Les Images de Documents

Numérisation, Dématérialisation, Analyse, Exploitation

Véronique EGLIN

LIRIS, INSA Lyon

Janvier, 2011

Objectifs du cours

- Comprendre ce qu'est le document numérique:
  - Pourquoi numériser ?
  - A quoi servent les images de documents?
  - Comment les numériser ?
  - Et après que faire avec ?
  - … Ensuite les choses se compliquent: il faut les organiser, les indexer, les exploiter, les analyser…

- Quelques applications de la gestion documentaire dans l'industrie (documents d'entreprises), dans les bibliothèques (documents du patrimoine).

Plan du cours

- Jeudi 6 janvier
  - Présentation générale: les enjeux de la dématérialisation et de la GED
  - Les bases du prétraitement des images (partie 1)
- Vendredi 7 janvier
  - Les bases du (pré-)traitement des images et de l'indexation par le contenu des documents multimédias (partie 2)
- Jeudi 13 janvier
  - Les bases de la Reconnaissance de formes et de la classification
- Vendredi 14 janvier
  - L'analyse de structures: physique et logique
  - Des applications cibles: WS, CBIR, TAO, Tri, e-Edition…

Contexte

- A l’origine: des besoins d’accéder à des ensembles de documents
  - Archives
  - Documents juridiques, administratifs, …
  - Documents du patrimoine

- Généralement, on dispose de collections hétérogènes
  - Différentes langues
  - Différentes structures/mises en page
  - Différentes provenances
  - Différentes modalités d’acquisition et qualités…

- Rarement, on dispose de métadonnées descriptives des contenues, homogènes et pertinentes

Du document “papier” au document électronique?

- Objectif: “sans papier”
  - Transfert efficace
  - Organisation facilitée
  - Reproduction “à la demande”…

- Accès à une plus grande variété de contenus
  - Echange de données (email, attachments, diffusion…)
  - Lorsque la “copie” rend un service identique à l’original

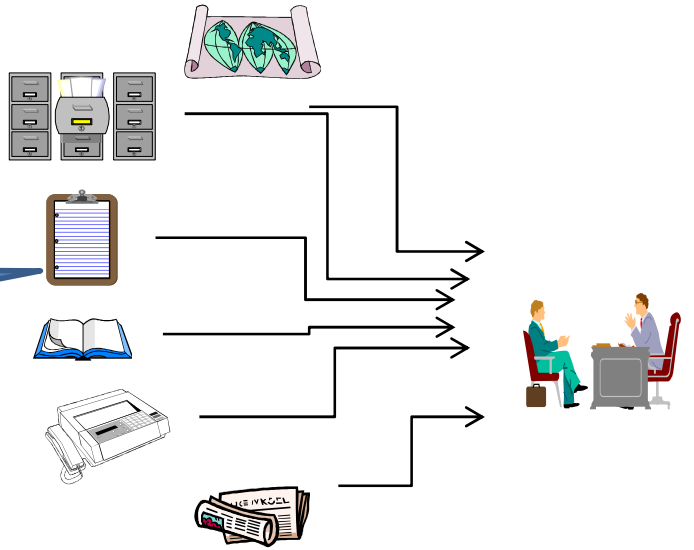
- Sauvegarde et préservation

- Edition électronique, édition critique

Quels usages?

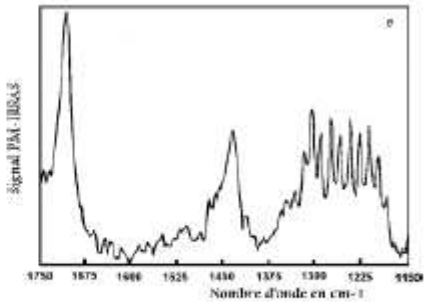
- Internet
- Email Attachments
- Online Proceedings
- Electronic Fax
- Mass Digitization Repositories…
- Sauvegarde du patrimoine

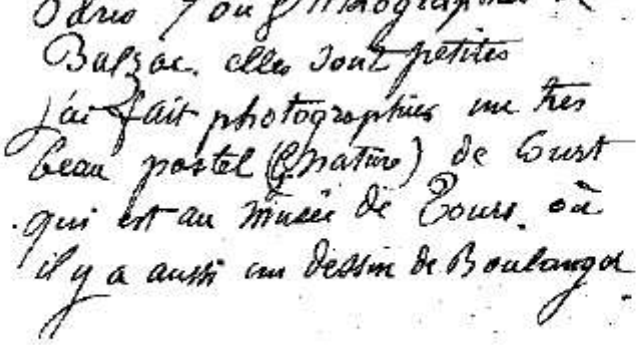
Une nécessaire organisation GED des documents



Quels usages?

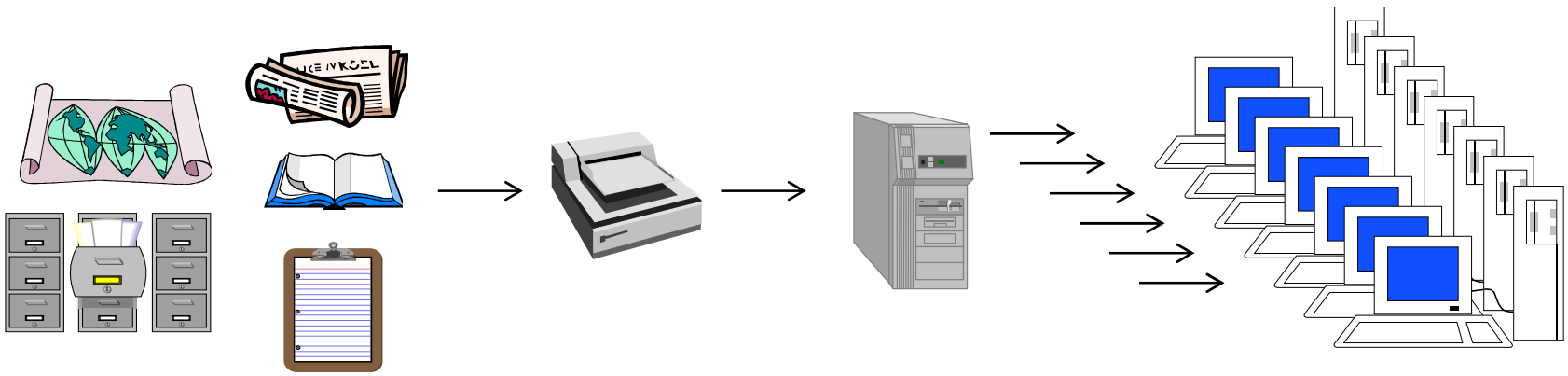
- Médical : détection de pathologie à partir d'ECG/EEG, imagerie : aide à la décision
- Militaire : détection de cibles
- Sécurité automobile : détection d'obstacles
- Traitement du signal : reconnaissance de la parole, de l'écriture
- etc.




Odrio 7 ou 8 lithographies de Balzac. elles sont petites J'ai fait photographier un très beau pastel (grandeur) de Gust. qui est au Musée de Tours. où il y a aussi un dessin de Boulanger

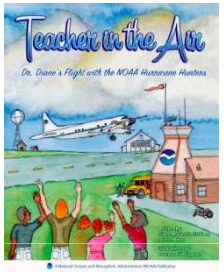
Bases d’images de documents

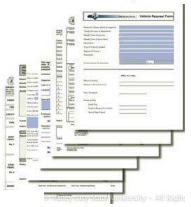
- Collection d’images numérisées
- Organisées pour rendre possible différentes actions tq: indexation, recherche, édition, dissémination, interprétation….

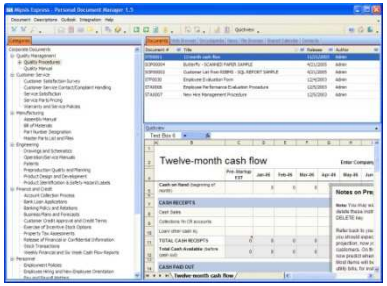


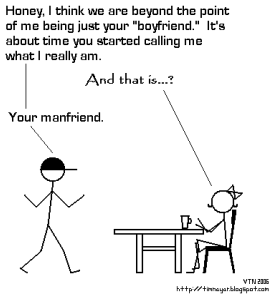

Stadt- und Universitätsbibliothek
A a
The Washington Post
After Zarqawi, No Clear Path In Weary Iraq
Difficult Questions Surround Legacy of Insurgent Leader
The Young Apprentice
How U.S. Forces Found Iraq's Most-Wanted Man
Kaine Delays Execution Of Inmate for 6 Months
FDA Approves Vaccine That Should Prevent Most Cervical Cancers
Code ACI
Expéditeur
Destinataire
2005104A
Boulogne
COURRIER INTERNE
Programming success

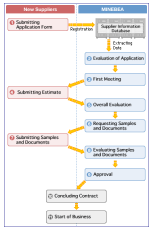
Autres types de “Documents”

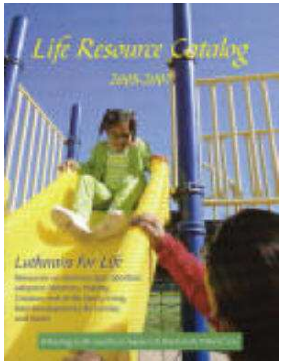












Définitions générales

Numérisation : *Conversion des documents sous la forme d'une image numérique à l'aide d'un scanner ; c'est la première étape de la chaîne de dématérialisation.*

OCR : *Reconnaissance Optique des Caractères, étape intermédiaire vers la dématérialisation; elle produit un texte ASCII au kilomètre, sans aucune structure, enrichi seulement des styles et des polices de caractères.*

ARD : *La Reconnaissance Automatique des Documents produit un document électronique enrichi de la structure logique et physique*

Dématérialisation ou Rétroconversion : *La dématérialisation consiste à convertir les documents papier dans un format logique pivot les rendant de nouveau rééditables et donc réutilisables. Elle va au-delà de la simple conversion ASCII du texte par un OCR, et propose une réelle réhabilitation de la structure interne du document, la plus proche possible de celle qu'il avait à l 'origine. On parle dans ce cas de rétroconversion [Belaid 97]*

Chaîne de traitement de l'ARD

1/ Numérisation

2/ Pré-traitement

prétraitements →
- Binarisation
- Débruitage
- Correction d'inclinaison
- etc.

3/ Extraction d'attributs et indexation par le contenu

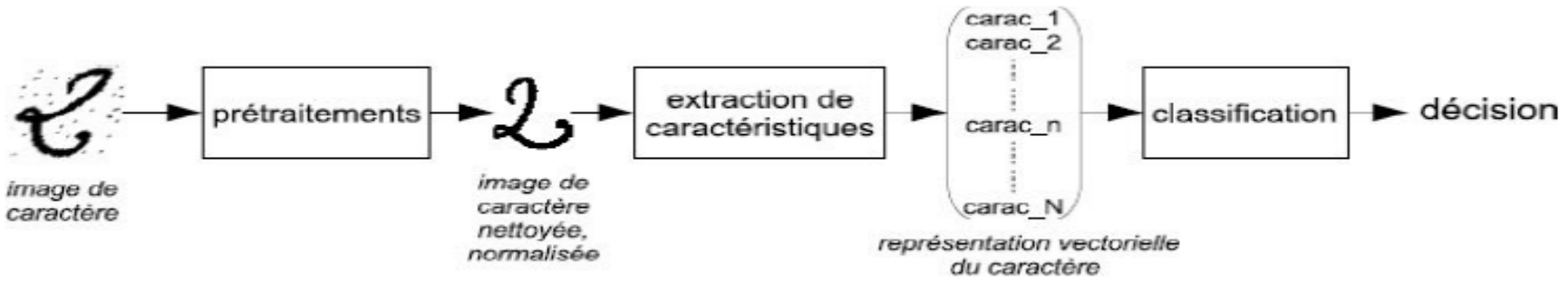
Extraction de caractéristiques →
- Augmenter la pertinence de la représentation
- Réduction de la quantité d'info.

4/ Classification, apprentissage et reconnaissance

Apprentissage automatique →

Idée = Apprendre la diversité et la variabilité des formes grâce en soumettant à la machine de (très) nombreux exemples

"machine learning"



Les tâches usuelles en ARD

- Analyse et reconnaissance des structures physique/logique
- Classification de zones d'intérêt (ROI)
- Recherche d'information (WS, Shape Spotting, CBIR…)
- Indexation et Reconnaissance des contenus des ROIs (OCR, transcription, identification de scripteurs, authentification….)
- Tri et applications industrielles

Partie 1

LA NUMERISATION

La révolution du numérique

La numérisation ne fait qu'une copie NUMERIQUE de l'ouvrage.

C'est un procédé coûteux mais avantageux à long terme !

Contrairement au procédé argentique, les images numériques peuvent être dupliquées INDEFINIMENT SANS PERTE

Sous forme numérique, les ouvrages peuvent être échangés, stockés, manipulés, traités et recopiés avec des coûts très réduits.

Les images numériques peuvent être reproduites sur de nombreux types d'affichage : papier (Book-on-demand), écrans (PDA, eBooks, poste PC…)

La numérisation est un projet INTERDISCIPLINAIRE

Nécessite des expertises (et donc des experts) sur

- Le choix des collections
- Les usages
- Le choix des métadonnées
- Les choix administratifs (accès par le web, poste de consultation sur place…)
- Les technologies informatiques (Matériel d'acquisition, logiciels d'analyse des images, formats de fichiers, les bases de données, l'interface du poste client, le langage de requêtes…)

Qui déterminent les fonctionnalités de la bibliothèque digitale

Quelles images pour quels usages ?

Pour être interprétées par une machine

 Prévoir l'évolution des logiciels de reconnaissance

Qualité des images pour la reconnaissance *comment éviter les pertes d'information ? Comment réduire l'accumulation des bruits ?*

Pour être lues sur un écran

 Confort visuels *quels type d'affichage ? Quels seront nos comportements demain (ex: e-book) ? Quels traitements pour améliorer la qualité des textes sur écran ?*

ATTENTION : une image bien lisible par l'homme n'est pas nécessairement bien interprétable par une machine et vice et versa !

Quelles images pour quels usages ?

Pour être lues sur le Web

 Consultation à distance Quelle compression adaptée au débit du réseau ? Quelle interface de navigation ? comment éviter la copie ?

Pour le stockage

 Conservation sur des supports compacts quels supports ? pérennité des supports et lecteurs ?

 Problème du volume des données quelle compression ?

 Qualité des images pour des traitements ultérieurs Quelle résolution ? Nombre de couleurs ? Quels traitements sont réversibles ?

Pour la reproduction (Sous d'autres formats)

 Nécessite une totale rétroconversion

Quelles images pour quels usages ?

Pour l'impression (livres rares à la demande, réimpression de titres épuisés, impression à domicile d'ouvrages achetés sur Internet)

- *Copyrights ? Marquage des images ?*
- *Résolution minimale requise pour l'impression ?*
- *Quelle restauration pour améliorer la qualité d'impression ?*

Pour l'enrichissement (Outils de recherche, navigation transversale par thème, auteur, date…, mise en commun des critiques, annotations, des différentes interprétations ou traductions d'un même ouvrage)

- *Rétro-conversion* *Quel niveau de rétro-conversion est possible pour un document donné ? Quel niveau d'automatisation ?*
- *Indexation* *Méta-Données ? Création automatique de liens ?*
- *Format de données* *Quel format est le plus adapté ?*

Numérisation et usages

Les choix techniques dépendent des usages que l'on fera de la bibliothèque numérique

Ces usages sont **multiples** suivant les époques et les sciences qui s'y rapportent :

- Philologie (science de l'interprétation des documents)
- Épigraphe (science des inscriptions)
- Paléographe (science des écritures anciennes)
- Codicologie (science des livres et de leurs supports)
- Linguistique (grammaire, lexicologie …)
- Histoire des textes
- …

Comment Numériser ?
Comment prétraiter les images ?

Acquisition, restauration
Traitements, compression
1ères étapes

Qu’est ce qu’une image numérique ?

10	27	33	29
27	34	33	54
54	47	89	60
25	35	43	9

Une **image numérique** est une représentation **discrète** des **rayonnements visibles**.

Image = tableau[x,y] de pixels y=1..nb_lignes
x=1..nb_colonnes

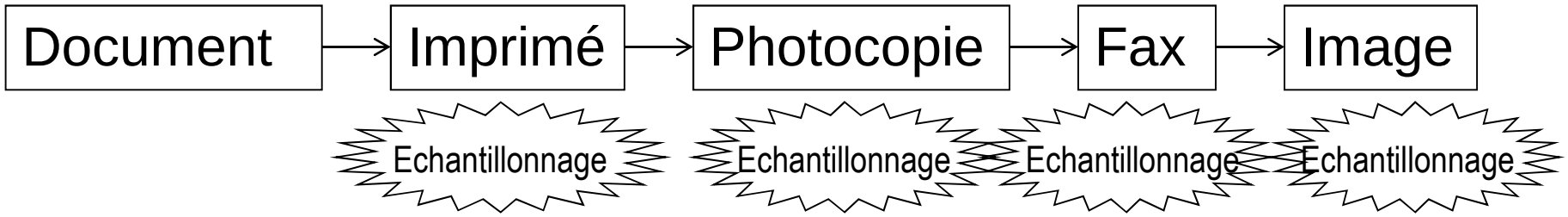
Chaque indice (X,Y) correspond à une information élémentaire de l ’image nommée **pixel**.

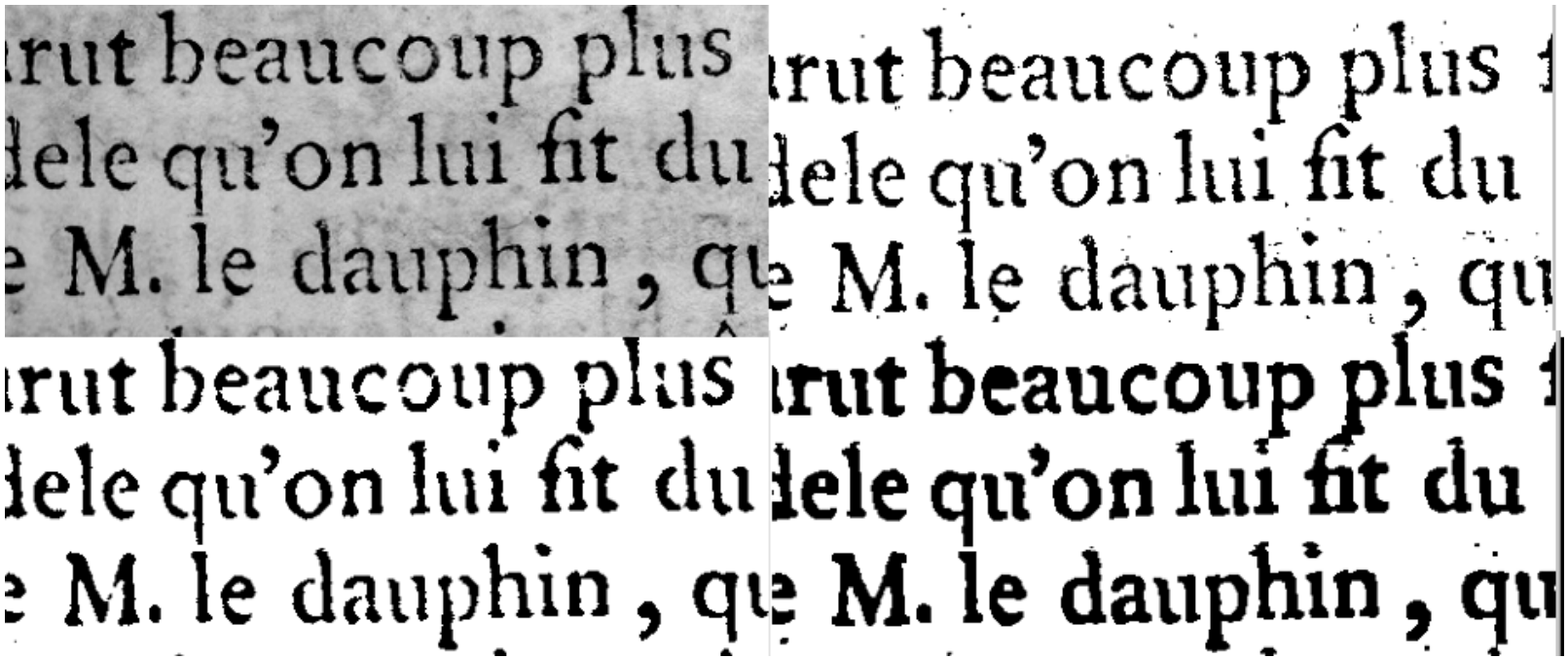
La valeur de chaque pixel mesure le rayonnement quantifié sur cette surface élémentaire par un capteur. Celle-ci est quantifiée sur M valeurs.

Remarque : le concept de pixel est utilisé pour représenter indifféremment sa valeur ou sa position.

Numérisation des documents

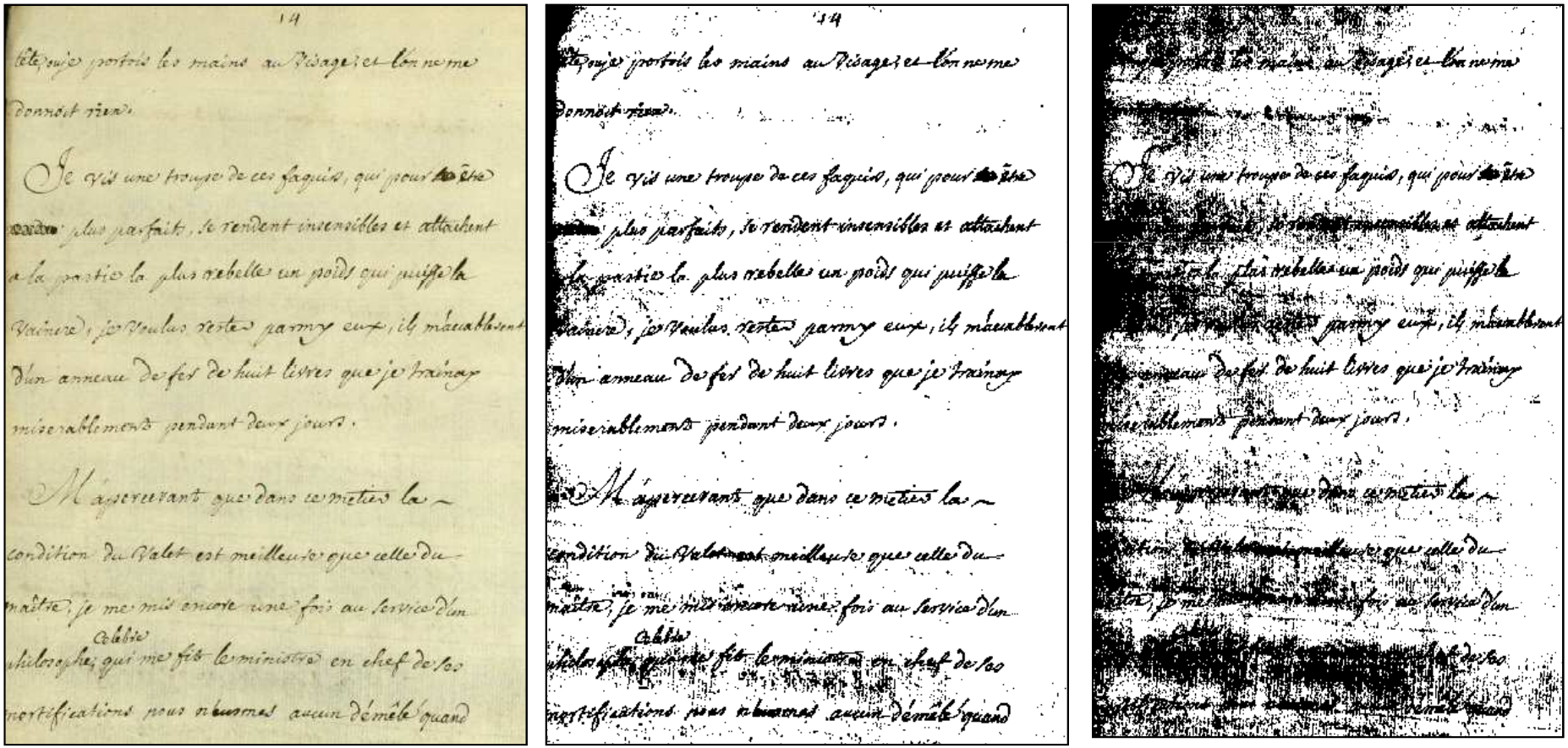
Une image numérique de document est une représentation plus ou moins fidèle suivant la résolution du scanner, le nombre de couleurs utilisé et le nombre de transformations qui précèdent.





Numérisation des documents

Image 24 bits **Photocopie** **Fax**

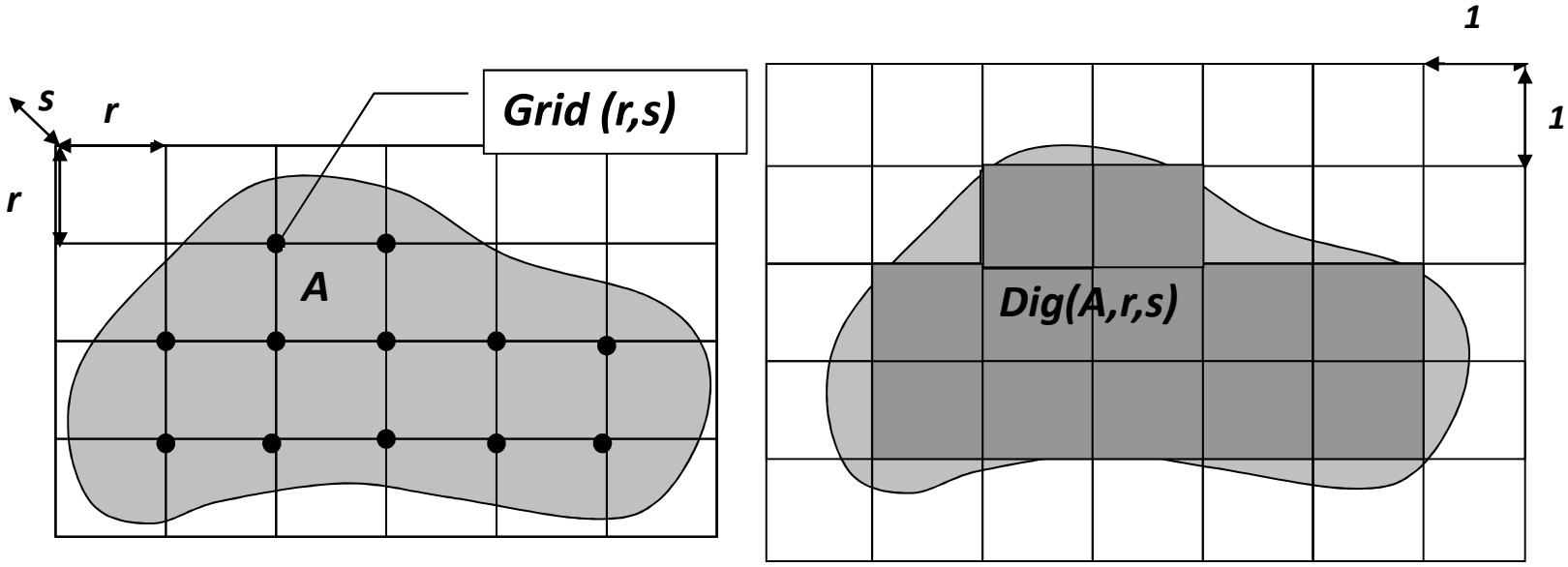


(Extrait collection Montesquieu, Histoire Véritable, 1750)

Géométrie discrète : du continu au discret

La numérisation est le résultat d'une discrétisation d'une forme continue ***A*** par une grille d'échantillonnage de taille ***r*** et qui se déplace avec un décalage ***s*** avec $|s|<r$

$$Dig\ (A,r,s\ ) = A \cap Grid(r,s)$$

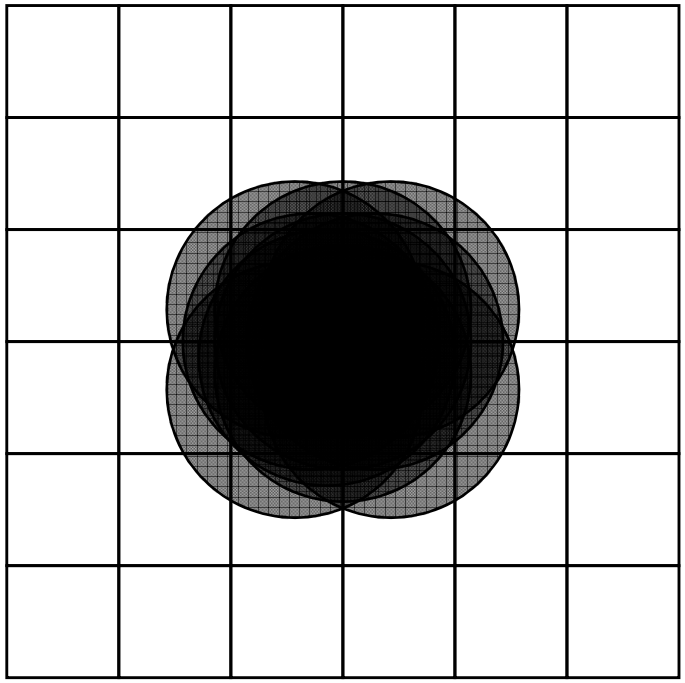


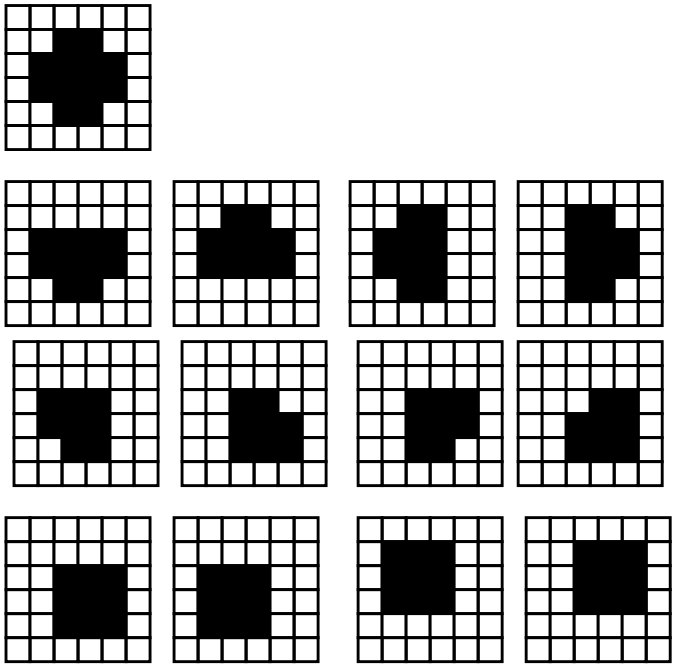
La grille d'échantillonnage correspond au capteur de la caméra

- La résolution ***r*** est la taille d'une cellule photosensible
- Le décalage ***s*** provient du placement imprévisible de la caméra

La théorie de la numérisation (1)

Combien de formes binaires obtient-on en déplaçant la grille de numérisation ?

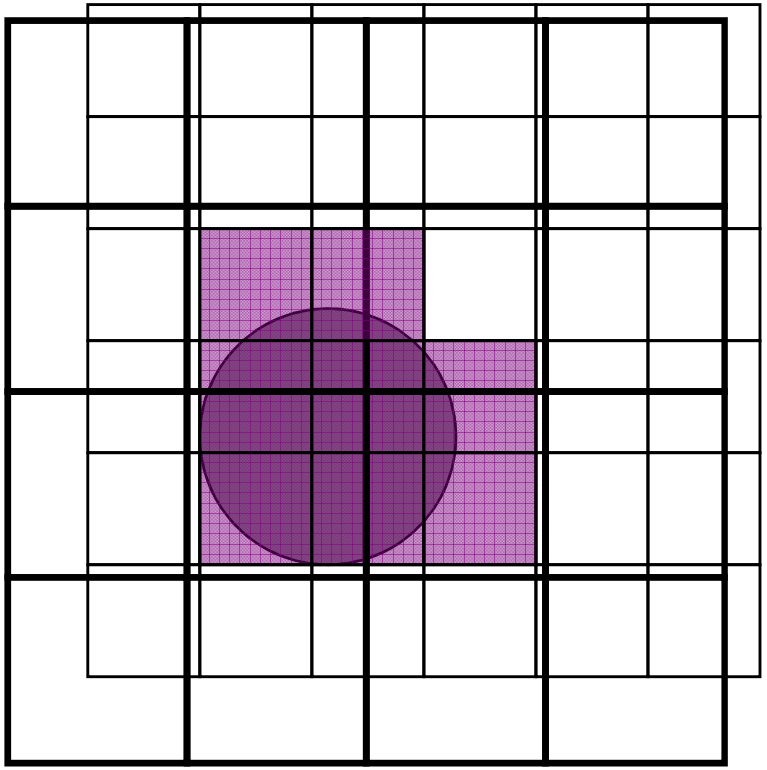


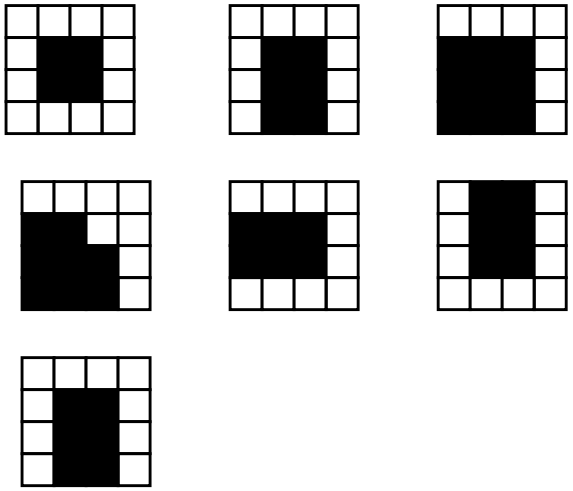


La théorie de la numérisation (2)

Pour une forme numérisée avec la grille précédente, combien de nouvelles formes trouve-t- on avec une seconde grille de numérisation qui a un pas d'échantillonnage différent ?

(exemple : scanner un document imprimé ou photocopier un fax…)





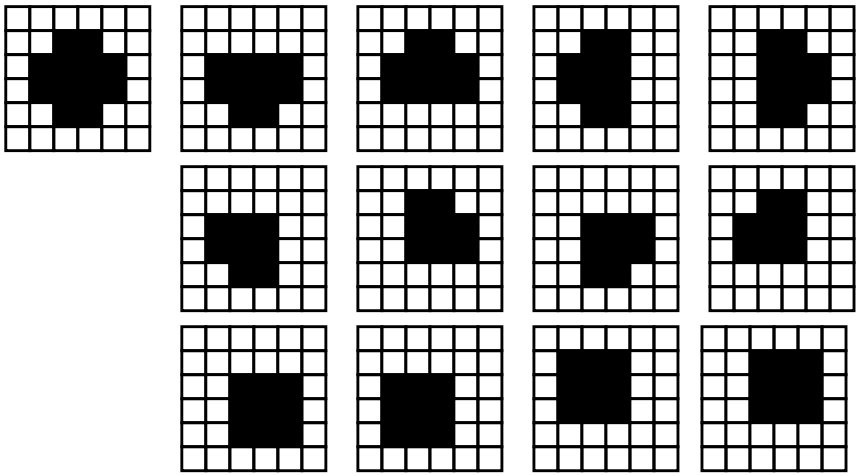
Pour chaque forme obtenue précédemment, on doit multiplier par le nombre de nouvelles formes trouvées avec la seconde grille de numérisation !

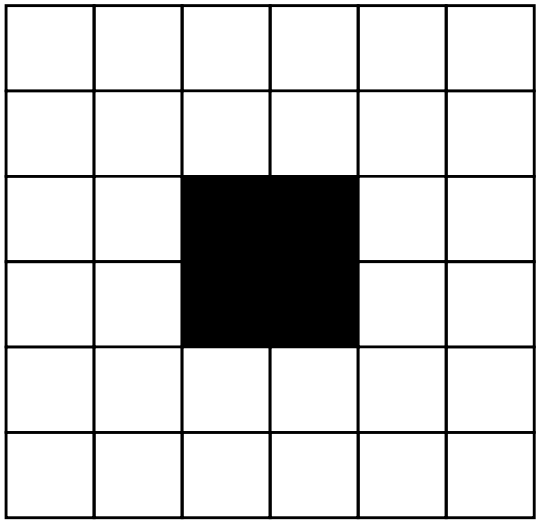
La théorie de la numérisation (3)

Le nombre de combinaisons de formes possibles augmente en fonction :

- de la complexité des formes,
- du pas d'échantillonnage de la grille
- du nombre de grilles superposées

Remarque : Certains pixels apparaissent toujours dans toutes les grilles !





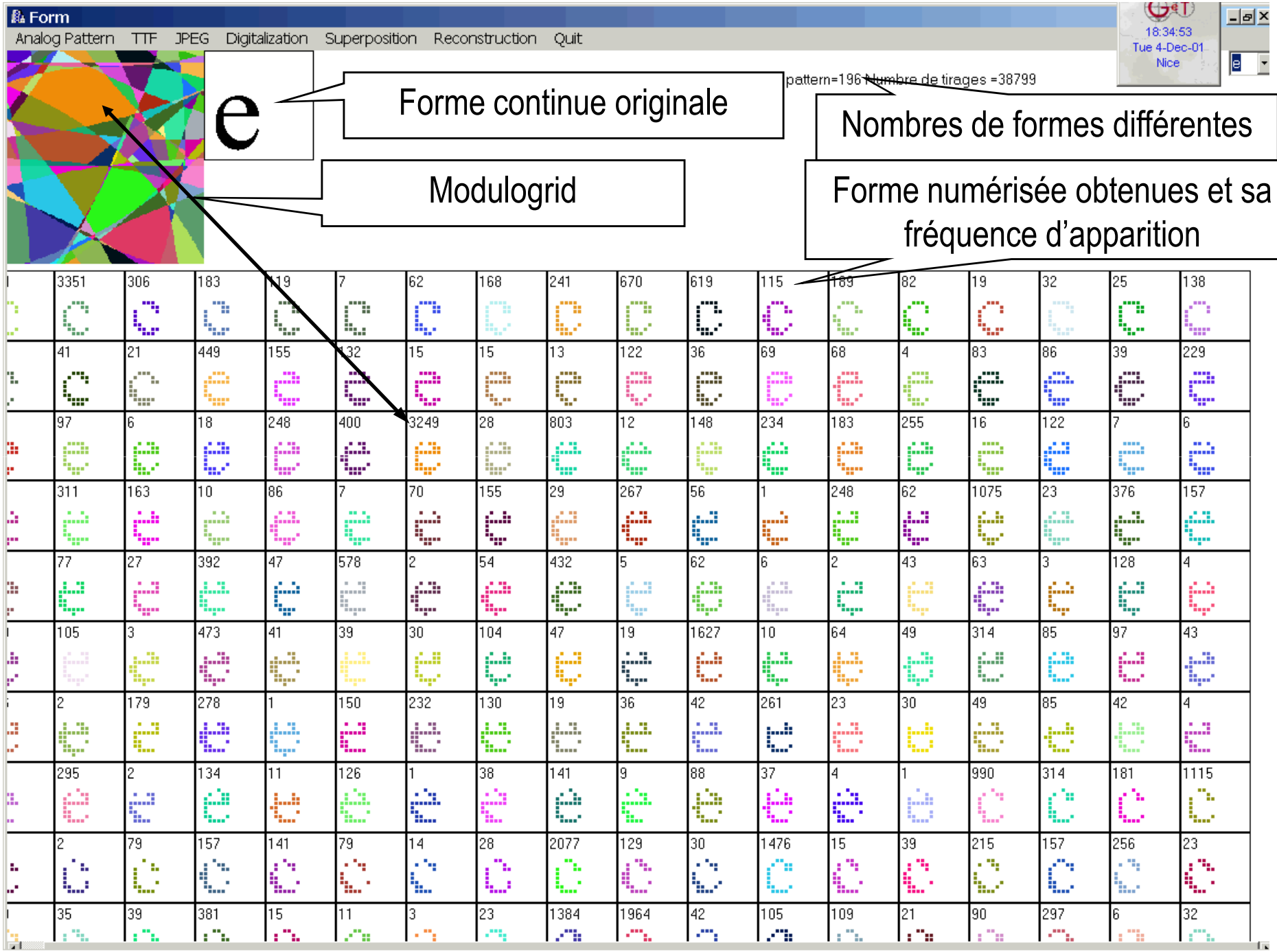
Pixels toujours présents

La théorie de la numérisation (4)

La numérisation ne conserve pas la topologie des formes :

- Le ‘e’ de police ‘Times’ dans une grille 10x10 produit
  - 196 formes différentes
  - 50% font apparaître une rupture de tracé ou un remplissage des boucles.
  - 20% ressemblent plus aux caractères ‘c’ et ‘o’ qu’au caractère ‘e’
- Fréquences d’apparition de chaque forme ?

Le Modulogramme permet à la fois de connaître le nombre de formes et la fréquence d’apparition de chacune d’entre elles


Form
Analog Pattern
TTF
JPEG
Digitalization
Superposition
Reconstruction
Quit
18:34:53
Tue 4-Dec-01
Nice
pattern=196 Numbre de tirages =38799
e
Forme continue originale
Nombres de formes différentes
Modulogrid
Forme numérisée obtenues et sa fréquence d'apparition

Images couleurs

Information Maximale (16 millions de valeurs possibles/pixels)

Pour un usage avancé et complet (codicologie, étude des miniatures, encres, décorations ; restauration&segmentation…)

Images couleurs (2)

- L'œil humain ne distingue qu'une centaine de milliers de nuances de couleurs différentes

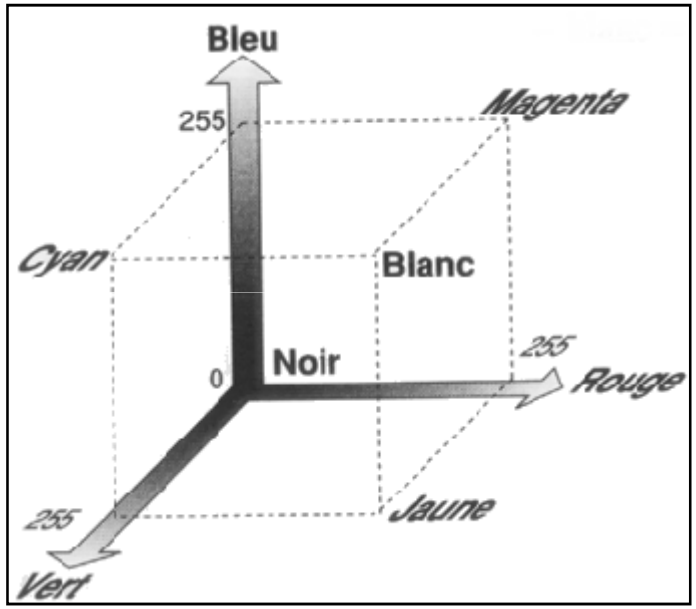
- On parle d'« intensité lumineuse » et de « chromaticité »

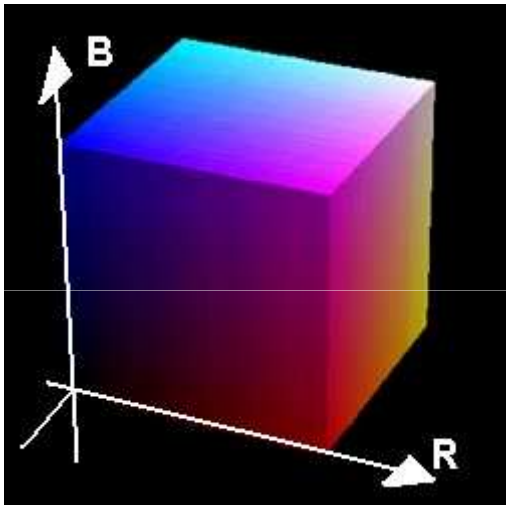
- Un pixel est codé par trois valeurs numériques
  - La signification de ces valeurs dépend du type de codage choisi
  - Le plus courant (stockage, photo et visu. écran) : RVB (ou RGB)
  - Quelques autres :
    CMJ (ou CMY), TSL (ou HSL en vidéo), YUV, YIQ, Lab, XYZ …
  - Exemple : Vraie Couleur RVB (True Color)
    chaque valeur code une composante couleur sur 8 bits

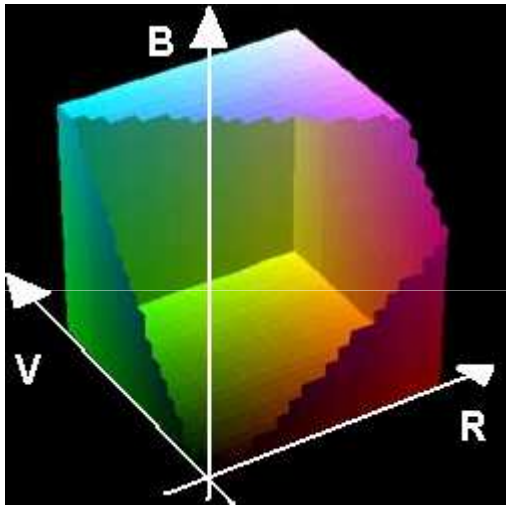
    donc $2^8 \times 2^8 \times 2^8 = 16\,777\,216$ couleurs possibles

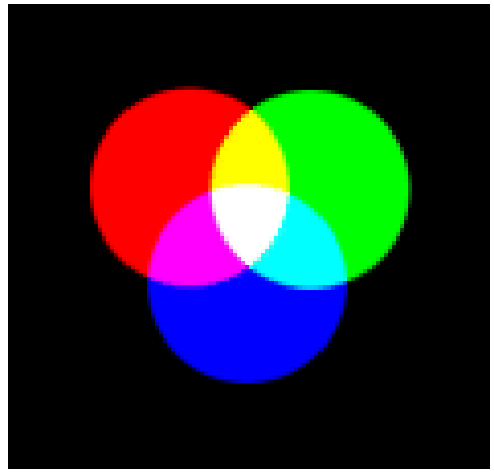
Images couleurs (3)

- L ’espace couleur **RVB (Rouge Vert Bleu)**





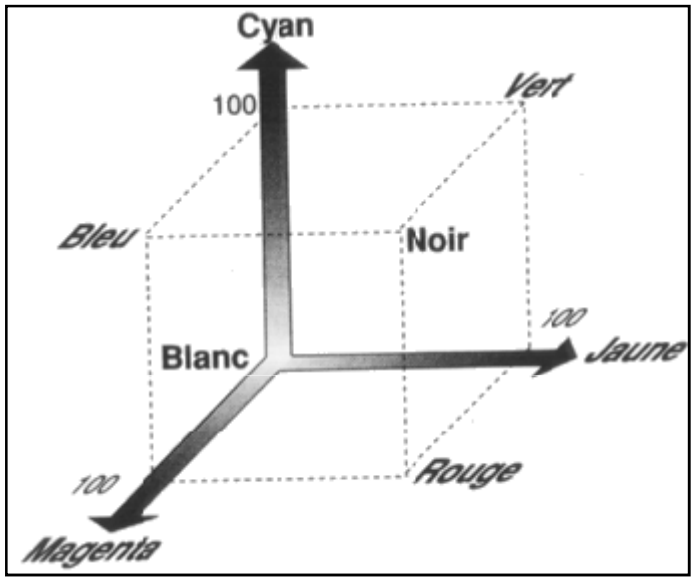


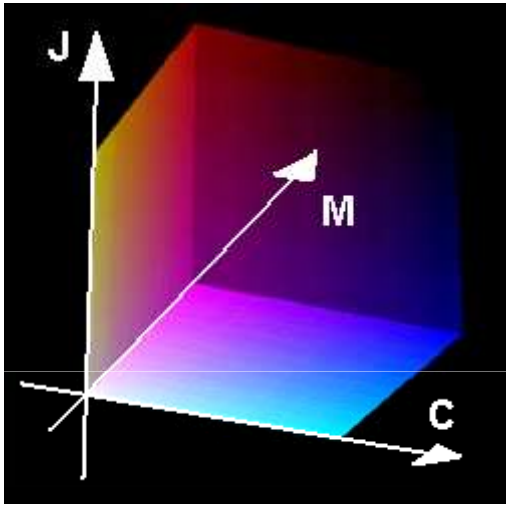


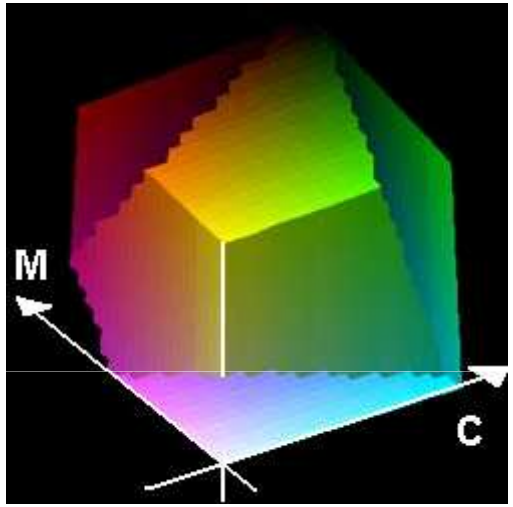
– Les couleurs sont contenues dans un cube
– La restitution des images sur écran utilise cette représentation RVB (synthèse additive)

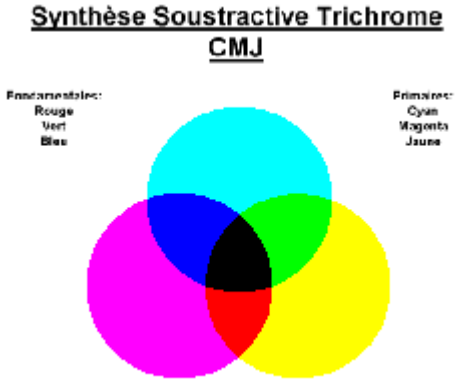
Images couleurs (4)

- L ’espace couleur **CMJ (Cyan Magenta Jaune)**





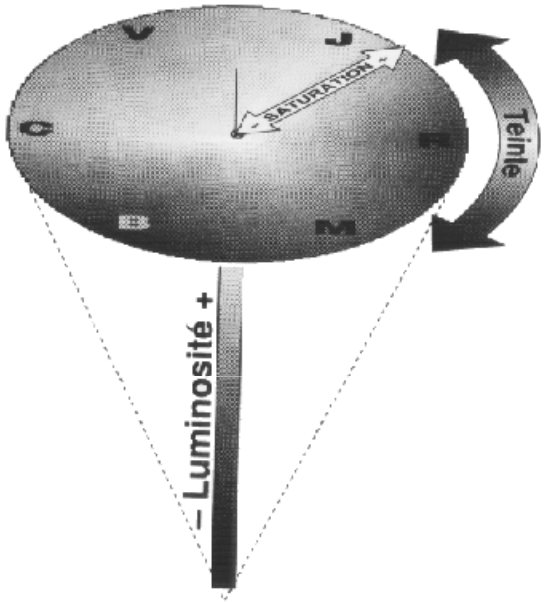


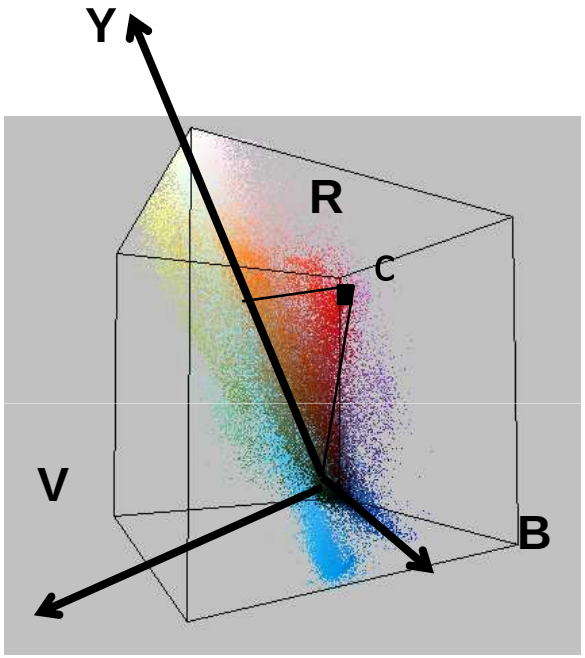


– Les couleurs sont contenues dans un cube
– Principe utilisé en imprimerie (par mélange d’encre)
– La restitution des images sur papier utilise cette représentation CMJ (synthèse soustractive)

Images couleurs (5)

espace couleur TSL (Teinte Saturation Luminance)





c=(R,G,B)

Y= l'axe luminance

« La couleur des peintres »

► **Luminance =** projection(c/Y) « quantité de noir »

► **Teinte =** angle(c,Y) « longueur d'onde de la couleur »

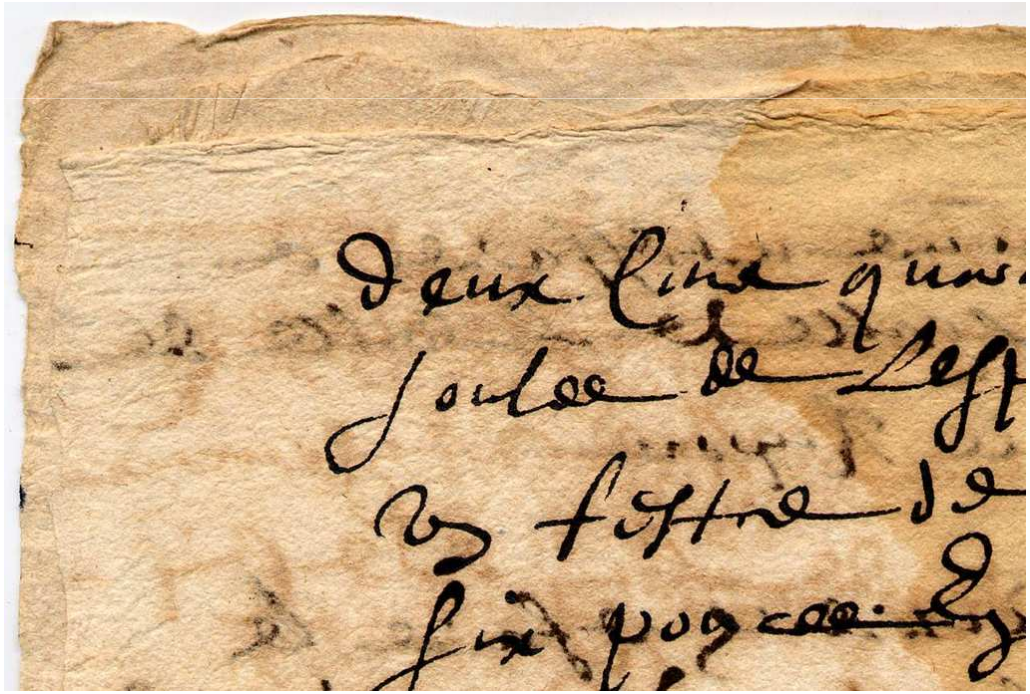
► **Saturation =** distance(c,Y) « quantité de blanc », une couleur à 100% est dite pure, ne contient pas de blanc

Numérisation directe
en niveaux de gris ou en couleur

Numérisation directe toujours meilleure que celle des microfilms

L’information niveaux de gris ou couleur est utile pour la segmentation

Suivant les usages , l’information couleur est indispensable



Exemple de la qualité d’image par une numérisation directe

La numérisation en Niveaux de gris (1)

Toujours plus d’information dans une image faible résolution en niveaux de gris que dans une image haute résolution en binaire

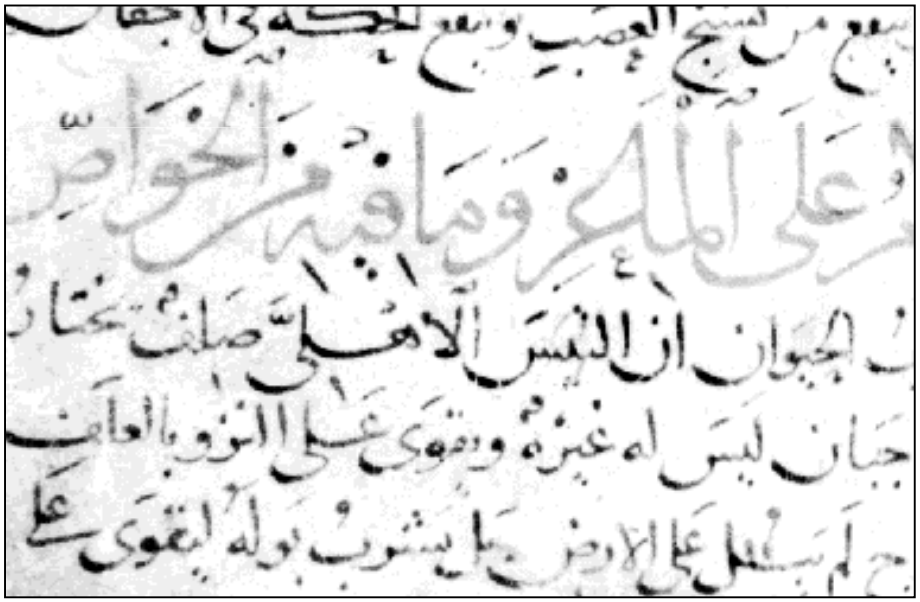


Image 16 niveaux de gris

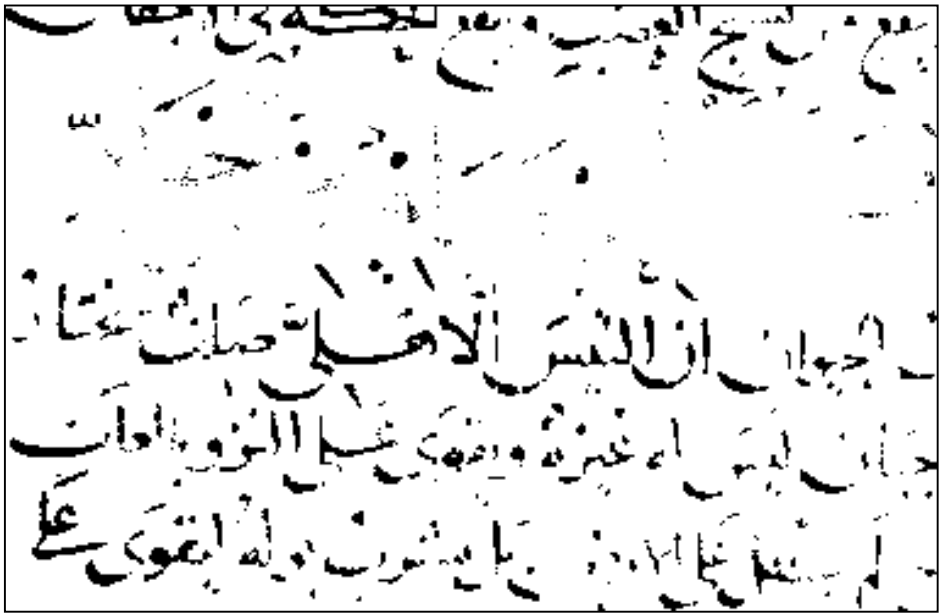


Image binaire

Images niveaux de gris (2)

information moyenne (256 valeurs possibles / pixel)

Pour un usage multiple (codicologie, étude des textes, restauration et segmentation limités des images …)

Images niveaux de gris (3)

Représentation 3D d’une image à niveaux de gris

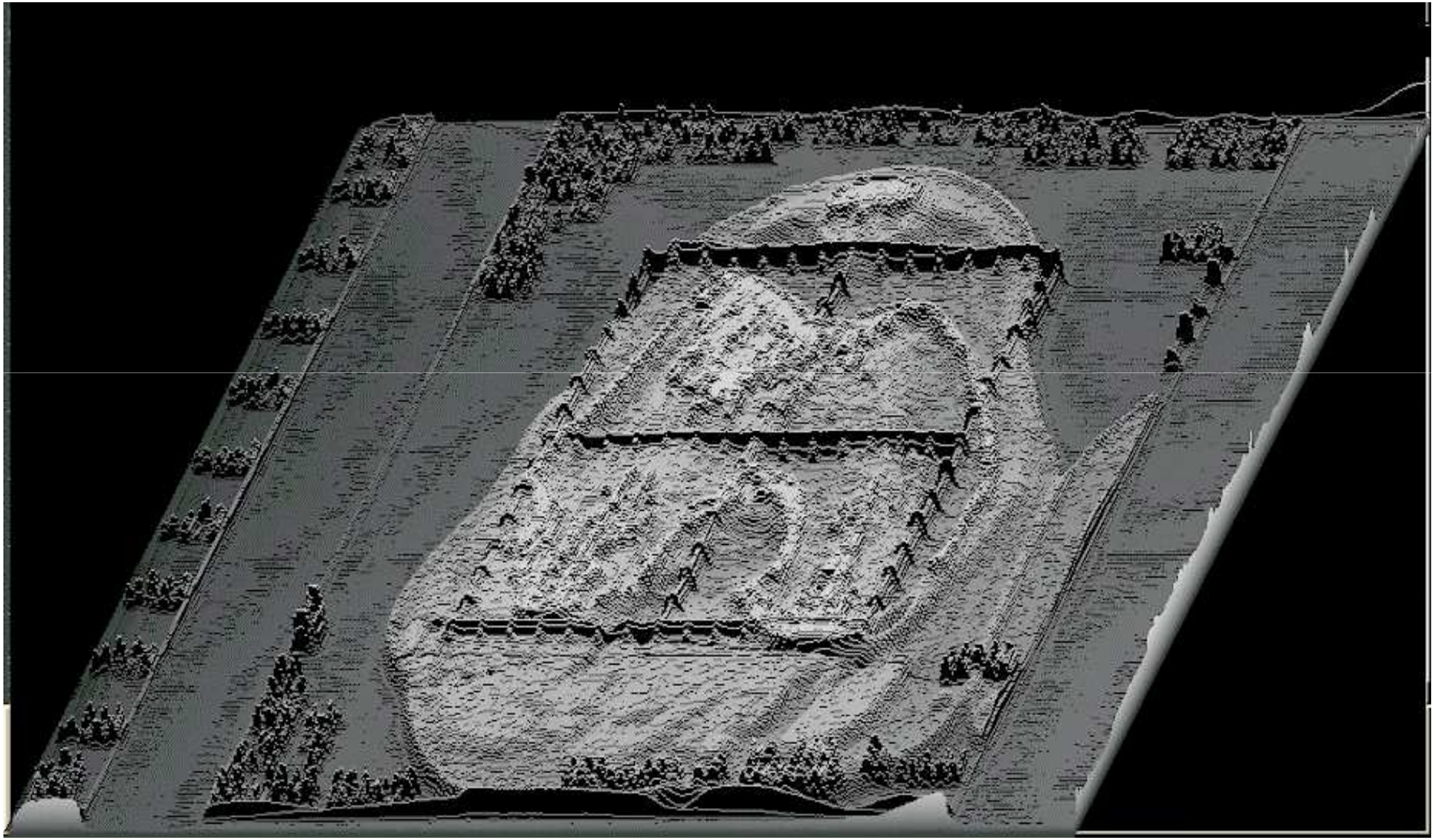
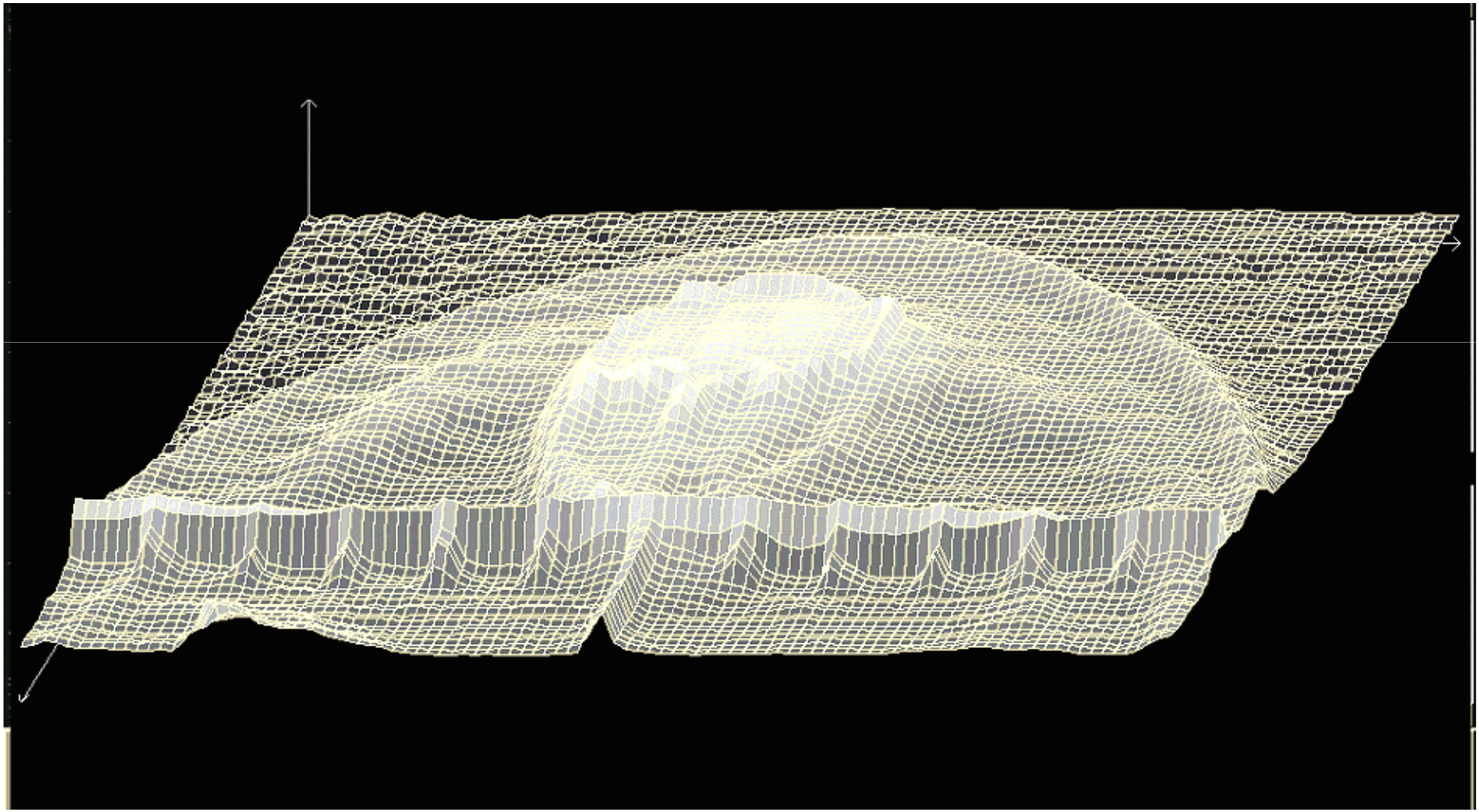


Image numérique vue représentée par une surface

Images niveaux de gris (4)

Mais c’est une surface discrète



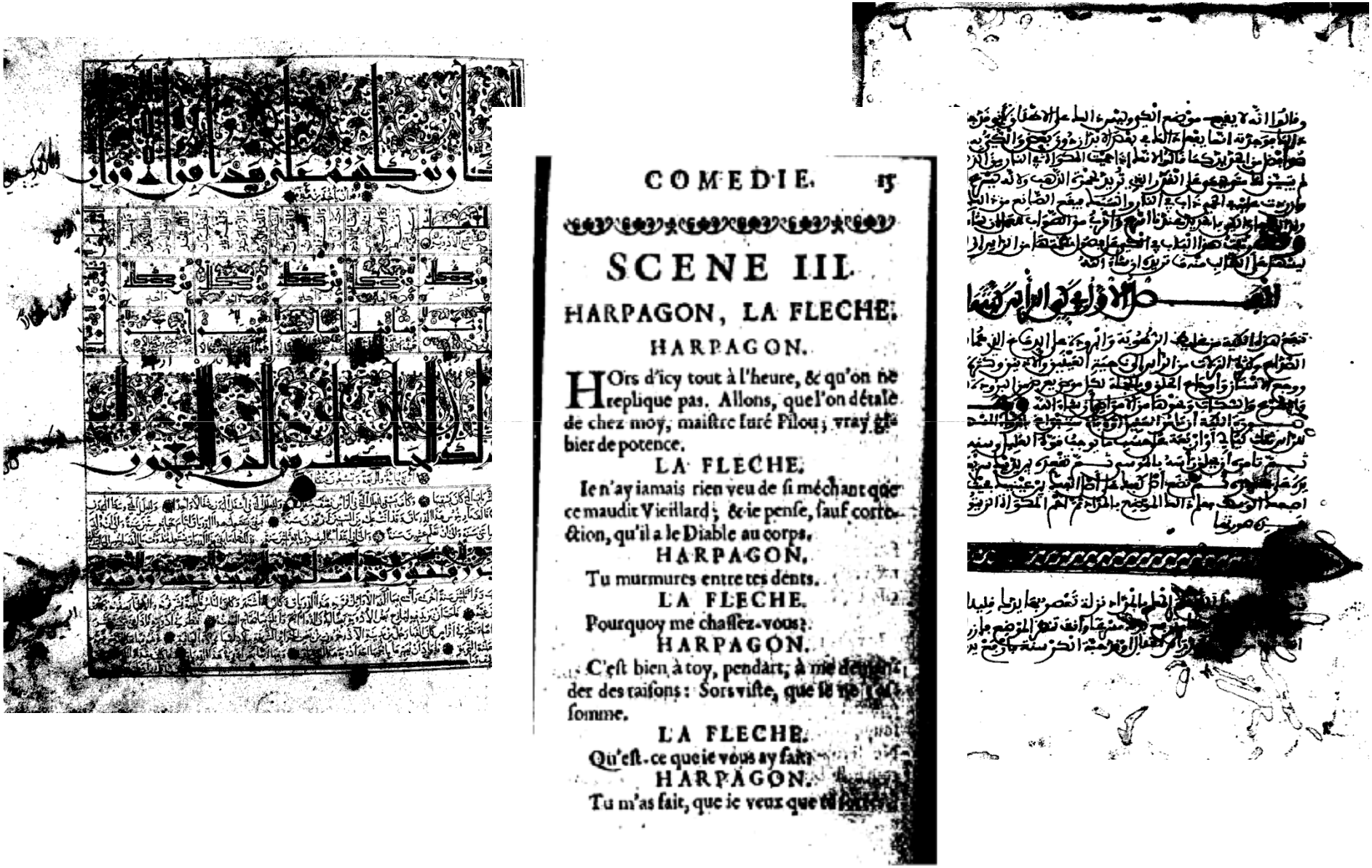
Gros plan sur la région supérieure de l’image

Images binaires (N&B)

information limitée : 2 valeurs possibles: 0 (blanc) 1 (noir)

Binaire pour un usage limité à celui des textes (dans les limites de la qualité de la méthode de binarisation)

La numérisation en binaire : Quelle utilité ?



COMEDIE. 15

SCENE III.

HARPAGON, LA FLECHE.

HARPAGON.

HOrs d'icy tout à l'heure, & qu'on ne replique pas. Allons, que l'on détale de chez moy, maistre furé Filou; vray gibier de potence.

LA FLECHE.

Ie n'ay iamais rien veu de si méchant que ce maudit Vieillard; & ie pense, sauf correction, qu'il a le Diable au corps.

HARPAGON.

Tu murmures entre tes dents.

LA FLECHE.

Pourquoy me chassez-vous?

HARPAGON.

C'est bien à toy, pendart, à me demander des raisons: Sors viste, que ie ne t'assomme.

LA FLECHE.

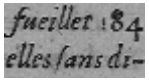
Qu'est-ce que ie vous ay fait?

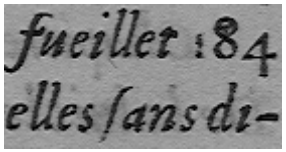
HARPAGON.

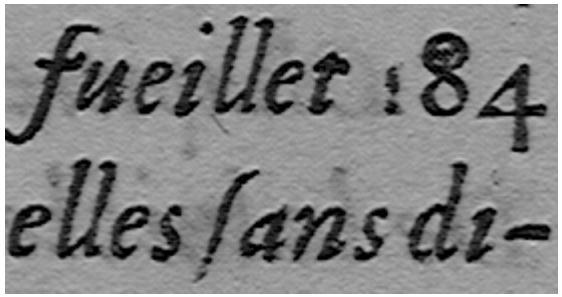
Tu m'as fait, que ie veux que tu

Numérisation de microfilms en binaire puis stockés en PDF !

Choix de la résolution

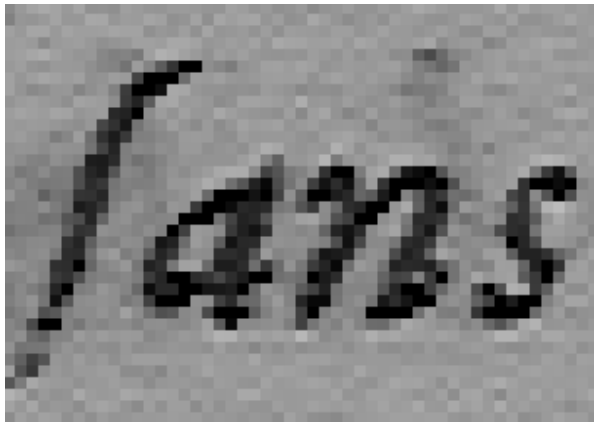


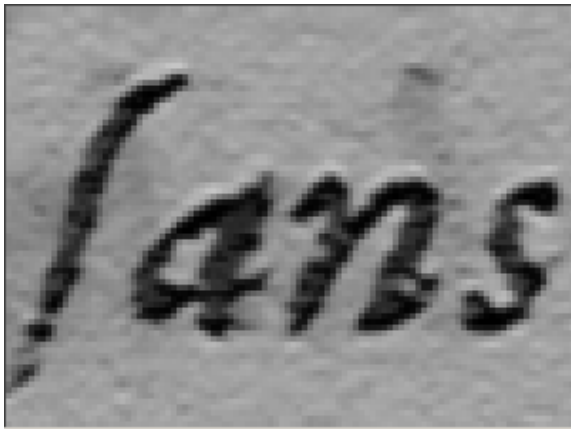




150 dpi **300 dpi** **600dpi**

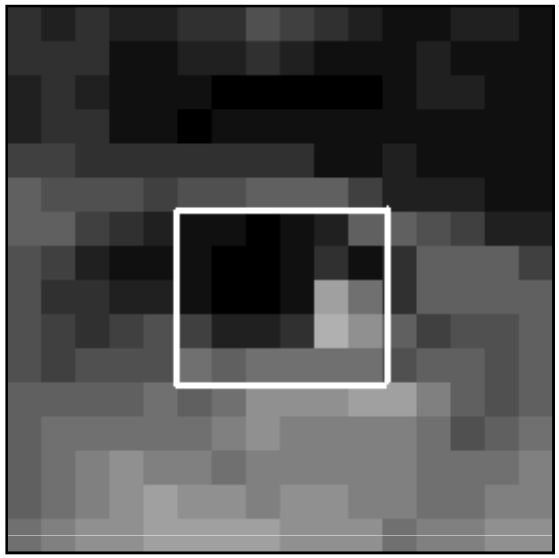
…Après harmonisation des tailles à l'écran par zoom :





DPI = Dot Per Inch (nombre de pixels réels par pouce)

Choix de la résolution



Exemple :

Résolution : 180x180 pixels
avec 256 niveaux de gris
(16 réellement utilisés ici)

16	16	0	16	32	96
16	0	0	16	48	16
16	0	0	16	160	112
64	32	32	48	176	144
112	96	112	112	112	112

Règles de conduite (1)

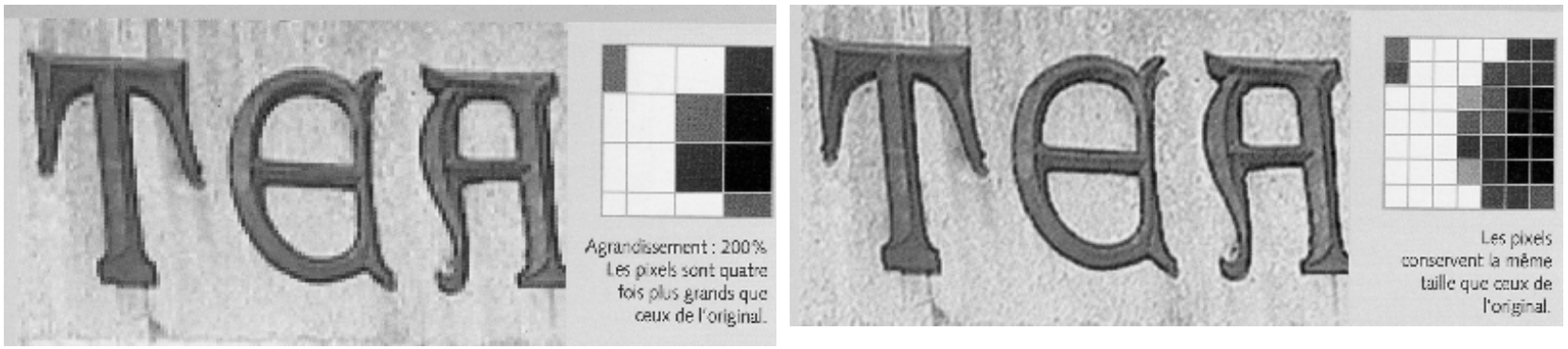
▶ **Toujours utiliser la résolution la plus élevée possible!**

▶ **C’est à dire plus d’information…**

Tous les systèmes d’analyse de documents numérisés *(reconnaissance des caractères, des structures, indexation…)* sont sensibles à la résolution des images

La vision humaine n’est pas sensible à la résolution des images !

Exemple :



Résolution : Donné en dpi ou ppi (nb de points par pouce) il définit le nombre de pixels réels par unité de mesure

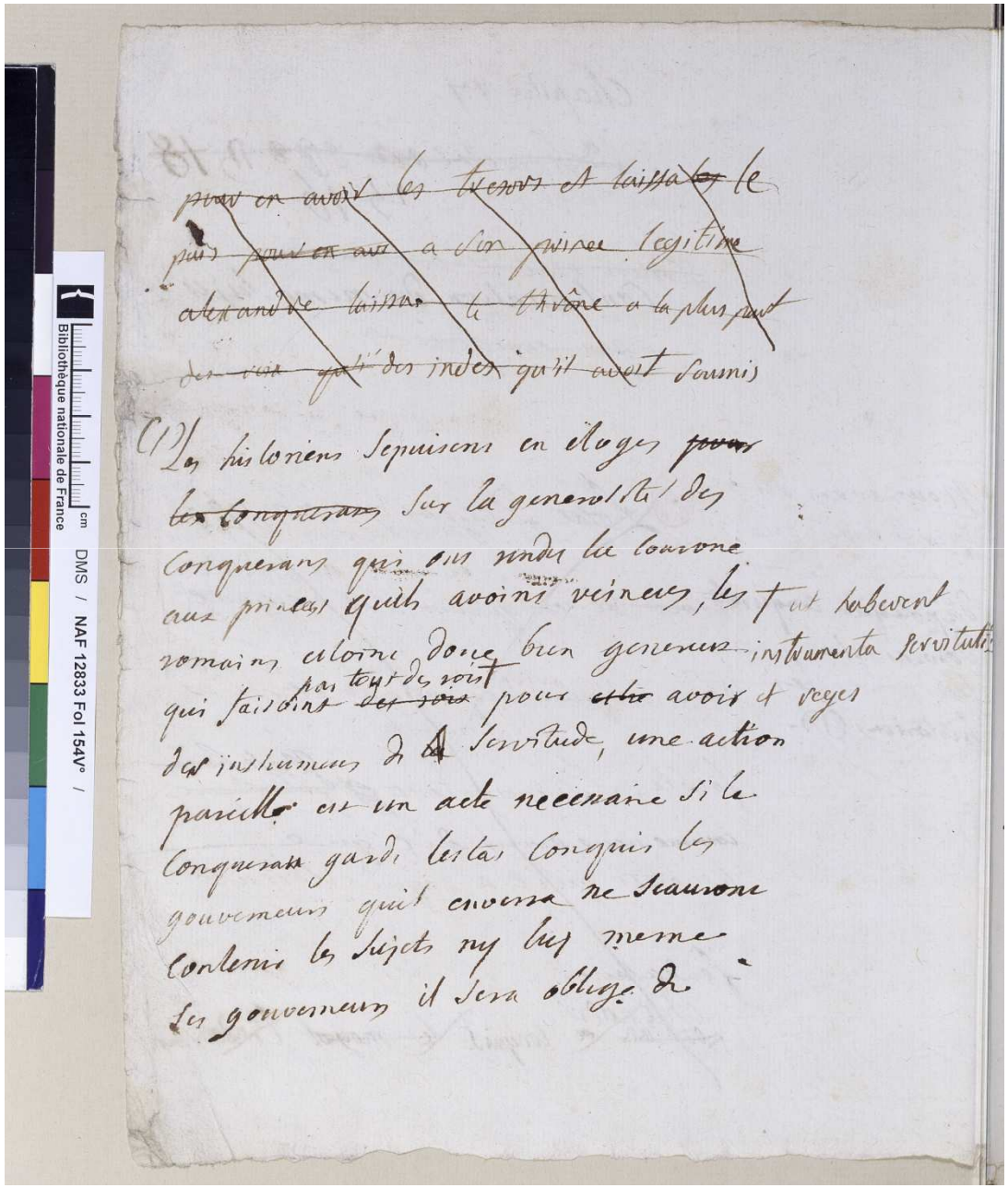
Règle de conduite (2)

- Numériser avec la mire couleur et une échelle
- Equilibre entre la qualité des images et la vitesse de capture

Bib. Mazarine

Règle de conduite (2b)

▶ Numériser avec la mire couleur et une échelle



Feuillets Montesquieu
BM Bordeaux

Règle de conduite (3)

- Fixer le facteur d’échelle inférieur ou égal à 100% (éviter les agrandissements lors de la capture)
- Ne jamais sur - échantillonner une image !!
- Seule la résolution optique d’un scanner est importante (pas la résolution interpolée)

Règles de conduite (4)

▶ **Eviter les étapes intermédiaires**

Mauvais exemples:

document papier → microfilm → scanner

document papier → photographie → scanner

document électronique → imprimerie → scanner

document papier → photocopie → scanner

Bons exemples :

document papier → scanner

document papier ou objet → photographie numérique

signer une convention pour récupérer les documents électroniques chez les éditeurs

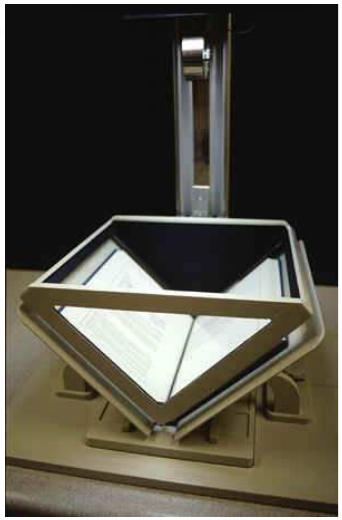
Quelle qualité d’images pour quels usages

Avant de commencer la numérisation, il faut faire une étude préalable

- ❑ Numérisation en interne ou sous-traitance ?
- ❑ Pour quelle utilisation numérise-t-on ces fonds ?
- ❑ Quelle résolution spatiale ?
- ❑ Numérisation en binaire ? Niveau de gris ? Couleur ?
- ❑ Quelle échelle ?
- ❑ Quels traitements pour restaurer les images ?
- ❑ Quel format de fichier (avec/sans compression avec/sans perte)
- ❑ Quel support pour les sauvegardes (durée du support dans le temps)

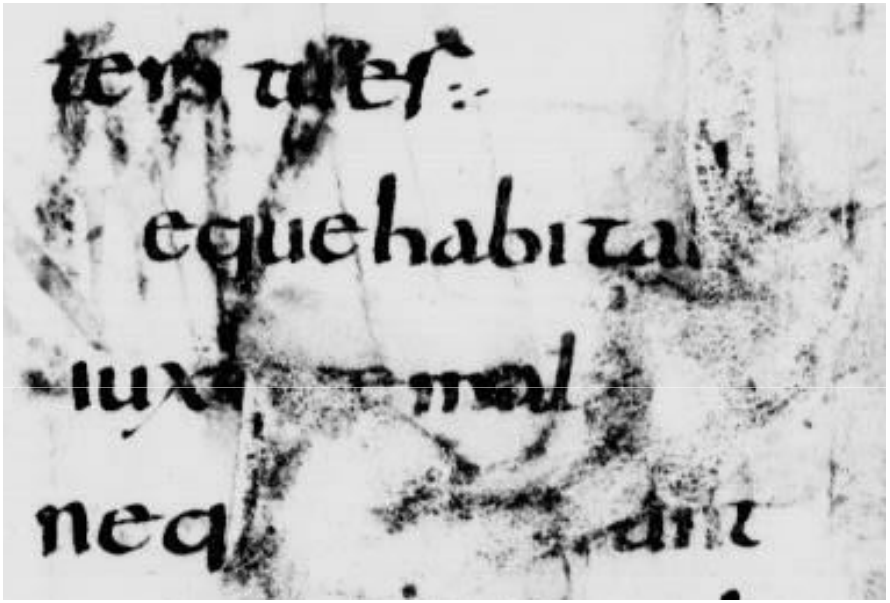
Critères de choix du matériel

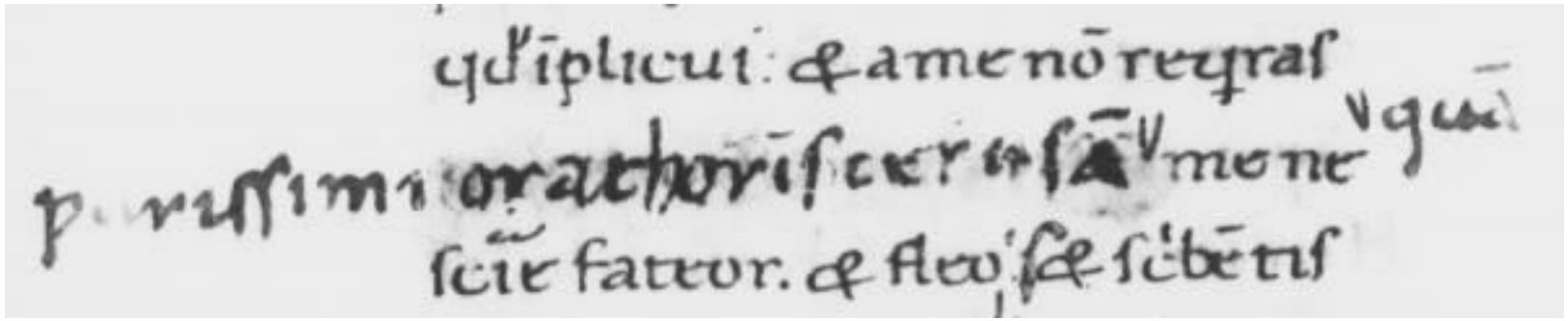
1) **Préservation des originaux** *(Scanners sans vitre, Lumière froide, plateau pour préserver la reliure du livre..)*
2) **Qualité des images** *(Résolution et fidélité de restitution des couleurs)*
3) **Vitesse** *(influence les coûts et le temps)*
4) **Automatisation, service et logiciels** *(Ergonomie zero-fatigue)*
5) **Coûts** *(amortissement , prix par page…)*



La qualité des images

Documents dégradés avant la numérisation

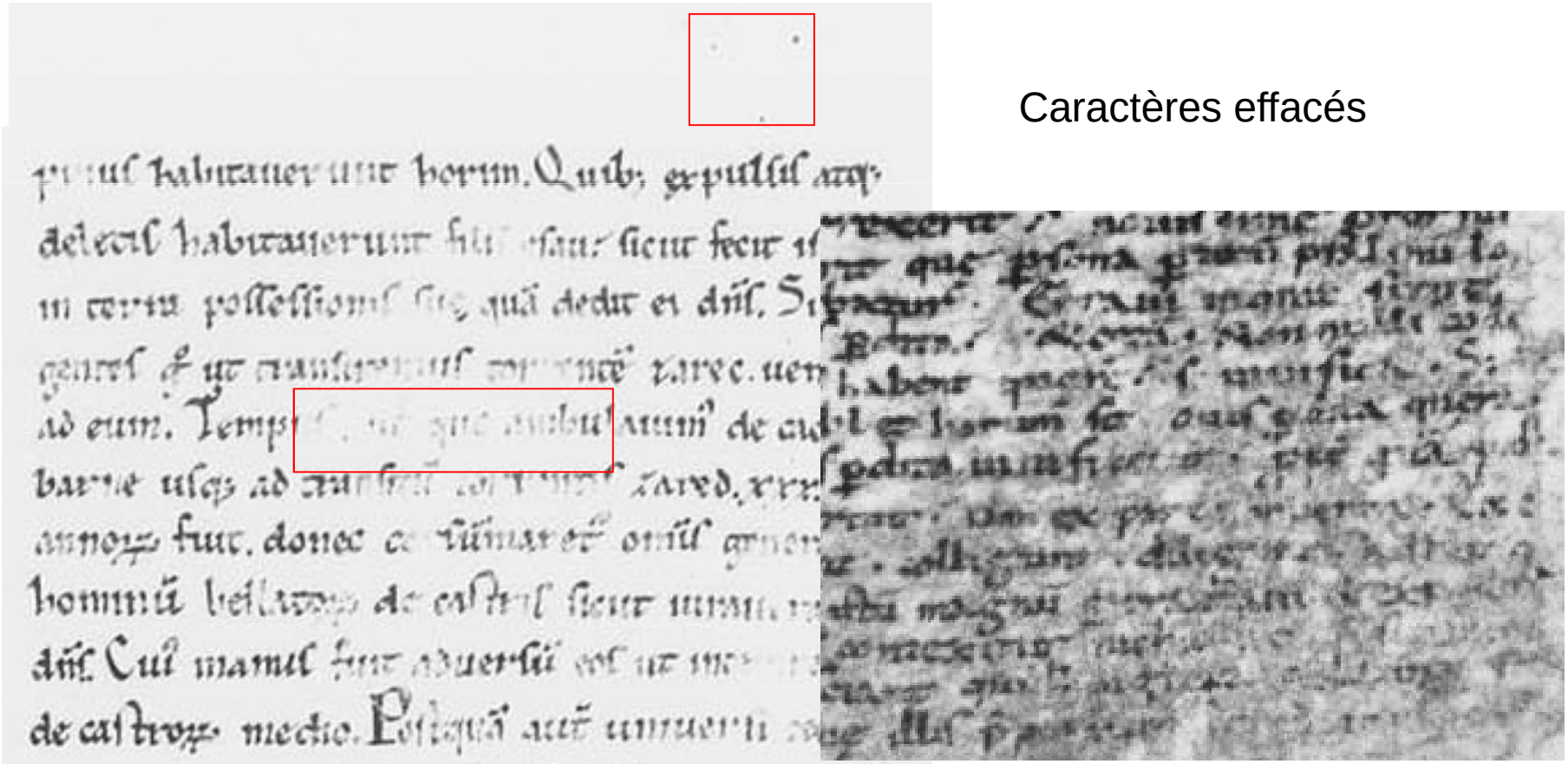




Numérisation de microfilms

Plus rapide et plus économique qu'une numérisation directe
…mais aux détriments de la qualité des images

Le microfilm nécessite le rehaussement du contraste qui efface les niveaux de gris intermédiaires

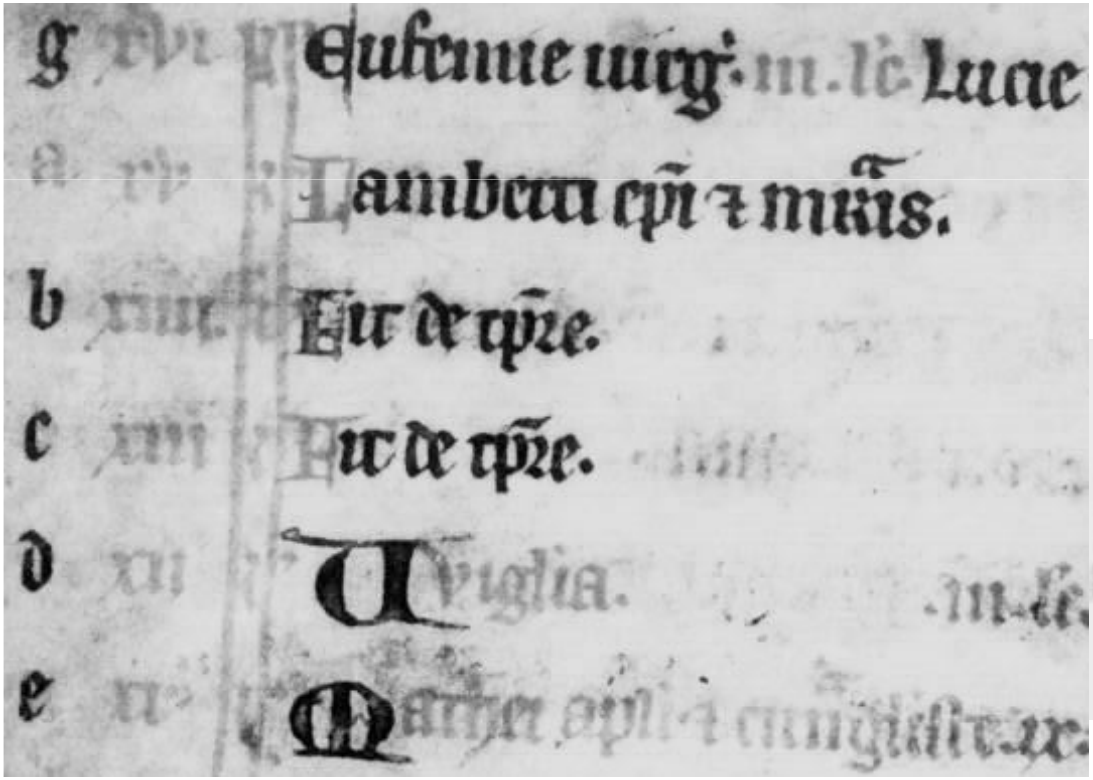


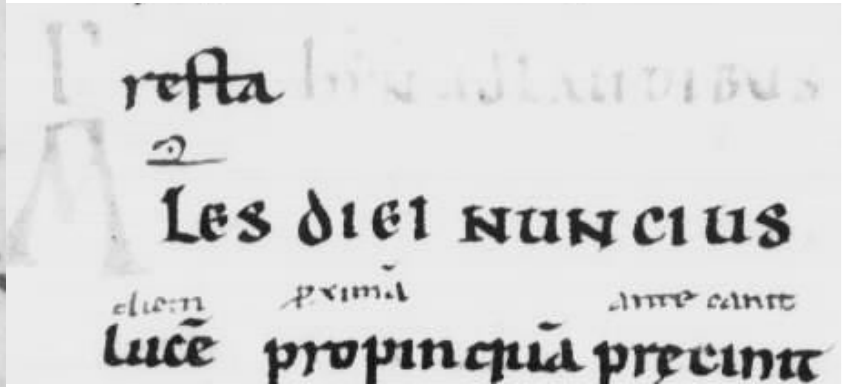
Caractères effacés

Numérisation de microfilms

Perte d’information sur la couleur des caractères

Le processus de microfilmage estompe les caractères rouges AVANT même la numérisation



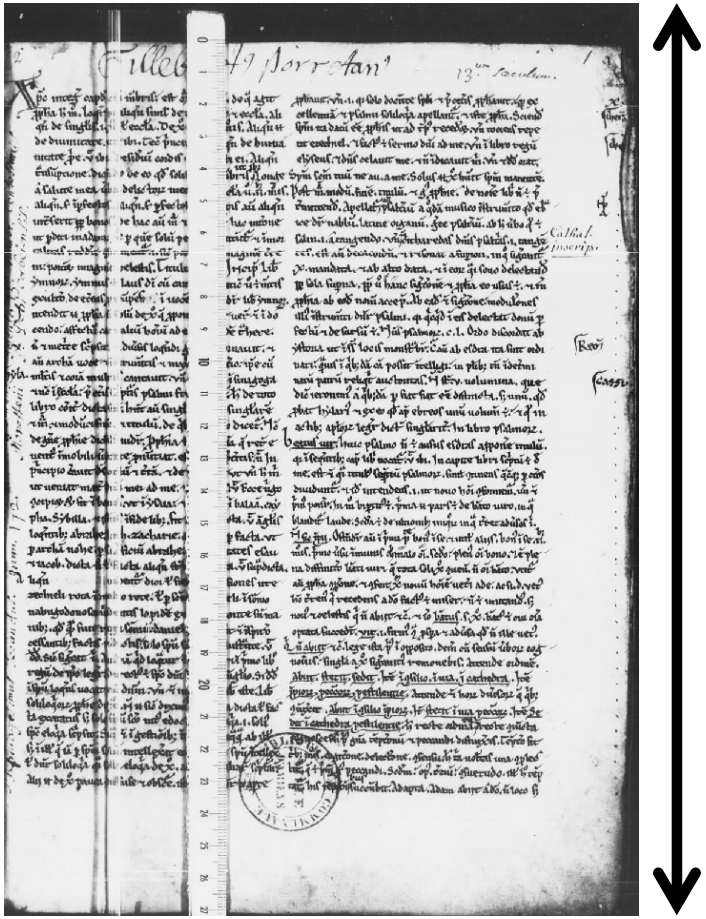


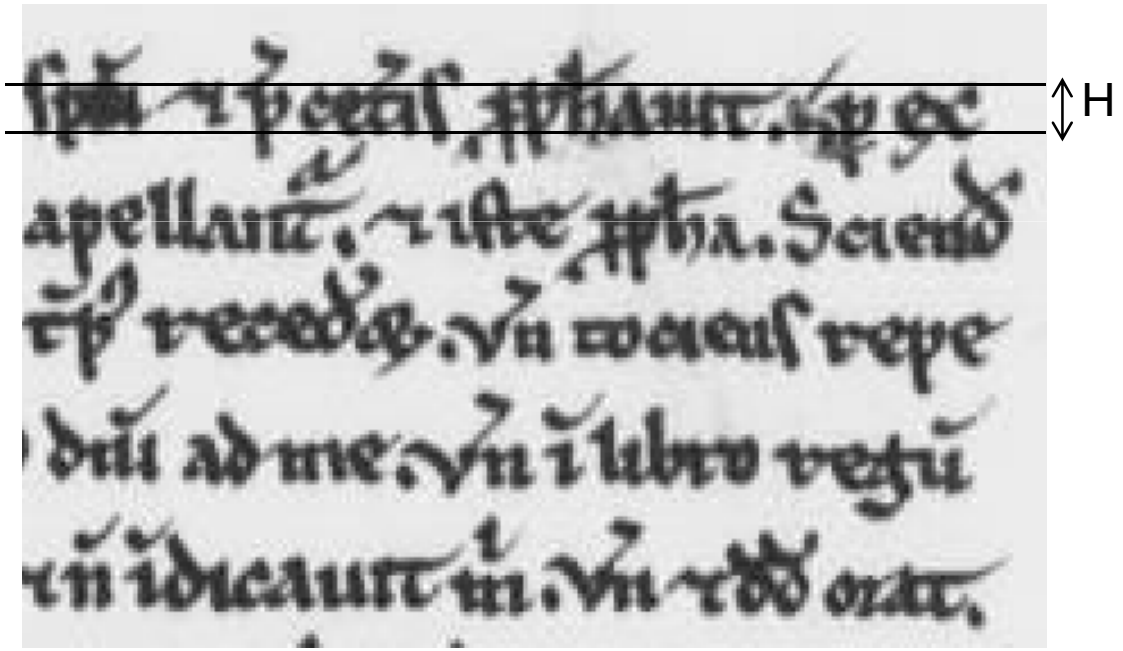
Numérisation et résolution

Choix de la résolution dépend de la taille (Hauteur) du plus petit caractère présent dans le document.

H>12 pixels pour la lisibilité

H>24 pixels pour une machine !





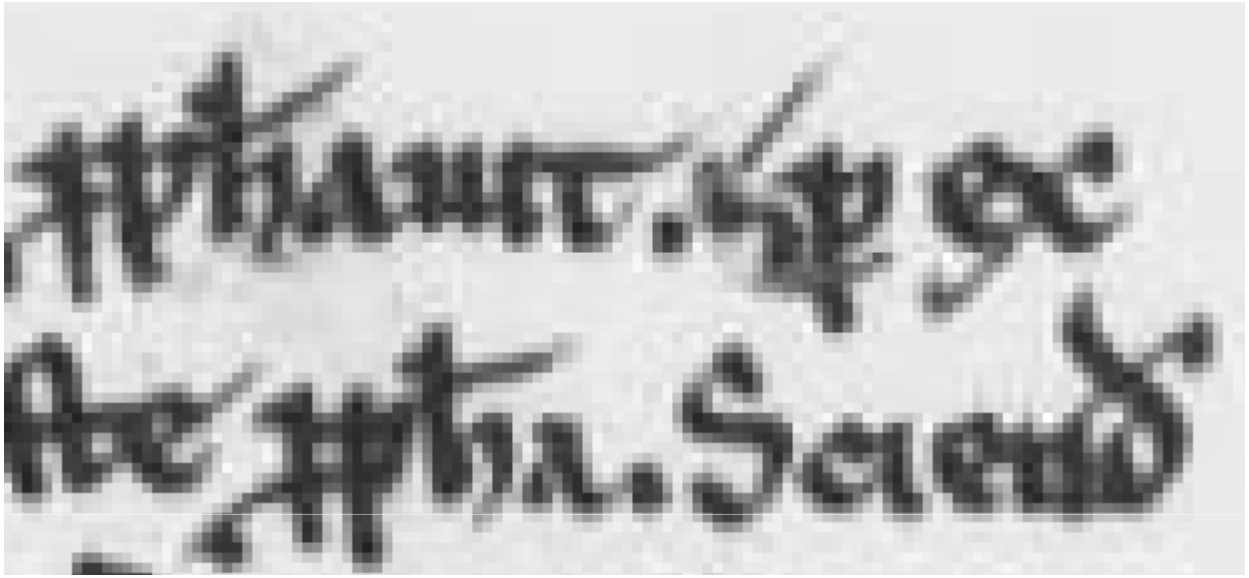
Résolution efficace

=

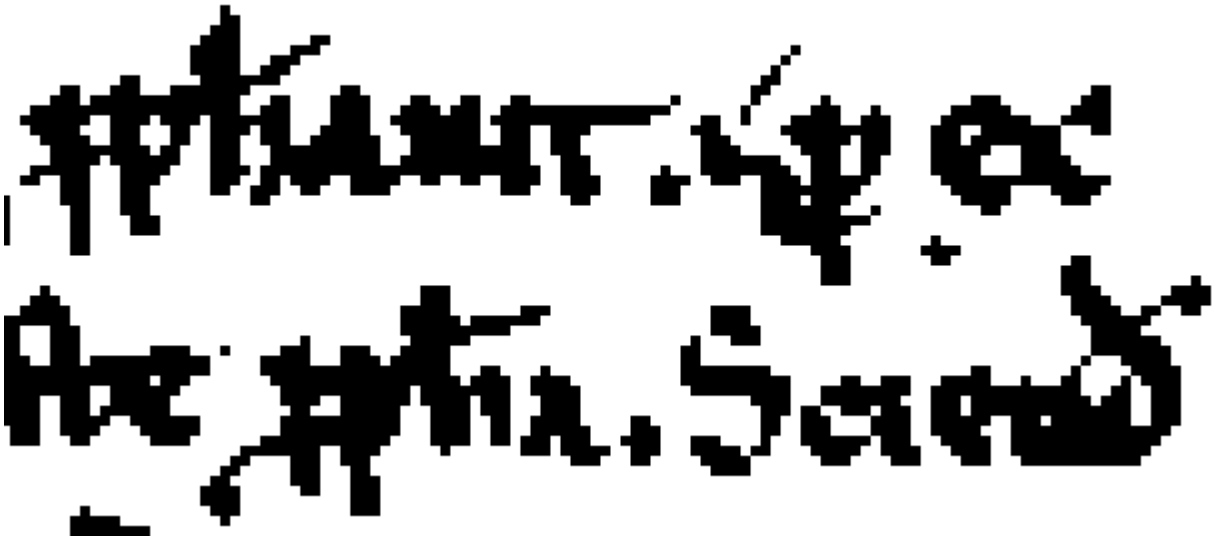
Nbr de pixels réels / Hauteur en pouces

En basse résolution

Il est impératif de numériser en niveaux de gris !



Résolution insuffisante mais reste lisible à l'Homme grâce aux niveaux de gris



En binaire le texte n'est plus lisible par l'Homme et par la machine

Le problème c'est que la machine binarise l'image pour l'analyser

Problèmes de résolutions

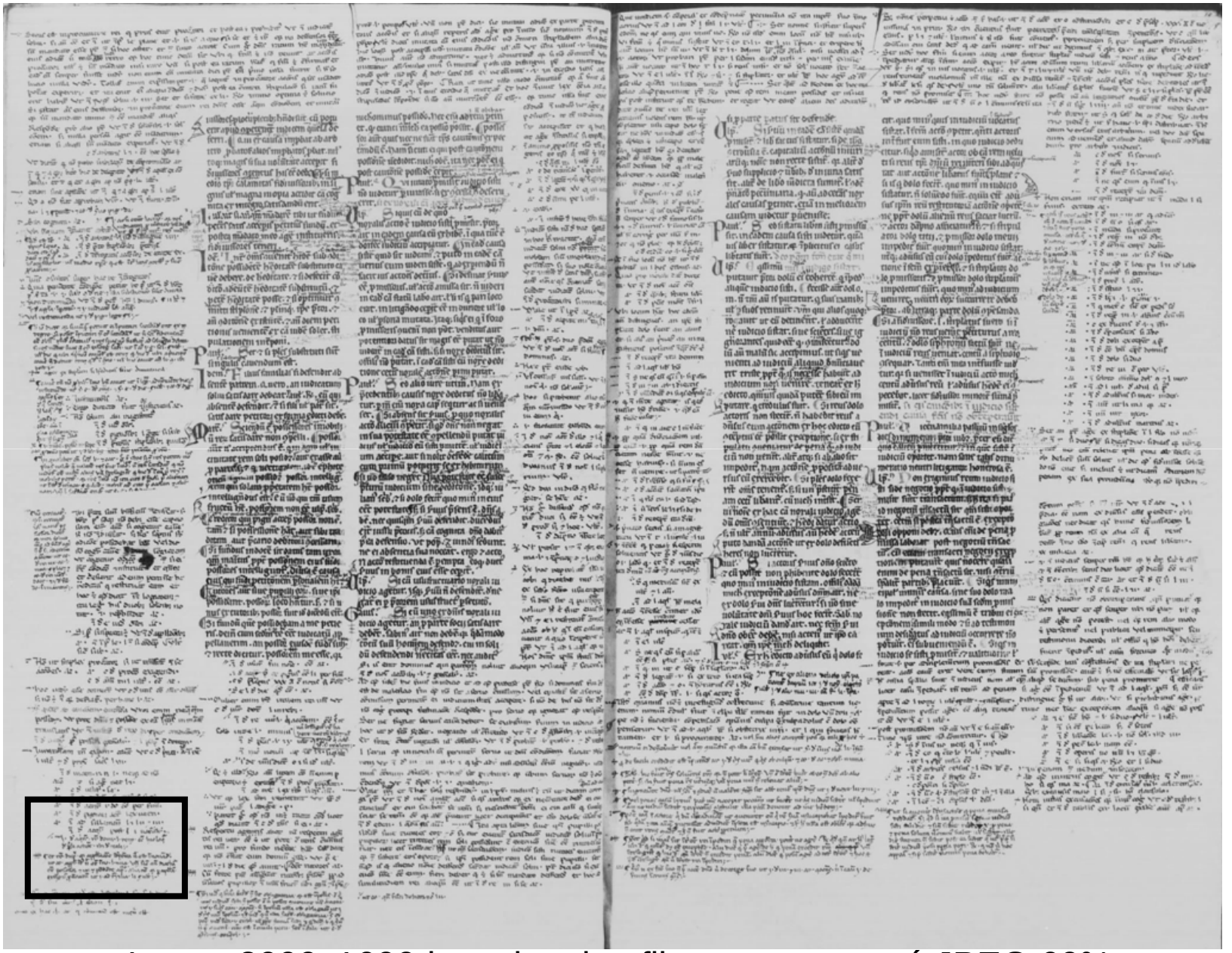
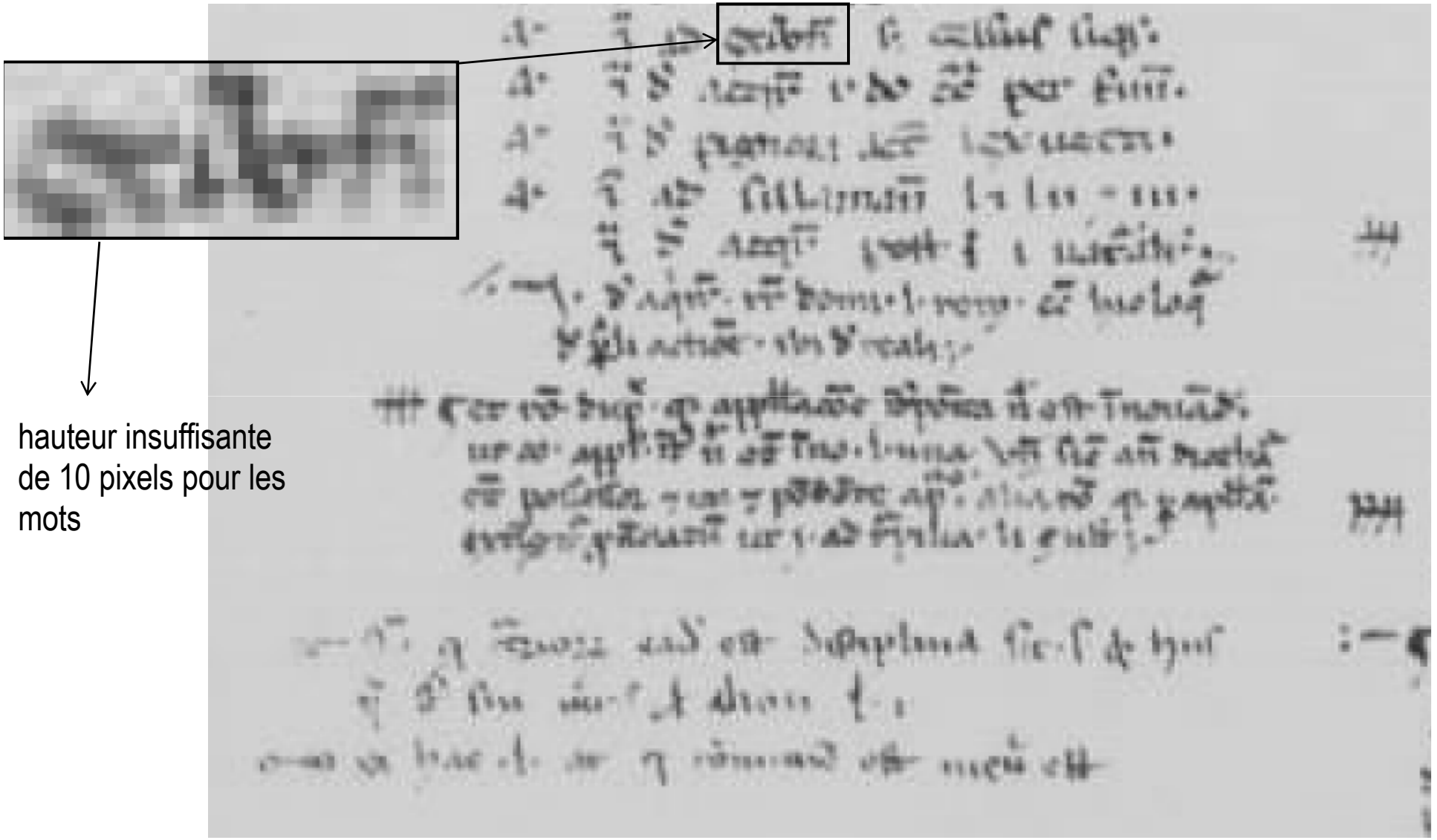


Image 2000x1600 issu de microfilm et compressé JPEG 60%

Problèmes de résolutions

Impossible à lire même pour un expert alors pour une machine…



hauteur insuffisante de 10 pixels pour les mots

Zoom de la portion d'image 330x220 du coin en bas à gauche

Exemple en vidéo ...

Que faire avec une image numérisée?

- Peut on y accéder?
  - Recherche
  - Navigation
  - Lecture

- Peut-on l'indexer et la récupérer sur simple demande?

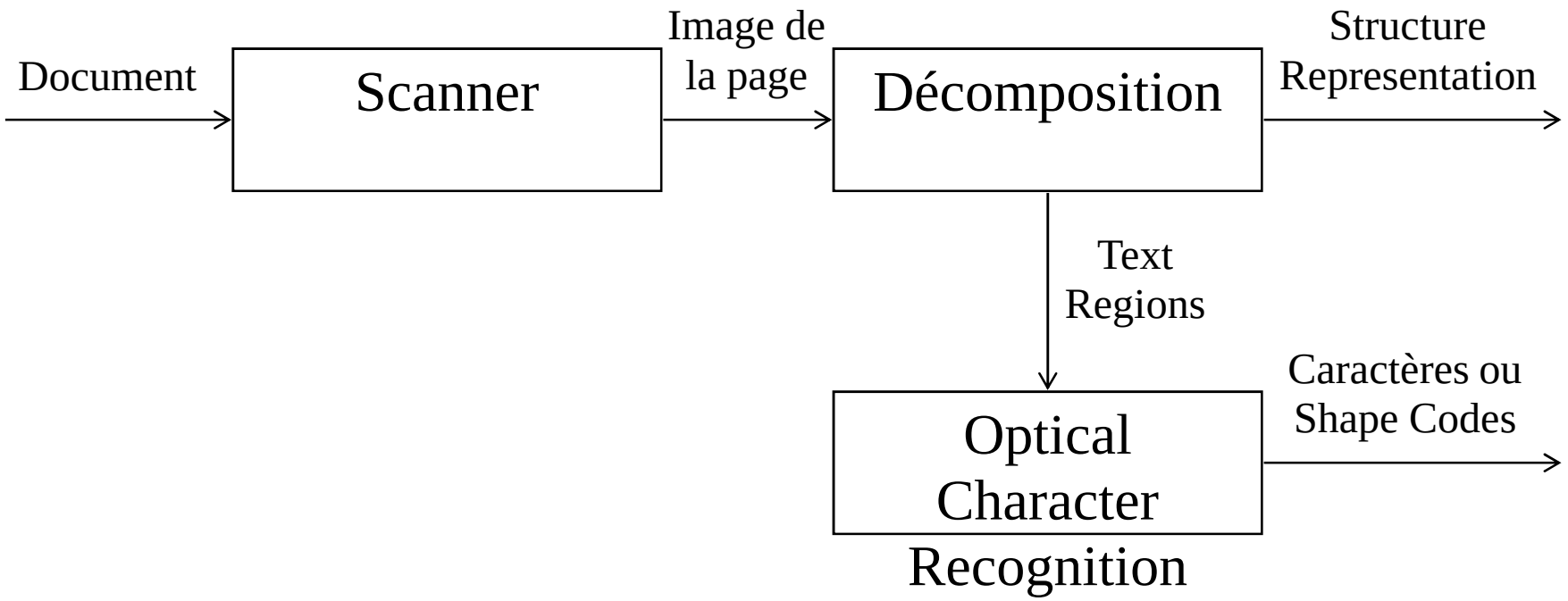
  Dans sa représentation initiale, pas réellement!

- Il est néanmoins possible de
  - La visualiser
  - L'imprimer
  - La recopier… mais pas beaucoup plus

Quels outils pour quels usages ?

Comment accéder aux textes des documents numérisés ?

Principe général



Comment accéder aux textes des documents numérisés ?

Imagerie

Acquisition Scanner / Caméra → Amélioration de l'image → Localisation (Blocs de texte, ligne, mot)

Structure physique

IA

Post-traitements Cohérence logique, Linguistique…

Décision / RF

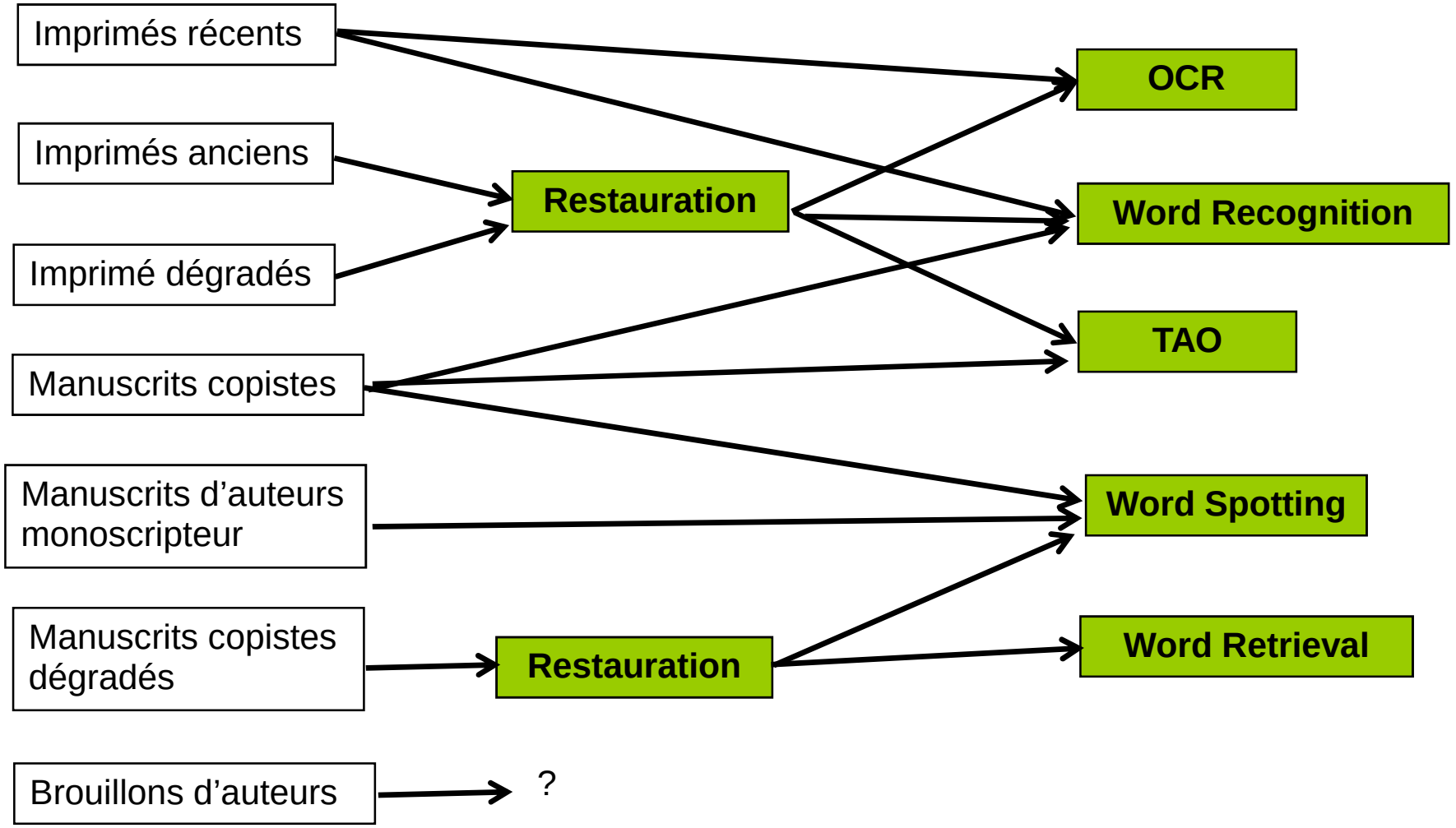
Reconnaissance (Titres, Paragraphes, mots, caractères)

Structure logique

Brouillons d'auteurs → ?

Comment accéder aux textes des documents numérisés ?

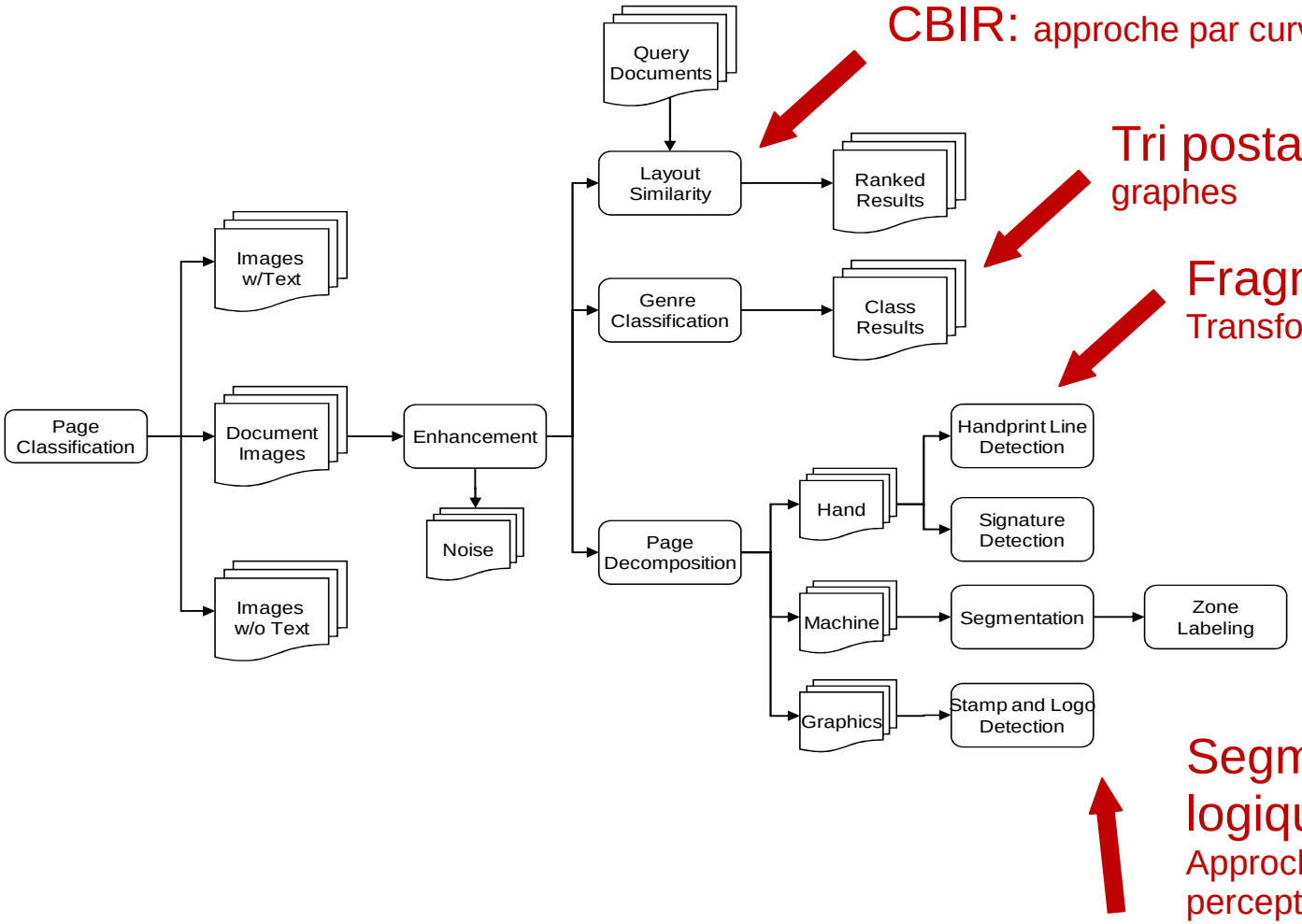
Cela dépend des images et de leurs contenus



Partie 1

L’ANALYSE DE DOCUMENT

Analyse de documents



CBIR: approche par curvelets

Tri postal: coloration de graphes

Fragmentation:
Transformée de Hough

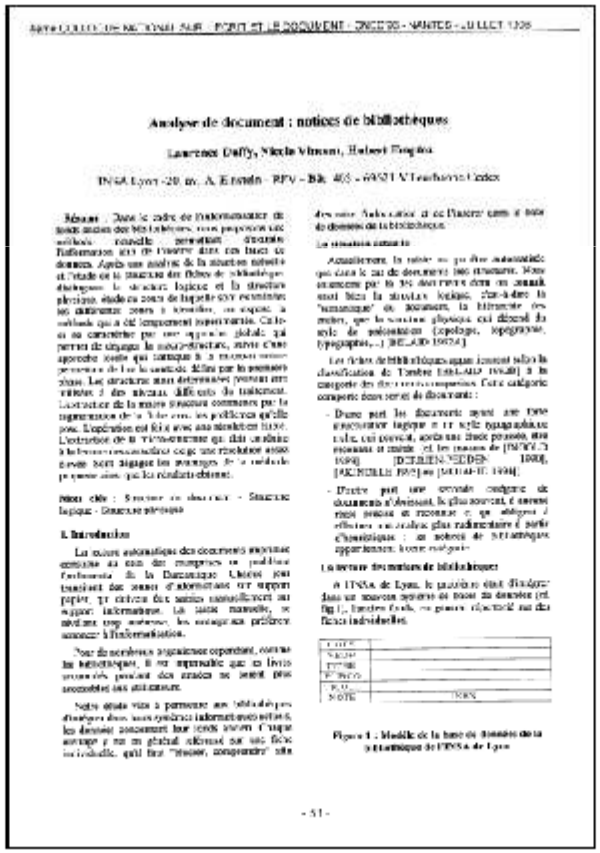
Segmentation logique:
Approche psycho-perceptive

Une grande variété de tâches pour une grande variété de contenus

Quelles sont les difficultés inhérentes à l'analyse de documents ?

Les différences de structures?

Manhattan layout



Rectangular layout

Random structure

Quelles sont les difficultés inhérentes à l’analyse de documents ?

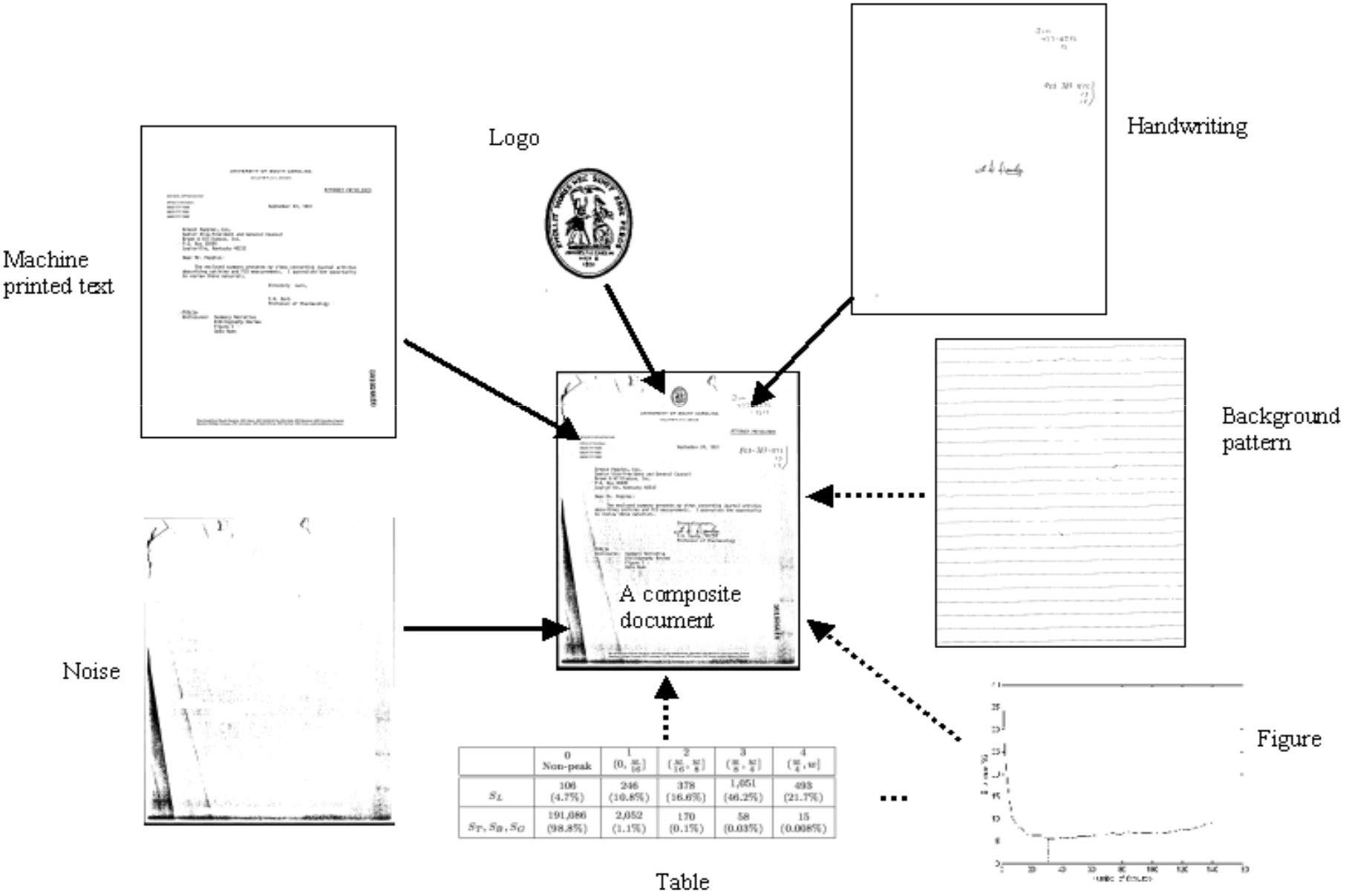
- Tout est représenté dans un tableau “2D”
- Tout est symbole
- Les variations de symboles sont très nombreuses: épaisseur, taille, orientation…

- … et lorsque l’on a affaire à du manuscrit?

Sur les documents réguliers composites, on sait assez bien faire

- Correction de l’inclinaison
  - *Orientation primaire des lignes de texte*
- Détection des images et des régions de texte
  - Texture et orientation dominantes
- Classification structurelle
  - Inférence de la structure physique sur la logique
- Classification des région de texte
  - Titre, auteur, notes entêtes, signature, etc.

Sur les documents composites bruités?



Principe général de la segmentation des documents de structures régulières

- Typiquement basé sur l'analyse des proximités spatiales
  - Plages blanches
  - Marges
  - Différences dans les contenus
- Peut être très sensible au bruit
- Distinction entre deux types d'approches
  - Top Down – du global au local
  - Bottom up – du local au global

Sur les documents de structure irrégulière de contenu hétérogène?

- On peut tenter des approches sur des régions localisées
- Il reste à localiser les régions !!
- On peut essayer de classifier les contenus: faut il parvenir à les caractériser initialement ?
- La tâche est rendue difficile, parfois impossible…

De la structure physique à la structure logique (ordre de lecture)

- Difficile à obtenir selon le cas: pas toujours stable
- Existence de guides explicites
  - “Continue à la page 4”
- Existences d’indices structurels
  - Continuité d’un texte sur deux colonnes
- Analyse du contenu sémantique
  - Statistique d’occurrences de mots, analyse de syntaxe

SEGMENTATION PAR L’EXEMPLE

Architecture d’un système : Segmentation

Image couleur RGB

Analyse Couleur

- **Luminance**
- **Classification automatique**
- **statistique**

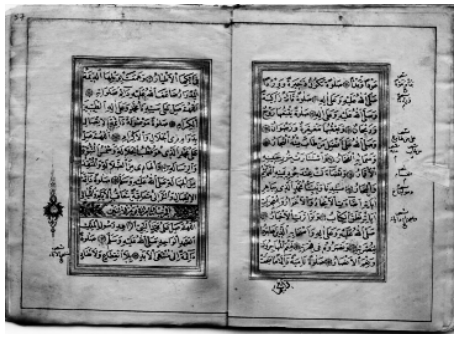


Image en 256 niveaux de gris

Binarisation

- **Seuil global automatique**
- **Seuil adaptatif**
- **Classification automatique**

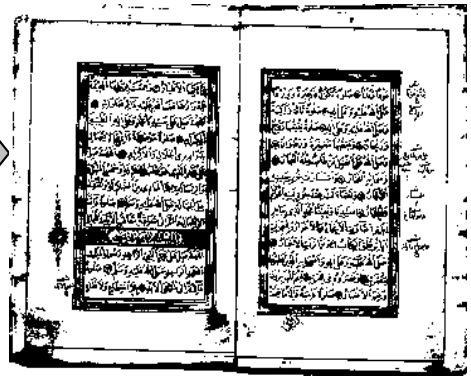


Image binaire

- **Suppression des cadres**
- **Segmentation des objets connexes**
- **Détection de la zone principale**

Segmentation

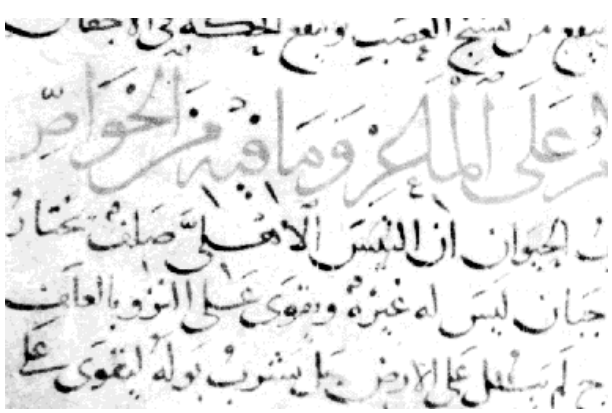
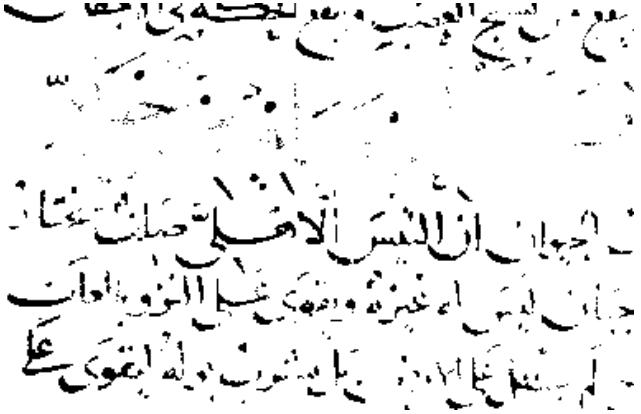


Image originale 16 niveaux de gris



Seuillage automatique globale

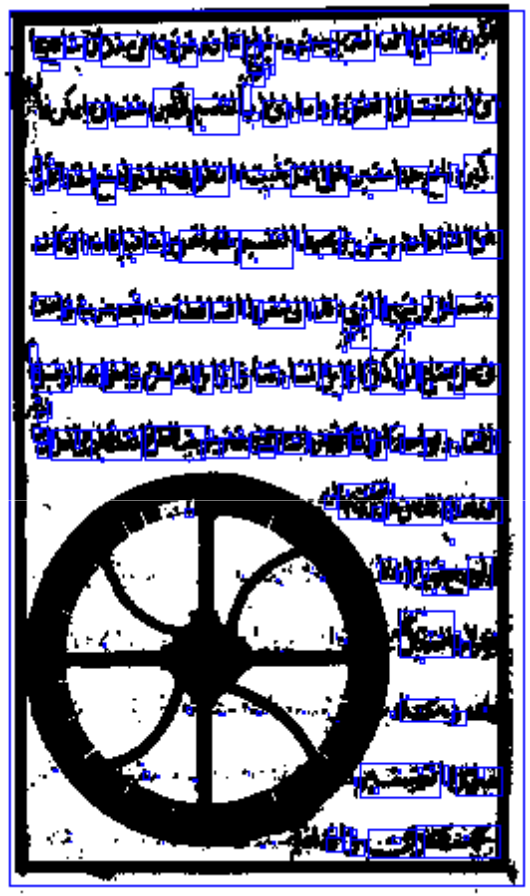
Seuillage automatique adaptatif

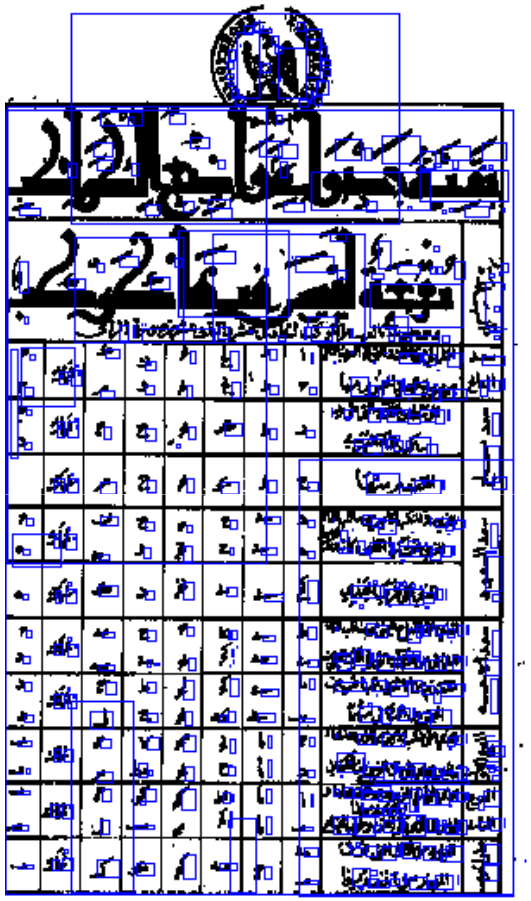
Seuillage par classification automatique

Recherche de tous les objets

Segmentation des objets connexes

Problèmes des cadres et lignes





Les cadres et tableaux forment aussi des objets connexes

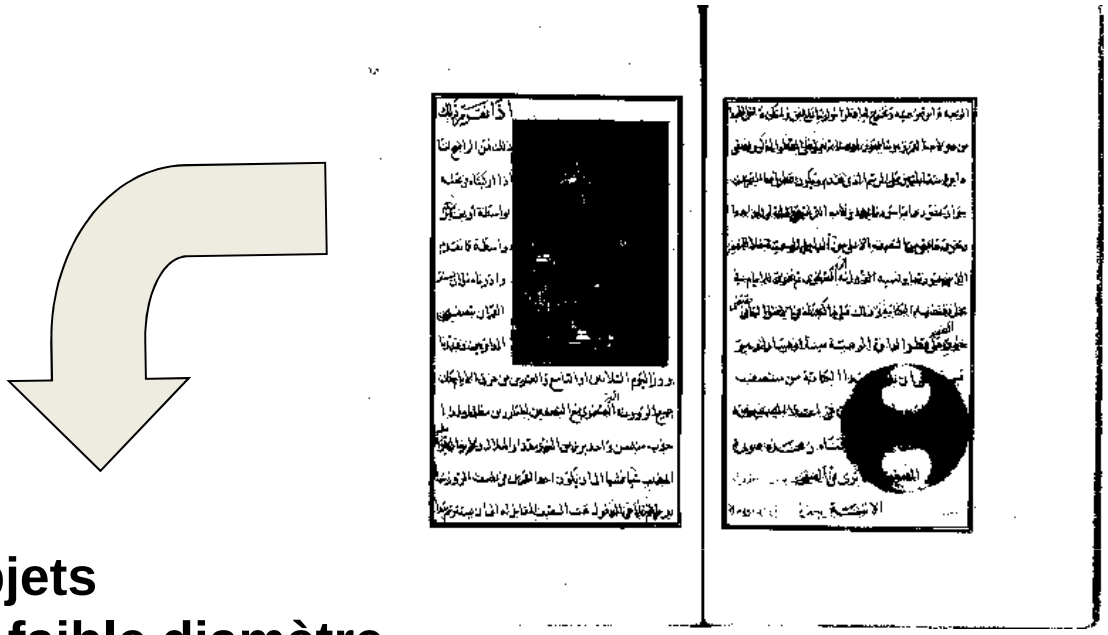
Suppression automatique des cadres

Pour simplifier l’image, il faut retirer les cadres, bordures et bords des ouvrages

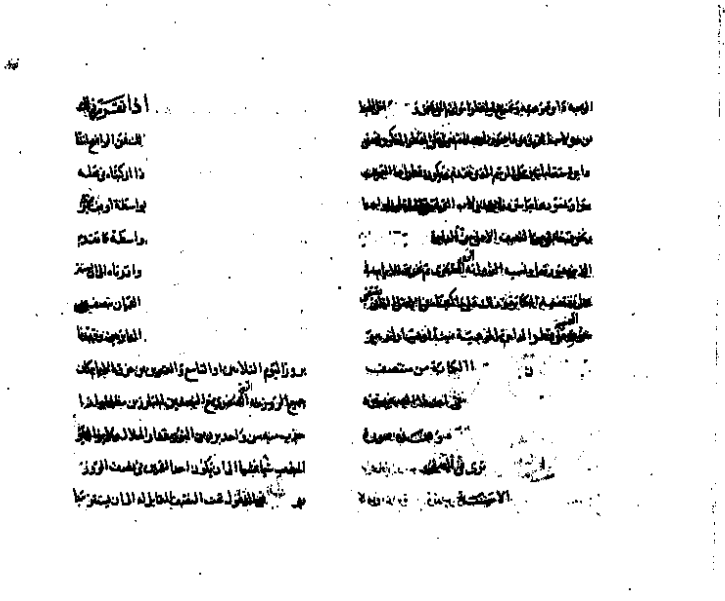
Mesure le diamètre vertical maximal des objets

Chaque objet est remplit d’un niveau de gris correspondant à son diamètre maximal en hauteur

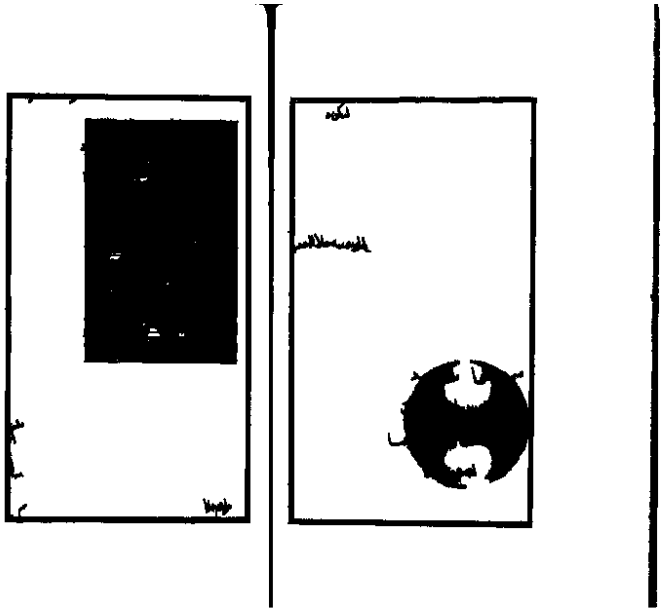
Classification Texte/graphique



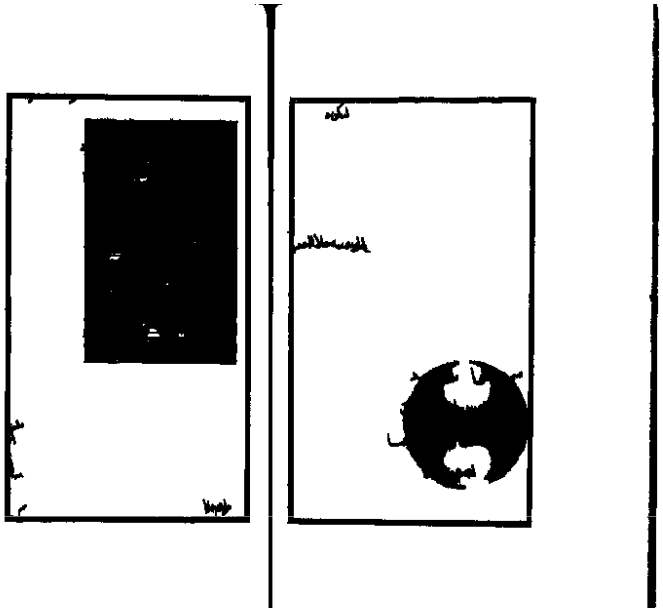
Objets de faible diamètre

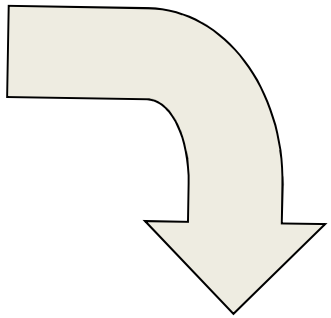


Objets de diamètre élevé



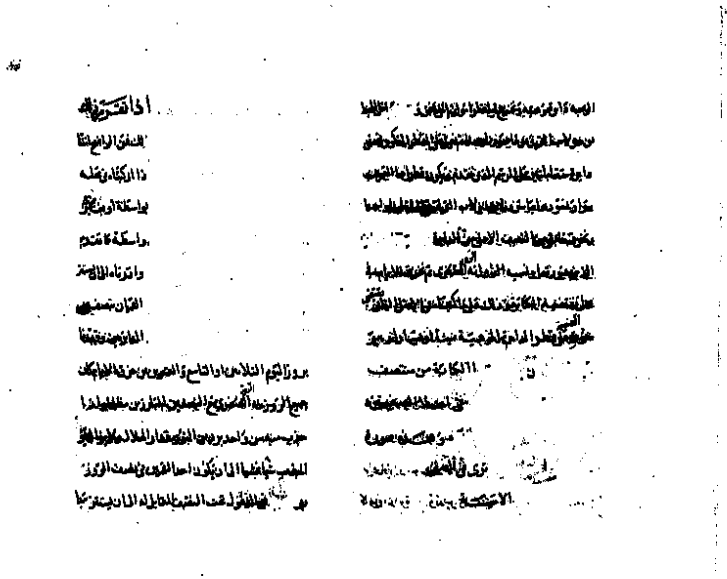
Suppression des objets de faible épaisseur



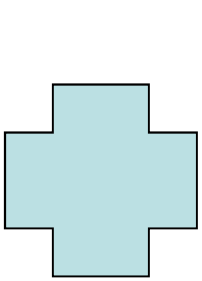


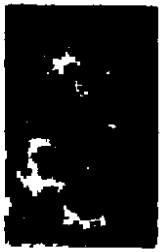
Suppression des objets de faible épaisseur

Image finale

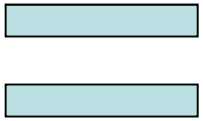


Objets de faible diamètre





Objets de grande taille



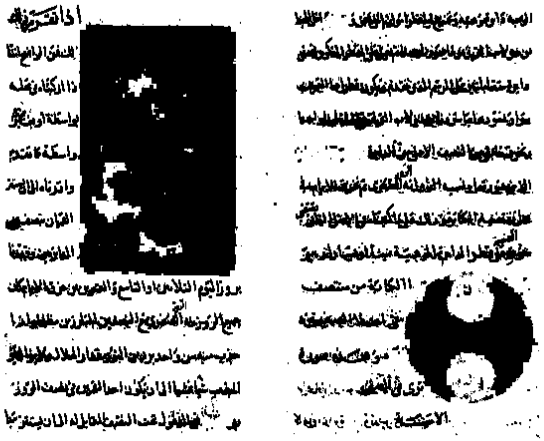
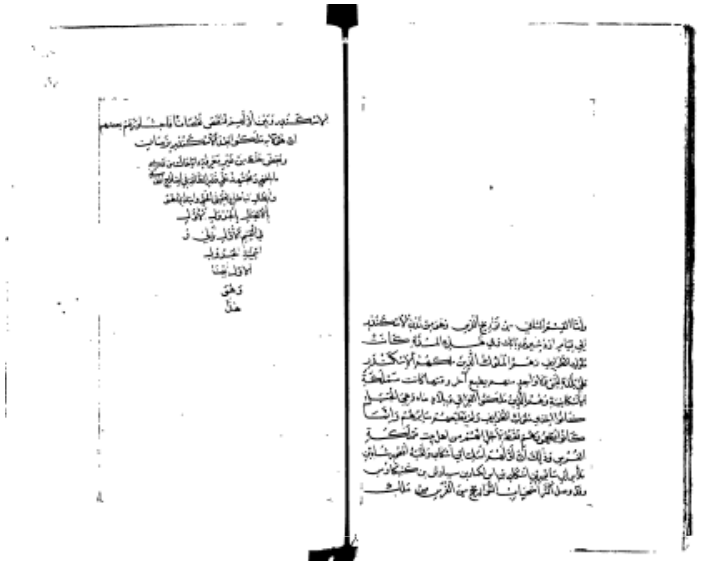
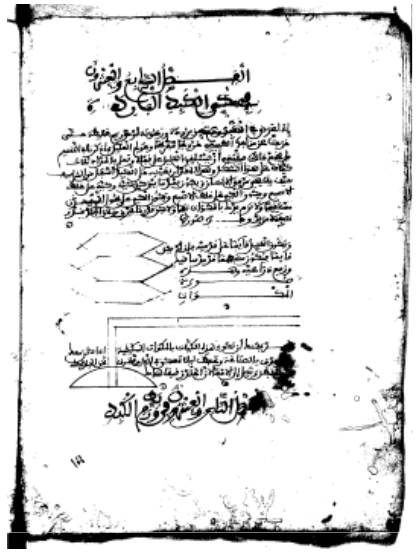


Image sans bordure, ni cadre

Problèmes de la localisation des corps





L’utilisation des cadres est difficile

La présence de textes non justifiés est un obstacle

L’absence de texte dans de larges zones vides de la page pose des problèmes de détection.

Partie 3

OCR
ET ALTERNATIVES PAR TAO ET WS

Optical Character Recognition

- Documents imprimés
- Segmentation
  - Blocs
    - Colonnes
    - Paragraphes…
  - Mots
  - Caractères
- Reconnaissance
  - Caractères
  - Guidée par le contexte
    - Probabilité d'apparition
    - Dictionnaire
- Vérification
  - Dictionnaire

Optical Character Recognition

- Les approches usuelles

- Approche type « Pattern-matching »
  - *Standard approach in commercial systems*
  - *Segment individual characters*
  - *Recognize using a neural network classifier*

- Modèle des chaînes de Markov cachées
  - *Segment into sub-character slices*
  - *Limited lookahead to find best character choice*
  - *Useful for connected scripts (e.g., Arabic)*

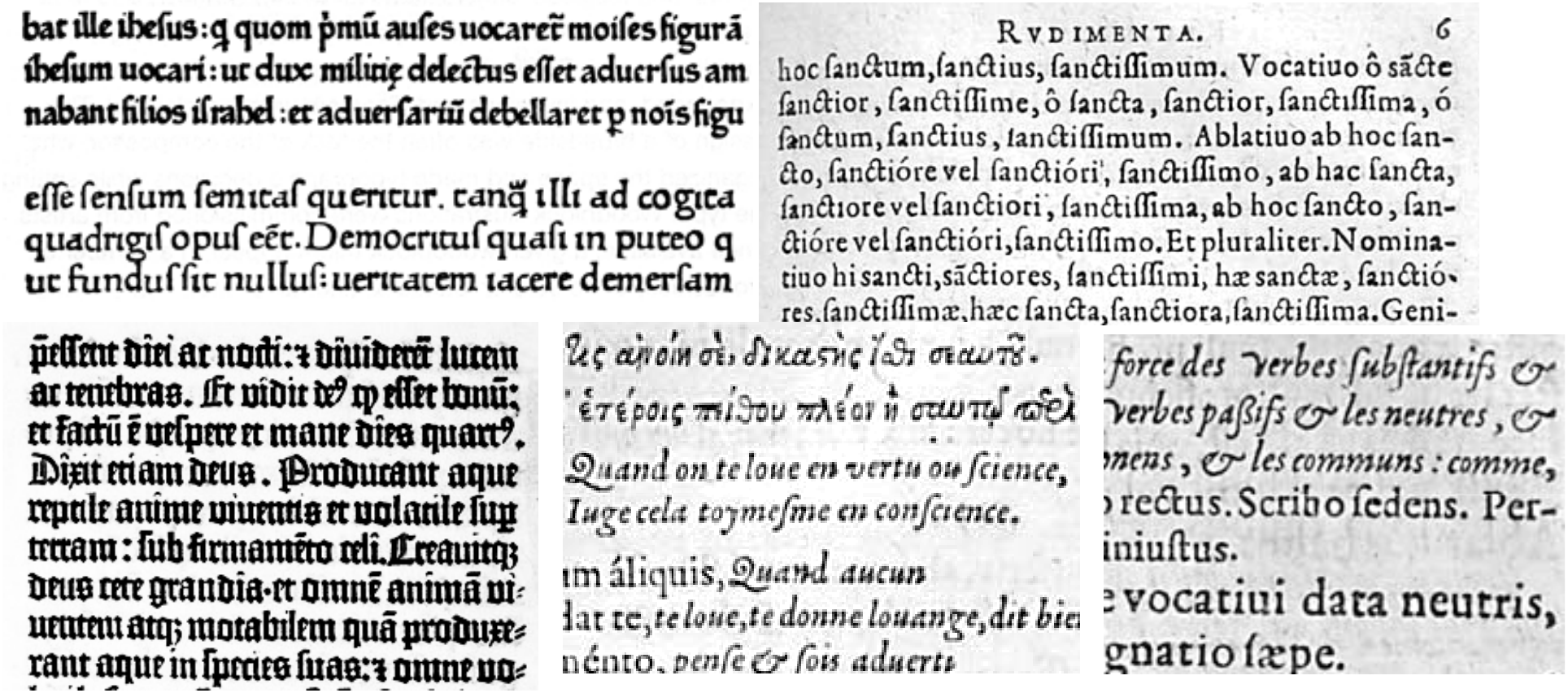
OCR Accuracy Problems

- Character segmentation errors
  - In latin scripts, segmentation often changes “m” to “rn”
- Character confusion
  - Characters with similar shapes often confounded
- OCR on copies is much worse than on originals
  - Pixel bloom, character splitting, binding bend
- Uncommon fonts can cause problems
  - If not used to train a neural network

Améliorer la précision des OCR

- Prétraiter les images
  - Mathematical morphology for bloom and splitting
  - Particularly important for degraded images
- Mettre les OCR en concurrences, les sérialiser
  - Individual systems depend on specific training data
- Employer des correcteurs linguistiques
  - Use confusion statistics, word lists, syntax, …
  - But more harmful errors might be introduced

Les limites des OCRs sur le 16ème



bat ille ihesus: q quom ꝑmū auſes uocarēt moiſes figurā
ihesum uocari: ut dux milicię delectus eſſet aduerſus am
nabant filios iſrahel: et aduerſariū debellaret ꝑ noīs figu

eſſe ſenſum ſemitaſ queritur. tanq̄ illi ad cogita
quadrigis opuſ eēt. Democritus quaſi in puteo q
ut fundus ſit nullus: ueritatem iacere demerſam

RVDIMENTA. 6

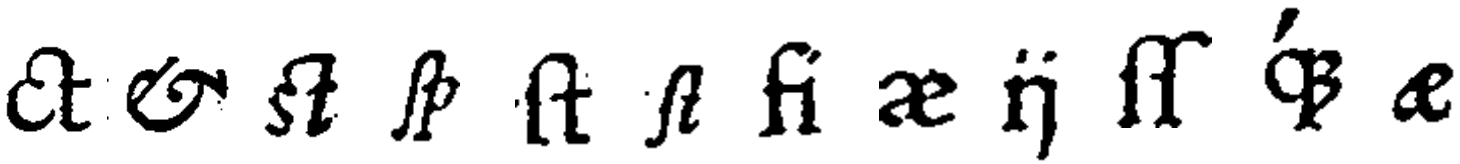
hoc ſanctum, ſanctius, ſanctiſſimum. Vocatiuo ô sācte
ſanctior, ſanctiſſime, ô ſancta, ſanctior, ſanctiſſima, ó
ſanctum, ſanctius, ſanctiſſimum. Ablatiuo ab hoc ſan-
cto, ſanctióre vel ſanctiór', ſanctiſſimo, ab hac ſancta,
ſanctiore vel ſanctiori, ſanctiſſima, ab hoc ſancto, ſan-
ctióre vel ſanctiór, ſanctiſſimo. Et pluraliter. Nomina-
tiuo hi ſancti, sāctiores, ſanctiſſimi, hæ ſanctæ, ſanctió-
res. ſanctiſſimæ, hæc ſancta, ſanctiora, ſanctiſſima. Geni-

pſſent diei ac nocti: ꝛ diuideꝛ lucem
ac tenebras. Et vidit dꝰ q eſſet bonū;
et factū ē veſpere et mane dies quart⁹.
Dixit etiam deus. Producant aque
repule anime viuentis et volatile sup
terram: ſub firmamēto celi. Creauitq;
deus cete grandia. et omnē animā vi-
uentem atq; motabilem quā produxe-
rant aque in ſpecies ſuas: ꝛ omne vo-

Ὡς αἰρεῖσθαι δικαίως ἑαυτόν.
ἑτέροις πείθου πλέον ἢ σεαυτῷ ὢν ἐλ
Quand on te loue en vertu ou ſcience,
Iuge cela toymeſme en conſcience.
im áliquis, Quand aucun
lat te, te loue, te donne louange, dit bie
nénto. penſe & ſois aduerti

force des verbes ſubſtantifs &
verbes paſſifs & les neutres, &
mens, & les communs: comme,
rectus. Scribo ſedens. Per-
iniuſtus.
e vocatiui data neutris,
gnatio ſæpe.

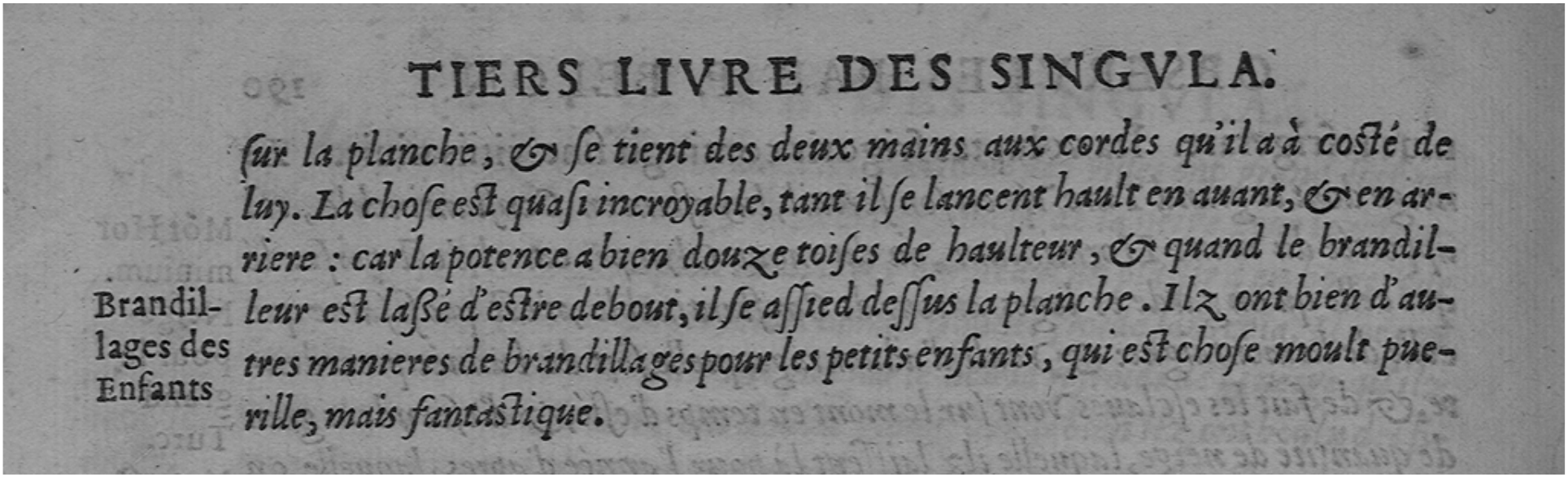
Polices et styles obsolètes que les OCRs ne reconnaissent pas



Caractères ligaturés aujourd'hui obsolètes

problème de codage UNICODE des caractères de la Renaissance

Imprimé de la Renaissance



TIERS LIVRE DES SINGVLA.

ſur la planche, & ſe tient des deux mains aux cordes qu'il a à coſté de luy. La choſe eſt quaſi incroyable, tant il ſe lancent hault en auant, & en arriere : car la potence a bien douze toiſes de haulteur, & quand le brandilleur eſt laſſé d'eſtre debout, il ſe aſſied deſſus la planche. Ilz ont bien d'autres manieres de brandillages pour les petits enfants, qui eſt choſe moult puerille, mais fantaſtique.

Brandillages des Enfants

Sortie OCR

TIERS HVRE „Es Si NG VL A~

tient dès. ðeùx; thé 5J 4UX: Cø?dt5 F COŠ16 de

Ìu,...Mb.ofieffquafiiiwiô414 tafIt-Itft læflççflî hatilt en ariant, 4?,

rièrè~ :t4tL4pOtCfltC4biÇ%åØU~~tO~ftS dehaulteur. quanS w brandÌ~

ira~hdS1a leur th Ìæßé d'ëftn debost, íljc affiedde] u~ laPl4ñChC •ÃiLt~~ ont ken .d'ij~.

lagçš çks tres mænie Tes: dc brandillages pour les peurs Èr?flmnu; 1m 4, ſt choſe ~móufr puca~

Enfants ñll4 mats

+ Récent : Dictionnaire XVIII (trévoux)

ACÈ, ſ. f. *Ace.* Ville de Phénicie, dans Strabon & dans Étienne. Ce fut depuis Ptolémaïs. Voyez ce mot.

ACÉE, ſ. f. Ce mot ſe diſoit autrefois pour *bécaſſe* : il vient d'*acceia*, qui vient d'*acus* à cauſe du long bec de la *bécaſſe*.

ACEPHALE, ſ. m. *Acephalus.* Proprement, qui n'a point de Chèf, de l'*a* privatif, & de κεφαλὴ tête, Chèf. On a donné ce nom, 1° à ceux qui dans l'affaire du Concile d'Éphèſe ne voulurent ſuivre ni S. Cyrille, ni S. Jean d'Antioche : 2° à des Héréticques du V. ſiécle qui ſuivirent d'abord Pièrre Mongus, ou Moggus : puis l'abandonnèrent, parce qu'il ſouſcrivit au Concile de Chalcédoine. Ils ſuivoient les èrreurs d'Eutichès. Et ſous l'Empire de Juſtin, les Sèctateurs de Sévère d'Antioche, & généralement tous ceux qui ne voulurent pas recevoir le Concile de Chalcédoine, furent appellez *Acephales*. Quelques uns prétendent que ce nom ſignifie *héſitant*, & que parce qu'ils tenoient la neutralité pour les decrèts du Concile de Chalcédoine,

Résultat de TexteBridge (Xerox)

A C ~, C f. *Ace.* VilI~ d~ Øhénicie~ dans Stra~, ou & dans t~tiei~ne. Ce fut depuis Ptôl~maVs. Voyez ce mot.

A *C _~.* E, fi *f.* Cé mot fe dif~ir autrefois pour *bJcAJJ'e* : il vient d'ac~ qui vient *d'acus* ~i c~ufe du long Lec *de1ab~4ffe4* C E P H A LE, fi n,. *Acephalus*& Proprement, qui n'a poit~t de Chef, de l'a privatif, &de ~ u4tête, Chéf. On ~i donné ce norti, 10 ~ ceux qui dans t'affaire du Concile 4~Éphèfe ne votl— lurent fiuivre ni S. Cyrille, ni S.Jean d'Antioche *2 z*^{0} ~ desH~r&. tiques ~Iu V~ fi~cle qui fuivirent d'abord Piêrre Mongus, oU Moggiis~ ~ùis i'abandonu~renr, parce qu'il foufcrivit au Condle de Çhalc~doi~ie, Ils fuivQîent les erreurs d'~utich~s.Et fous l'Empire de Ju<tin ,les S~ateurs de S~vêt~ d'Ant~oche,&g6-4~êralemenit tous ceux qui ne voplurenr pa~ recevo~de Concile de Ch~lêédoin&, furent appdlle~ *4cephales.;* Qpelque~ uns pr&~ndenr ~uec~noni fwnifie *hé/itant,* & que parce qu'il~te~ hoient la heutralitd pour l'es decr~rs ditConcile deCh~1cédoine~ qu'ils ne *fed&êrrù~nd~entd* rien qu'ils h~fitoienrquancipn les préflbit , ~1s **furent appeliez** *Acephales* c^{3}ea-~-dire, *hófltatis.* Mais Vautre opinion ea ph~ *vraiè, & Ac~pbale* n'~ pointce fens. Voyez Bolland. T~ I. Anaaafûs le ~ibItothéca~r~ appe!Ie cette ~emption de la frdfd~aion ~u P.îtiiarche, Atinocéphalie, *~4utocepha1~a; ;* on a ~ppe1 1~ *Acephat~s* le~ Clercs *qw* ne vivoient pàs fous'la Difcipline Écci~fiafUqÙe d'un ~v~queAfidore *PcEcc!ef off. La;. 11J.* LesConciles de Majtence, Can. *22.*. de Meaux I an~~.y .Can.~7. de PaAsCan~ *i* o~ de Pavie cri 8yo.Can~ *i* ~ .&c. ont fait différens r~glernens c~n~re ces *ClercsAcephales.* 'On *cii* trônve ericore dans *tes* Capitul aires de Ch~trlcs le Chais. Ve, **L. VI. C.5** 7. dari~ 13urchard, **L. Il. C.** z z6. dans kegin~n ~ l'an de Jefus-Chrift S .~ 13aroniùs ~ U~ntiée *i* o~o~ ~iucbert frère dc Thi~rb~rge Concubine dc. Ï~oçhaire, fut appellé *4ce-~ eha7é,* parce q~xe <coinrne di~n(1~ Araiales de Mets û1~

Résultat de Omnipage (Caere)

A C 1:

AC , Γ *f. Àce. 'Ville de* fhénicic;dans ~trabdn & dans 4'tierine. Ce fut depuis Ptôlémaïs. Voyez ce mot:

r C $ E, f f. Cë tnot fe diroit autrefois pour *hécaffe : il* vient d'ac= *cela, qui* vient *d'acus â esufe du* long bec *delahécaJTe.*

â~ C E P H A L E, fm. *Acephâlusi* Proprement, qui n'a point *de* Chef, de l'a privatif, & de macp,~À~ tête, Chè£, 011,1 donné ce 110111, i ° â ceux qui dans l'affaire du Concile 'Éphéte ne voulurent fuivre ni S. Cyrille, ni S.Jean dAntioche' '~ z' â des Hçré~: tiquesdu V; fiécle qui fuivirent d'abord Pierre Mongus, ou 1VIoggus: puis fabandonnérenr, parce qu'il toufcrivit au Concïle de GIzalcédoiàe: Ils fuivoient les erreurs d'putichés.$t fous l'Empire de juftin, les Sédateurs de Sévérë dAntioche , & gé'héralementtoûsceuxquinevoulurémtpas rècëvoii~IéConcile de Châltédoinè, furent appellez *Acephatès.: Q_uclque, â uns* prétendent que-cd *nom fignifie h~ritant , & que parce* qu'ils,tei

Résultat de Finereader (Abby)

A C E .

A C E, f. f. *Ace.* Ville de Phénicie, dans Strabon & dans Etieline. Ce fut depuis Ptolémaïs.^ Voyez ce mot.

A C E E, f. f. Ce mot fe diloit autrefois pour *bécœffe :* il vient *d'acceïa, qni vkntd'acusàcàute* du long bec *àehèéca/e*$_t$

A C E P H A L E, f; m. *Acephalusi* Proprement, qui n'a point de Chef, de 1*4 privatif, & de te^Ai} tête, Chef. On a donné ce nom, i°(à ceux qui dans l'affaire du Concile d'Ephèfe ne voulurent fuivre ni S. Cyrille, ni S.Jean d'Antioché i *2,* ° a' desHéré"-tiqnesdu Vi fiécle qui (uivirenc d'abord Pierre Mongus, ou Moggus S puis l'abandonnèrent, parce qu'il foufcrivit au Concile de Cnalcédoifie, Ils fuivpient les erreurs d'Butichès.Et fous l'Empire de Juffin, les Sèclateurs de Sévère d'Antfioche » & généralement tous ceux qui ne voulurent pas recevoir le Concile de Chsilèédoine, furent appeliez ^*cefialts.* Quelque^ uns prétendent que ce nom fignifie *béfitant,* &: que parce qu'ils te^ hoient la neutralité pour les décrets dttConcile deChalcédoine>

Transcription Assistée par Ordinateur (TAO)

- *Documents imprimés*

- Segmentation
  - Lignes
  - Mots
  - Caractères

- Prototypage / Groupement (clustering)
  - Caractères

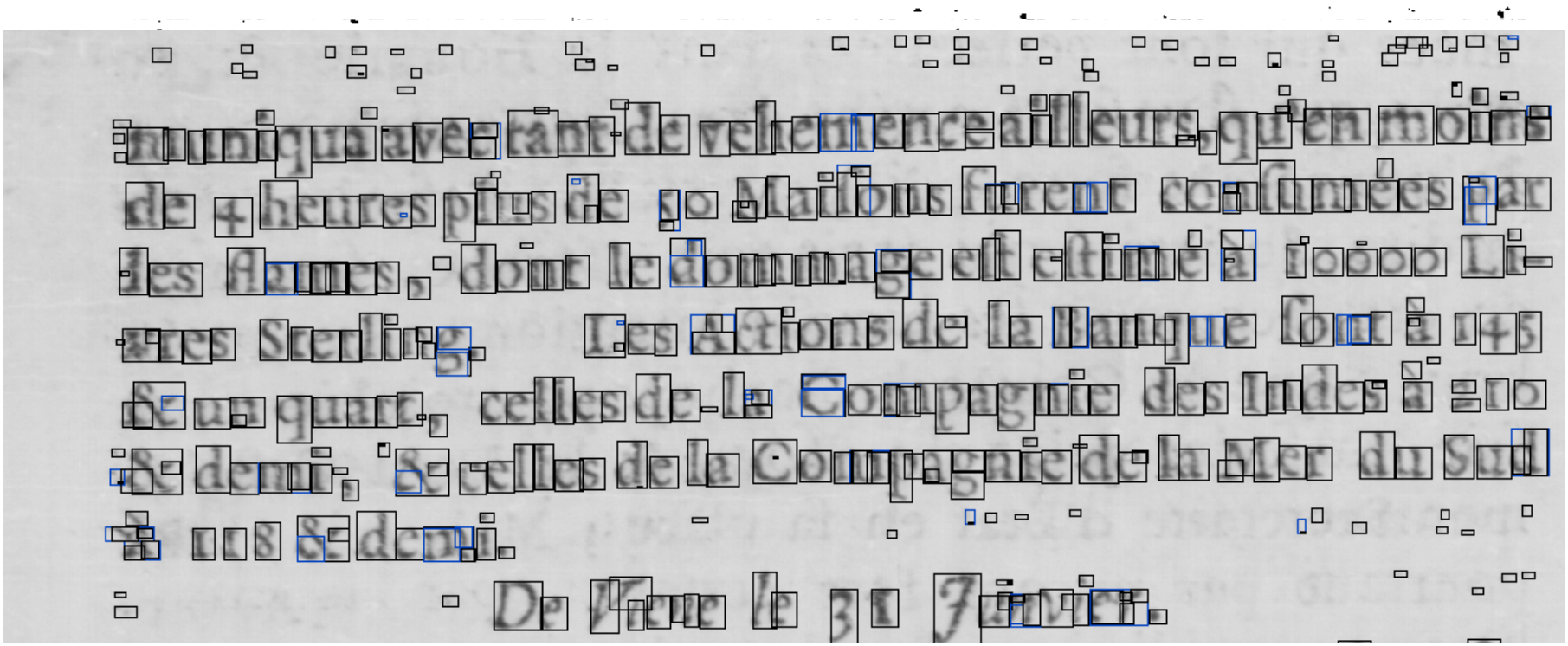
- Étiquetage manuel des groupes

Principes du prototypage

Nouvelle segmentation en caractères guidée par la redondance des formes basé sur les hypothèses suivantes :

H1 : pour un caractère fragmenté, il existe au moins une forme correctement imprimée du même caractère

H2 : pour deux caractères collés, il existe au moins une occurrence de ces caractères qui sont imprimés isolément.



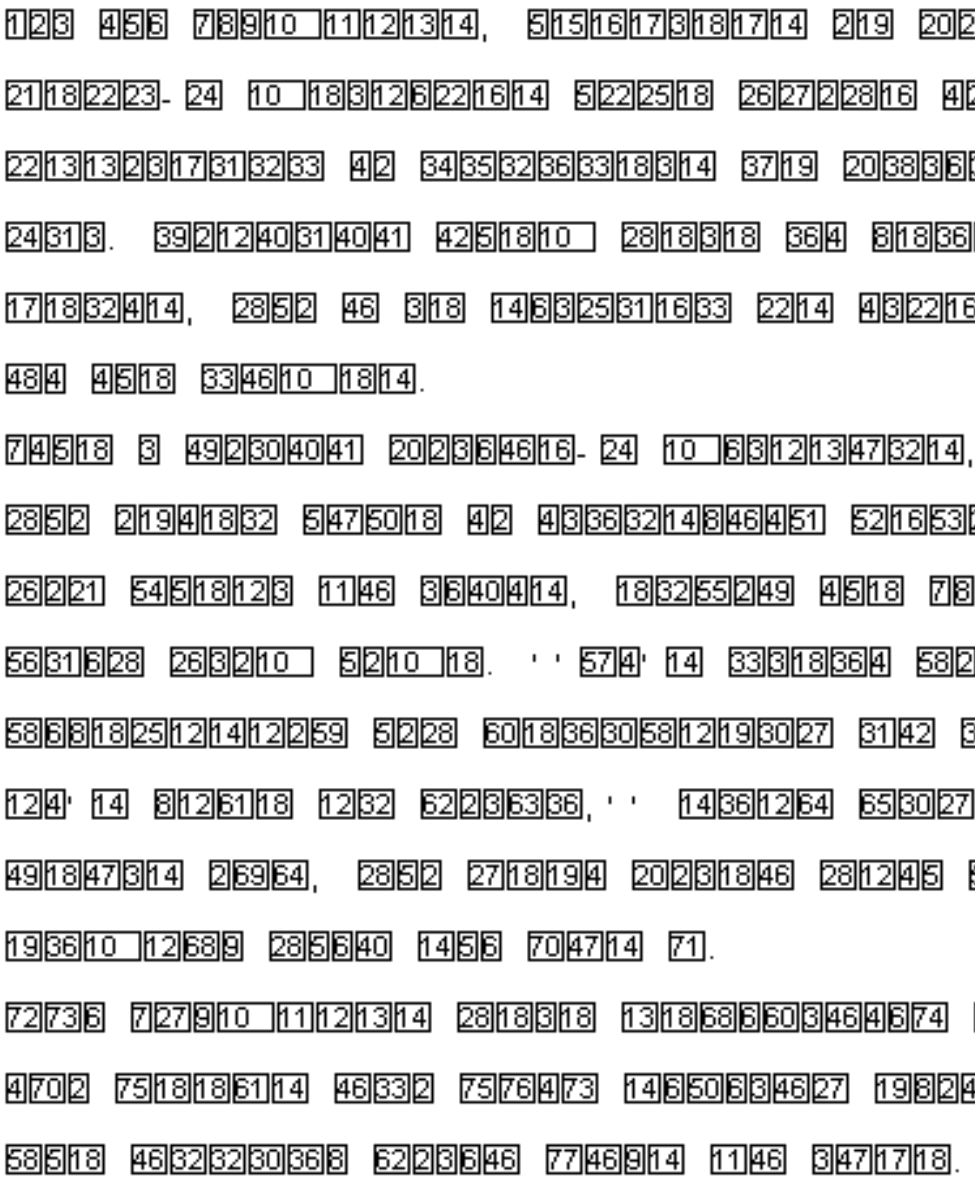
Transcription Assistée par Ordinateur (TAO)

500 caractères | **80 prototypes**

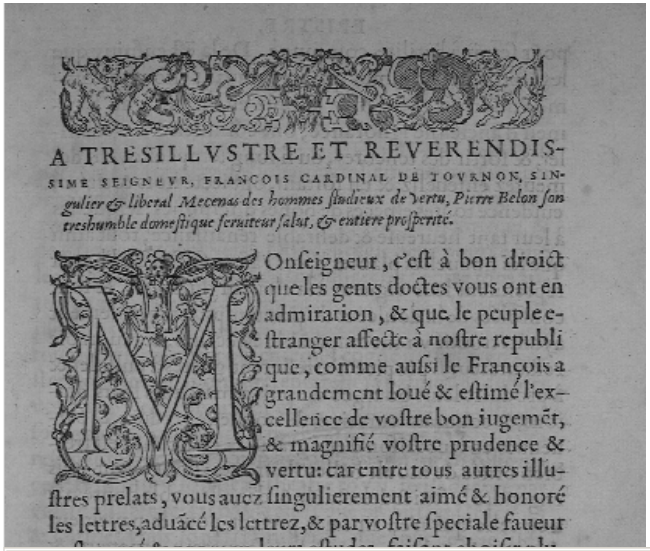
For the Olympics, hundreds of Korean-Americans have flown to Seoul, according to managers of Korean Air. Joining them were at least 50 students, who are serving as translators at the games.

Other young Korean-Americans, who often have to translate English for their parents, enjoy the Olympic view from home. "It's great to see on television how beautiful it is and what it's like in Korea," said Julie Hong, 21 years old, who left Korea with her family when she was 4.

The Olympics were celebrated here two weeks ago with several floats in the annual Korea Days parade. A

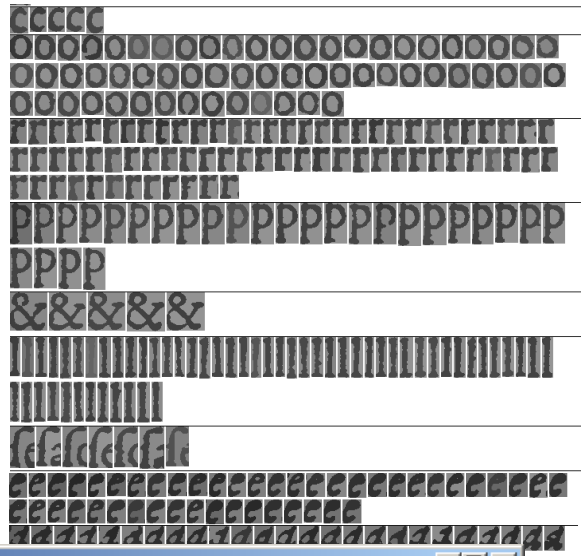


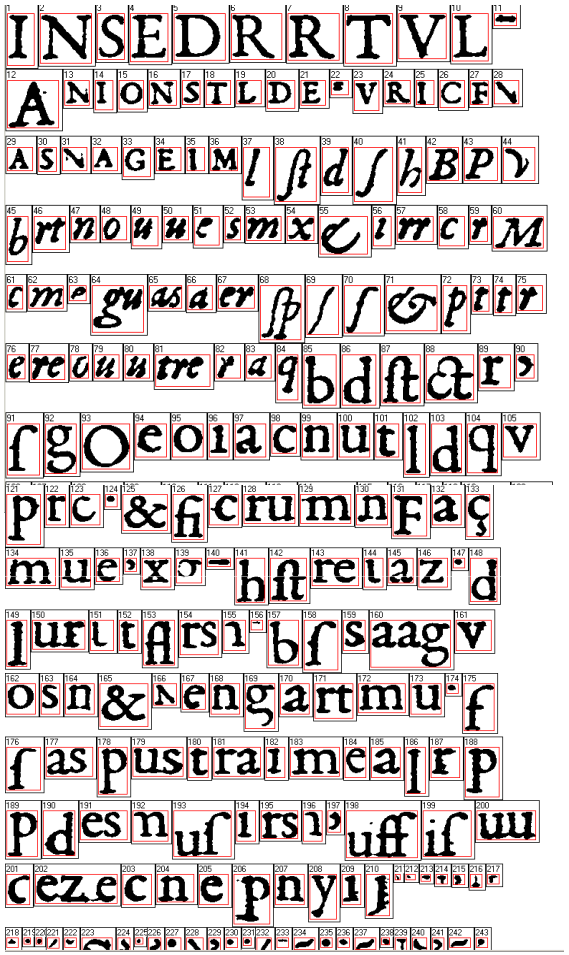
Transcription Assistée par Ordinateur (TAO)



A TRESILLVSTRE ET REVERENDIS-
SIME SEIGNEVR, FRANCOIS CARDINAL DE TOVRNON, SIN-
gulier & liberal Mecenas des hommes ſtudieux de vertu, Pierre Belon ſon
treshumble domeſtique ſeruiteur ſalut, & entiere proſperité.

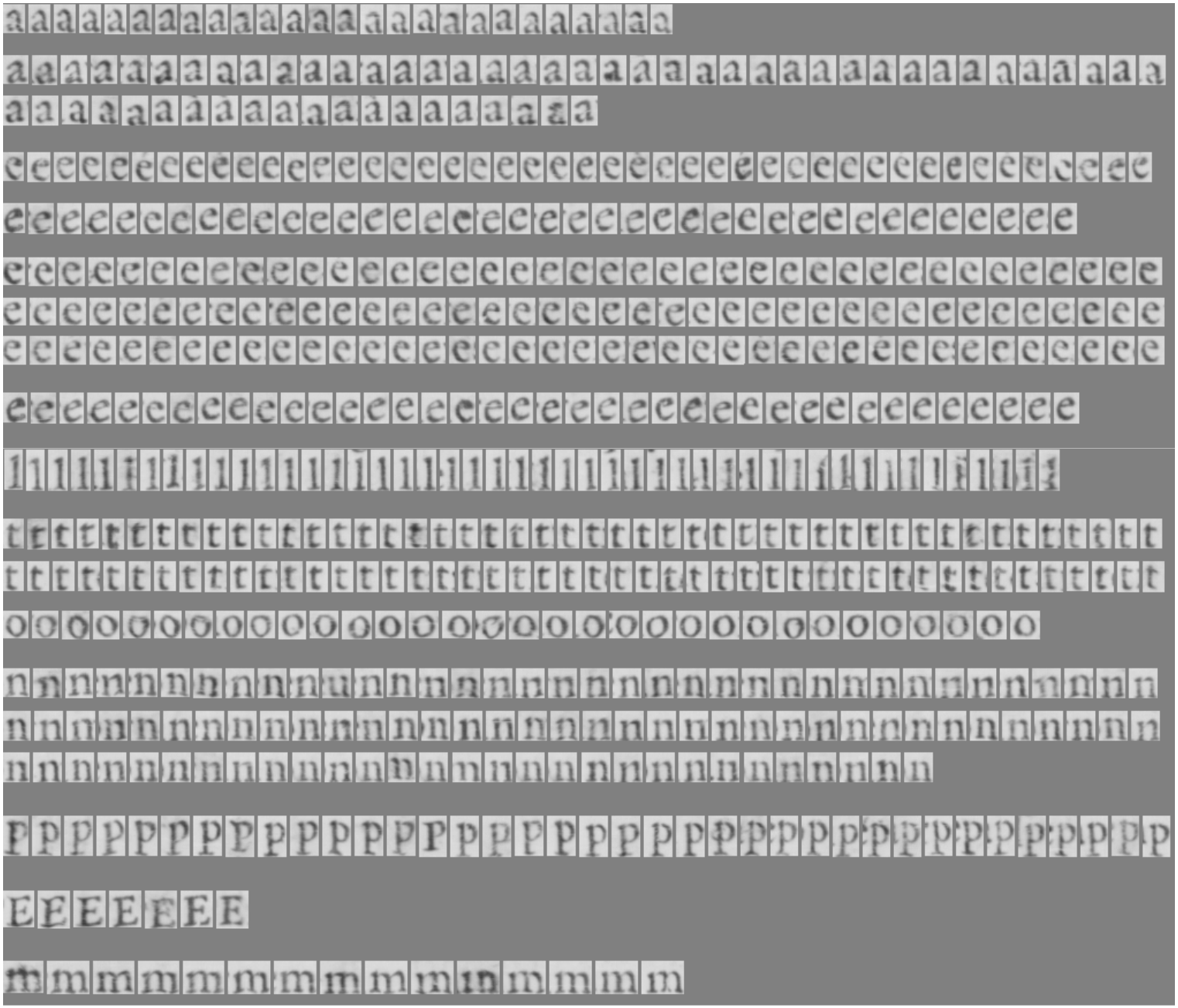
Monſeigneur, c'eſt à bon droict
que les gents doctes vous ont en
admiration, & que le peuple e-
ſtranger affecte à noſtre republi
que, comme auſsi le François a
grandement loué & eſtimé l'ex-
cellence de voſtre bon jugemẽt,
& magnifié voſtre prudence &
vertu: car entre tous autres illu-
ſtres prelats, vous auez ſingulierement aimé & honoré
les lettres, aduãcé les lettrez, & par voſtre ſpeciale faueur





La saisie de 2% des caractères, réalisée en 6 heures, suffit à reconstruire un livre du XVI transcrit pour une recherche par mot

TAO vs OCR (TAO + robuste)

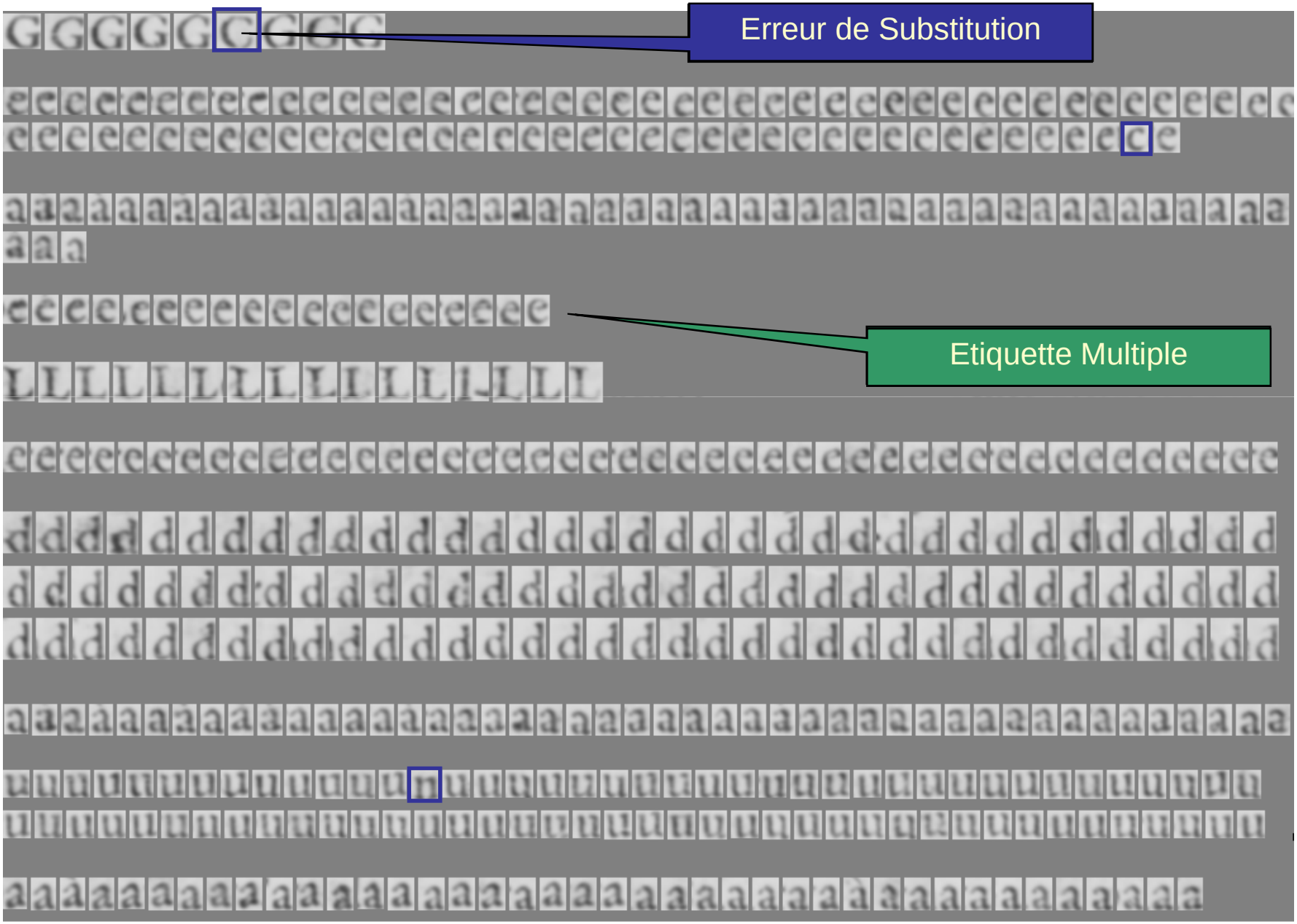


Transcription parfaite en quelques heures

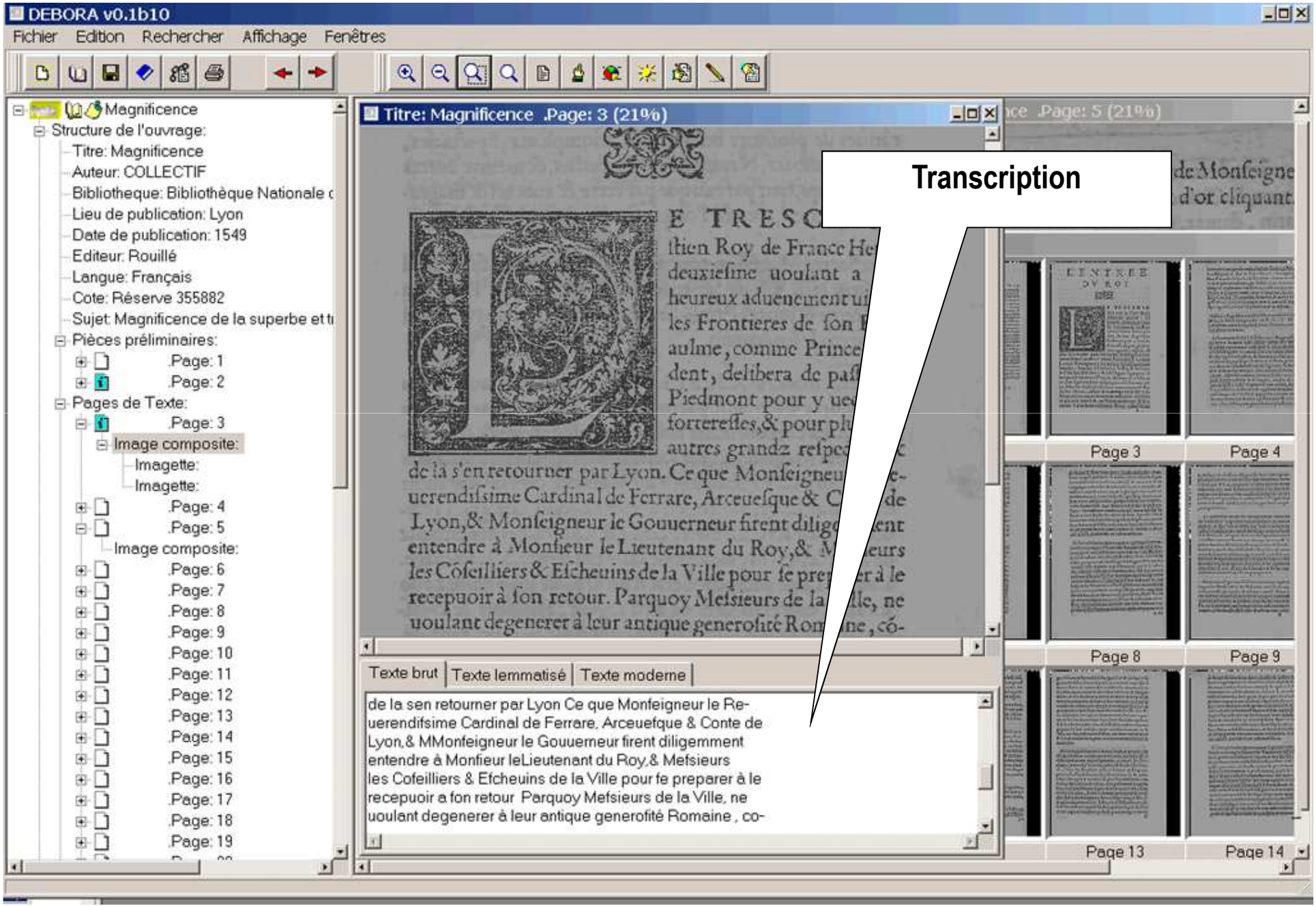
- TAO implantée dans DEBORA (projet Européen 1998-2003)

Comment étudier les erreurs de la TAO ?

En affichant toutes les formes par prototype

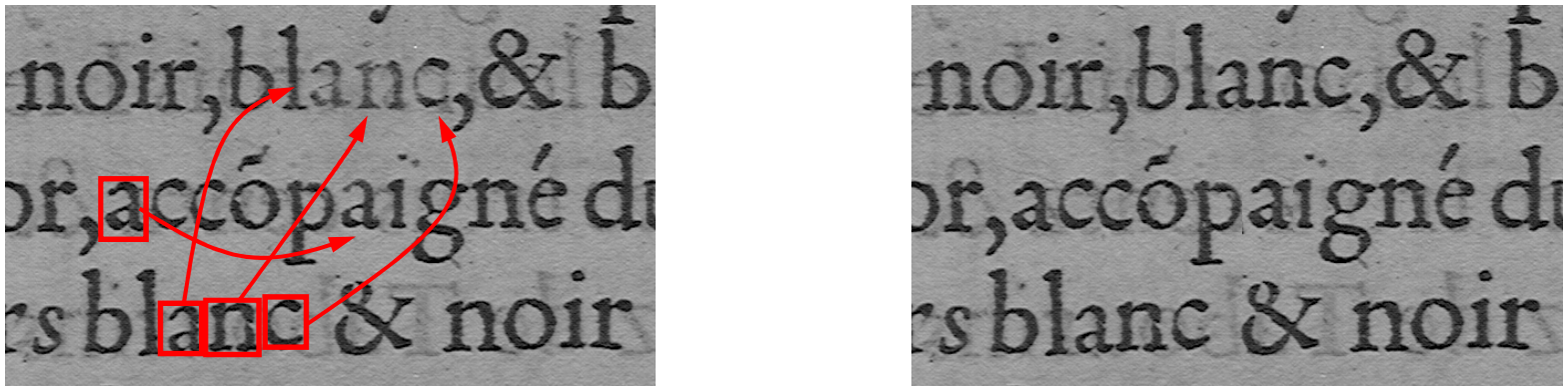


Transcription parfaite en quelques heures

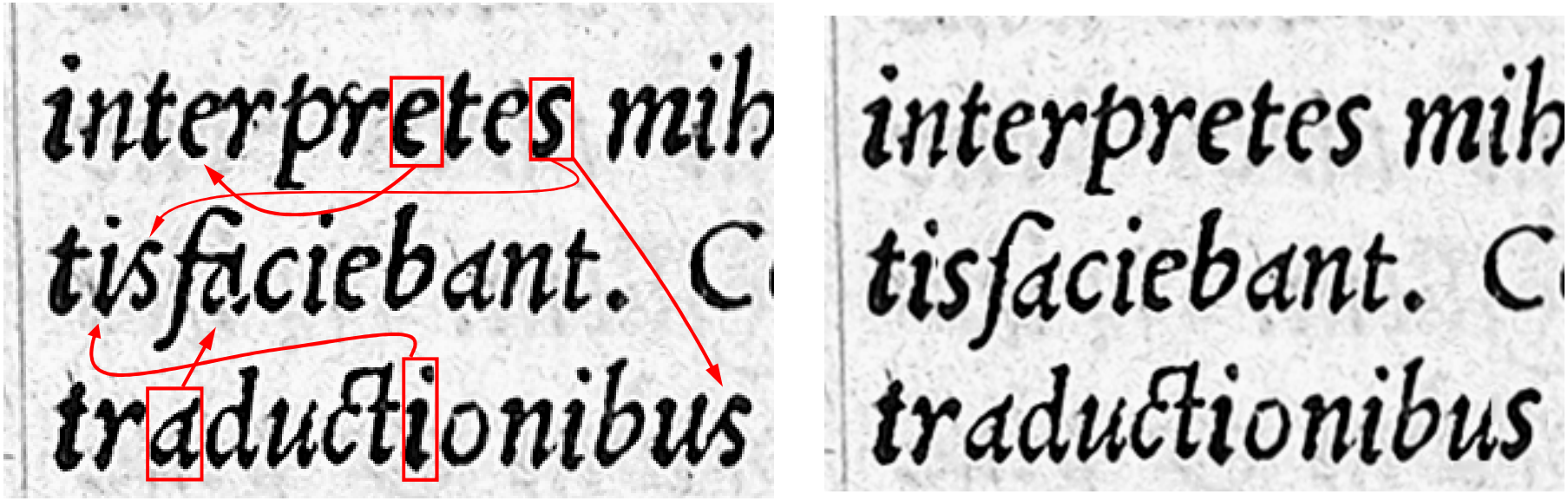


Principes de reconstruction

Phase 1 : Recouvrir les caractères cassés par la forme entière similaire.



Phase 2 : Recouvrir une forme collée par des formes isolées similaires

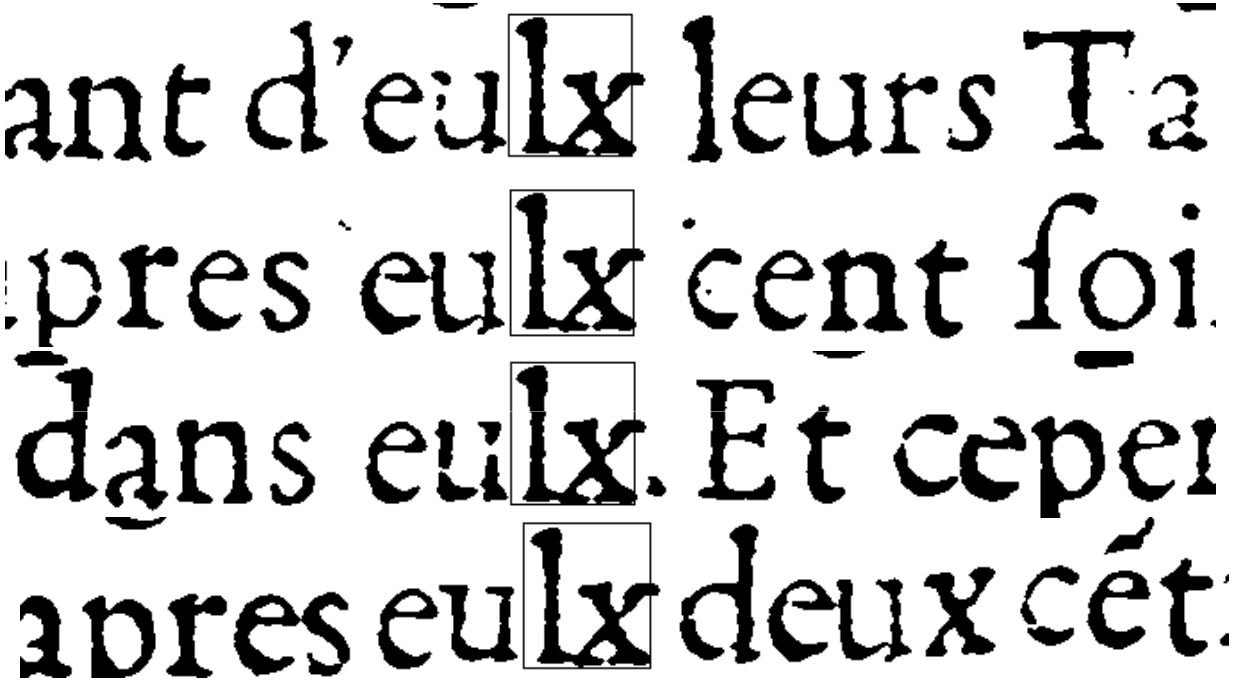


TAO: ouverture

- Compression
  - Dictionnaire de formes
  - Compression JPEG du fond (papier)
- Restauration
  - Substitution des caractères par leur prototype
- Classement des mots par typographie
  - Les polices et différentes typographies (gras, italiques) génèrent des prototypes différents
- Reconnaissance des mots par décodage
  - Dictionnaire

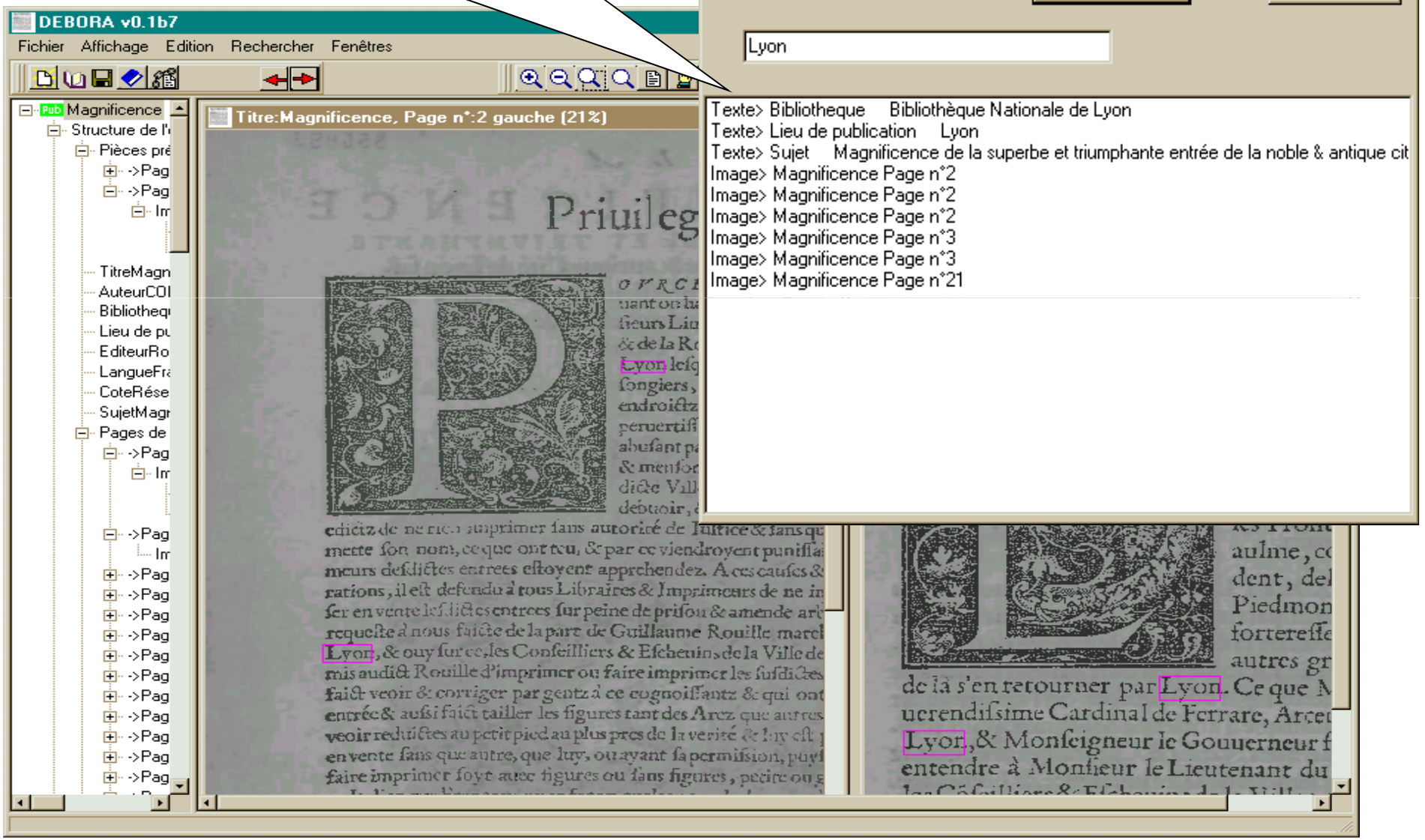
Découvertes grâce à la TAO

1) Les typographes de la Renaissance composaient les mots fréquents avec les mêmes plombs



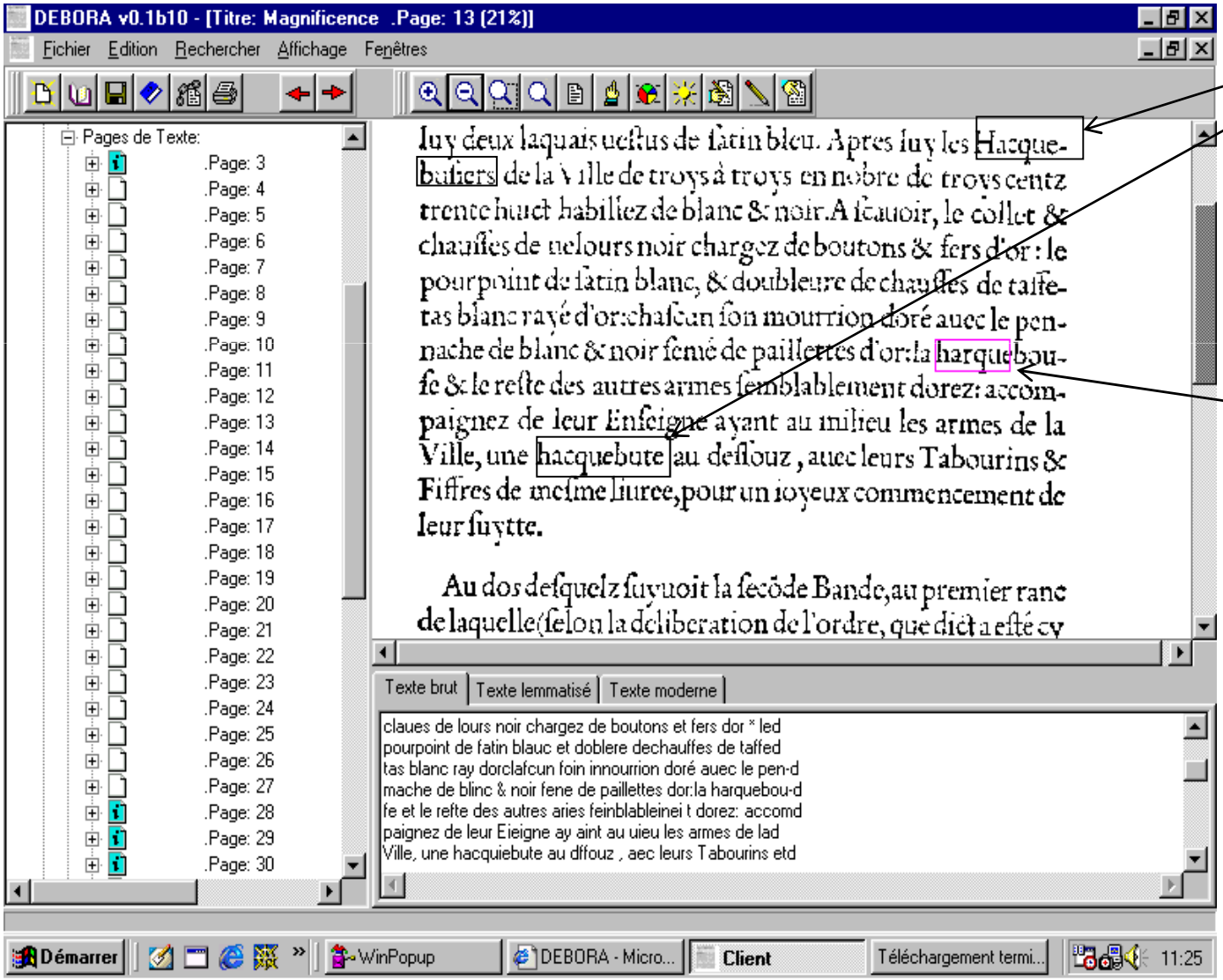
2) Une augmentation soudaine du nombre de nouveaux prototypes de caractères témoigne d'une « rupture » dans la fabrication comme l'insertion de nouvelles pages crées dans d'autres ateliers et nous renseigne sur la fabrication des livres

Résultat d'une recherche par mot dans les données et les métadonnées



Limites d’une recherche de mots dans une transcription non lemmatisé

Recherche de "harque" ramène « harquebouse » mais pas « hacquebute » ni « hacquebusiers ». Pas de recherche imprécises avec substitution de lettres ou de jokers !



Mots non retrouvés par le logiciel de recherche

Mot retrouvé par le logiciel

Recherche de mot-image (WS)

Sur les documents non transcriptibles

- Proposée par la communauté de reconnaissance vocale « Tohlicek et al, 1989 », « Paul et Rause, 1990 ».
- Utilisée comme une alternative à l'OCR
  - Pour des expériences sur l'indexation des documents manuscrits « Kuo et Agazzi, 1994 », « Manmatha et al, 1996 », « Rath et Manmatha,2003 » ,« Rath et Manmatha, 2007 ».
  - Pour l'extraction d'informations des documents imprimés « chen et al, 1993 », « Chen et al, 1995 », « Yue et al, 2004 », « Jawahar et al, 2004 » et manuscrits « Zhang et al, 2004 », « Srihari et al, 2005 », « Srihari et al, 2006 ».
- Plus efficace que l'OCR sur les documents dégradés et manuscrits.

Recherche de mot-image (WS)
Sur les documents non transcriptibles

- Principe général:
  - Prétraitement:
    - Enlever le bruit, corriger l'inclinaison.
    - Dépend de la qualité du document sur lequel on travaille
  - Analyse (segmentation et extraction de caractéristiques):
    - Etape de segmentation du document:
      - Segmentation du document en lignes.
      - Segmentation du document en mots.
      - Segmentation du documents en mots puis en caractères.
    - Extraction des caractéristiques:
      - Chaque mot est décrit par un vecteur de caractéristiques ou une description structurelle.
  - Appariements:
    - Avec/ sans apprentissage
    - Appariement d'images (similarité, métrique)

Recherche de mot-image (WS)
Sur les documents non transcriptibles

- Les techniques de segmentation se divisent en deux groupes:

SANS SEGMENTATION: Le document tout entier est considéré comme une seule image et aucune segmentation n'est appliquée (Leydier et al, 2005), (Leydier et al, 2007).

- Les images de documents sont analysées globalement par une recherche de zones d'intérêt, de guide, d'éléments structurants: SIFT, Gradient cumulés, orientations par Curvelets…) puis une recherche séquentielle est produite.

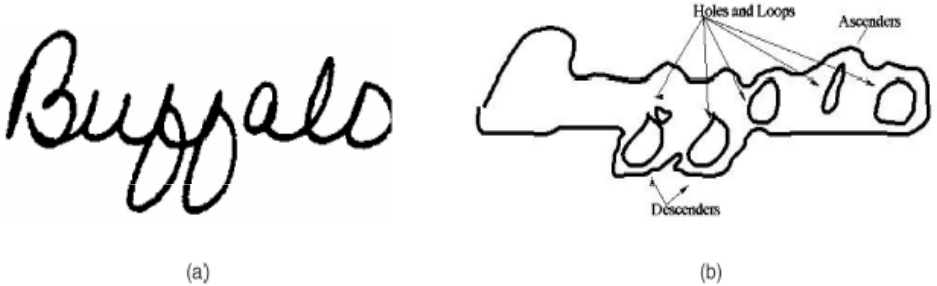
AVEC SEGMENTATION: le document est considéré comme une collection de sous-images (mots et caractères).

- Les documents sont segmentés en sous images (CCs).
- Méthodes top-down vs bottom up
- Méthodes globales (holistiques: ligne et mot) vs locales (analytiques: caractère)
- Appariement séquentiel entre mot –requête et mot(s)-image

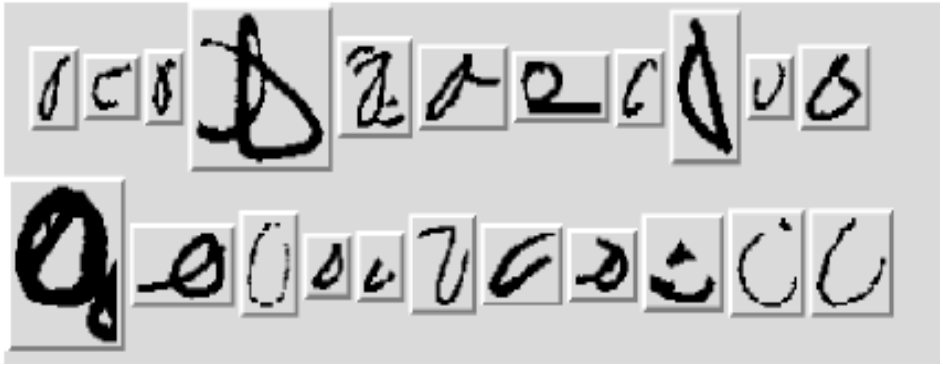
Recherche de mot-image (WS)

Sur les documents non transcriptibles

- Le mot vu dans sa globalité (*Madhvanath et al, 2001, Lavrenko et al, 2004, Madhvanath et Govindaraju, 2001)*



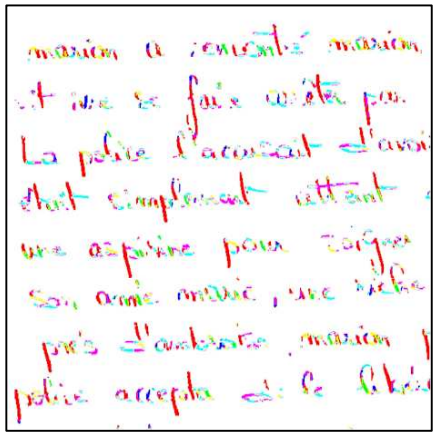
- Le mot vu comme une collection de sous unités (utilisation d'un lexique), *Gatos et al, 1997, Madhvanath et al, 2001.*

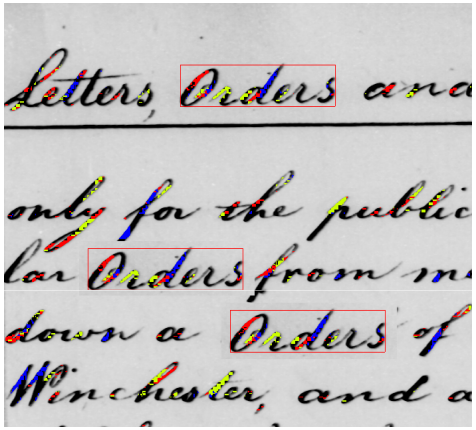


Recherche de mot-image (WS)

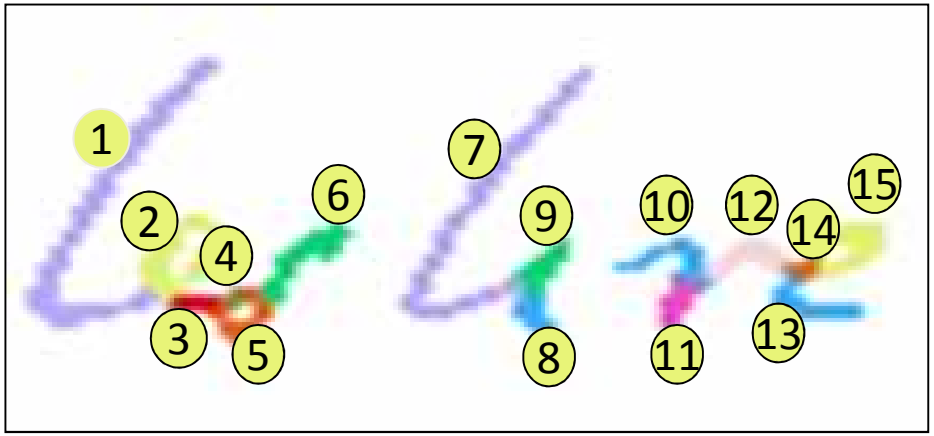
Sur les documents non transcriptibles

- Le mot vu comme une collection d'orientations *(Joutel-eglin, 2008)*





- Le mot vu comme un ensemble de graphèmes, (Tayeb, Daher 2010).



Recherche de mot-image (WS)

Sur les documents non transcriptibles

Techniques:

- Word Shape/HMM - (Chen et al, 1995)
- Word Image Matching - (Trenkle and Vogt, 1993; Hull et al)
- Character Stroke Features - (Decurtins and Chen, 1995)
- Shape Coding - (Tanaka and Torii; Spitz 1995; Kia, 1996)
- Gradient cumulés (Leydier, 2006)
- Code Book (Tayeb 2007, Daher 2010)

Applications:

- Navigation rapide de documents non transcrits (pas de version “texte- ASCII” disponible)
- Accès aux documents manuscrits

Evaluation de systèmes :

- Gribouillis et OCR (DeCurtins, SDIUT 1997, Yue et al, 2004): comparaison système de reconnaissance de mots avec système OCR, le système de word spotting est plus robuste.

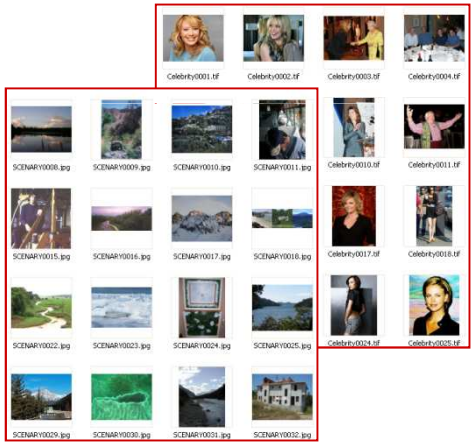
Partie 3

RECHERCHE D’INFORMATION

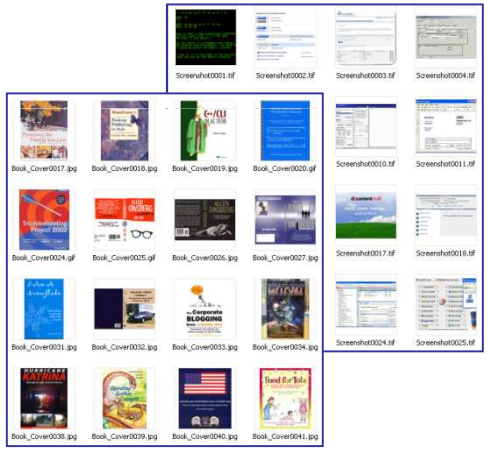
CLASSIFICATION, INDEXATION, SIMILARITÉ
APPARIEMENT ET RECHERCHE D’INFORMATION

Classification d’images par similarité

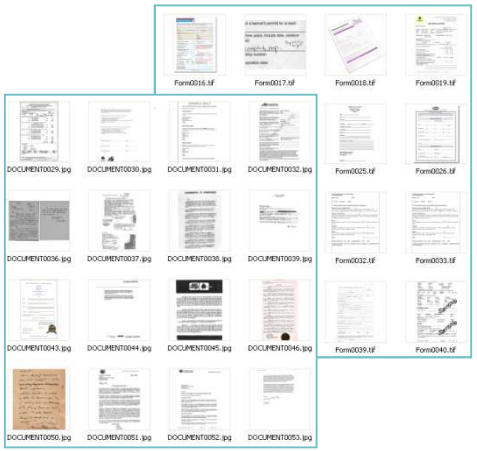
- Une grande variété de contenus
- Distinction: documents contenant ou pas du texte



Images



Images et texte



Images de Document

Indexer les images selon quels critères?

- La structure: indicateur important de contenu
  - Courriers d'entreprises, documents administratifs, financiers
  - Journaux, sommaires…

- Présence d'informations: connaissances a priori des contenus
  - Présence de manuscrit?
  - Bruit, résolution spécifique?
  - Couleurs, niveaux de gris?
  - Informations implicites de structures?
  - Quelles caractéristiques discriminantes?

Quels liens entre les processus d'indexation: images naturelles et images de documents

Images de traits

Les images de traits =
- une sous famille des images naturelles ?

- un cas particulier ?

- des données différentes pas nature ?

Images naturelles

Quelles caractéristiques ?

Images de traits

- *Accumulation de motifs et détails structurés*
- *Géométrie, topologie, résolution, échelle…*
- *Des bandes de fréquences ciblées (en particulier, images de texte)*
- *Une information plus localisée liée au tracé*
- *Séparation forte forme / fond (information / environnement)*
- *Prédominance du segment (1D)*
- *Intention de l'auteur*

Images naturelles

- *Couleurs ou niveaux de gris répartis avec cohérence spatiale sur des voisinages locaux*
- *Couleur, texture, forme…*
- *Accumulation de fréquences équiprobables*
- *Information plus diffuse*
- *Tout peut être potentiellement information*
- *Prédominance de la surface (2D)*

Quelles caractérisations ?

- **Dépendantes de l'objectif :** ***CBIR, catégorisation, restauration***

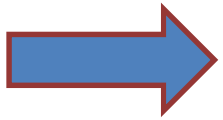
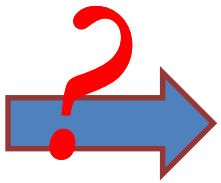
- **Images de traits**
  - ***Caractérisation locale de formes élémentaires (lettres, graphèmes, segments, divers traits…)***
  - ***Agencement global des formes locales (textures par textons)***
  - ***Caractérisation par blocs***
  - ***Forte composante géométrique***
  - ***Importance des détails, de la résolution …***
  - ***Invariance à la rotation selon contexte***

- **Images naturelles**
  - ***Informations locales, globales, par régions …***
  - ***Caractérisation fréquentielle, ondelettes***
  - ***Information plus diffuse , souvent moins sensible à la résolution***
  - ***Invariance à la rotation très fréquente***

Approches et objectifs

	Locales	Par régions ou spécifiques	Globales
	Analyse de l'image au niveau pixel ou voisinage	Agglomération de caractéristiques locales Analyse par région	Analyse statistique de la multitude Transformées de l'image entière
Images de Traits	Analyse des contrastes locaux, de points caractéristiques…	Approches par régions d'intérêt, et fenêtre glissante Analyse de la géométrie, squelette WordSpotting, logos, LBA…	Agglomération de critères bas niveau, CodeBook de motifs, Fréqu., Ondelettes, Curvelets, Similarité de scripteurs, catégorisation de documents (mise en page)
Images Naturelles	Points d'intérêt, descripteurs et invariants locaux…	Recherche d'objets, Contours (CSS), pattern matching… Visages, bâtiments, plaques…	Analyse fréquentielle, ondelettes, statistiques, texture… Classification sémantique, catégorisation d'images…

Quelques pistes …

- Images naturelles  Images de traits
  - Extraction de points d'intérêts type SIFT (Word Spotting)
  - Caractéristiques Curvelets, Hermite (Reco. Scripteurs / Classe)
  - Analyse de texture / Texte (Reco. Scripteurs / Classe)
- Images de traits  Images naturelles
  . Analyse par composantes connexes élémentaires (Word Spotting, Reco. Scripteurs / Classe)

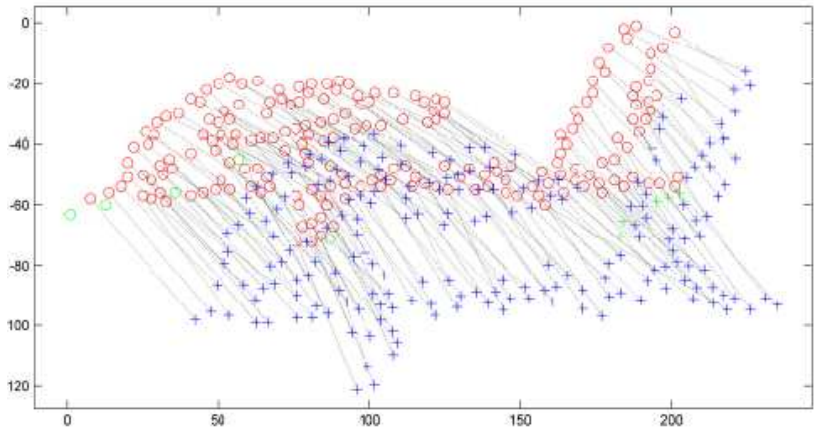
Mais le passage de l'un à l'autre n'est pas systématique !!! Il faut s'adapter au contexte qui constitue un environnement contraignant et impose de nombreuses limitations

QUELQUES EXEMPLES

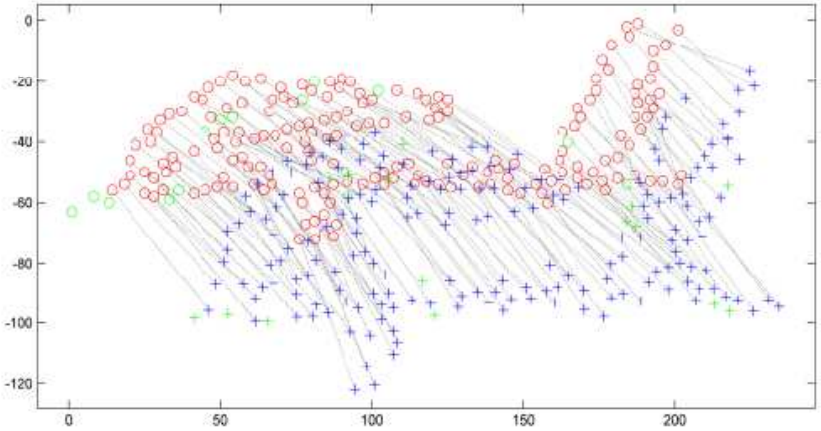
Appariement de formes

(a) (b)

(d) (e)



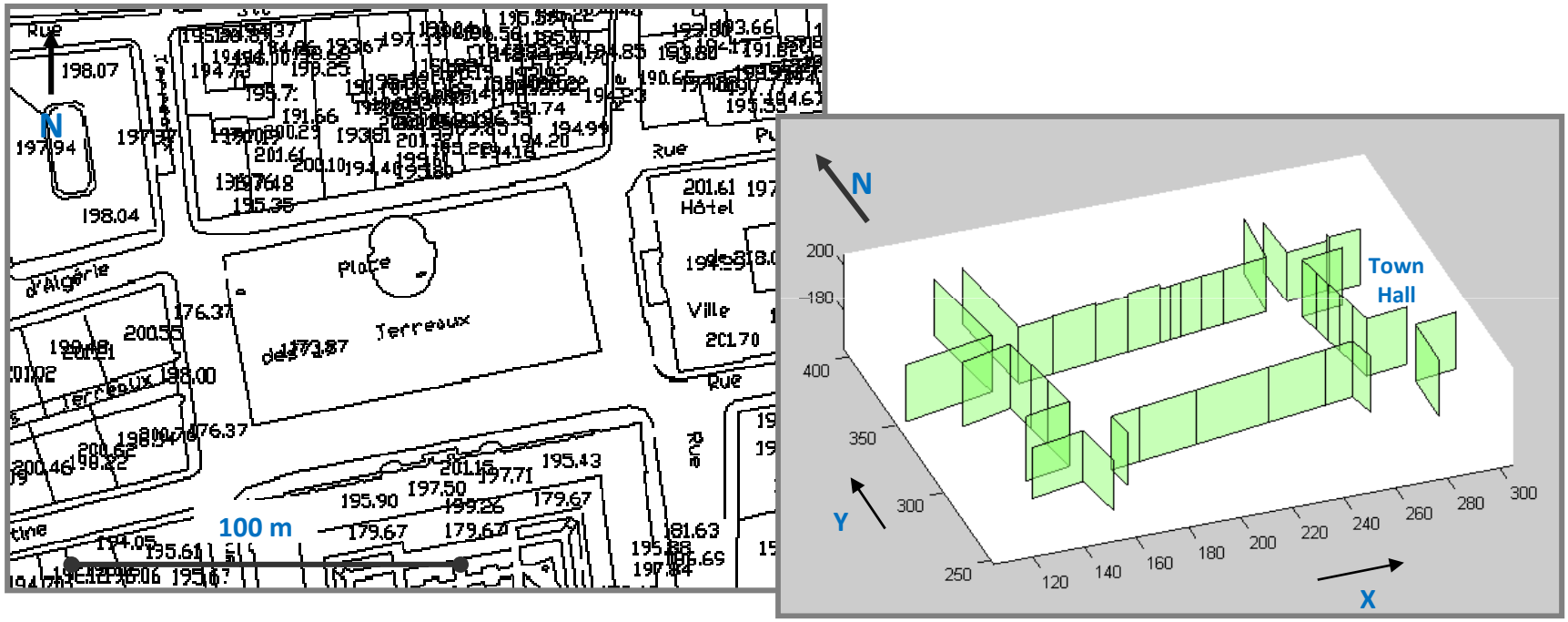
(g)



(h)

Illustration of signature matching using shape contexts and local-neighborhood-graph

Points d’intérêts type SIFT

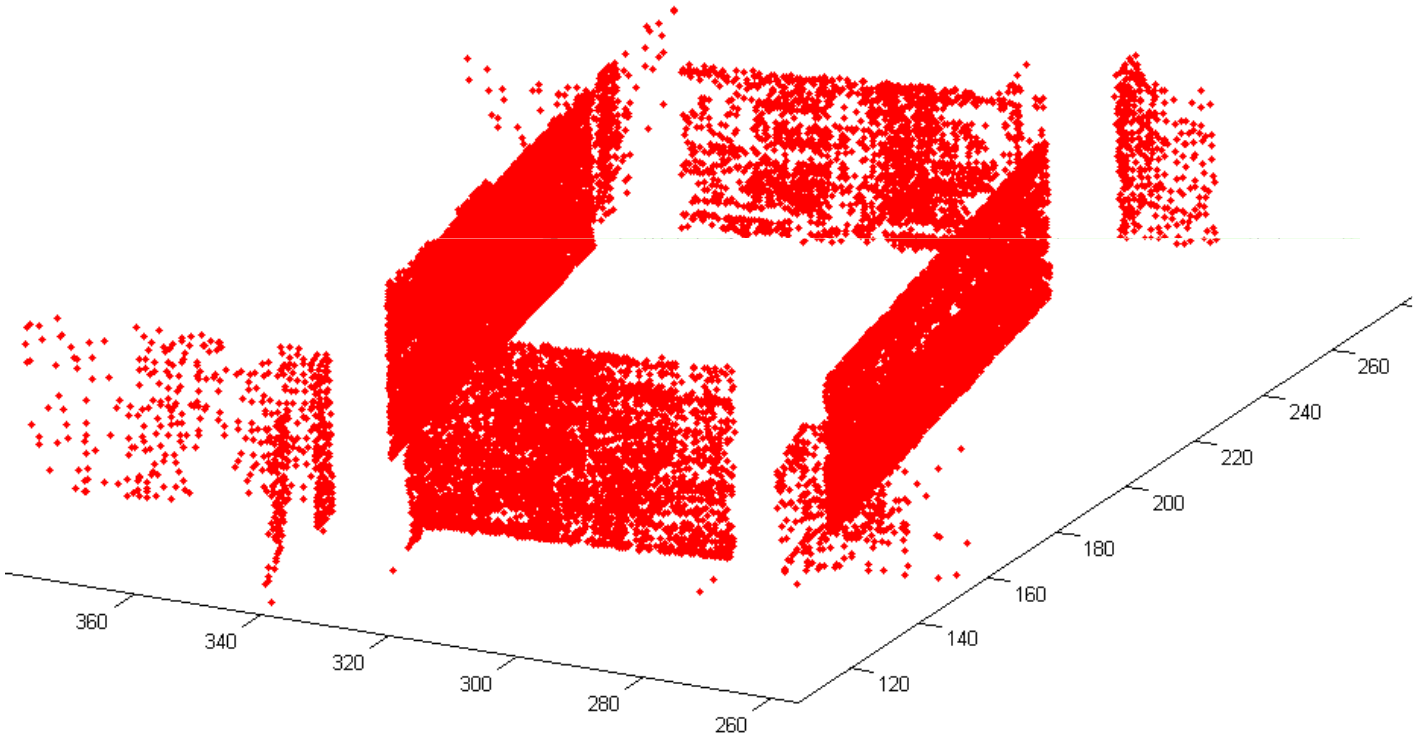


Points d'intérêts type SIFT

36.000 points

Chaque point a :

- une coordonnée 3D
- un déscripteur / signature



Points d’intérêts type SIFT

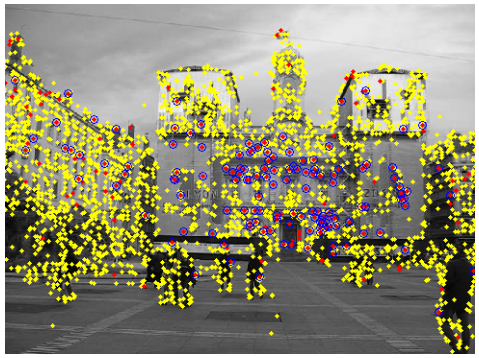
Application sur de la video :

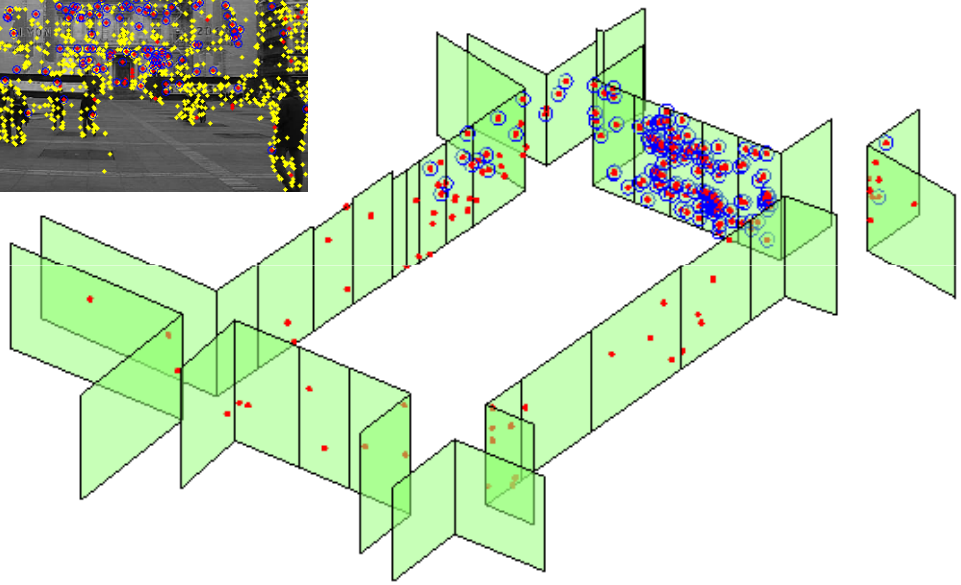
Extraction des points SIFT

Matching avec l‘ensemble de réference

Filtrage des résultats

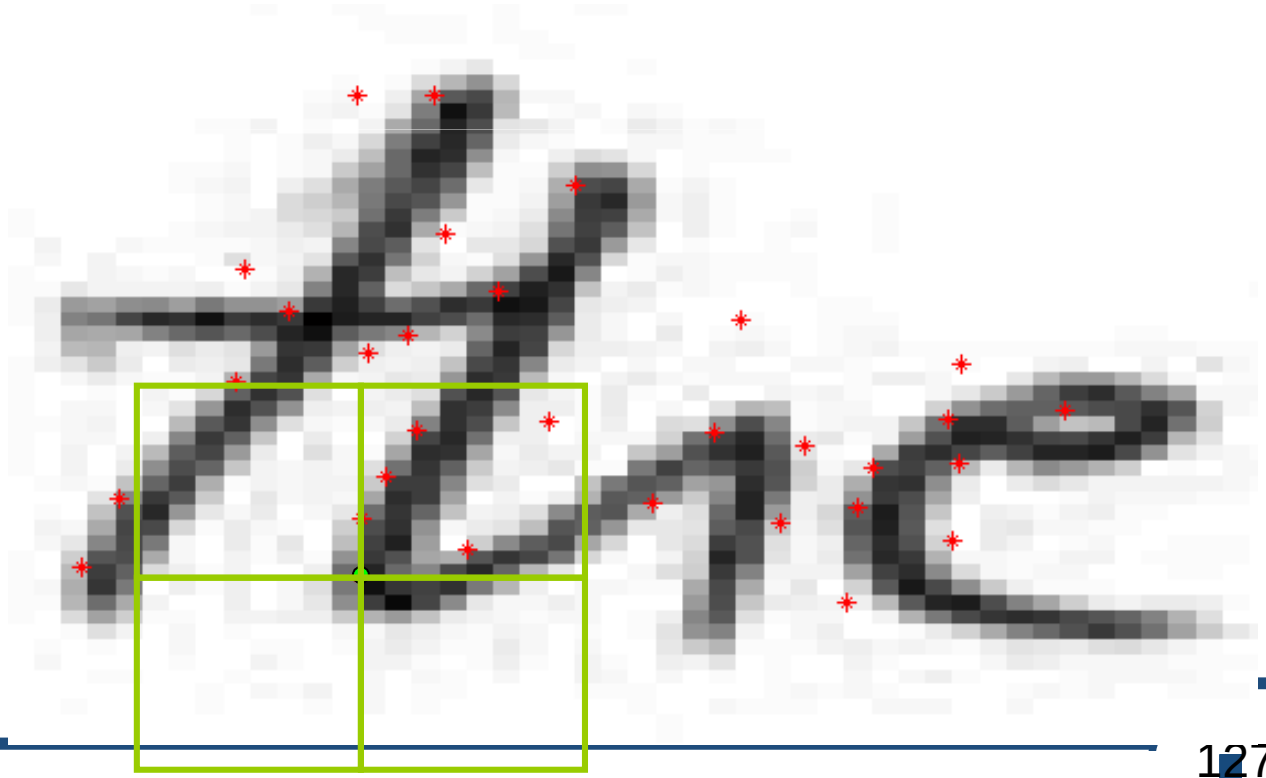
Appariement avec l‘image modèle



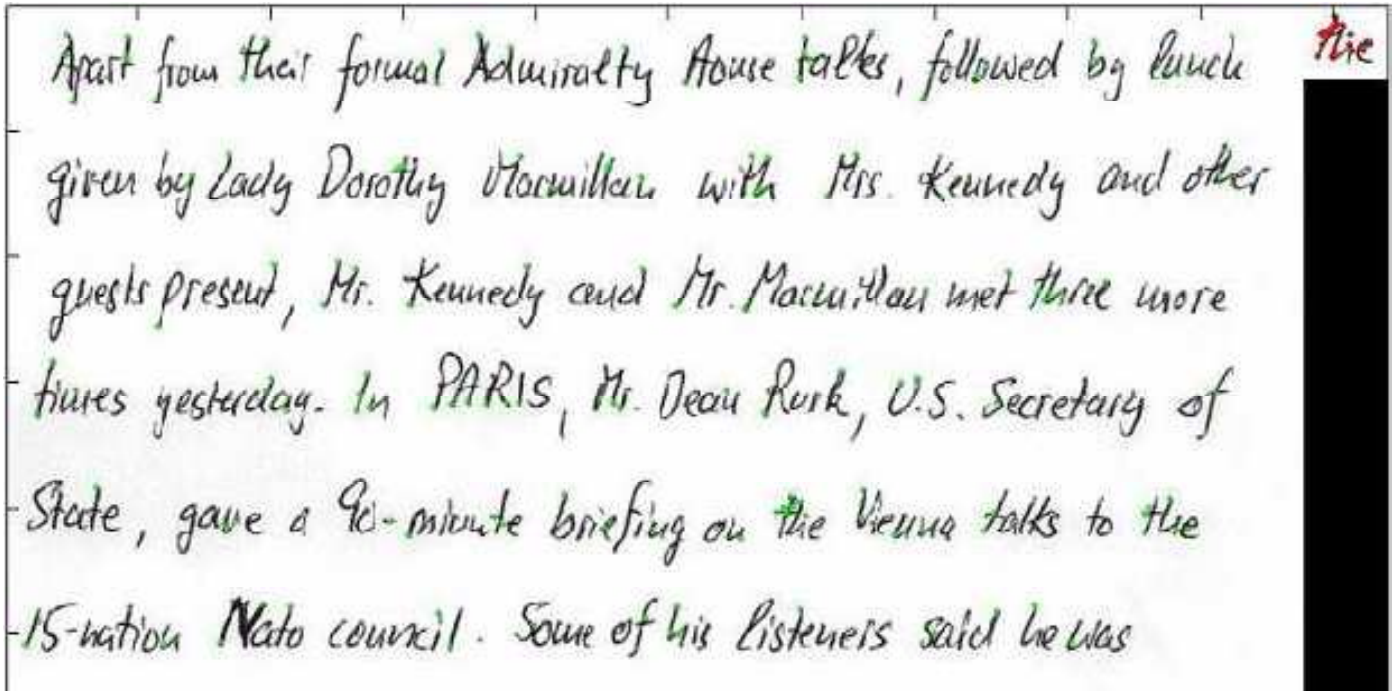


Points d'intérêts type SIFT

- Adaptation aux textes :
  - Limitation du nombre d'échelles, de l'invariance en rotation …
  - Extraction d'information de niveau de gris dans le voisinage des points

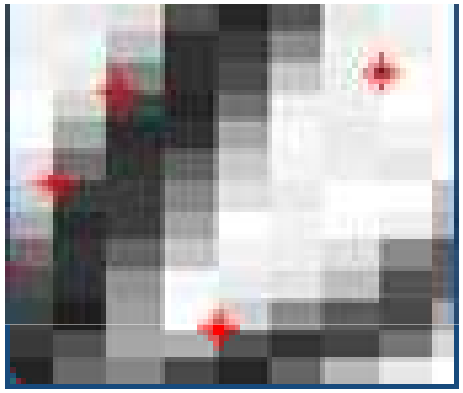


Points d'intérêts type SIFT

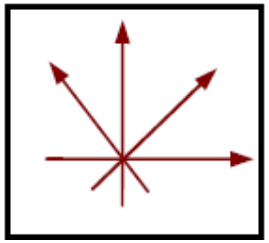


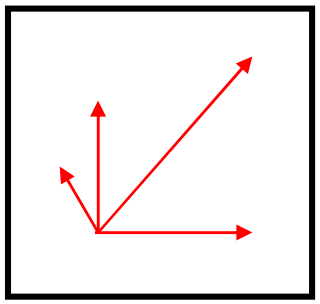
- Limitations :
  - Localisation des points liée aux niveaux de gris et moins à la topologie, géométrie locale
  - Insuffisance du nombre de points d'intérêts

Points d'intérêts type SIFT



Calcul de l'importance des quatre directions par auto-corrélation



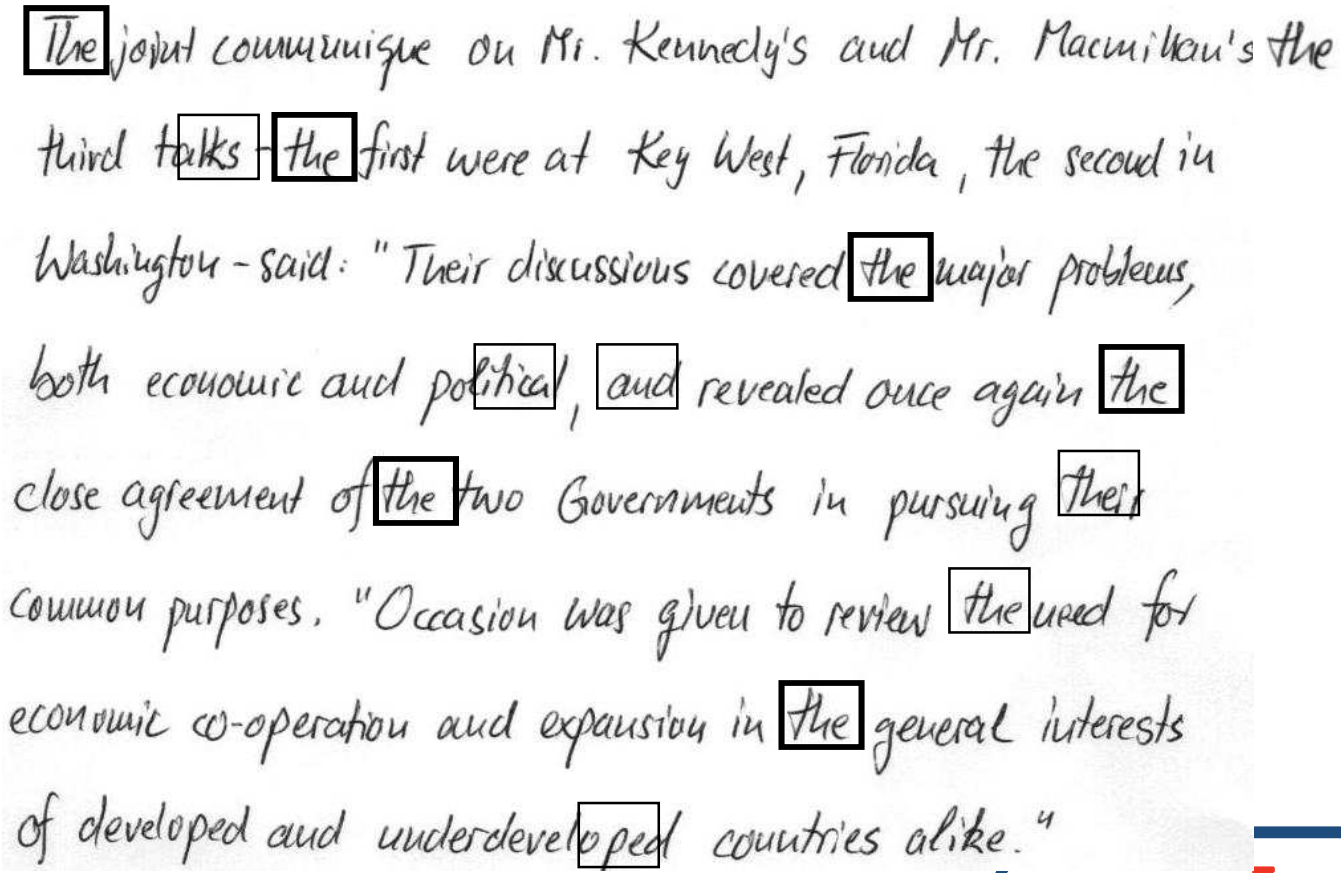


Adaptation de la signature :

Un vecteur de descripteurs
de dimension 16 pour chaque point d'intérêt.

Points d'intérêts type SIFT

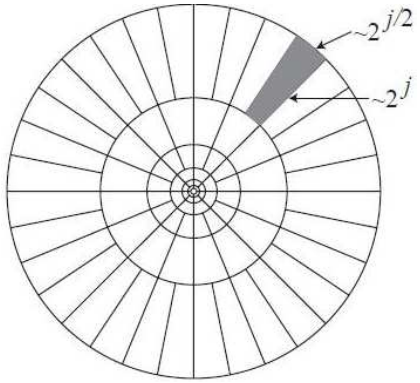
- Résultats intéressants sur la recherche de mots: recherche selon le nombre de points dans l'accumulateur



Analyse par Curvelets

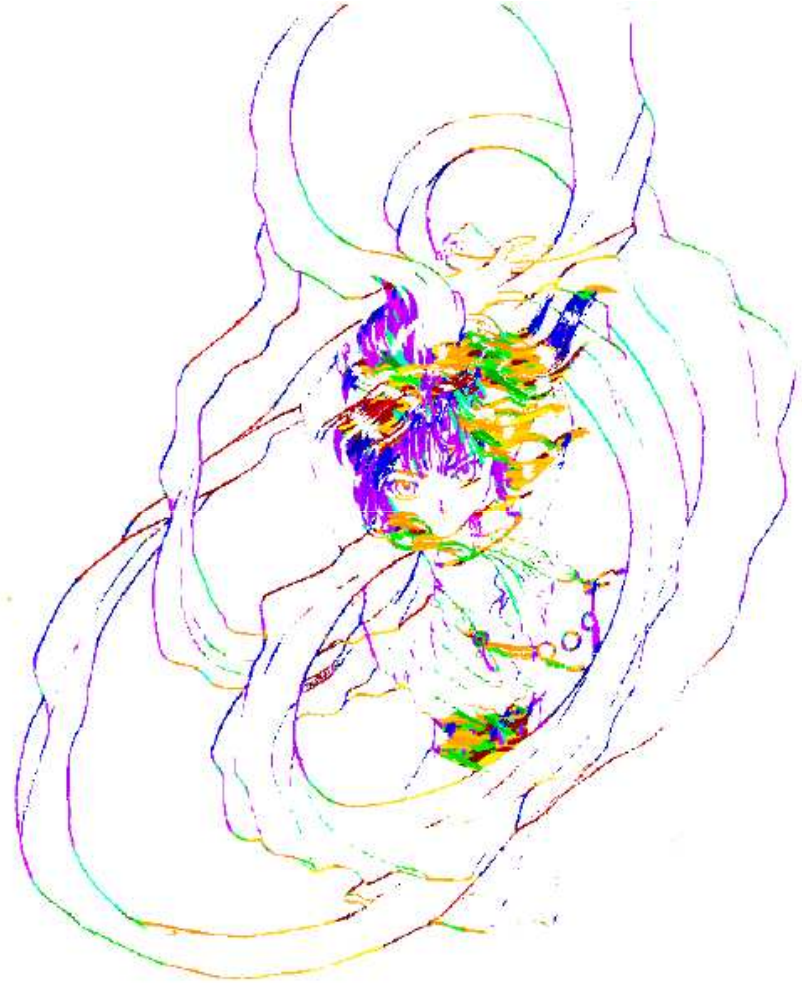
- ***Intérêt pour les informations structurelles des traits***
- ***Une caractérisation intégrant à la fois:***
  - ***Une information locale*** *reflétant les structures internes géométriques et anisotropes des traits : courbure et orientation*
  - ***et une information globale*** *par accumulation d'infos locales*

- ***Un choix porté vers un outil de caractérisation mixte: l'analyse par Curvelets***
  - *Un pavage du plan fréquenciel*
  - *Une transformée multi-échelle multi-directionnelle à 3 paramètres: position, échelle et direction*
  - *Des fonctions de base de transformées localisées à la fois en espace et en fréquence.*



Analyse par Curvelets

Original image



Représentation des orientations

Analyse par Curvelets

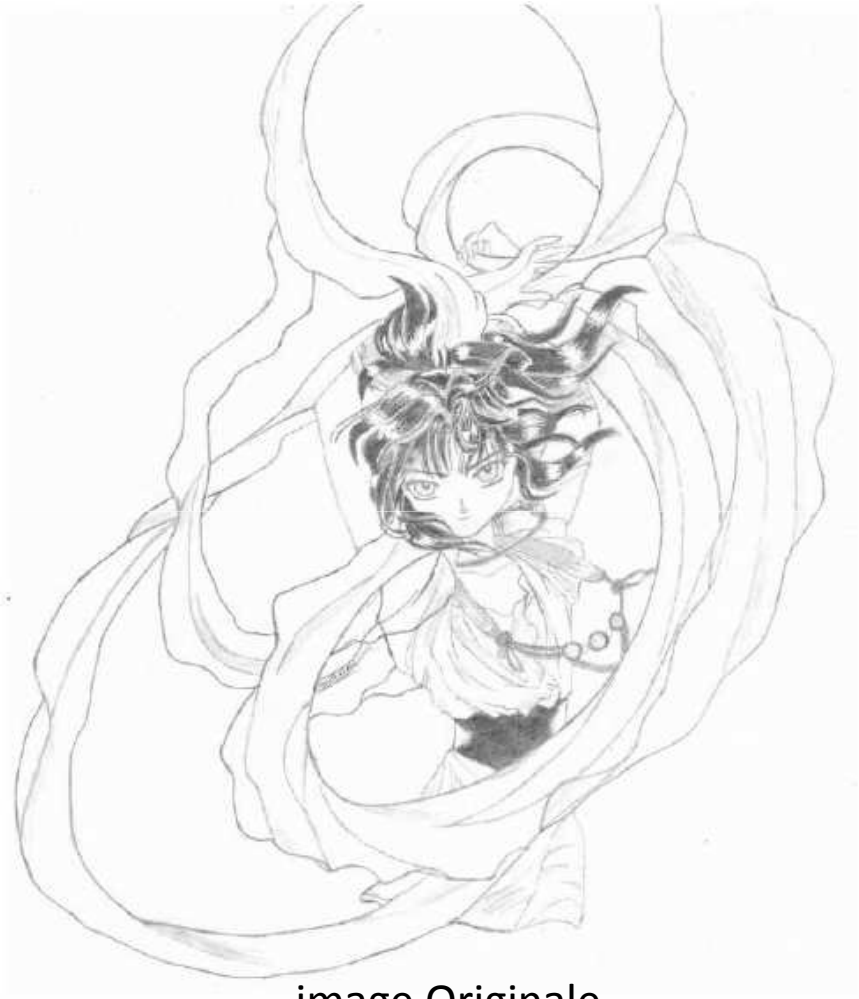
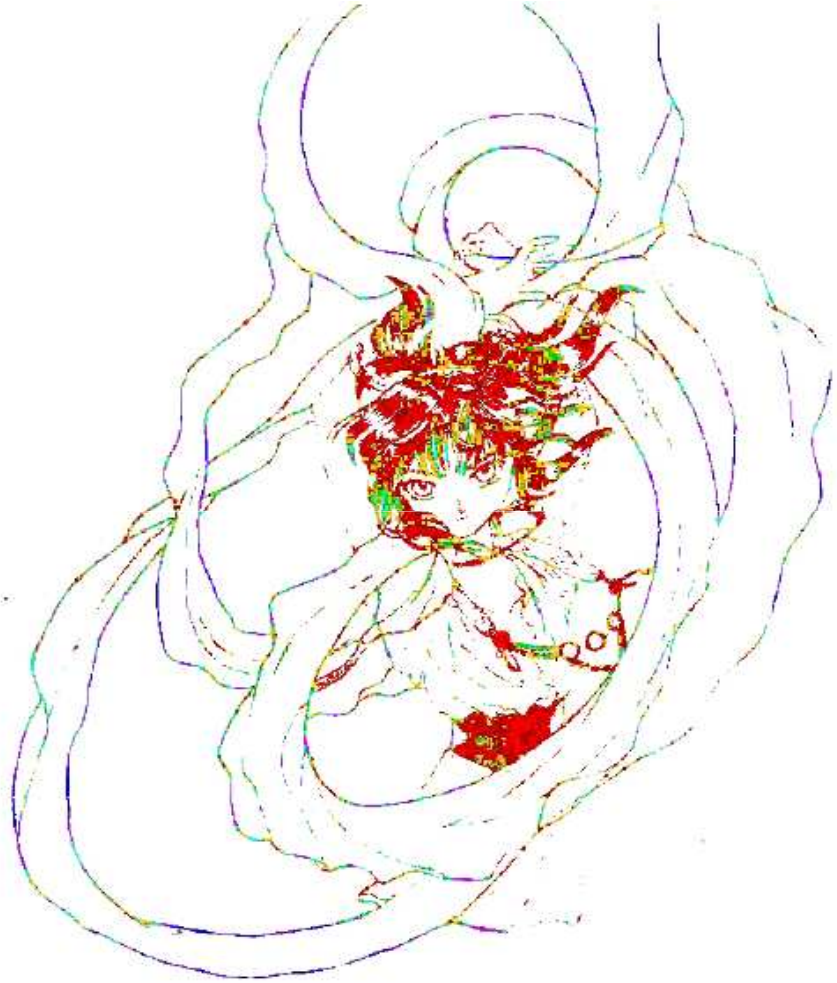


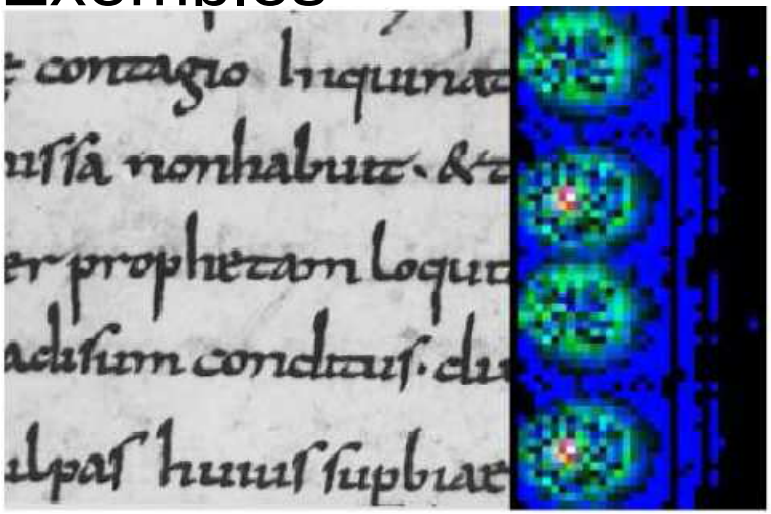
image Originale



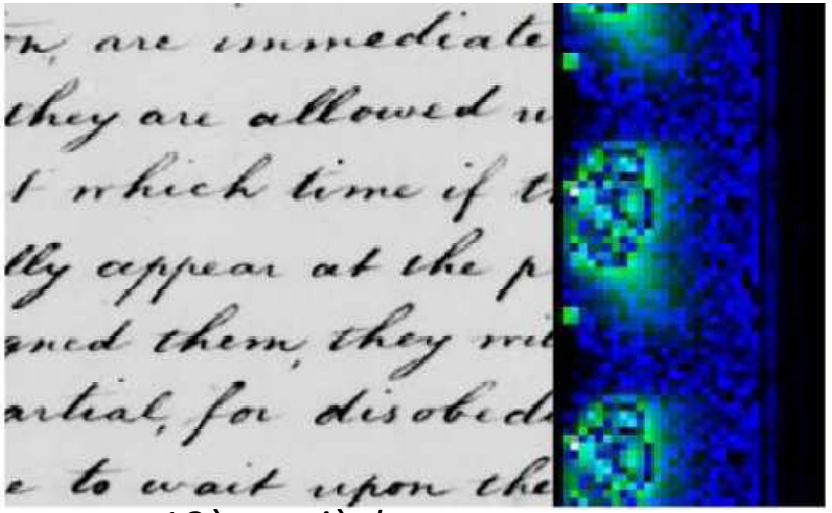
Représentation des courbures

Analyse par Curvelets

- Une signature 2D = Matrice d'accumulation orientation / courbure en tous points
- Exemples



14ème siècle



18ème siècle

Orientation

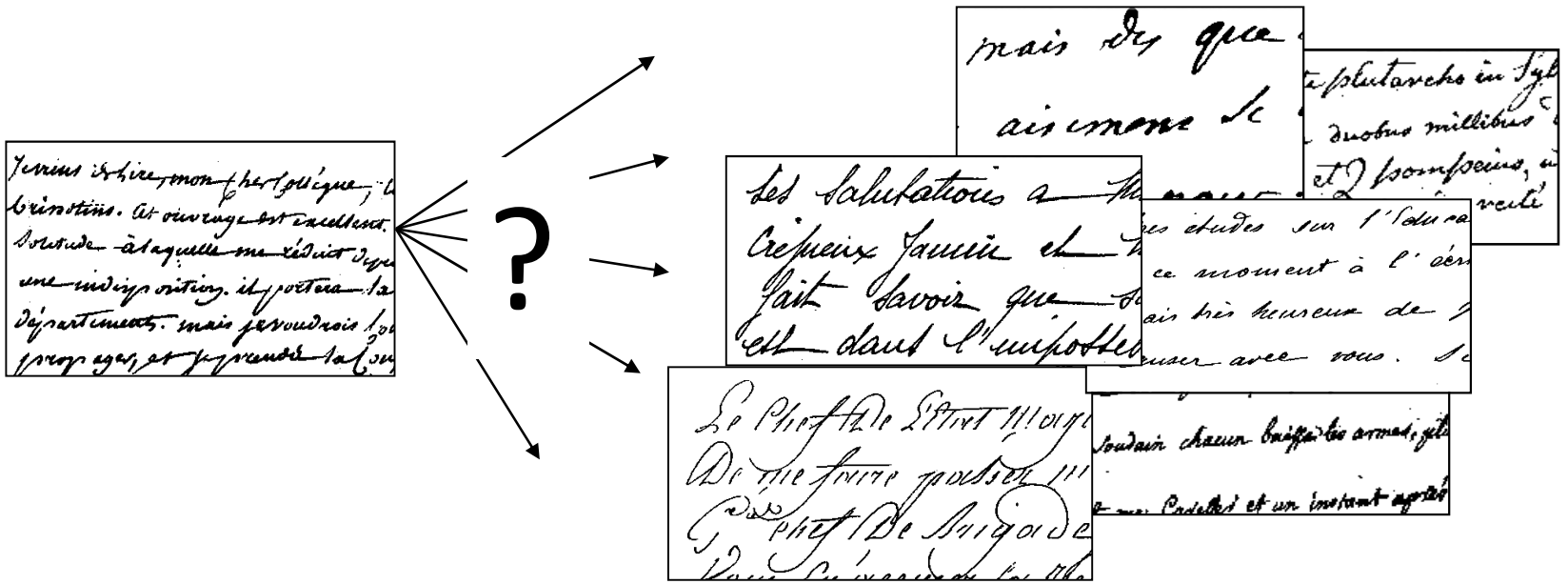
Courbure

⇨Mise en évidence des propriétés macroscopiques de la géométrie des formes (présence de courbures fortes en des points d'orientations dominantes)

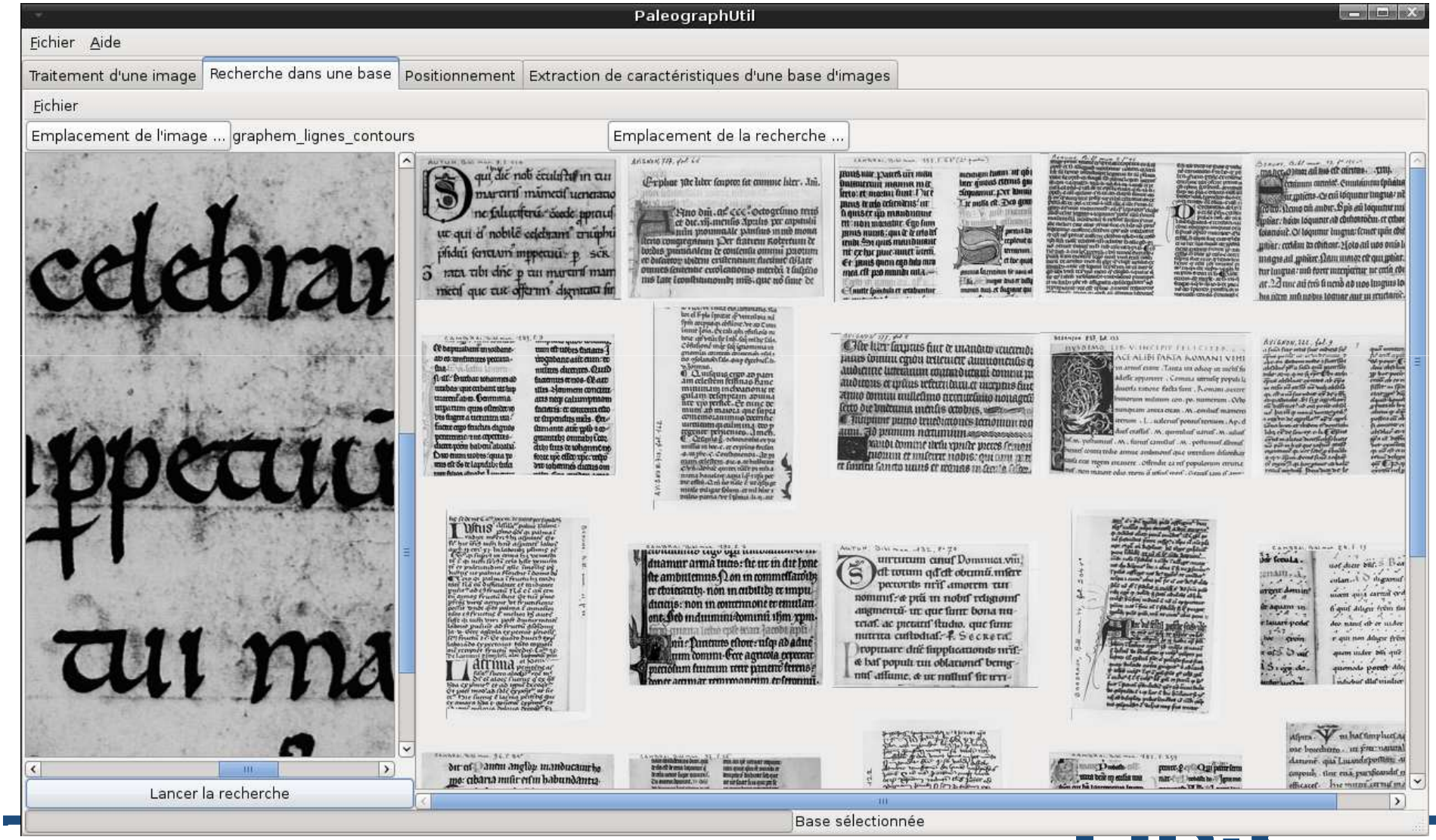
Analyse par Curvelets

- Deux similarités dédiées à deux objectifs: CBIR et groupement perceptif :
  - Similarité générique: corrélation directe entre signatures
  - Similarité « cognitive » (Tversky) non métrique: pondération différente des caractères communs ($Cor_{directe}$) et distinctifs ($Cor_{residus}$)

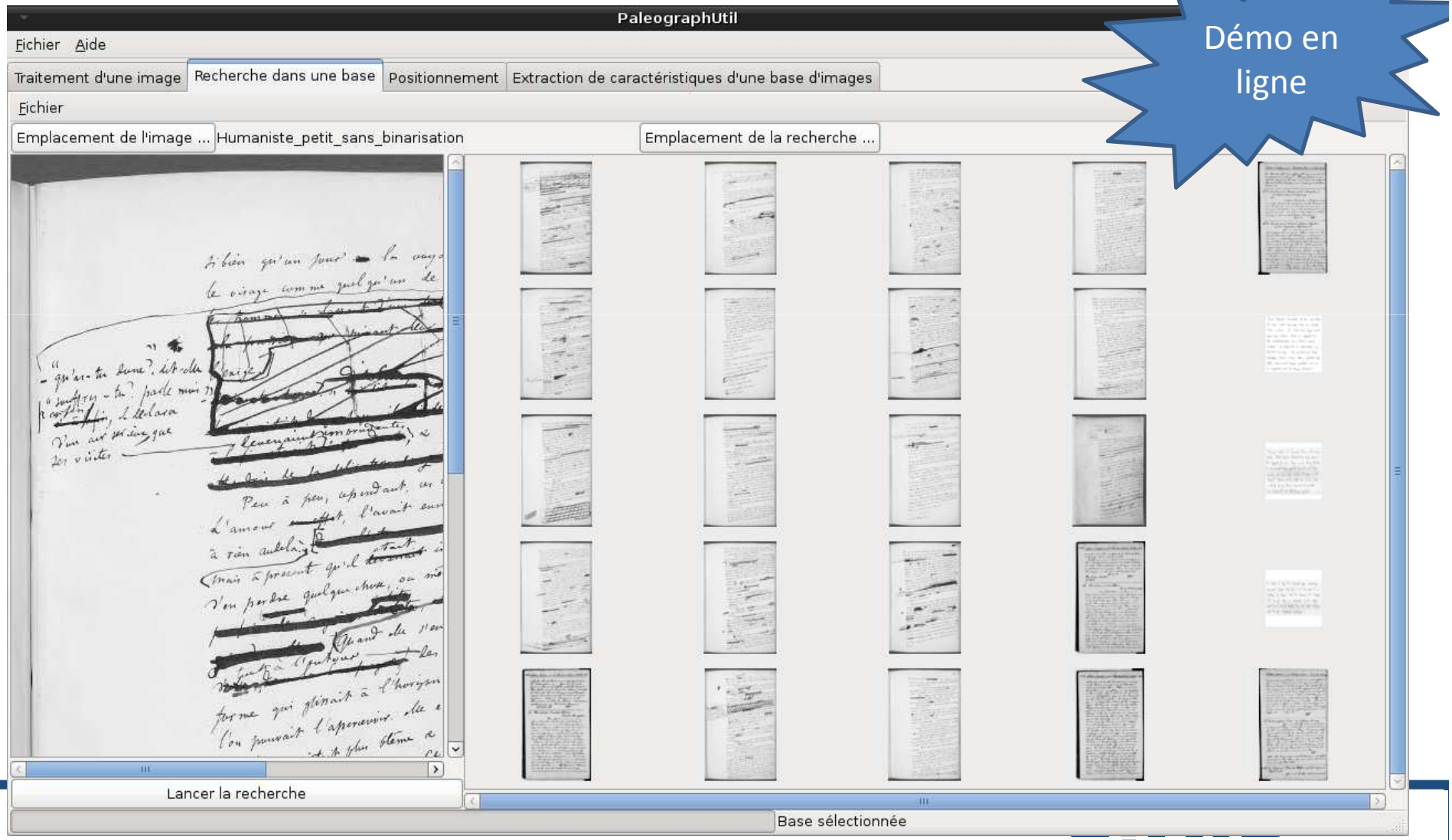
$$Sim = \gamma \times Cor_{directe} + (1 - \gamma) \times Cor_{residus}$$



Analyse par Curvelets



Analyse par Curvelets



Démo en ligne

ET LES APPLICATIONS A VENIR…

Vers une gestion électronique des documents complètes et enrichies

E-édition et Chaîne de dématérialisation :

- capteurs intelligents: caméras (tous types de documents), mobilité
- techniques d'indexation (sur les structures, sur les contenus),
- compression des données,
- accessibilité et visualisation des données,
- interactivité numérique,
- pérennité et hébergement des données…

Archives
Administration
Pédagogie…

Des besoins spécifiques:

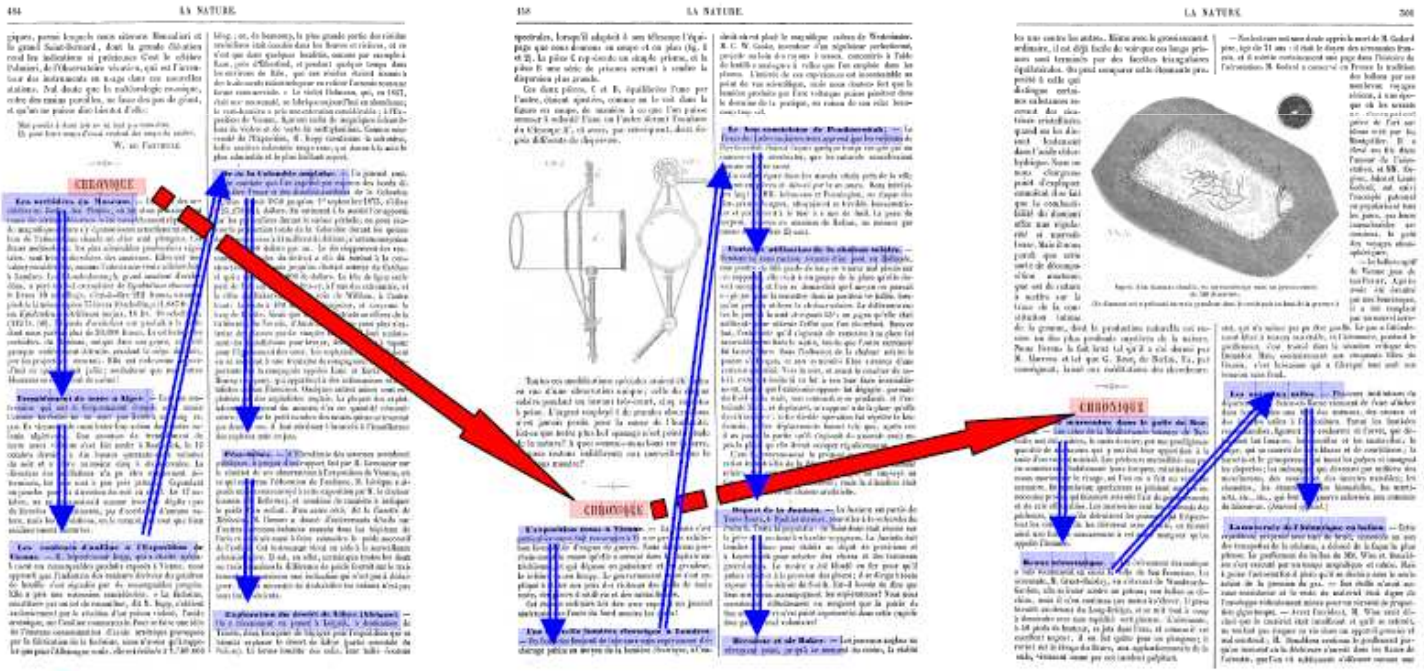
- infrastructure dimensionnée
- interfaces de visualisation interactives
- développement de logiciels ad hoc

Google books
Digital Libraries
!!!

Vers une gestion électronique des documents complètes et enrichies

Quelques applications à venir:

Indexation sémantique, navigation personnalisée et visuelle (intuitive) et résumé de contenu



navigation through a structured collection (from title to title for example)

navigation (browsing) in a page (*La Nature 1873*)

Techniquement, comment évolue le domaine ?

Vers plus de pluridisciplinarité: nouvelles correspondances entre communautés

- ***Images**: passerelles et correspondances d'objectifs, caractérisations locales / globales, recherche objets/mots…*
- ***Cognition**: modèles cognitifs, interprétation/reconnaissance*
- ***Perception visuelle**: bas niveau et accès aux données émergentes*
- ***SHS**: expression des besoins, attentes*

Quelles pistes pour aller plus loin ?

- *Privilégier les approches sans binarisation ?*
- *Choix des caractéristiques ?*
- *Espaces de représentation ?*
- *Taille des descripteurs ?*
- *Similarité ?*
- *Passage à l'échelle ?*
- *….*

Résumé

- Rien de magique dans l’accès aux images, aux contenus

- L’image seule est une chose, les métadonnées et caractéristiques une autre!

- Rien ne remplace l’œil, ni le cerveau de l’homme. Un modèle pour les systèmes actuels