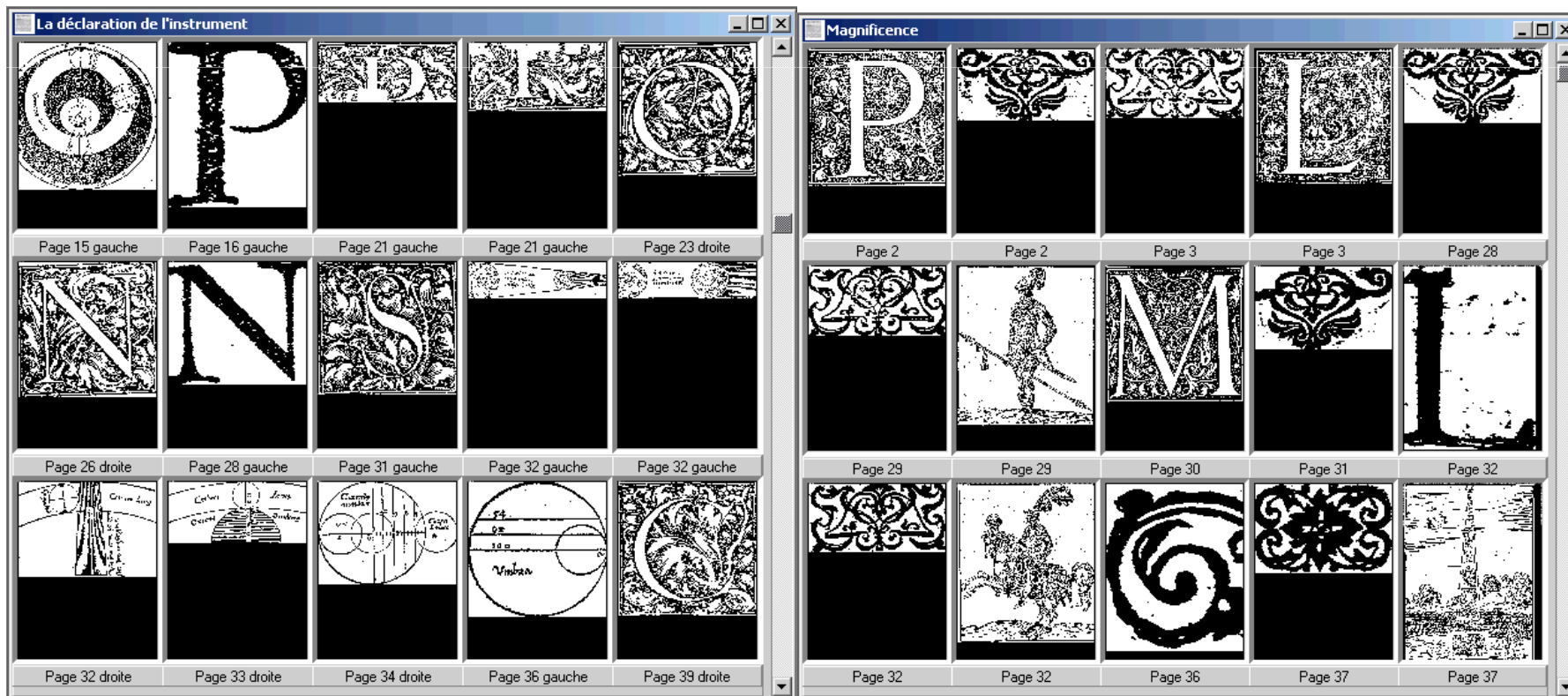
lemmatisé Texte moderne
 ur le Lieutenant du Roy, & Messieurs
 chevins de la Ville pour se preparer à
 our Parquoy Messieurs de la Ville, ne
 y à leur antique generosité Romaine, co
 nelle, se resolurent unanimement d'est

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

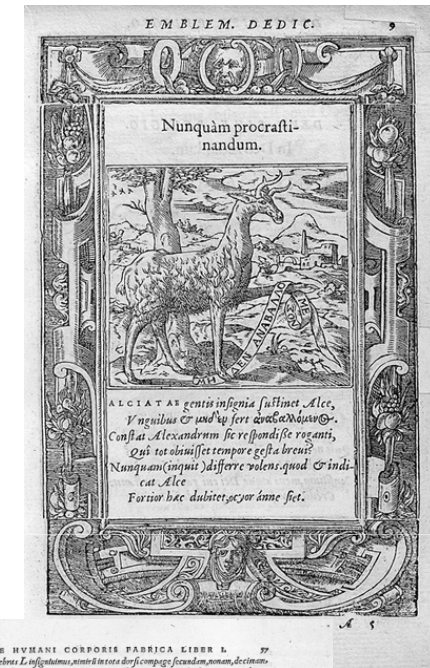
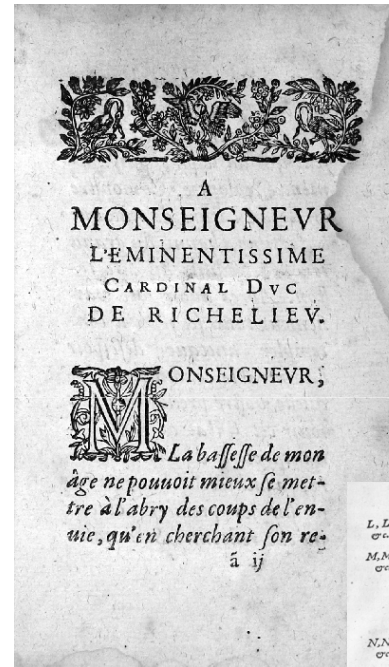
MADONNE

Que peut-on indexer ?

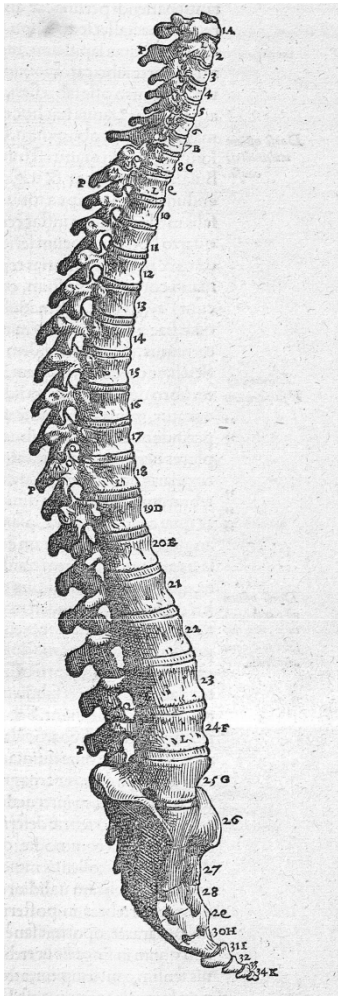
- structure de l'ouvrage
- Séparation texte/image
- Construction d'index d'images

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



Gravure



Index de lignes

N° de page

O, O O c.	M, M O c.
P, P O c.	L, L O c.

57

Notes

In triū subse-
quentiū ca-
figuris id fo-
ramen ob-
uium est.

Naturæ in
dorsi creatio-
ne industria.
Vt id susti-
neat.
Vt dorsalis
medullæ uia ef-
ficiatur, & in-
terim mobile
sit.

Paragraphe fonte normal

ERVM parens Natura homini dorsum instar carinæ cuiusdam, funda-
mentiq; machinata est. Dorsi enim operecti ambulare, & erecti consiste-
re ualemus. Quamquam neq; illa in hoc dorsum homini duntaxat dedit,
sed perinde atq; aliâs in unius membri cōstructione, eo ad uarios simul
usus abuti consuevit, ita nec hîc quoque minus industriam ipsius osten-
dit. Primum enim omnibus uertebriis ad suorum corporum posteriorum
regionem, * foramen exculpit, idoneam dorsali medullæ per ipsas
cœcensuræ uiam præparans. Secundò, non ex incomposito simpliciq; ossè uniuersum confit-
tuit

Lettrine



Titre

DE HVMANI CORPORIS FABRICA LIBER I.



- **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

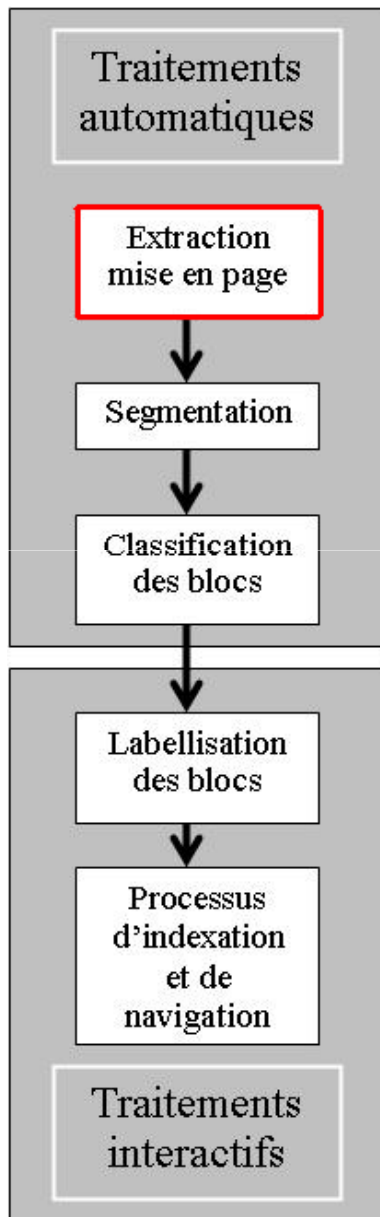
- Données accessibles?

- Pas d'ontologie

- Utilisateur non spécialiste du traitement des images et de l'ARD

Première étape: Extraction de la mise en page

MADONNE



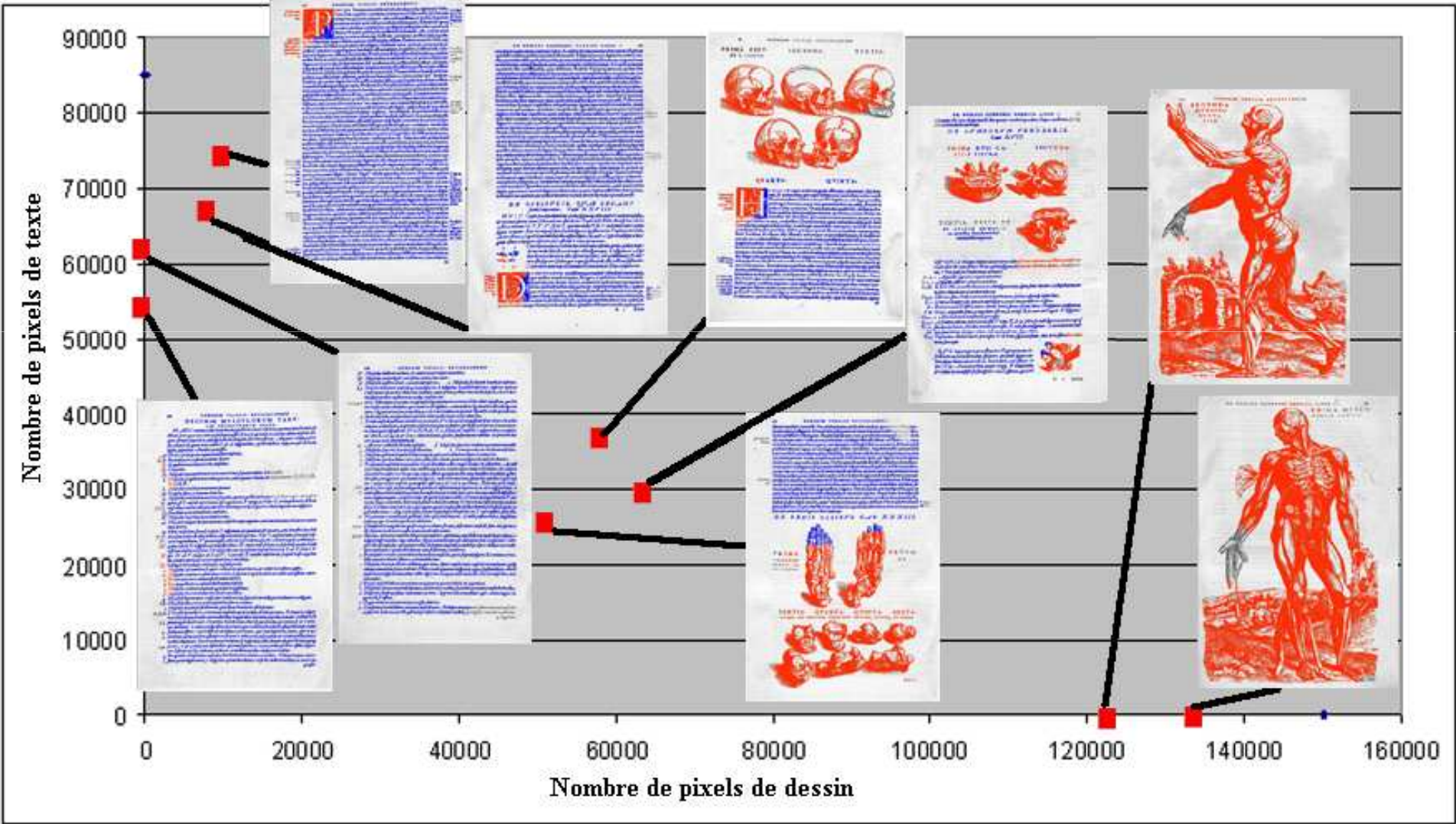
Pourquoi?

- 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)

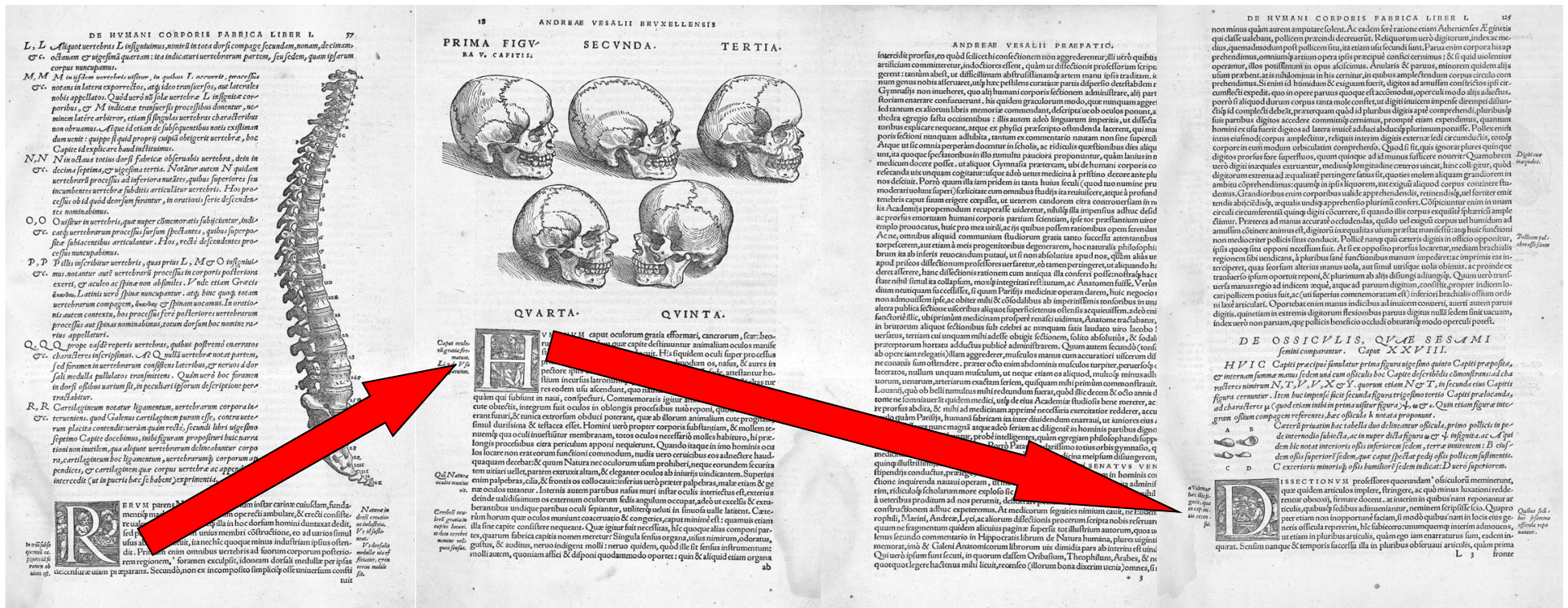
Résultats



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)



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

