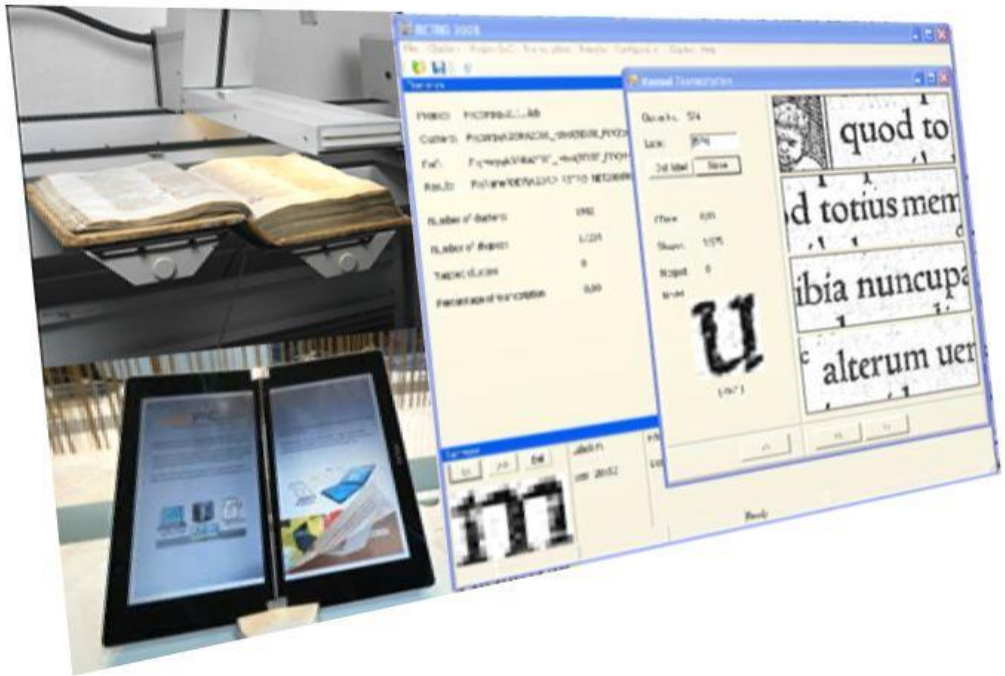
POLYTECH TOURS

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

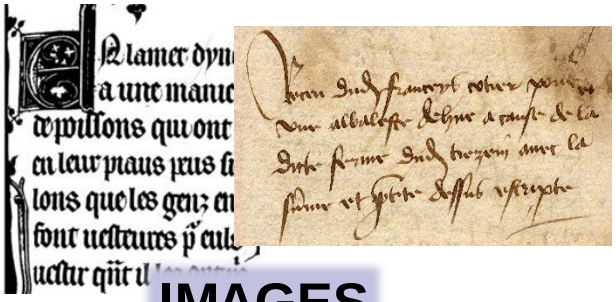
From pixels to content: An overview of the main techniques used in DIA

Jean-Yves RAMEL



LI
Laboratoire d'Informatique
EA 6300

From Pixels to Contents
Introduction



IMAGES

Meta-Data?

CONTENTS

Document Image Analysis
Machine Learning
Content based images Retrieval

Tools to automatically **extract information** inside the images / documents

Generation and utilization of meta data

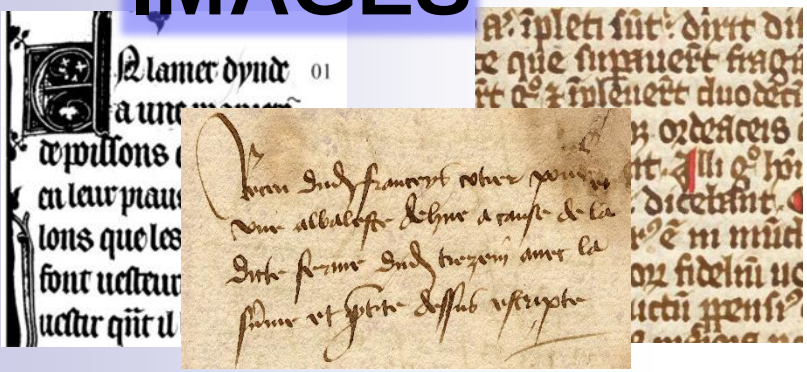
From Pixels … to contents

Outline

- From pixels…
  - What is an image?
  - Image (pre-)processing
- … to Text
  - Transcription and Layout analysis
  - Segmentation and content extraction
  - An overview of Pattern Recognition
- … but also to non-Text
  - Content characterization and signatures
  - Content retrieval and spotting
- Back to meta-data?
  - From descriptive to perceptual meta-data
  - Is there adequate encoding formats?
- Conclusions and perspectives

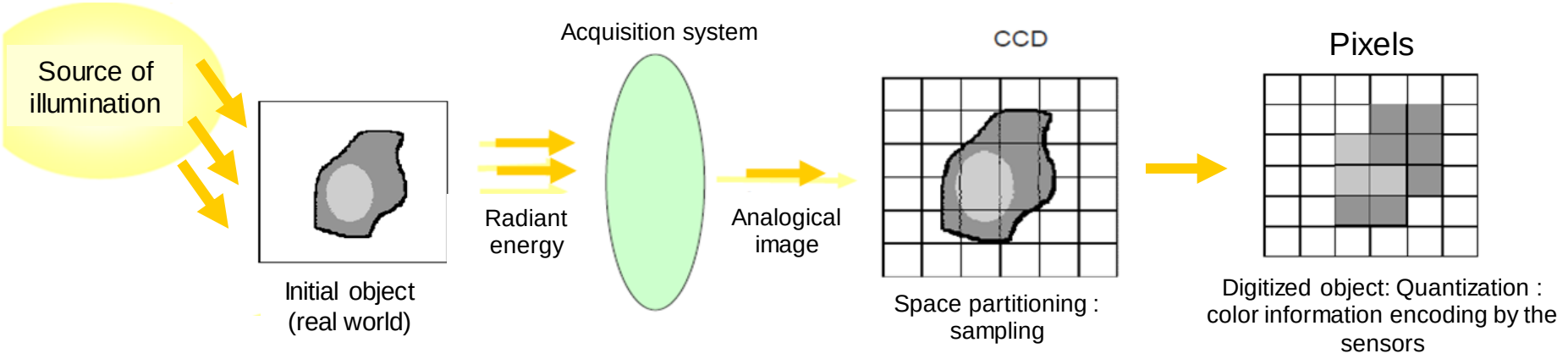
From Pixels...

IMAGES



From Pixels…
Digitization ➔ Set of pixels ?

- Images come from a grid of microscopic photosensitive cells called **PIXELS**
- **Sampling**



- **Quantization**
  - Assignment of a numerical value drawn from the received lighting energy / pixel (grid unit)
  - Continuous value (xi,yi) → Discrete value (xi,yi) → Pixels
  - The range of colors that each pixel can take

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

From Pixels…

What is an image?

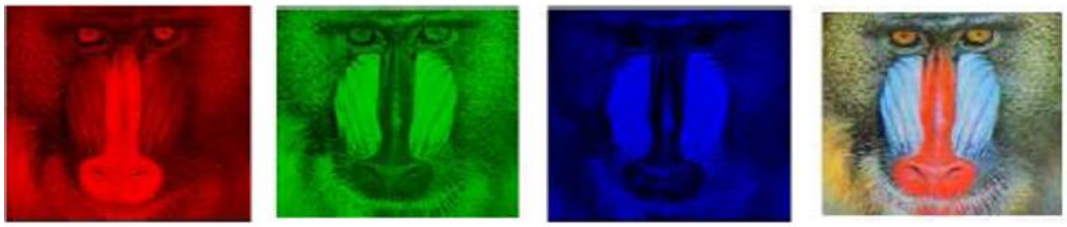
Image Quantization

Binary images: *I(i,j)* = 0 black or *I(i,j)* = 1 white

Gray level (8 bits/pixel) images:
I(i,j) = 0…..255 from the lighter to the darker.

Color images (24 bits/pixel):
3 values of lighting intensity Red, Green, Blue

$I_1(i,j)$ = 0…..255 – $I_2(i,j)$ = 0…..255 – $I_3(i,j)$ = 0…..255



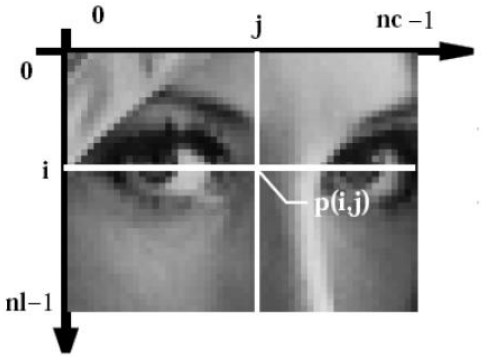
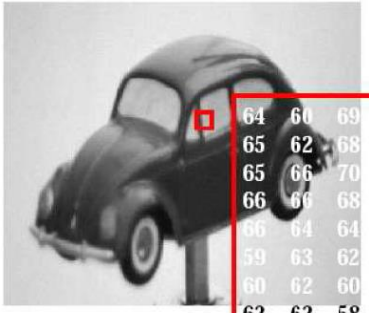


Image Representation & Processing

Image = Array(s) of pixels = Matrices of values
1 pixel = A position inside the image (i,j) + 1 color (1 to 3 values)

The values *I(i,j)* associated to each pixel *s(i,j)* represent their brightness intensity



Niveaux de gris - 8 bits:
0 - noir
255 - blanc

64	60	69	100	149	151	176	182	179
65	62	68	97	145	148	175	183	181
65	66	70	95	142	146	176	185	184
66	66	68	90	135	140	172	184	184
66	64	64	84	129	134	168	181	182
59	63	62	88	130	128	166	185	180
60	62	60	85	127	125	163	183	178
62	62	58	81	122	120	160	181	176
63	64	58	78	118	117	159	180	176

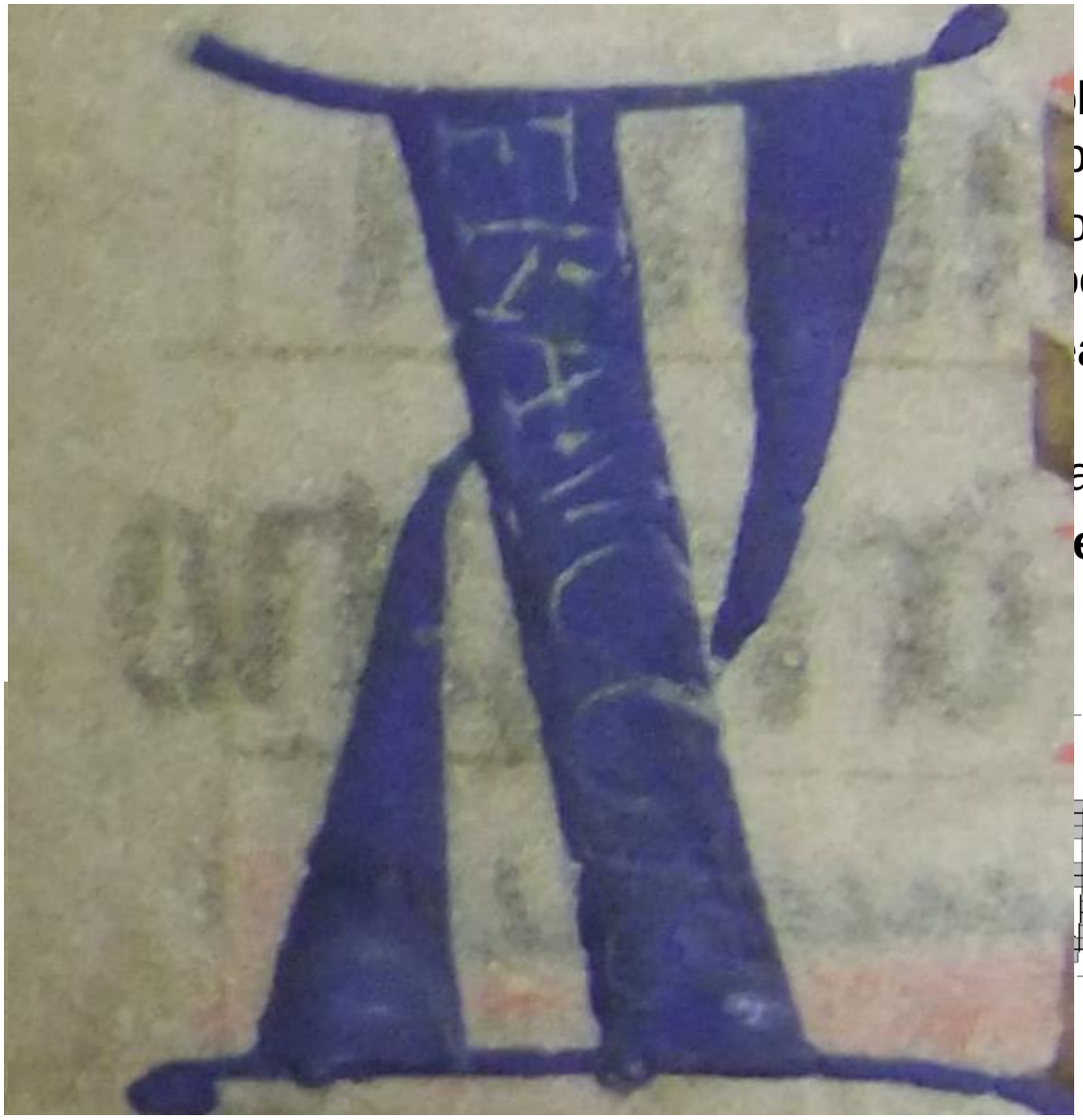
From Pixels…
What is an image?

- **Sampling ➔ Image resolution**
  - Number of pixels per lenght unit
  - In dpi (dots per inches) or ppp (points par pouce)
  - When the resolution decrease, the precision decrease

- **VF Image processing**
  - Page A4 = 21x29.7 cm
  - 200 dpi : 1650 x 2340 pixels = 3 861 000 pixels
  - 300 dpi : 3500 x 2480 pixels = 8 680 000 pixels
  - 16M colors, 1 pixel = 3 octets ➔ 10 à 25 Mo/page !
  - **A trade-off between quality-quantity/time is mandatory**
  - Fidelity of the numerical version
  - Mass of storage size – Transmission / Processing time

From Pixels…

Why few pixels are so important?



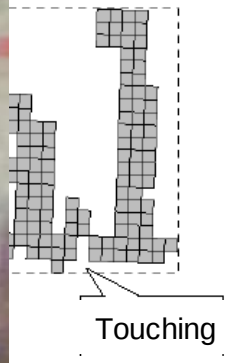
lated patterns can correspond
ixels) !

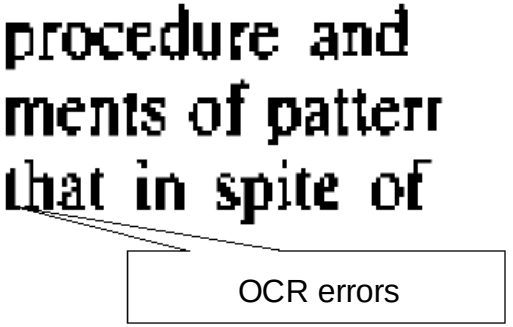
n the boundaries of the shapes
cess

arance of characters until the

acters or touching characters

e resolution



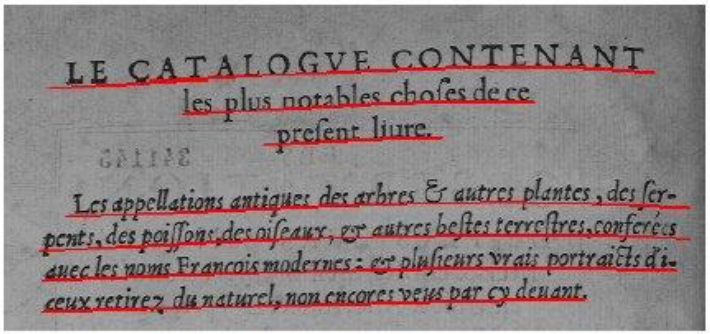


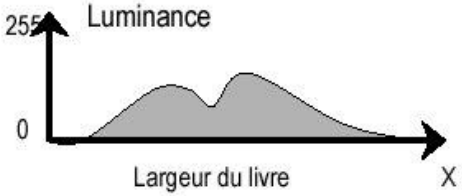
From Pixels…

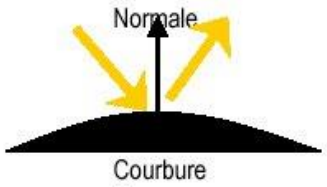
What is Image (pre-)processing?

After the digitization, the images usually still have a lot of defaults

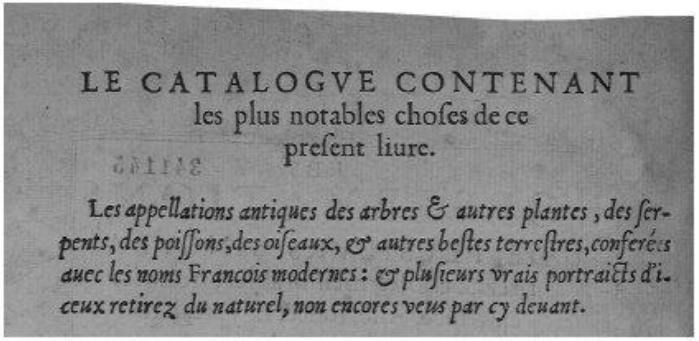
- Curvature and skew due to scanning
- Noise on boundaries, dots, blur, …

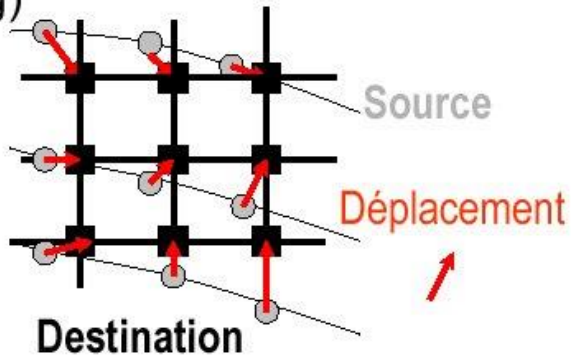


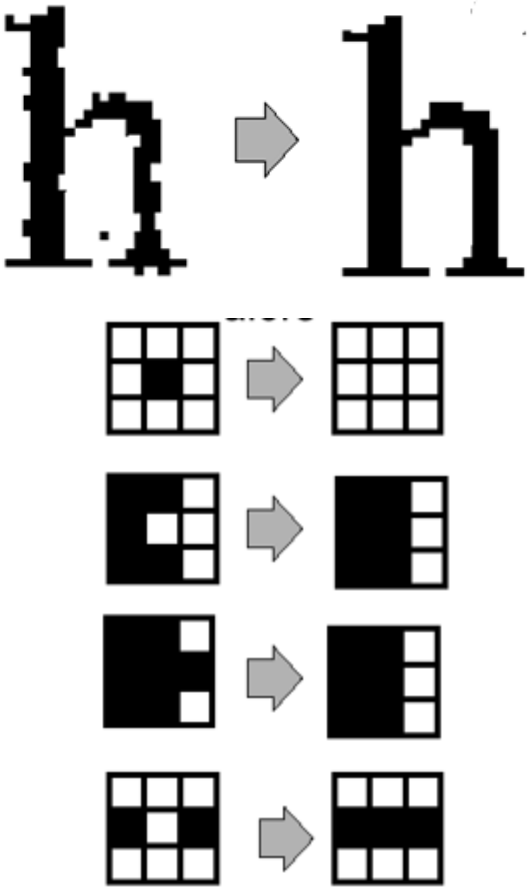




Corriger l'image (rotation, wrapping)



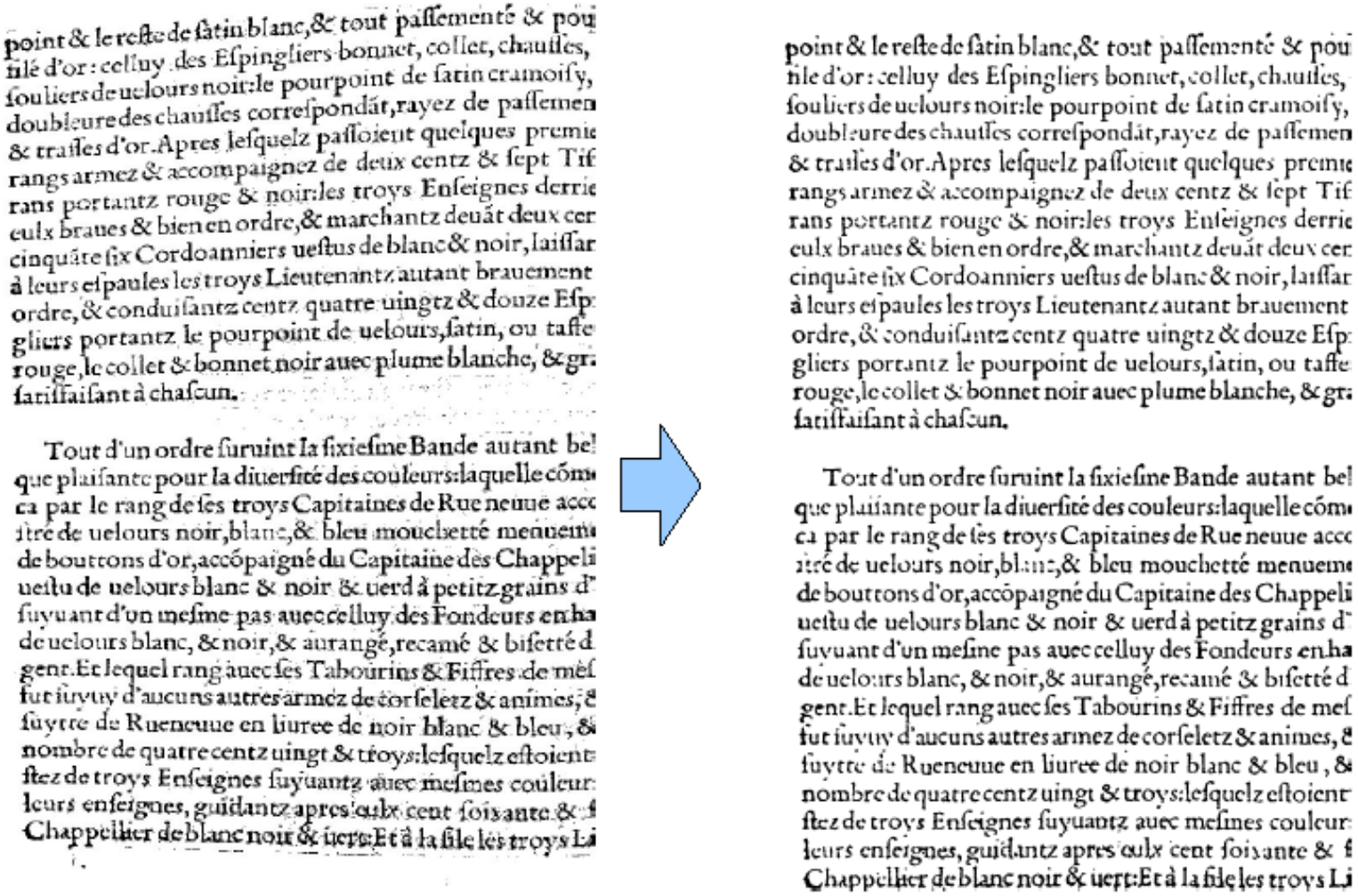




From Pixels…

Image (pre-)processing

Curvature and skew correction is possible on text images



point & le reſte de ſatin blanc,& tout paſſementé & pou
filé d'or: celluy des Eſpingliers bonnet, collet, chauſſes,
ſouliers de uelours noir:le pourpoint de ſatin cramoiſy,
doubleure des chauſſes correſpondãt,rayez de paſſemen
& trailles d'or.Apres leſquelz paſſoient quelques premie
rangs armez & accompaignez de deux centz & ſept Tif
rans portantz rouge & noir:les troys Enſeignes derrie
eulx braues & bien en ordre,& marchantz deuãt deux cer
cinquãte ſix Cordoanniers ueſtus de blanc & noir, laiſſar
à leurs eſpaules les troys Lieutenantz autant brauement
ordre, & conduiſantz centz quatre uingtz & douze Eſp
gliers portantz le pourpoint de uelours,ſatin, ou taffe
rouge,le collet & bonnet noir auec plume blanche, & gr
ſatiſfaiſant à chaſcun.

Tout d'un ordre ſuyuint la ſixieſme Bande autant bel
que plaiſante pour la diuerſité des couleurs:laquelle cõm
ca par le rang de ſes troys Capitaines de Rue neuue acc
ſtré de uelours noir,blanc,& bleu mouchetté menuem
de bouttons d'or,accõpaigné du Capitaine des Chappeli
ueſtu de uelours blanc & noir & uerd à petitz grains d'
ſuyuant d'un meſme pas auec celluy des Fondeurs en ha
de uelours blanc, & noir,& aurangé,recamé & biſetté d
gent.Et lequel rang auec ſes Tabourins & Fiffres de meſ
fut ſuyuy d'aucuns autres armez de corſeletz & animes,&
ſuyte de Rueneuue en liuree de noir blanc & bleu, &
nombre de quatre centz uingt & troys:leſquelz eſtoient
ſtez de troys Enſeignes ſuyuantz auec meſmes couleur
leurs enſeignes, guidantz apres eulx cent ſoixante &
Chappellier de blanc noir & uert:Et à la file les troys Li

point & le reſte de ſatin blanc,& tout paſſementé & pou
filé d'or: celluy des Eſpingliers bonnet, collet, chauſſes,
ſouliers de uelours noir:le pourpoint de ſatin cramoiſy,
doubleure des chauſſes correſpondãt,rayez de paſſemen
& trailles d'or.Apres leſquelz paſſoient quelques premie
rangs armez & accompaignez de deux centz & ſept Tif
rans portantz rouge & noir:les troys Enſeignes derrie
eulx braues & bien en ordre,& marchantz deuãt deux cer
cinquãte ſix Cordoanniers ueſtus de blanc & noir, laiſſar
à leurs eſpaules les troys Lieutenantz autant brauement
ordre, & conduiſantz centz quatre uingtz & douze Eſp
gliers portantz le pourpoint de uelours,ſatin, ou taffe
rouge,le collet & bonnet noir auec plume blanche, & gr
ſatiſfaiſant à chaſcun.

Tout d'un ordre ſuyuint la ſixieſme Bande autant bel
que plaiſante pour la diuerſité des couleurs:laquelle cõm
ca par le rang de ſes troys Capitaines de Rue neuue acc
ſtré de uelours noir,blanc,& bleu mouchetté menuem
de bouttons d'or,accõpaigné du Capitaine des Chappeli
ueſtu de uelours blanc & noir & uerd à petitz grains d'
ſuyuant d'un meſme pas auec celluy des Fondeurs en ha
de uelours blanc, & noir,& aurangé,recamé & biſetté d
gent.Et lequel rang auec ſes Tabourins & Fiffres de meſ
fut ſuyuy d'aucuns autres armez de corſeletz & animes,&
ſuyte de Rueneuue en liuree de noir blanc & bleu, &
nombre de quatre centz uingt & troys:leſquelz eſtoient
ſtez de troys Enſeignes ſuyuantz auec meſmes couleur
leurs enſeignes, guidantz apres eulx cent ſoixante &
Chappellier de blanc noir & uert:Et à la file les troys Li

From Pixels…

Image (pre-)processing

The problem is more complicated in case of heterogeneous content

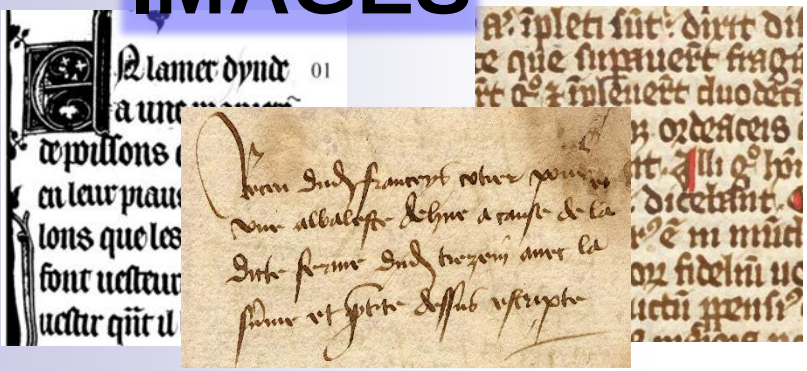
From Pixels…

Image (pre-)processing

The problem is more complicated in case of heterogeneous content

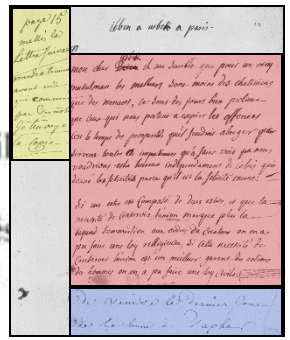
… to text and layout

IMAGES



CONTENTS

Dans la Mer d'Inde
il y a une espèce
de poissons qui ont
dans leur peau des poi
longs que les gens,
quant il les ont attrapés
en font des vêtements
pour s'habiller.

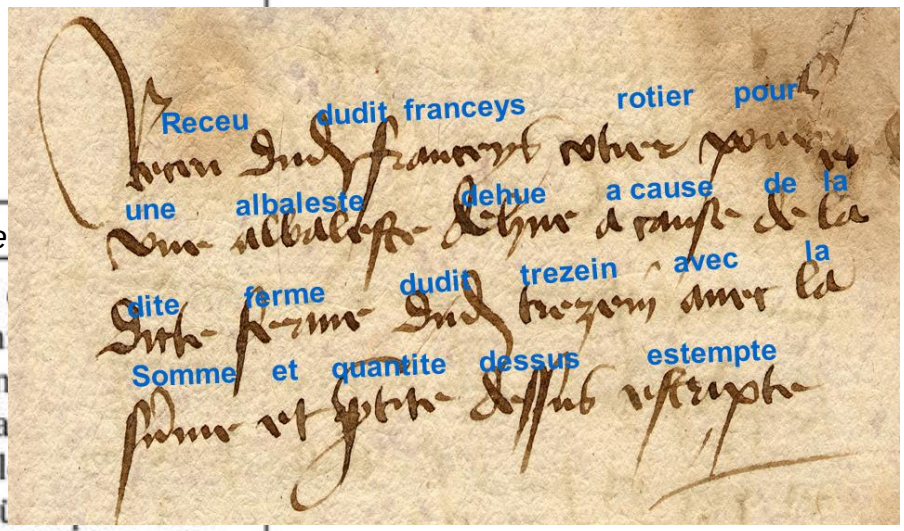


From pixels … to text

Automatic Trancription

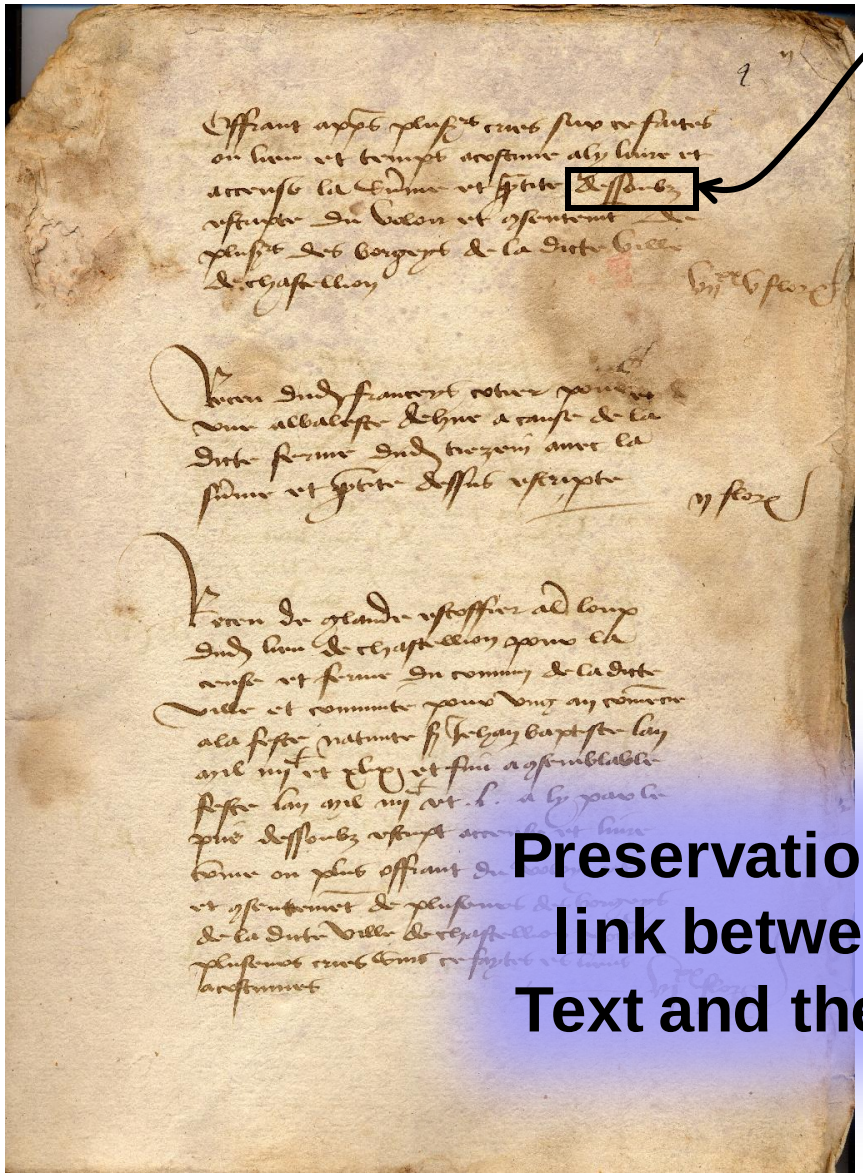
Manuscrit	*Translation*
	Dans la Mer d'Inde il y a une espèce de poissons qui ont dans leur peau des poils si longs que les gens, quant il les ont attrapés, en font des vêtements pour s'habiller.
Diplomatic Transcription	*Modernize*
EN lamer dynde a une maniere de poiſſons qui ont en leur piaus peus ſi lons que les genz en font ueſteures p̃ euls ueſtir qñt il les ont p's	En la mer a une man de poisson en leur pia lons, que l font veste vestir quant il les ont pris.

So many possibilities for transcriptions !



From pixels … to text

Automatic transcription but also layout Analysis



1-Offrant appres plusieurs rues sur ce faites
2-au lieu et temps acostume aly livre et
3- accenste la somme et quantite dessoulz
4-estempte du voloir et consentement de
5-plusieurs des borgeys de la dite ville
6-de chastellion _____IIIIxXX V
flor(ins)

7-receu dudit franceys rotier pour
8-une albaleste dehue a cause de la
9-dite ferme dudit trezein avec la
10-somme et quantite dessus estempte ?
11- _____II flor(ins)

12-receu de Alexandre ? escoffier al loup ?
13- dudit lieu de chatellion pour la
14-rense et ferme du commun de la dite
15-ville et communaute ? pour ung an commettre ?
16-albaleste nativite fi jehan baptiste lan
17-mil quatre cent et cinquante? A ly par le
18-pris dessouls esempt accense et hure ?
19-e au plus offrant du voloyr et
20-et consentement de pluseures des bourgeys
21-de la dite ville de chastellion apres
22-pluseures rues ….refaytes es lieus ?

Preservation of the link between the Text and the Image

From pixels … to text
An overview of OCR mechanisms

- As a first step, the segmentation result will have high impact on the final results provided by the OCR !!!!

« it is always the segmentation that raises problems »

- The segmentation is an irreversible processing because it is the result of an analysis of the image according to a criterion and a method.

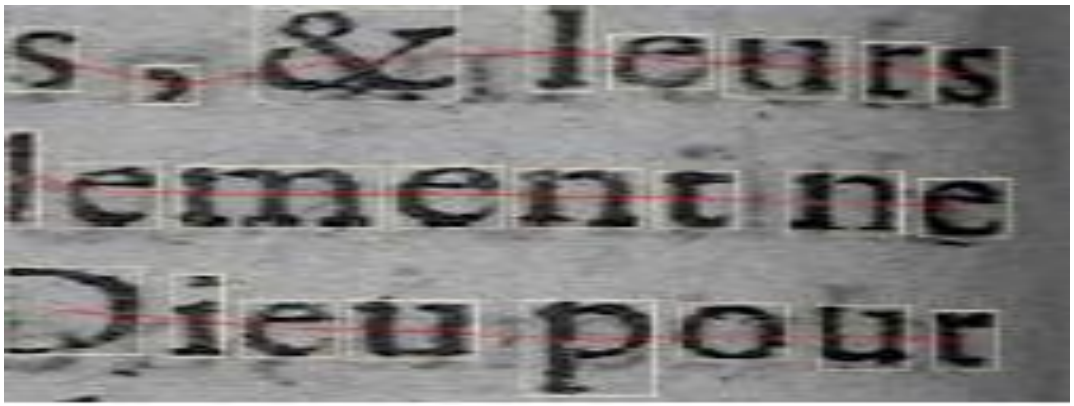
- Many segmentation methods have been proposed:
  - Top-down methods (image decomposition into smaller and smaller elements) ➔ Layout first – characters after
  - Bottom-up methods (from pixels tp higher and higher level elements) → Characters first – layout after
  - Etc…

From pixels … to text

An overview of OCR mechanisms

First step : Image segmentation

- Transformation of the image (set of pixels) into patterns (regions of interest) of higher level **(EoC)**
- These EoC could be very simple (part of characters) or more sophisticated ones (paragraphs, illustrations, …)
- EoC extraction: Background (white) / Foreground (black) separation
- **Color Image → Grayscale → binarisation**

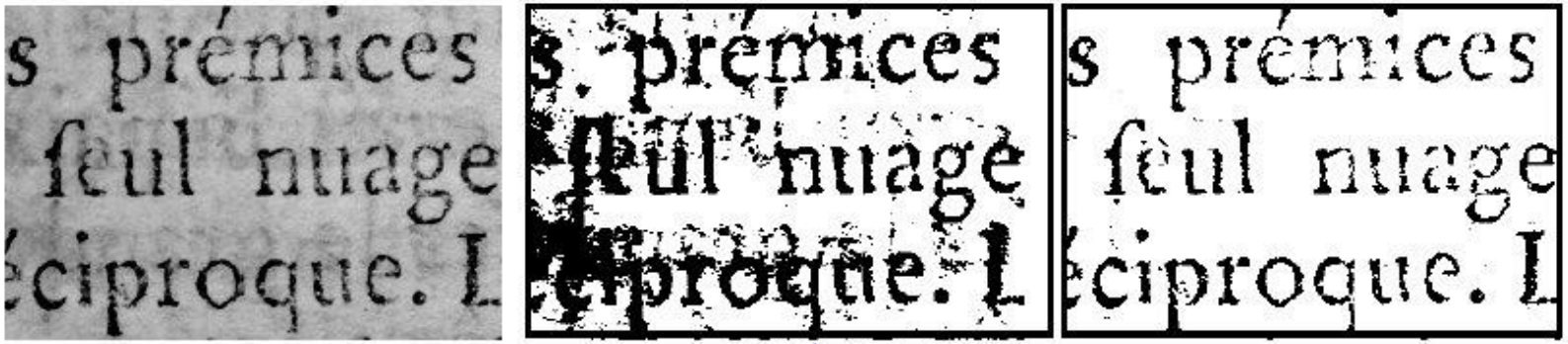


EoC

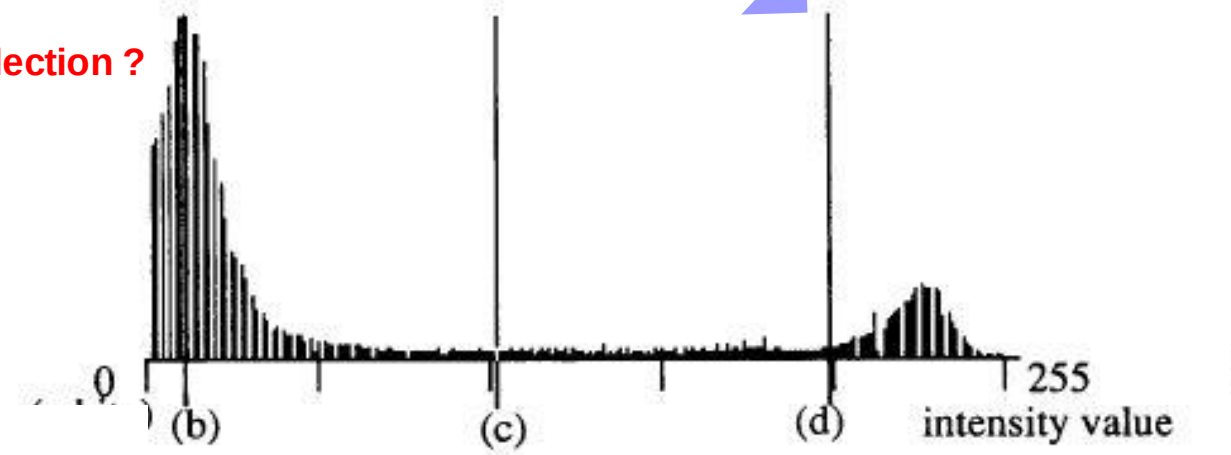
From pixels … to text
An overview of OCR mechanisms

Just to illustrate the difficulties…

- Most of the segmentation methods need a binarisation



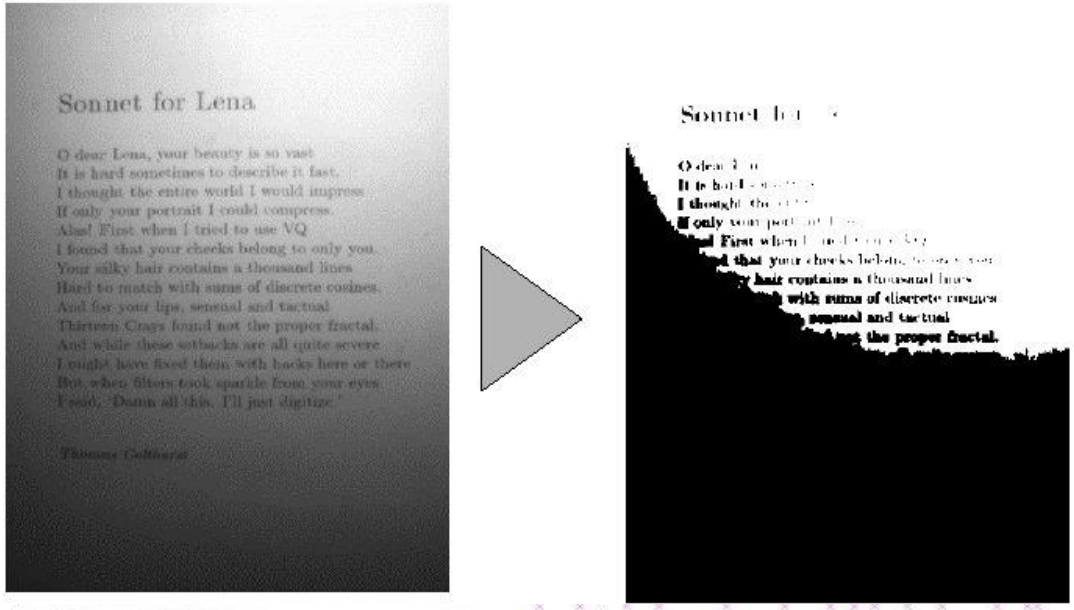
Threshold selection ?



From pixels … to text
An overview of OCR mechanisms

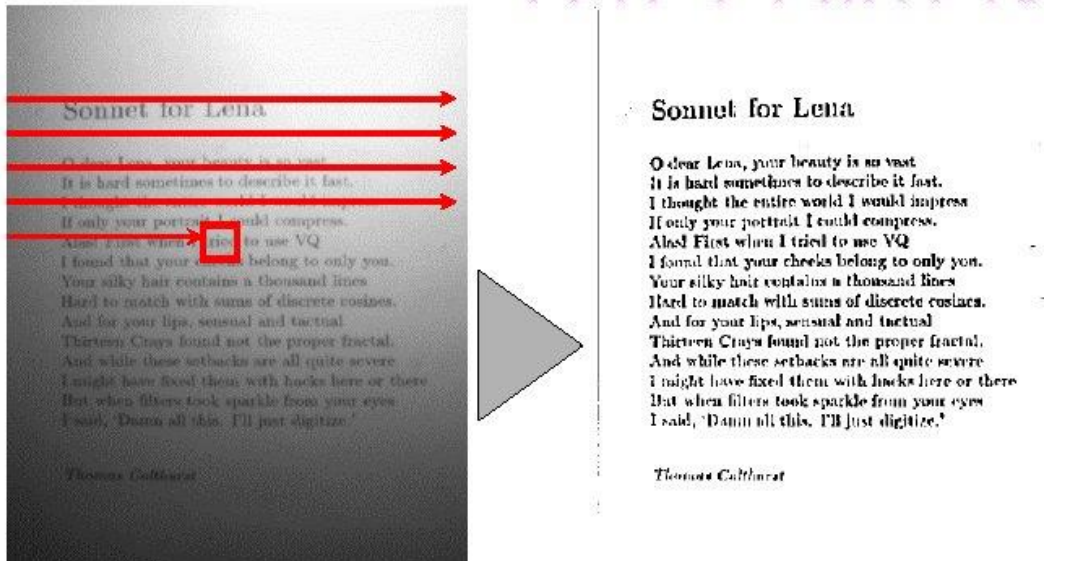
Just to illustrate the difficulties…

- Most of the segmentation methods need a binarisation
- Global threshold ➔



- Local thresholds ➔

Niblack : $S = m + ks^2$ avec k= -0,2
| *m* : mean et s : standard deviation

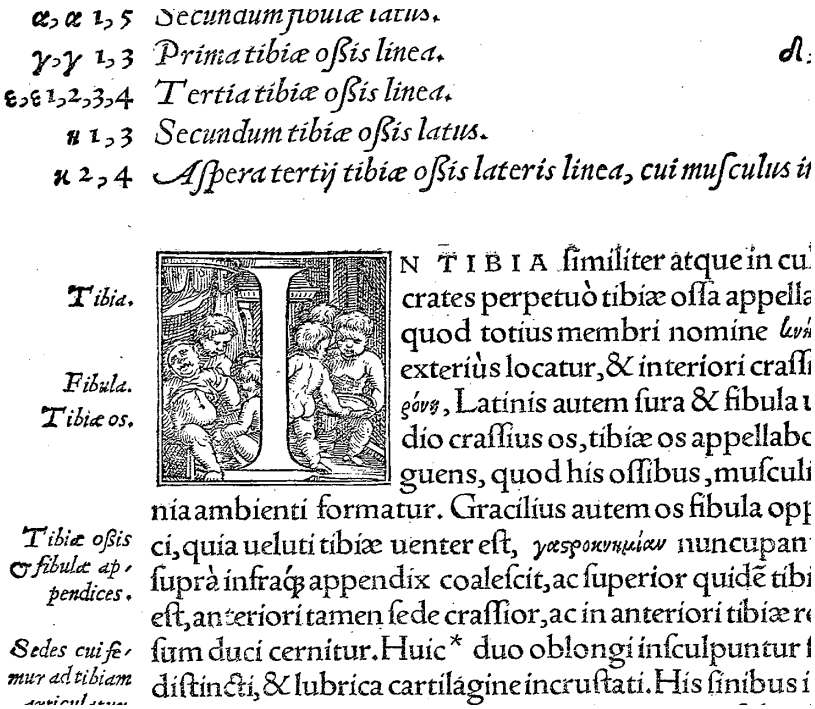


From pixels … to text

An overview of OCR mechanisms

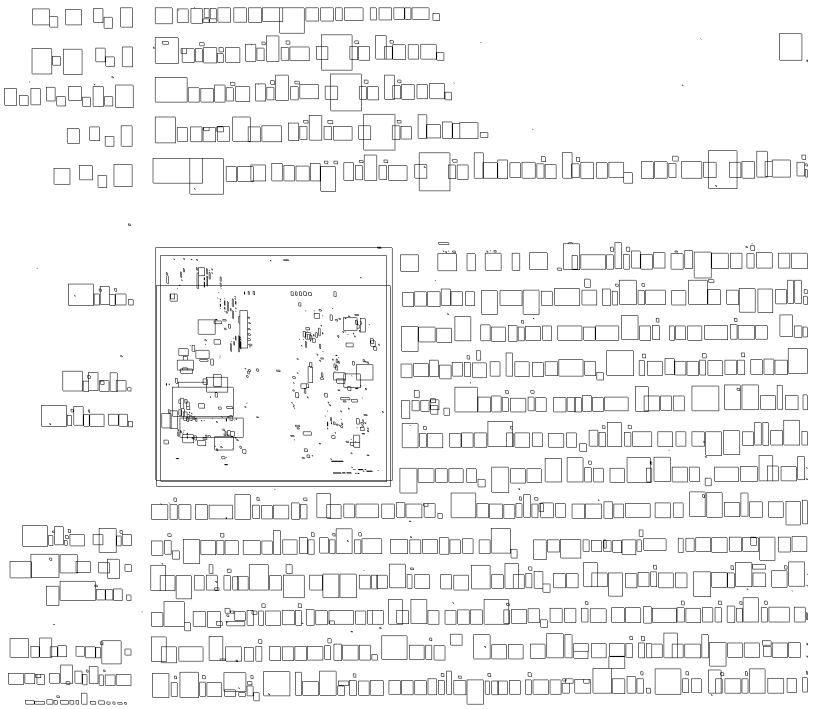
First step: Image segmentation / Connected components

- Then, we can try to group black pixels together to **localize** and **recognize** higher level Element of Content **(EoC)**


α, α 1, 5 Secundum fibulæ latus.
γ, γ 1, 3 Prima tibiæ oſſis linea.
ε, ε 1, 2, 3, 4 Tertia tibiæ oſſis linea.
ϰ 1, 3 Secundum tibiæ oſſis latus.
ϰ 2, 4 Aſpera tertij tibiæ oſſis lateris linea, cui muſculus i

Tibia.
Fibula.
Tibiæ os.
Tibiæ oſſis & fibulæ appendices.
Sedes cui femur adtibiam

N TIBIA ſimiliter atque in cu
crates perpetuò tibiæ oſſa appella
quod totius membri nomine
exteriùs locatur, & interiori craſſi
, Latinis autem ſura & fibula
dio craſſius os, tibiæ os appellabc
guens, quod his oſſibus, muſculi
nia ambienti formatur. Gracilius autem os fibula opp
ci, quia ueluti tibiæ uenter eſt, nuncupan
ſuprà infraq; appendix coaleſcit, ac ſuperior quidē tibi
eſt, anteriori tamen ſede craſſior, ac in anteriori tibiæ r
ſum duci cernitur. Huic* duo oblongi inſculpuntur
diſtincti, & lubrica cartilágine incruſtati. His ſinibus i

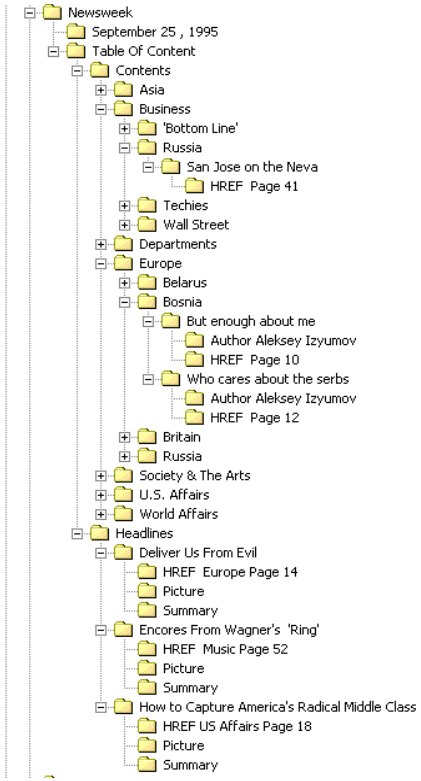


From pixels … to text

An overview of OCR mechanisms

Next step : Layout analysis

- Connected Components ➔ Words ➔ Lines ➔ Paragraphs ➔ Page
- The results have to be saved in XML format (Alto, …)
- Choosing how to organize the XML tree (physical / logical) is not so easy…



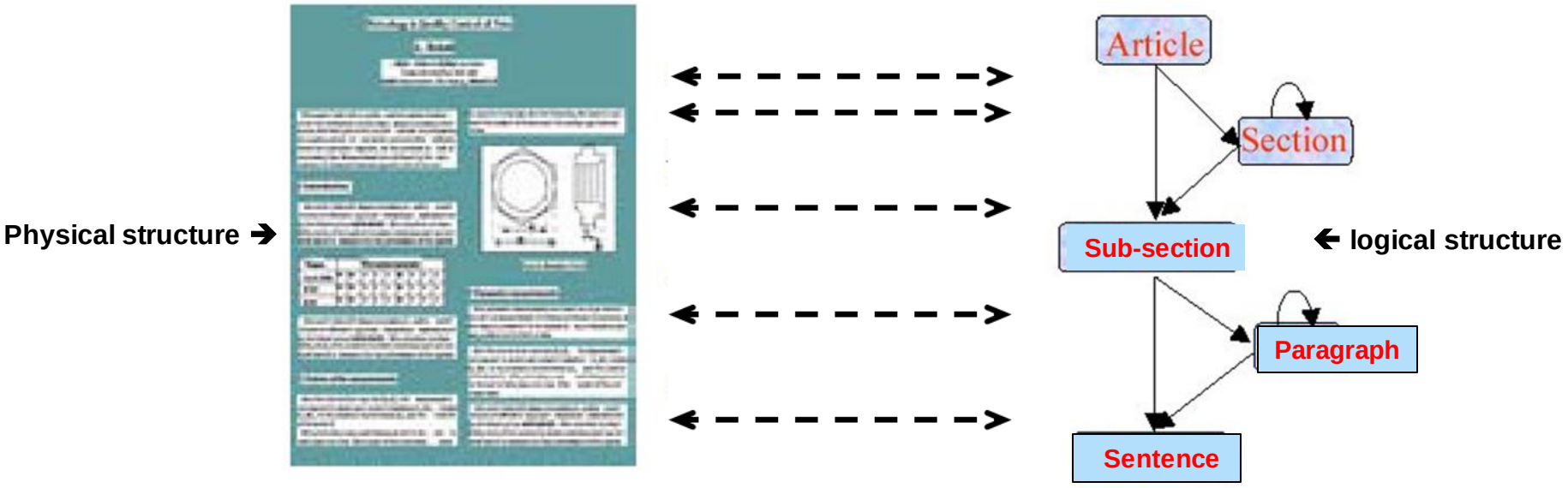
From pixels … to text

An overview of OCR mechanisms

Next step : Layout analysis

Two kind of structures have been identified by researchers in DIA:

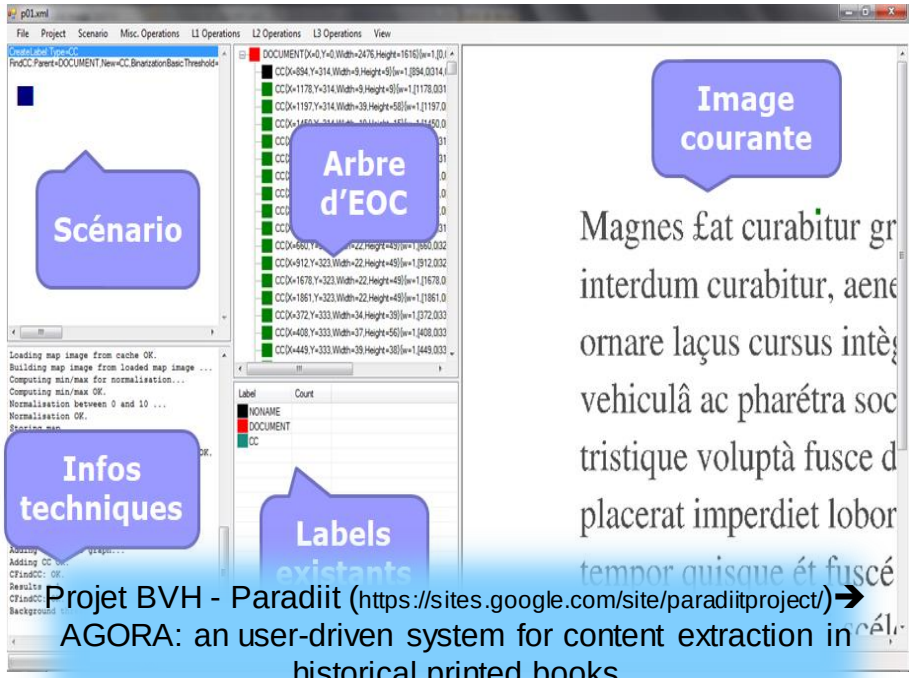
- The logical structure ➔ the generic one corresponding to a priori knowledge about the content of the document
- The physical structure ➔ the analysed instance corresponding to the extracted EoC inside the image, each one associated to descriptive features (size, position, number of sub-patterns, …)
- Layout analysis tries to recognize these 2 structures (EoC identification)



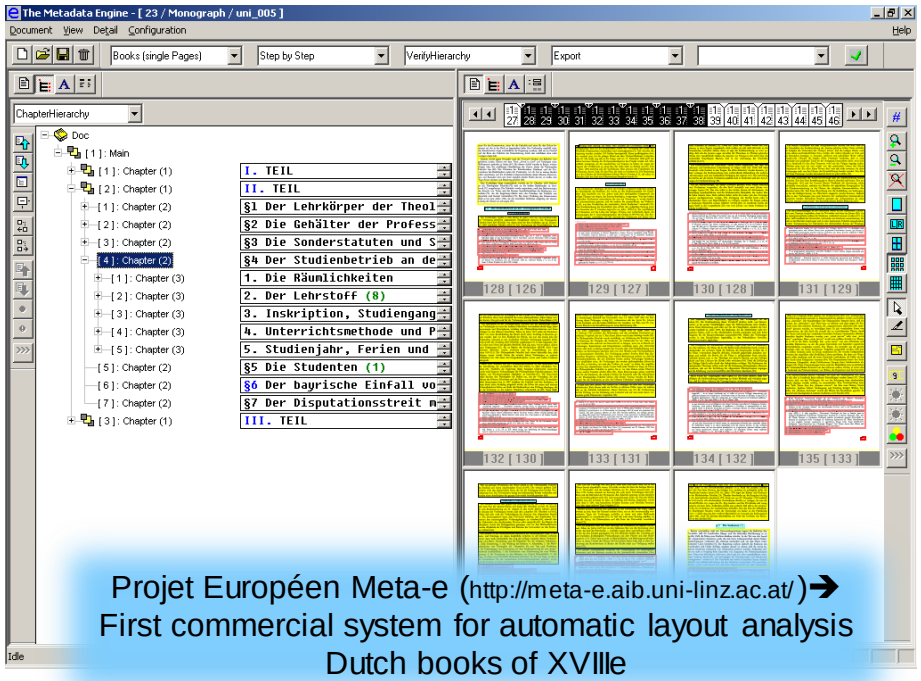
From pixels … to text
An overview of OCR mechanisms

Next step : Layout analysis

- The analysis / identification of the EoC is usually achieved based on a rule based system defined through a grammar (static one) or defined interactively by the users



Projet BVH - Paradiit (https://sites.google.com/site/paradiitproject/)➔
AGORA: an user-driven system for content extraction in historical printed books



Projet Européen Meta-e (http://meta-e.aib.uni-linz.ac.at/)➔
First commercial system for automatic layout analysis Dutch books of XVIIIe

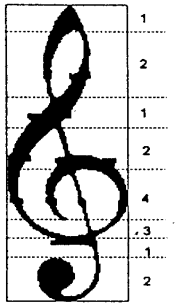
From pixels … to text

An overview of OCR mechanisms

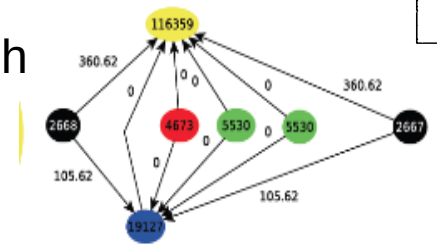
Many possible choices and techniques

- For selection of discriminative features

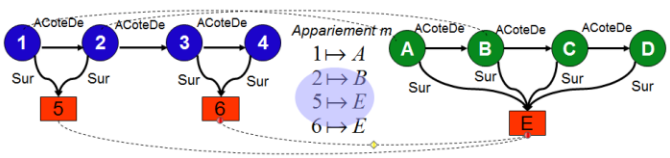
1 EoC ➔ 1 Vecteur $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{bmatrix} \in IR^n$

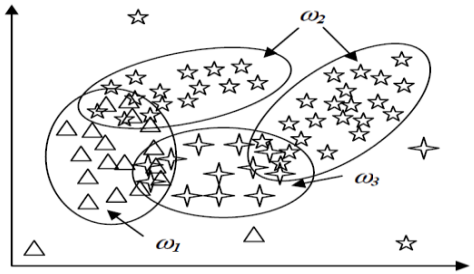


1 EoC ➔ 1 Graph

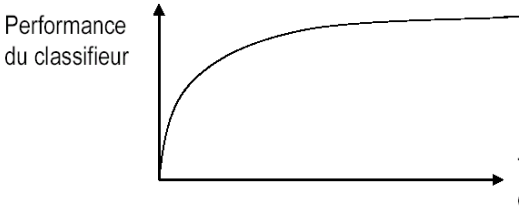


- Machine Learning models and tools

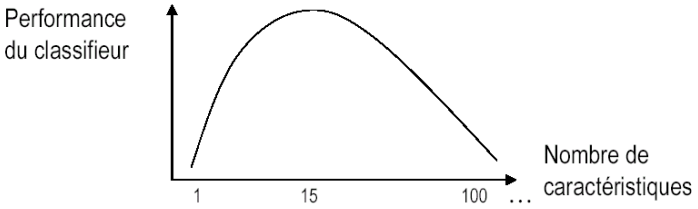




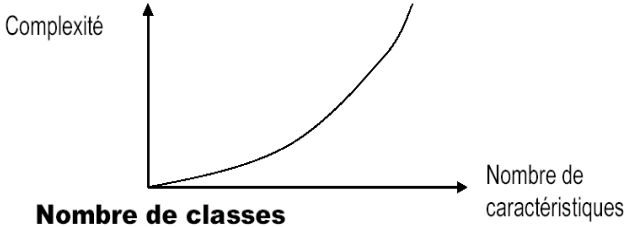
Statistique suffisante



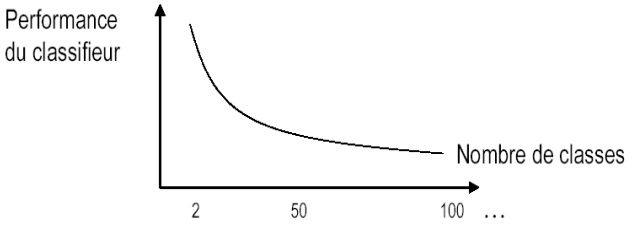
Malédiction de la dimensionalité



Complexité vs nombre de caractéristiques



Nombre de classes



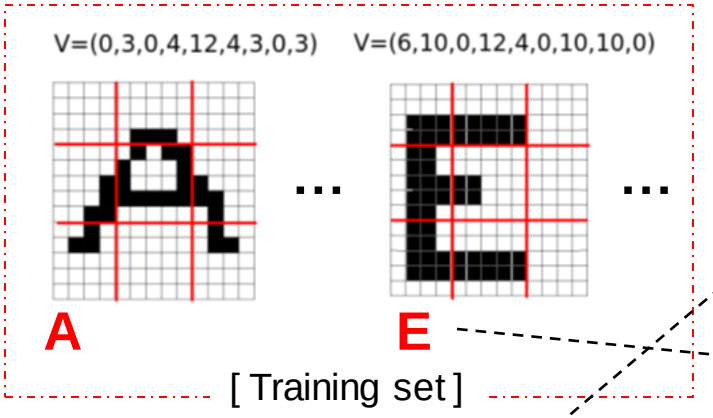
From pixels … to text

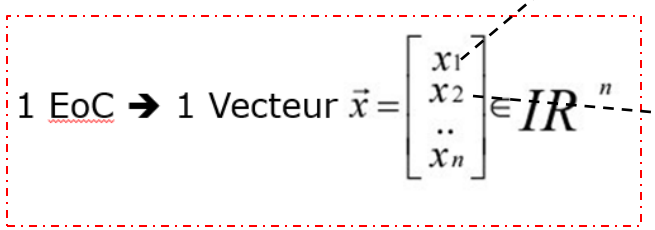
An overview of OCR mechanisms

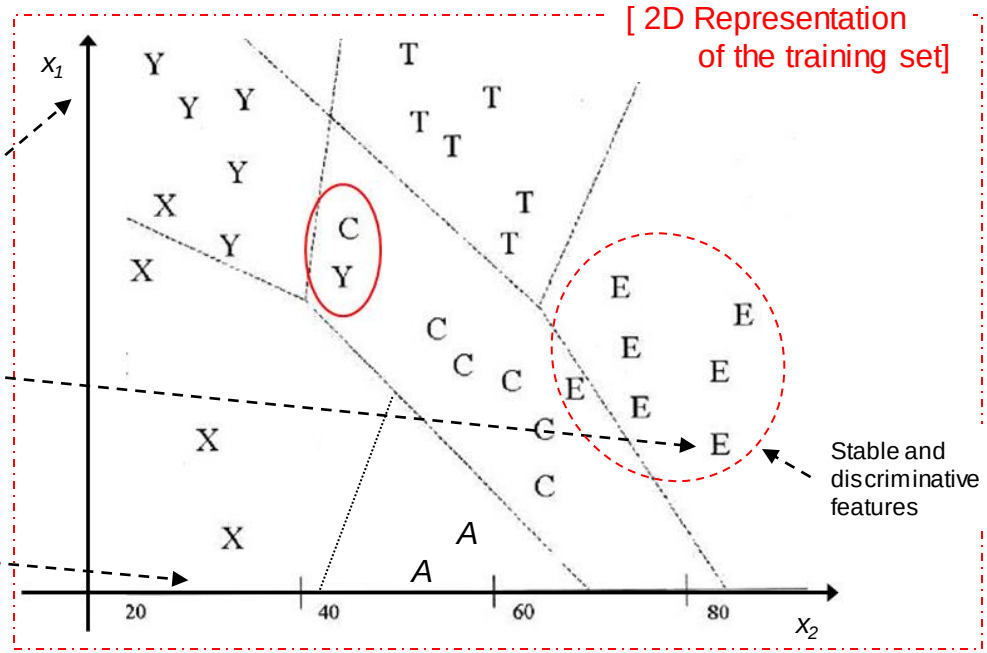
Next step : Pattern / EoC recognition (toward Machine Learning)

How computers can recognize objects?

- We need a large set of (labelled) examples similar to the patterns to be recognized ➔ **a training set**
- We need a list of stable and discriminative **features** (shape, color, size,…) used to describe the patterns (labelled ones and unknown one)







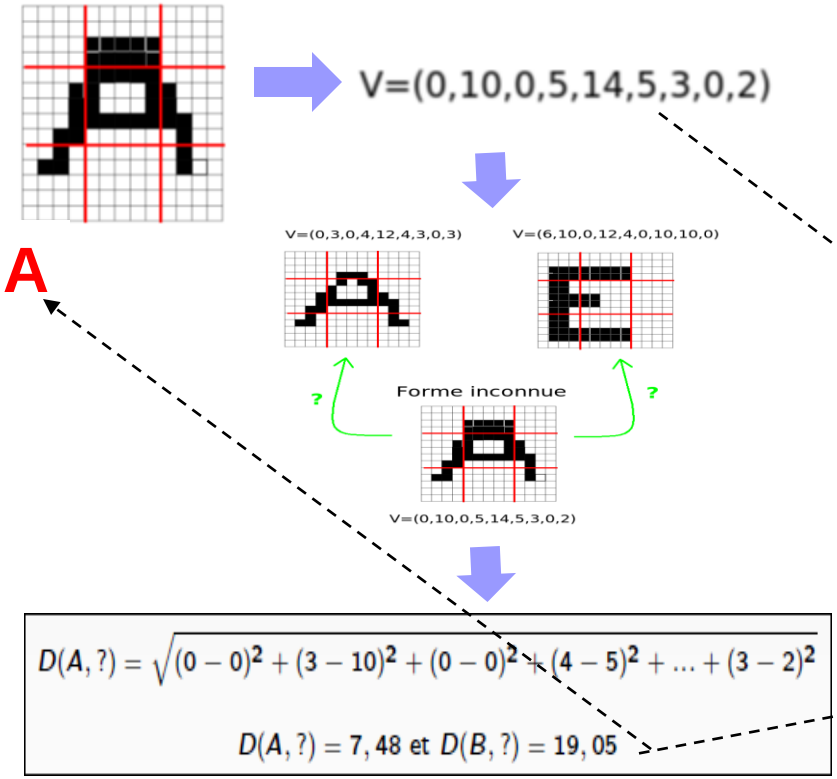
From pixels … to text

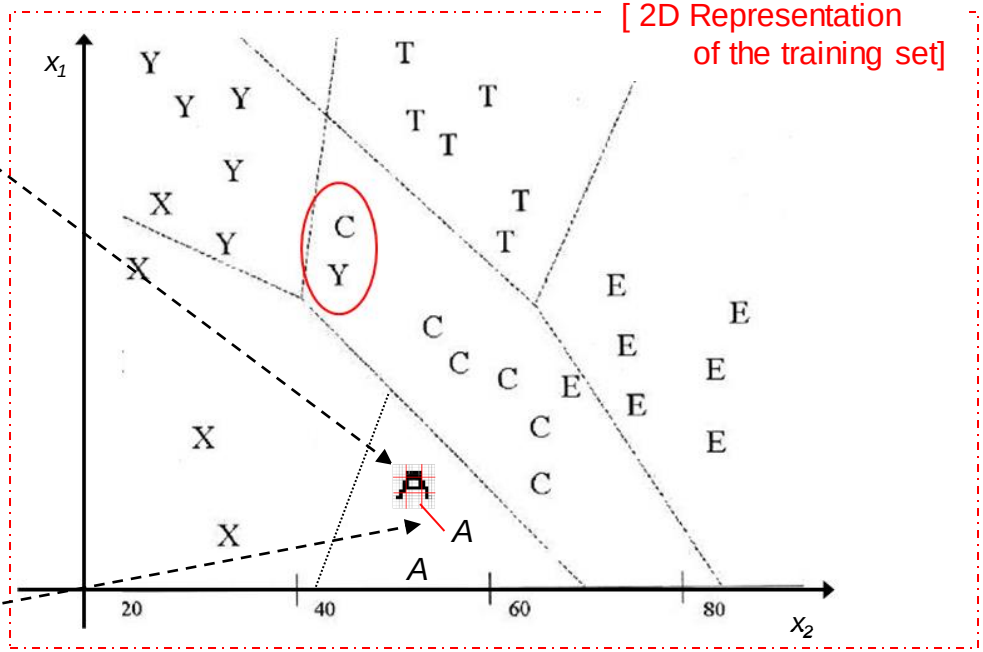
An overview of OCR mechanisms

Next step : Pattern / EoC recognition (toward Machine Learning)

How computers can recognize objects?

- When an unknown EoC arrives, we compute its features and compare it with the content of the training set (associated built models)

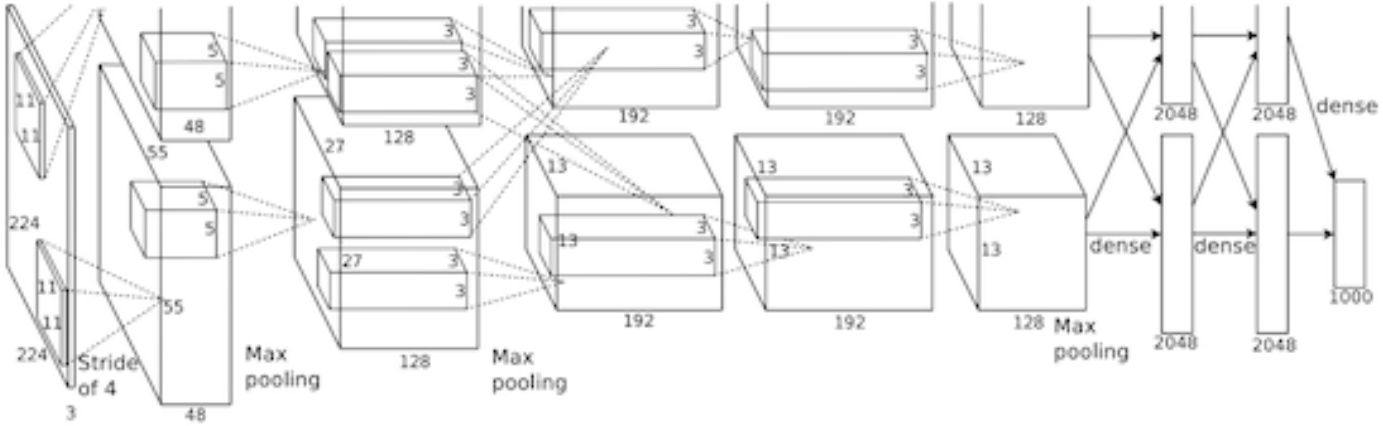


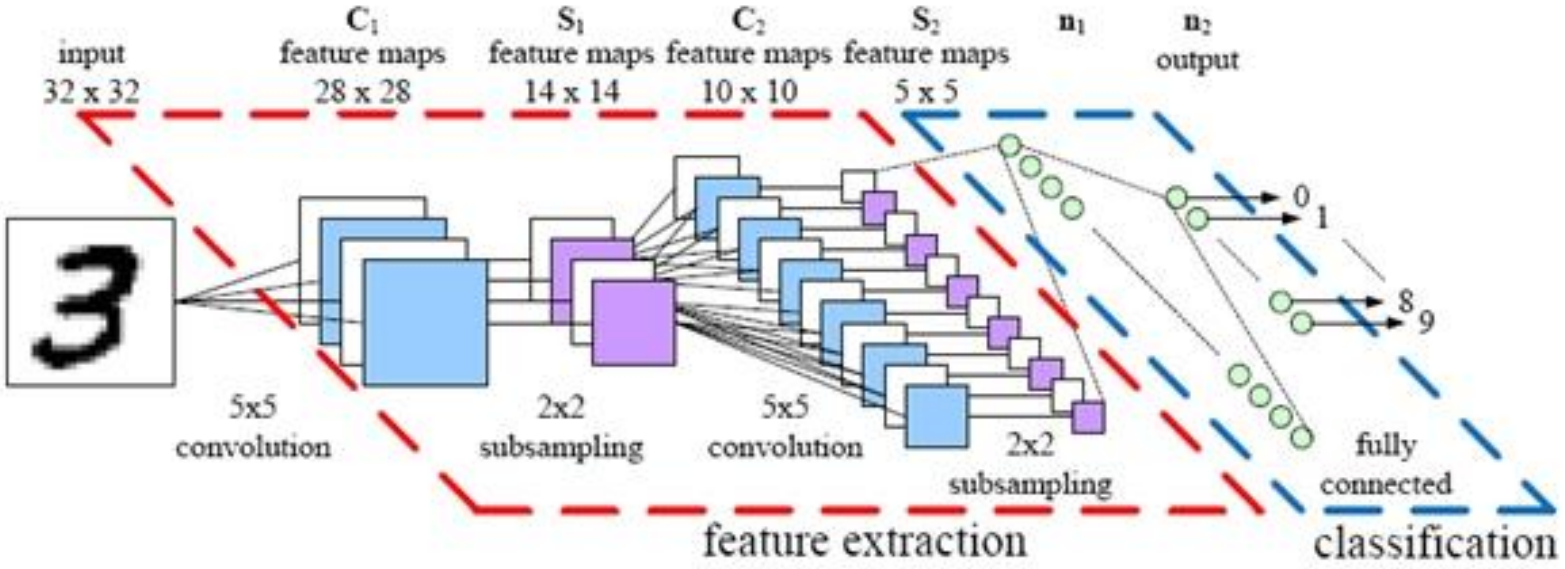


From pixels … to text

An overview of OCR mechanisms

Deep Learning (Conv. Neural Net)

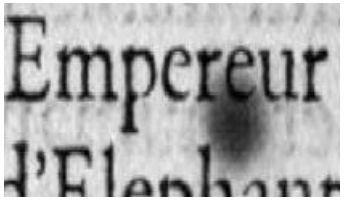


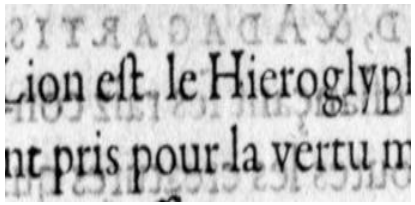


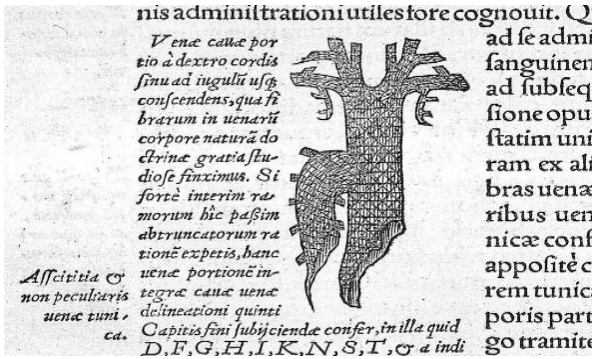
From pixels … to text

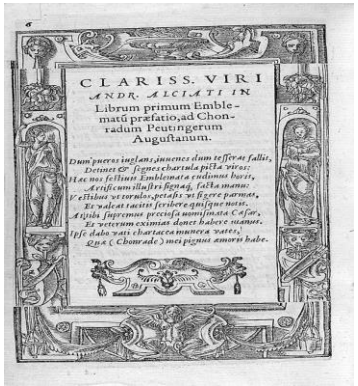
An overview of OCR mechanisms

- **Why commercial OCR are not working well on historical documents?**
  - Noises and degradations
  - Unusual layout
  - Unsuited training set

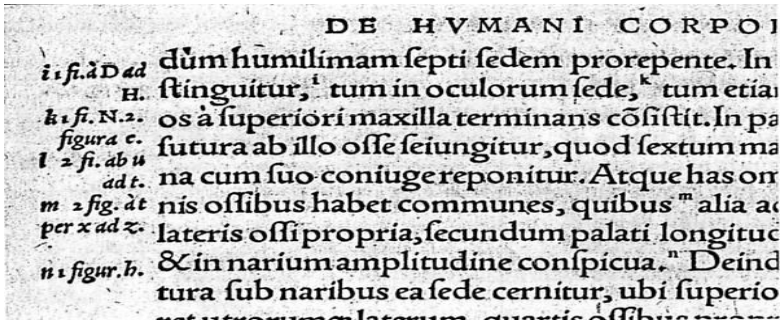


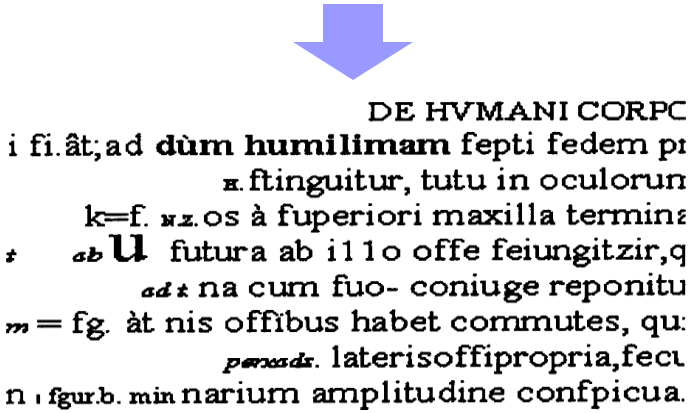






Fine Reader

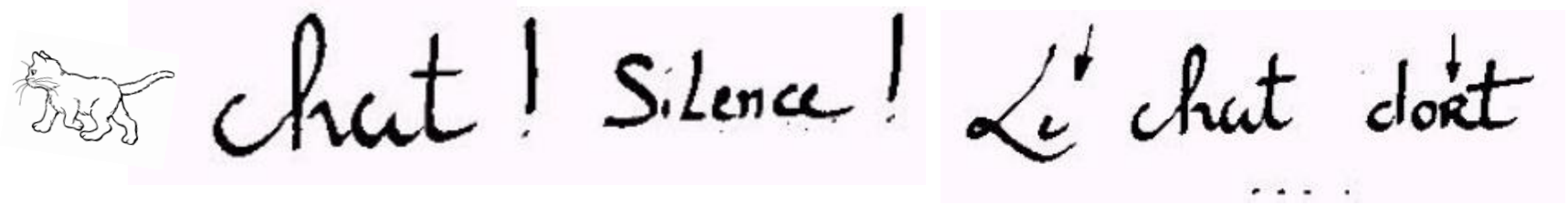




UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

From pixels … to text
An overview of OCR mechanisms

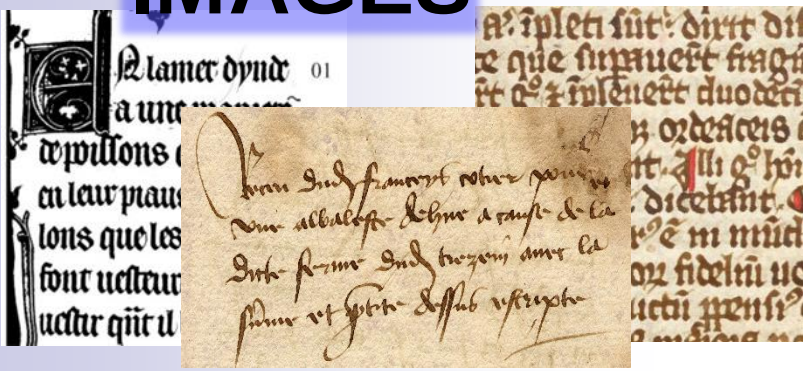
- **Why commercial OCR are not working well on historical documents?**
- Lack of data, knowledge and experiences
  - Unusual fonts and characters ➔ training data needs to be created
  - Unusual languages➔ Lexicons, dictionaries and language models need to be created
- Context often allows to modify our understanding of what is perceived by our senses
  - Until now, we tried to recognized EoC without using their context
  - The same EoC could be interpreted differently according to its surrounding context
  - Results of OCR are highly correlated to the adequacy of the used **word dictionary**



- Is there methods that need less a priori knowledge?
- Processing non-Textual parts can be good source of inspiration?

… to non-text

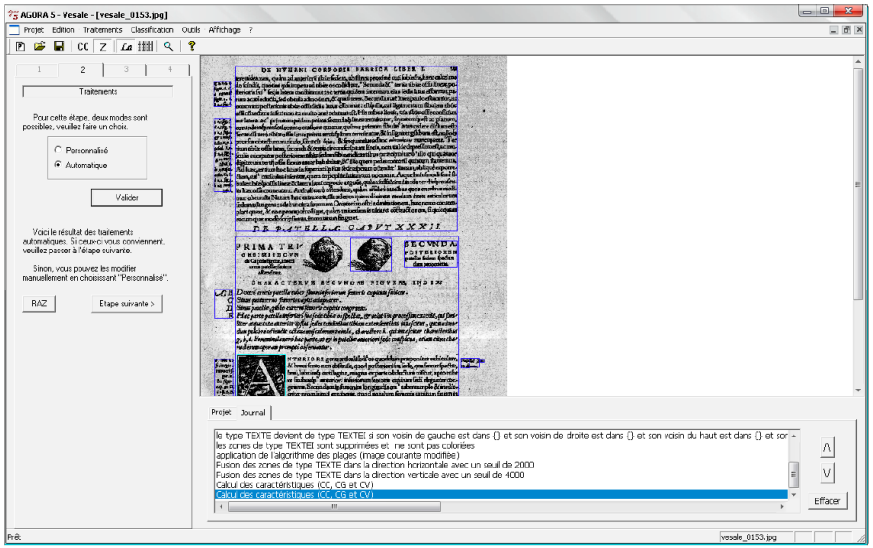
IMAGES



CONTENTS

From pixels … to non-textual contents

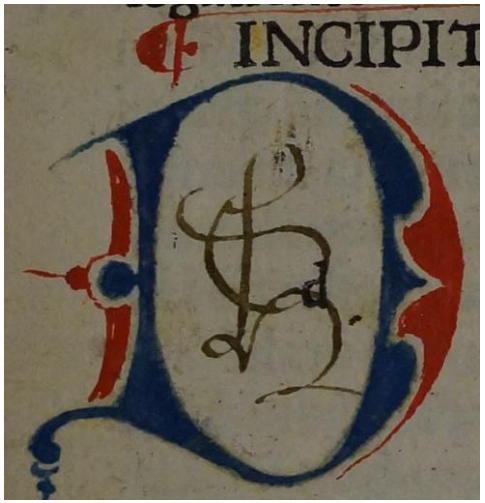
Pictorial Content is also of high interest



- **Ornamental letters (+of 20000)**

- **Figures (+ de 1500)**

BATYR: http://www.bvh.univ-tours.fr

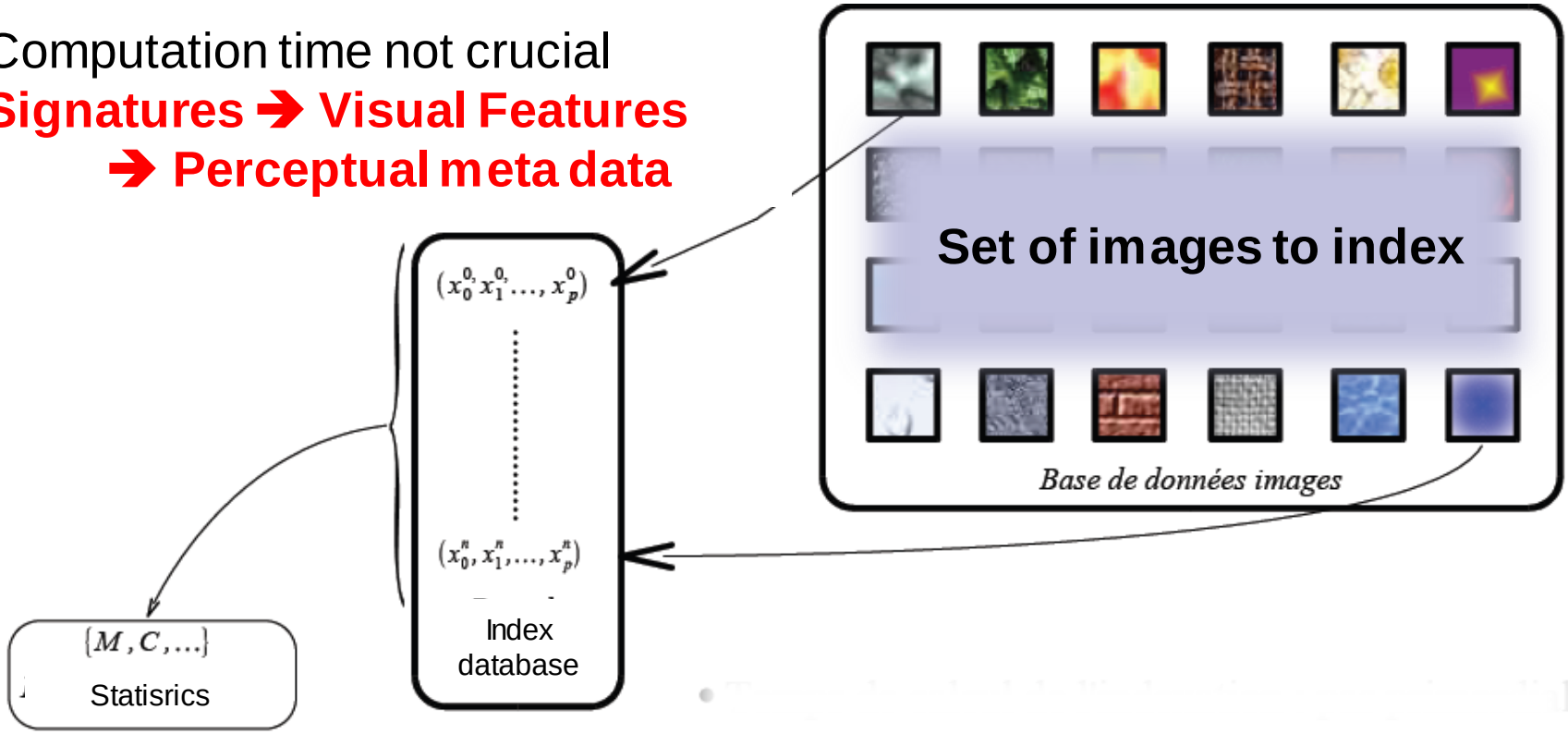


From pixels … to non-textual contents

An overview of Content Based Image Retrieval

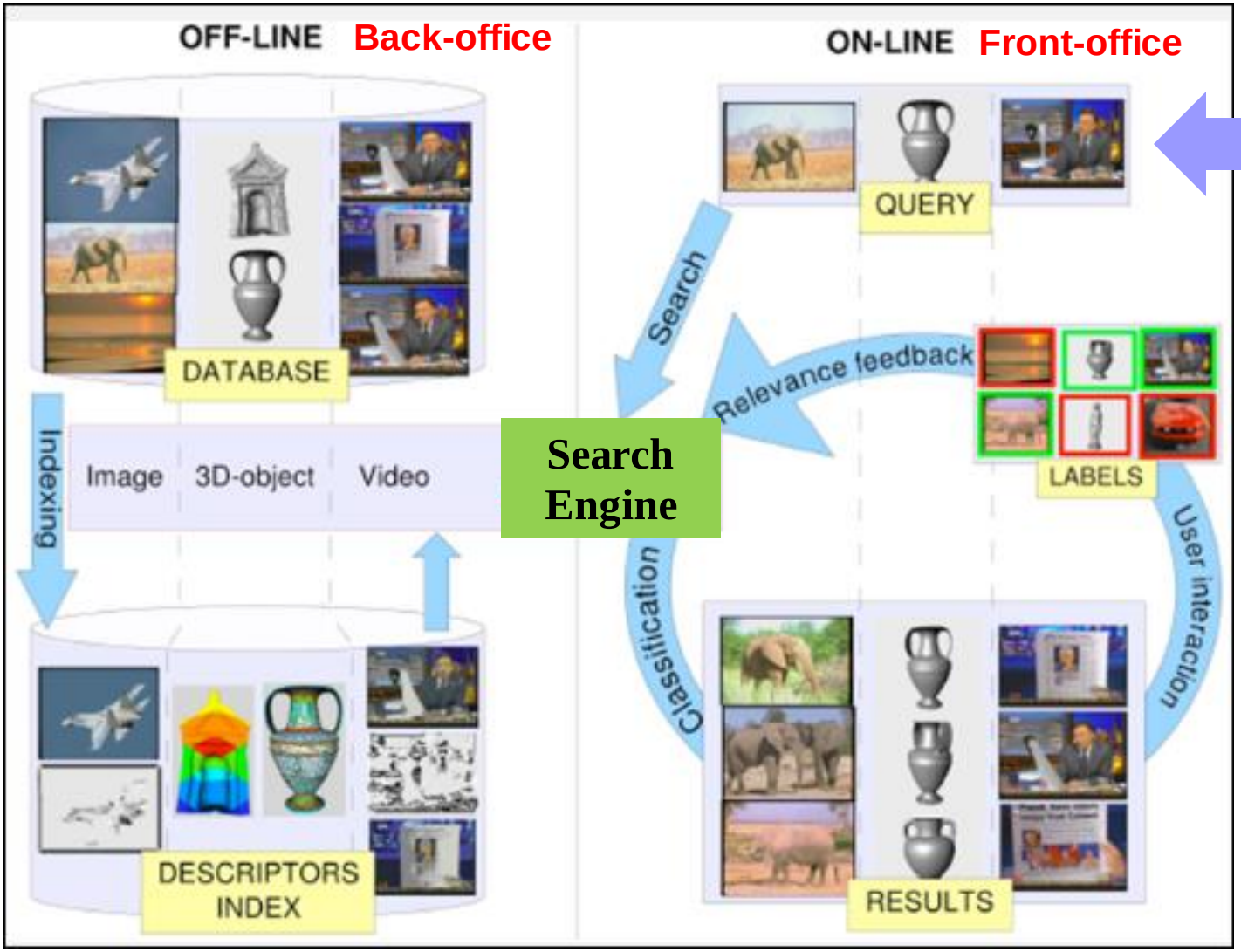
Perceptual meta data instead of classical meta data

- Computation of signatures for all the images or even **sub-parts of the images (EoC)**
- Computation time not crucial
- **Signatures ➔ Visual Features**
  ➔ Perceptual meta data



From pixels … to non-textual contents

An overview of Content Based Image Retrieval



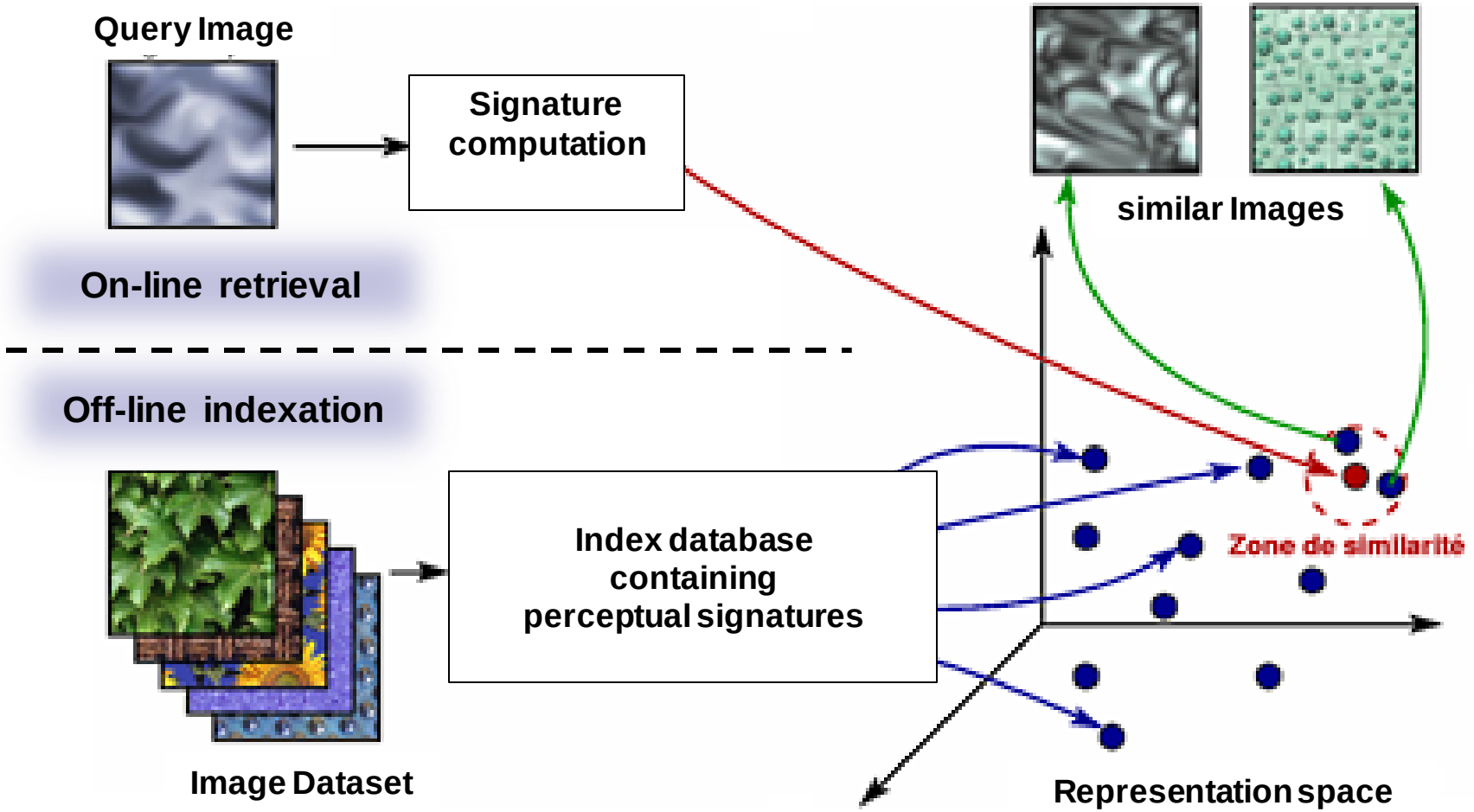
Using images as query instead of words

From pixels … to non-textual contents

An overview of Content Based Image Retrieval

It is again a question of features…

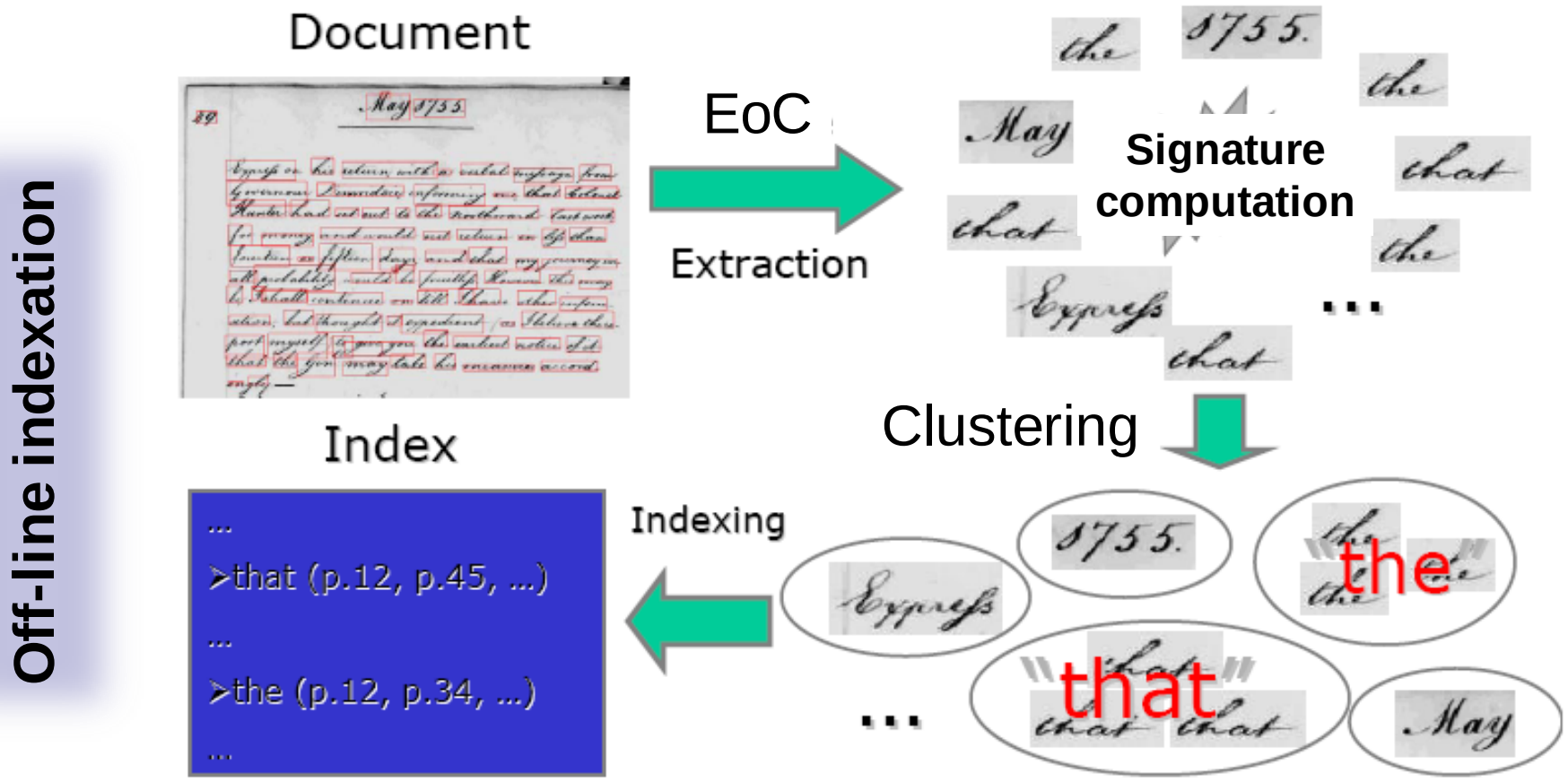
➔ We speak about signatures



From pixels … to non-textual contents

An overview of Content Based Image Retrieval

- From CBIR to Word spotting



UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

From pixels … to non-textual contents

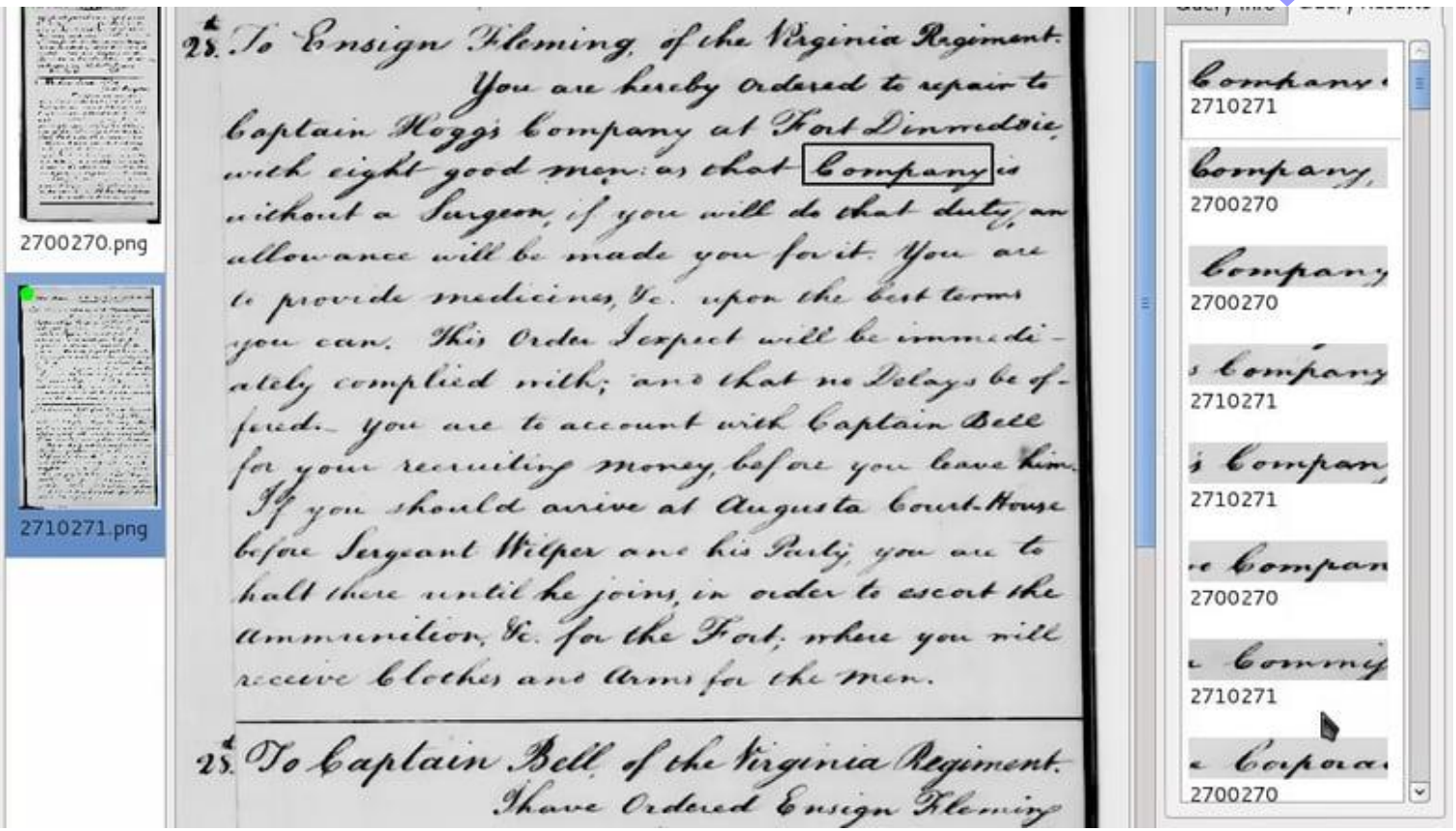
An overview of Content Based Image Retrieval

- From CBIR to Word spotting

Query images

Results: retrieved images

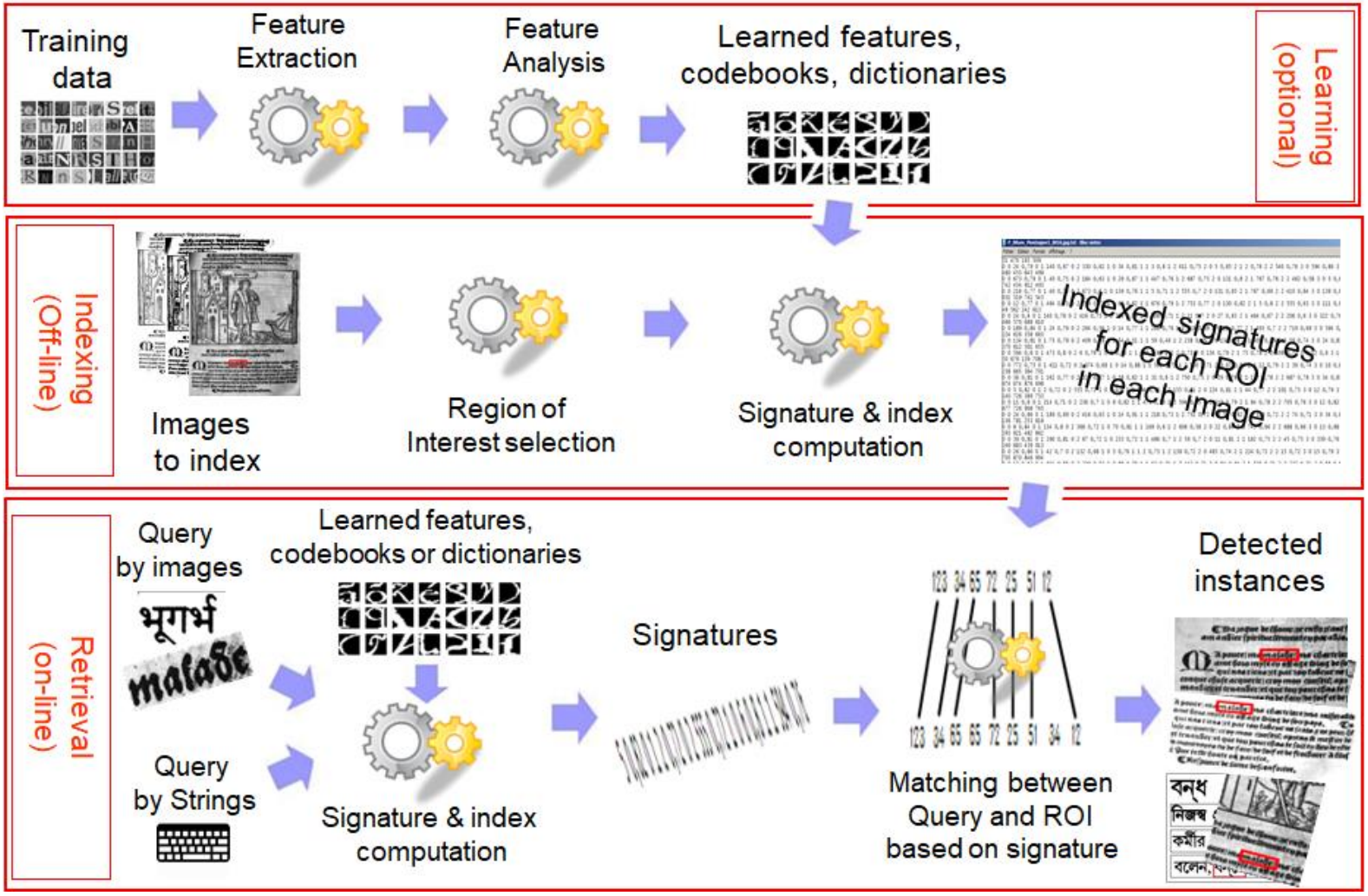
On-line Retrieval



From pixels … to non-textual contents

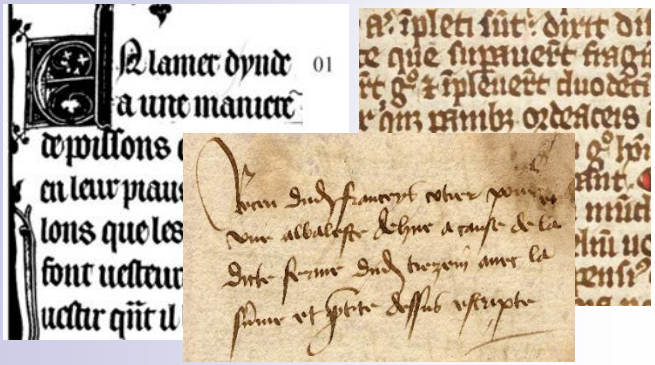
An overview of Content Based Image Retrieval

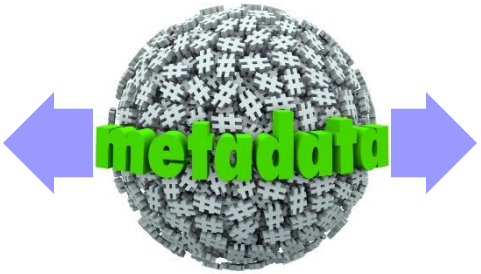
- From CBIR to Word spotting



Is it just a question of meta-data?

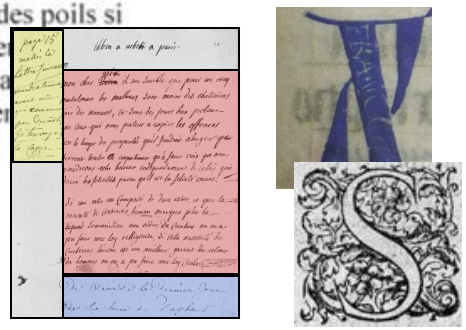
IMAGES





CONTENTS

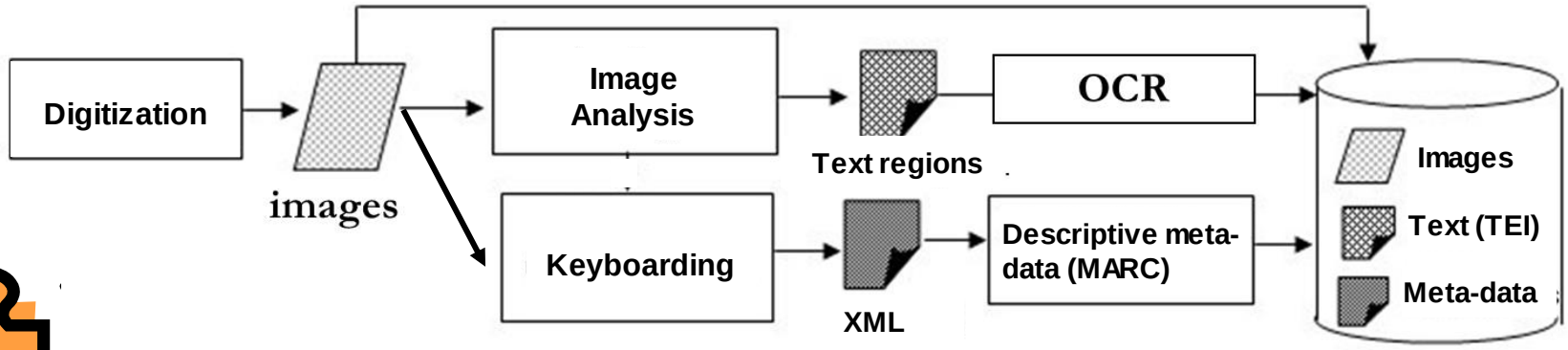
Dans la Mer d'Inde
il y a une espèce
de poissons qui ont
dans leur peau des poils si
longs que les ge
quant il les ont a
en font des vête
pour s'habiller.

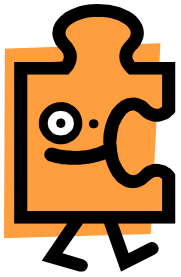


The standard model
Descriptive meta-data + transcription

We have usually

- Descriptive meta-data in standard formats **(MARC, EAD, Dublin Core, MODS, …)**
  - Edited manually
  - “Semantical” information
- Text transcription associated to additional meta-data (TEI)
  - Semi-automatic transcription or manually edited
  - “Semantical” information

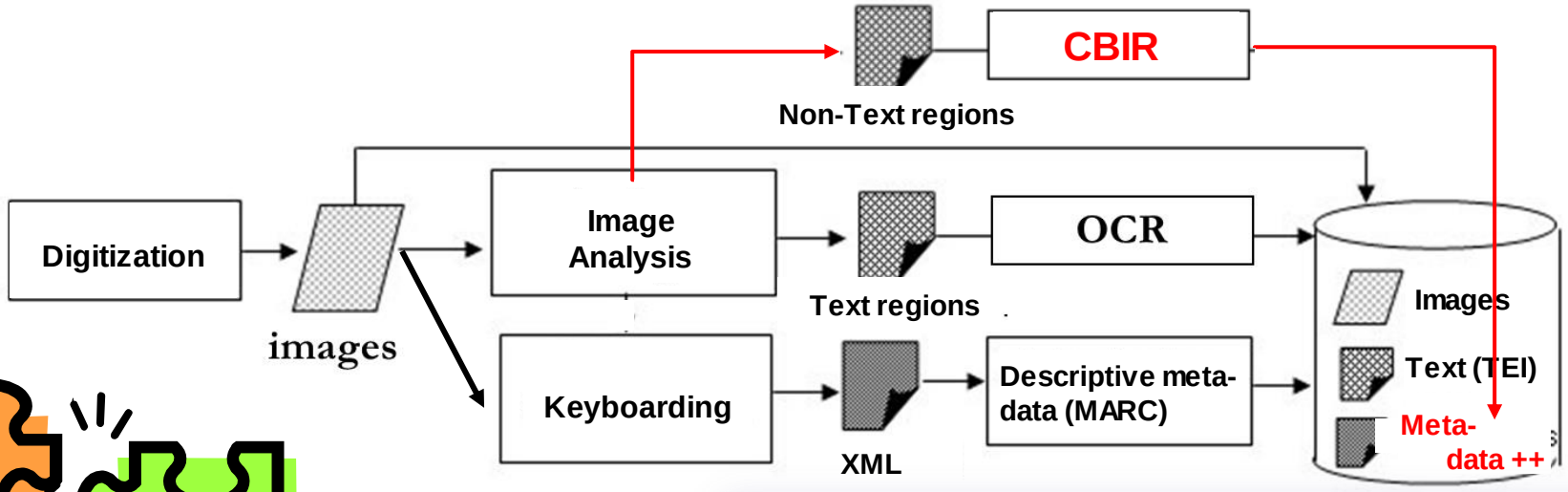


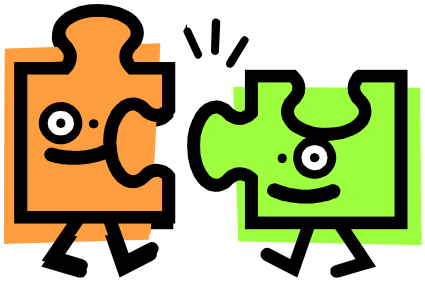


“Semantical” meta-data

The standard model
Perceptual meta-data

- It seems that **CBIR can help** to extract and save supplementary information about image content (EoC) without going to the semantical aspect (recognition)
  - Regions of interest
  - Visual features ➔ Perceptual **signatures and index**
  - Shapes, positions, colors, textures, … ➔ Numerical values (vectors)





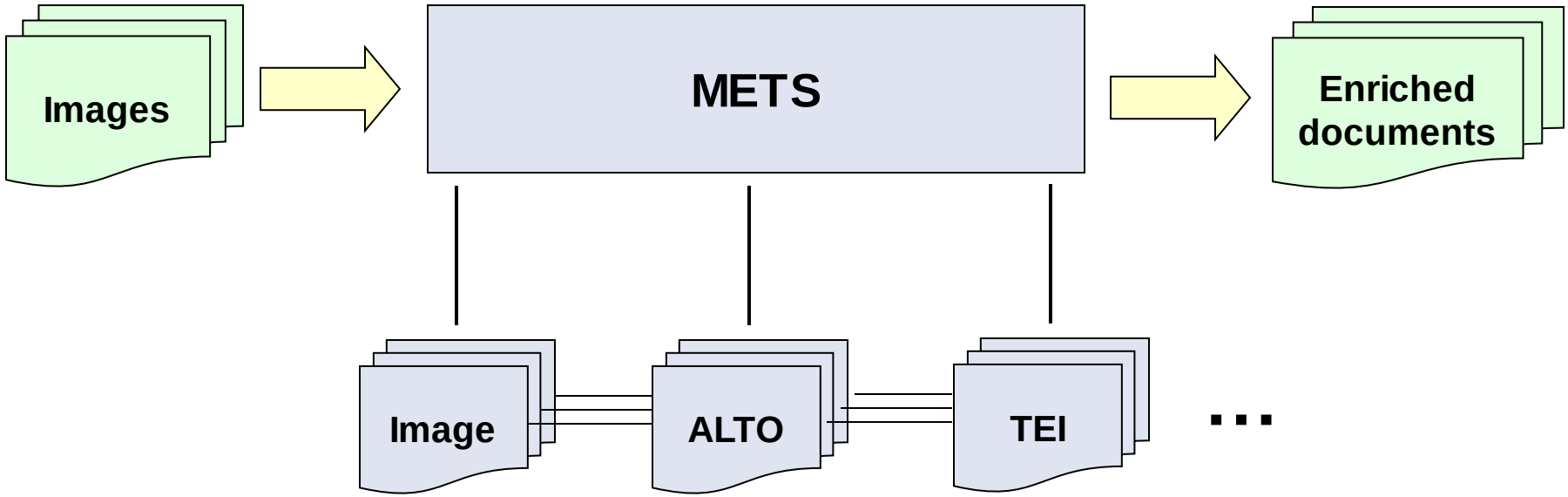
+ “Perceptual” meta-data

Is it just a problem of Meta-data?

What about the encoding formats?

Structuration of data and file formats is a difficult problem

- Data architects are needed
- Some interesting formats linked with previous discussions
  - **METS – Metadata Encoding and Transmission Standard**
  - ALTO – Analyzed Layout and Text Object
  - TEI – Text Encoding Initiative

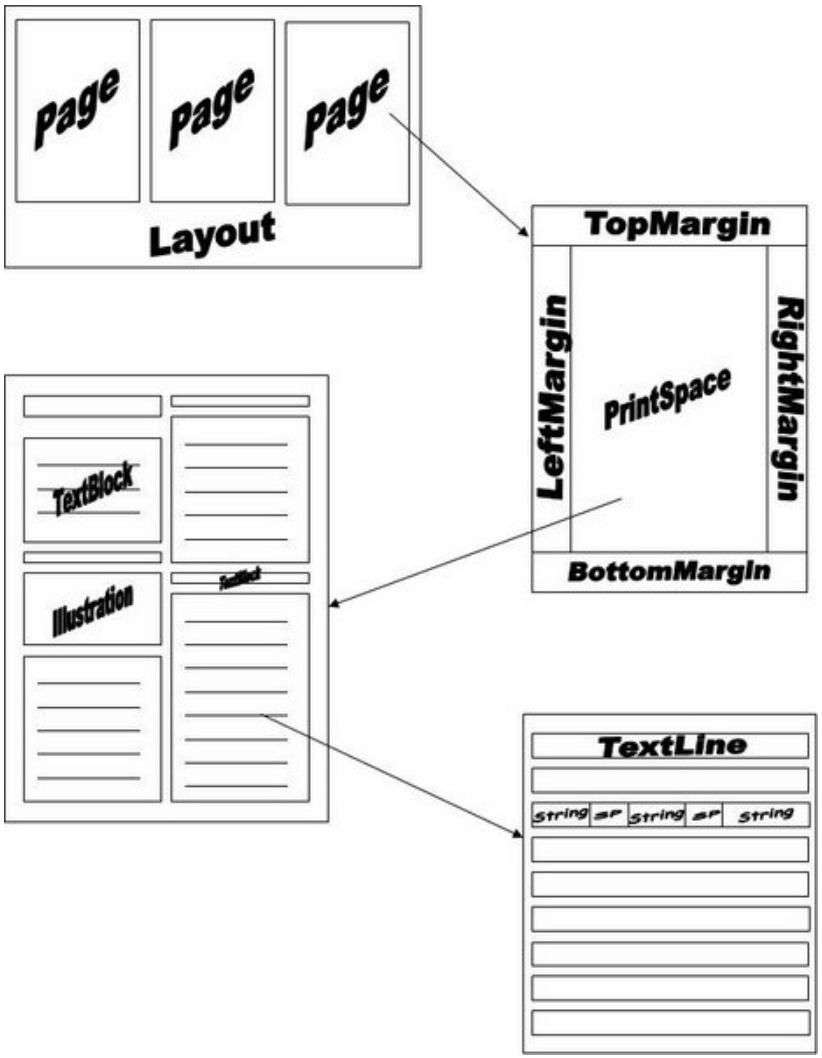


Is it just a problem of Meta-data?

ALTO for OCR

ALTO = Analyzed Layout and Text Object

- Standard XML
- Created in 2003 during METAe project
- Developed by Graz, Linz, Innsbruck universities
- **Description of the content and the physical layout of one page**
- Used by several OCR software
- Adapted and used by the BNF and other libraries
- Drawbacks: huge / static

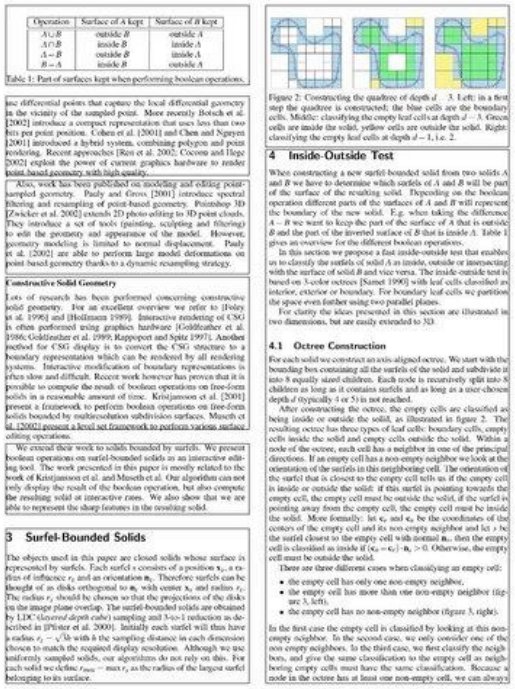


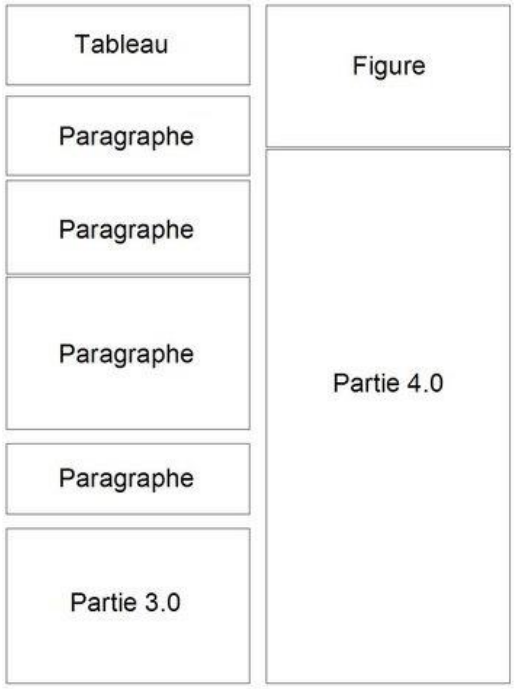
Is it just a problem of Meta-data?

TEI for transcription and enriched contents

- **TEI = Text Encoding Initiative - Standard XML**
  - Content tagging and logical structure encoding (full document)
  - Used a lot by libraries
  - Too much « open »? ➔ Quit « complex »

```
<TEI>
   <text>
      <body>
         <table rows="5" cols="3"/>
         <p xml:id="par0"/>
         <p xml:id="par1"/>
         <p xml:id="par2"/>
         <p xml:id="par3"/>
         <div n="3" type="partie"/>
         <figure/>
         <div n="4" type="partie"/>
      </body>
   </text>
</TEI>
```





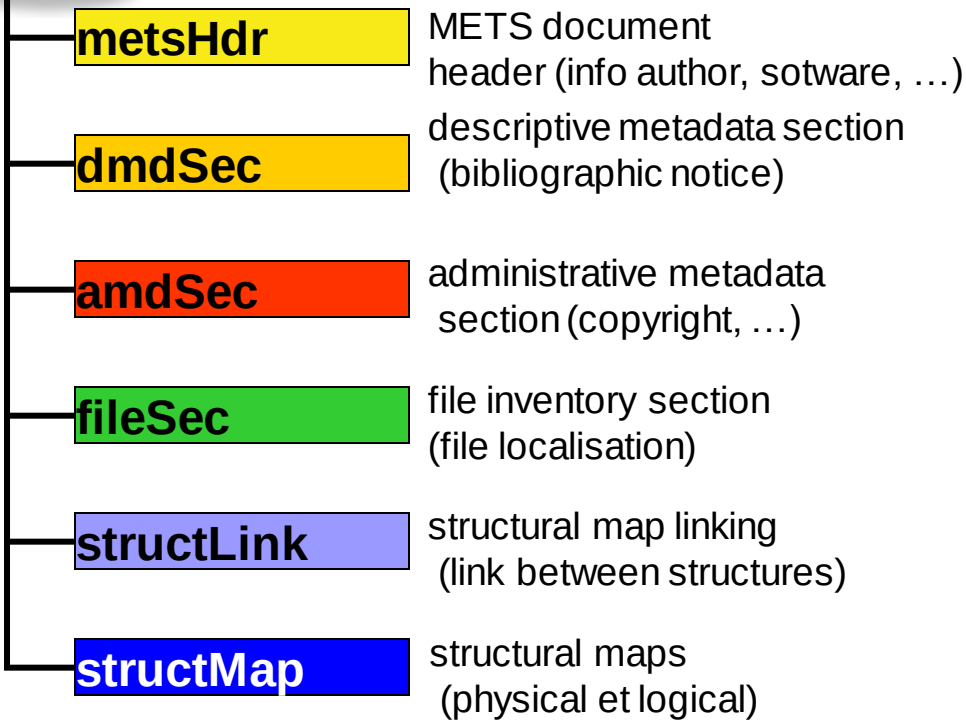
Is it just a problem of Meta-data?

Link between meta data?

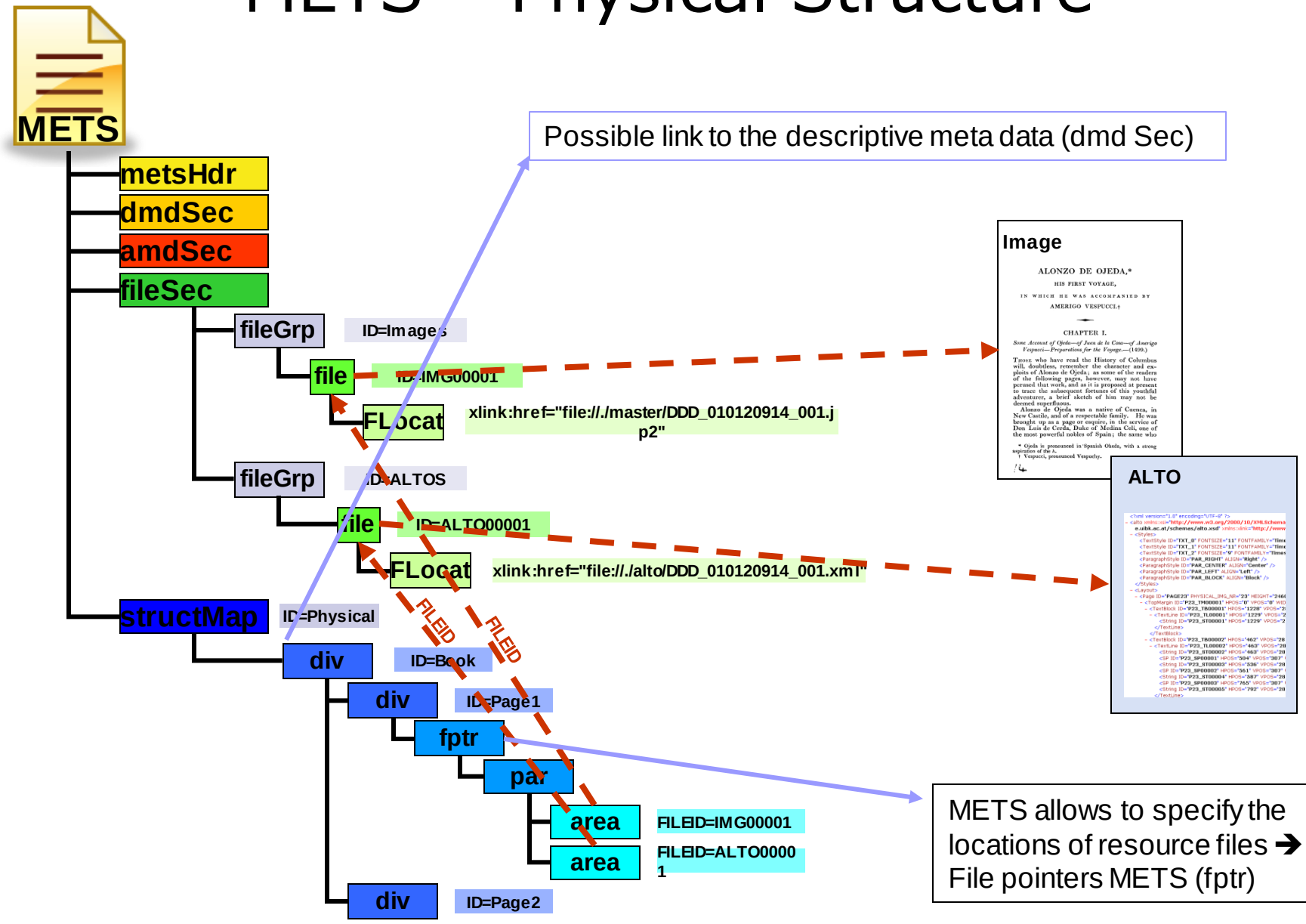
METS – Metadata Encoding and Transmission Standard

- Open XML Standard created in 2001 by the Digital Library Federation maintained by METS Editorial Board
- XSD-Schema
- **Linking between multimedia objects**
- Complete Description of digitized content (images, texts, audio, sculptures, …)
- Physical / logical structures
- Descriptive Meta data (DC, MODS, MARC, …)
- …

METS – First level Elements

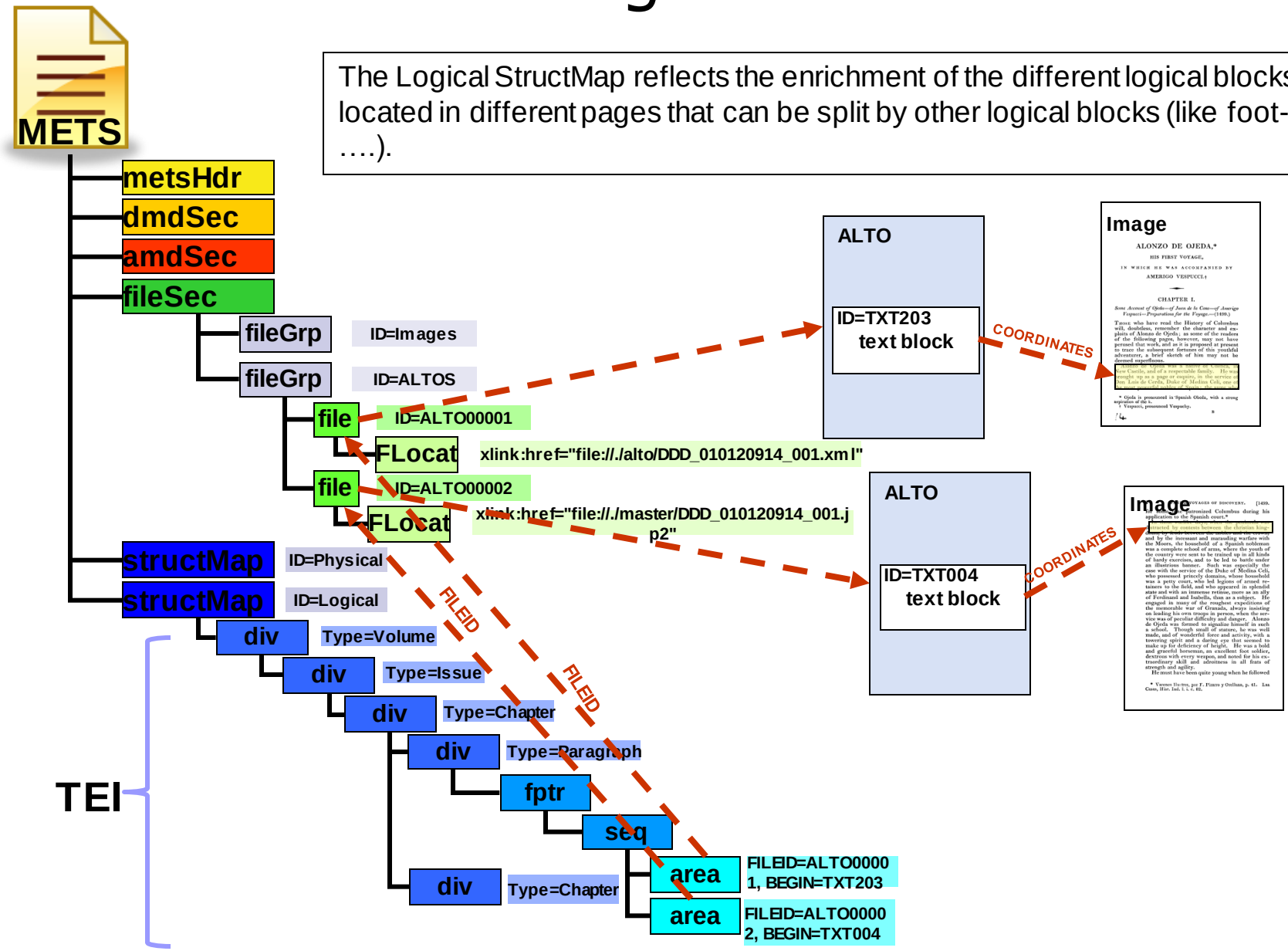


METS – Physical Structure



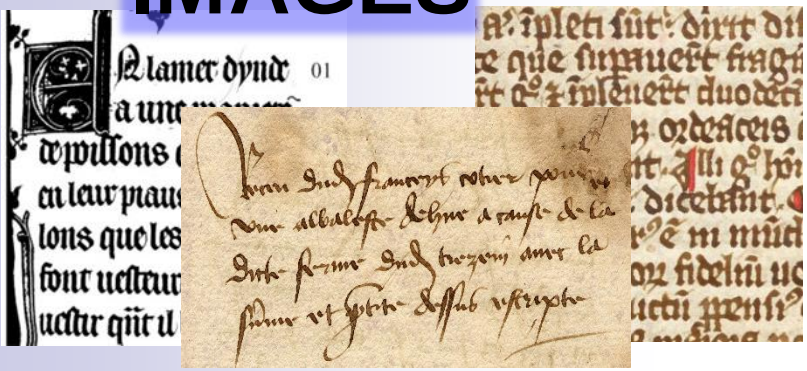
METS – Logical Structure

The Logical StructMap reflects the enrichment of the different logical blocks located in different pages that can be split by other logical blocks (like foot-notes, ….).



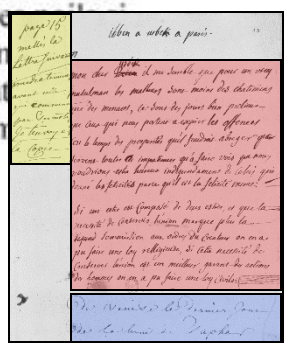
Conclusions

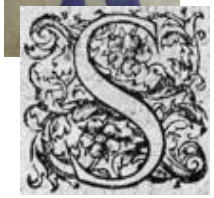
IMAGES



CONTENTS

Dans la Mer d'Inde
il y a une espèce
de poissons qui ont
dans leur peau de
longs que les gen
quant il les ont at
en font des vêten





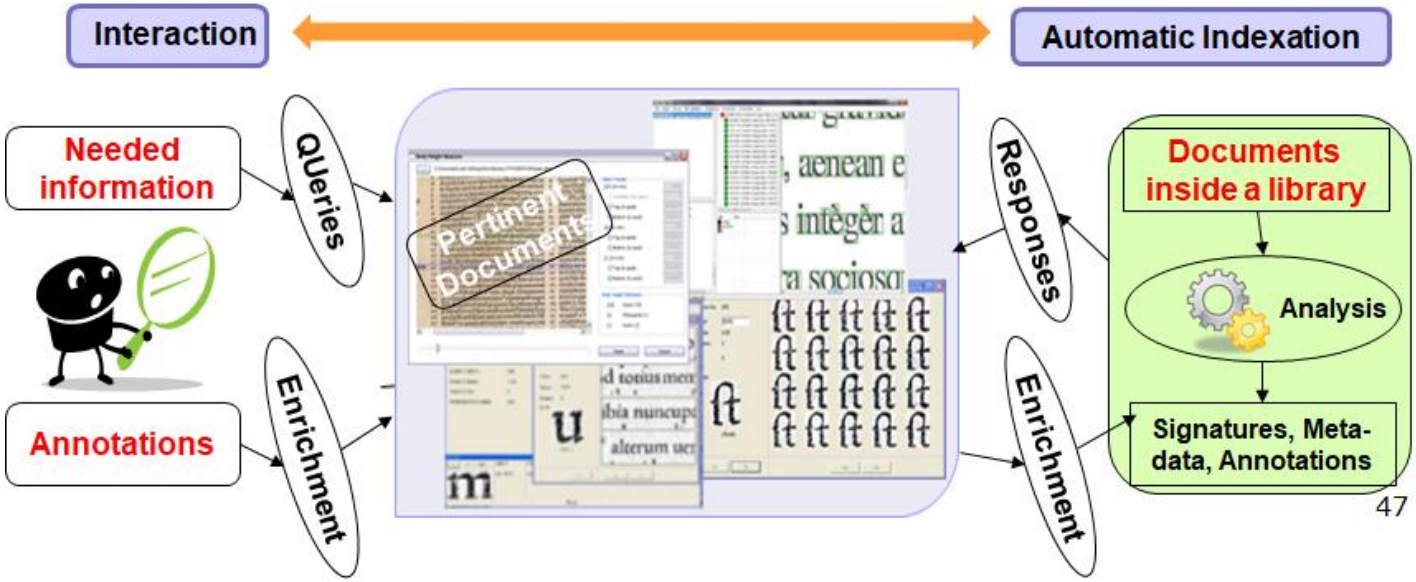
UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

From Pixels to Contents
Conclusions

- **Building tools for the valorisation of digitized historical content is a pluri-disciplinary task**
  - Meta-data production ➔ Experts of the domains
  - Selection and verification of the data ➔ Experts + Data accuratist
  - Structuration of the data and system ➔ Data / system architect
  - Computer vision, Machine learning ➔ Data scientist
- **Manual indexing is needed**
  - Descriptive meta-data ➔ Semantical meta data
  - Standard formats for data encoding
  - Annotations could be seen as supplementary meta-data?
- **Operational methods and tools are available**
  - Acquisition devices
  - Automatic tools: low level image processing, OCR
  - Perceptual meta-data should be added : CBIR

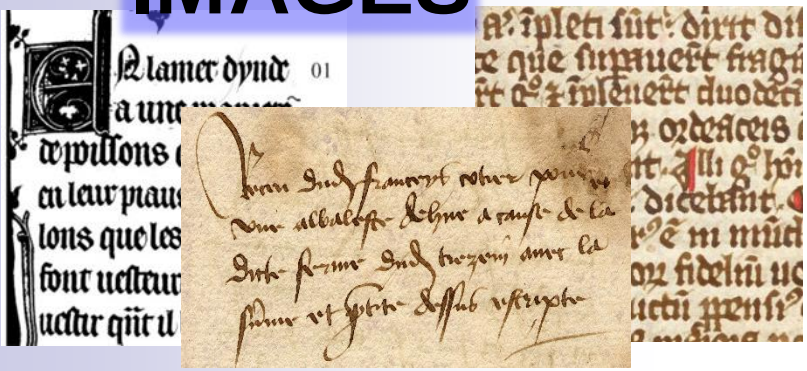
From Pixels to contents
Conclusions - Perspectives

- **The actual context: big data and heterogeneous collections**
  - Connexion between data, mutual enrichment, interoperability
  - Introduction and management of additional knowledge
  - Facing the diversity of the types of contents and usages
- **Quality of the interaction instead of only the quantity**
  - Semantic Web : queries reformulation, smart crawlers, automatic categorisation



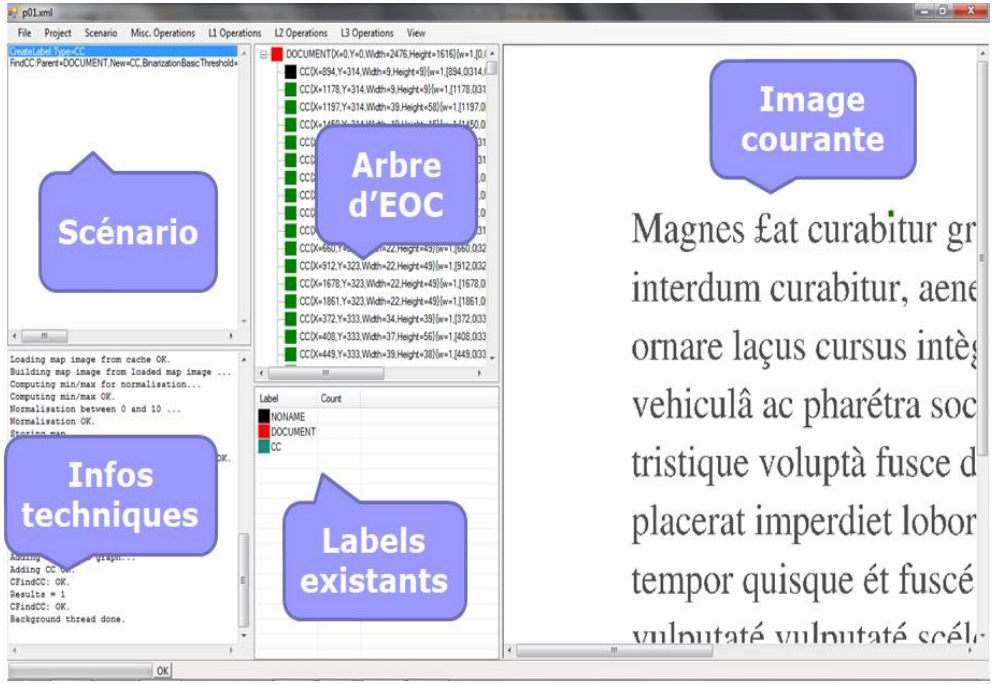
A walkthrough around interactive systems

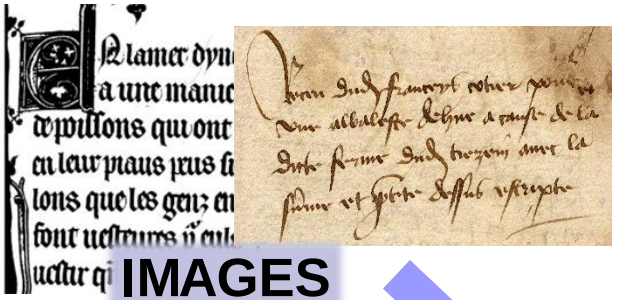
IMAGES



CONTENTS

A walkthrough around interactive systems
AGORA





Meta-Data ?

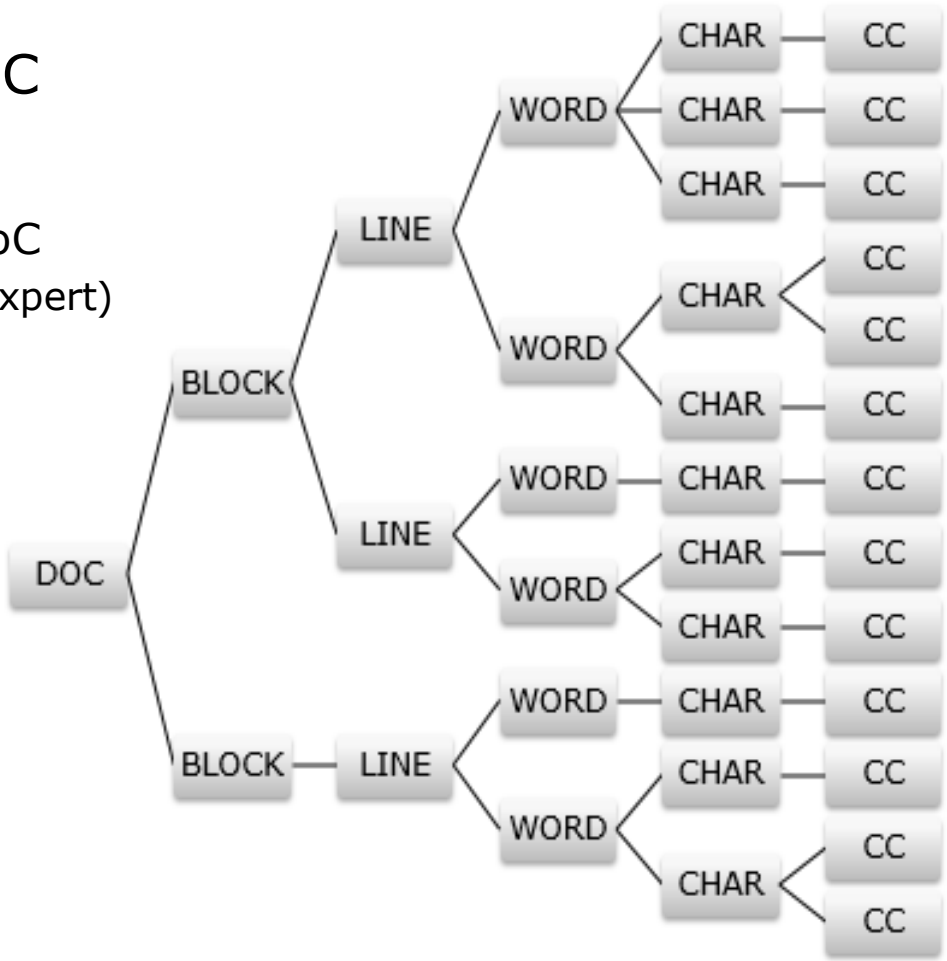
CONTENTS

Tools that let the users decide what to **extract / annotate** inside the images / documents

- **Utilization but also generation** of meta data
- **Variety of types** of meta-data

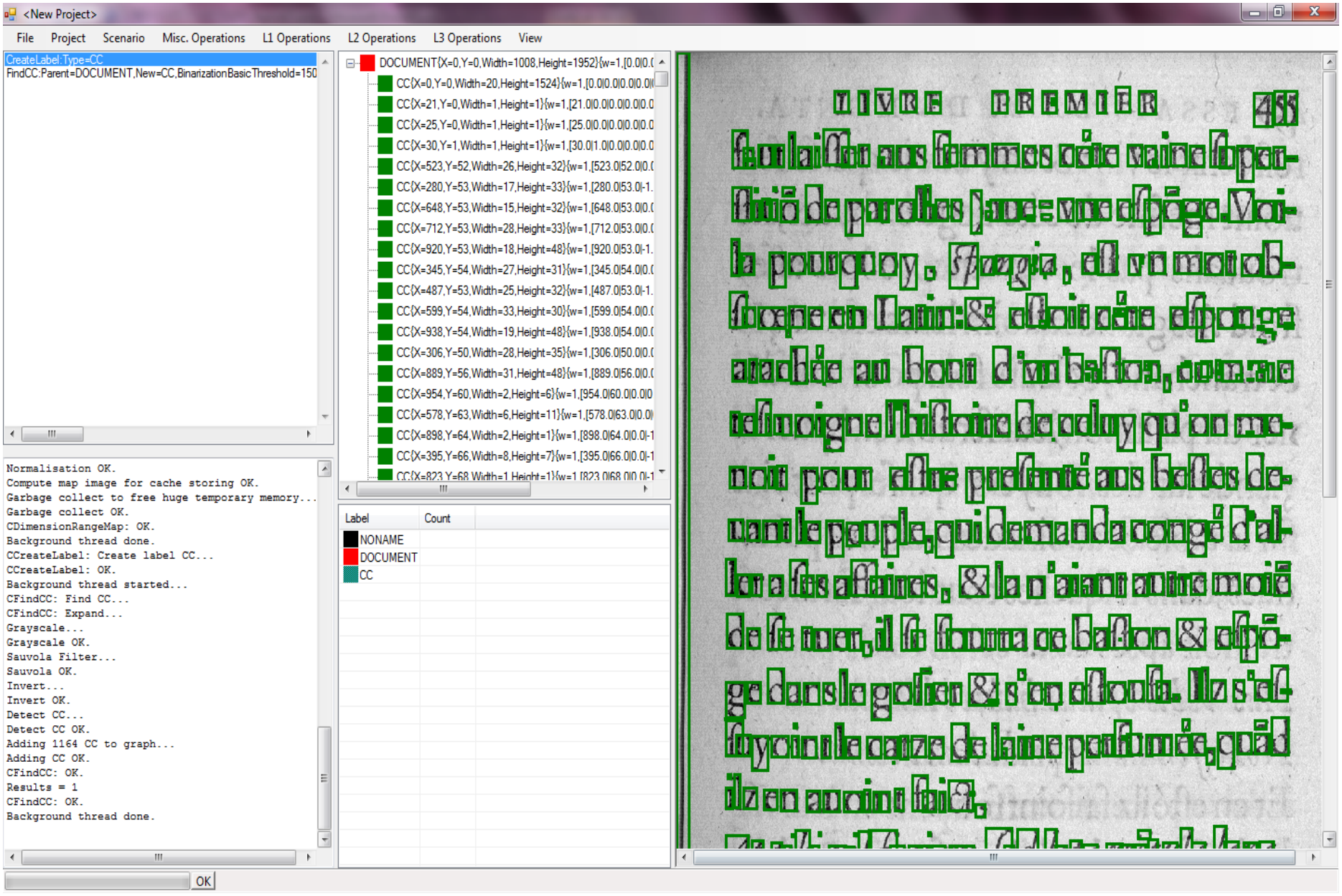
A walkthrough around interactive systems
AGORA

- Manipulation of a Tree of EoC
- 3 possible types of operations
  - Tagging, Fusion, deletion of EoC
  - 3 levels of operators (novice – expert)
- 3 types of criteria
  - EoC position
  - EoC neighbourhood
  - EoC characteristics (shape)
- Analysis scenarios
  - Built by the users
  - According to goals and books
  - A scenario repository
- Nothing pre-defined
  - Document models
  - EoC



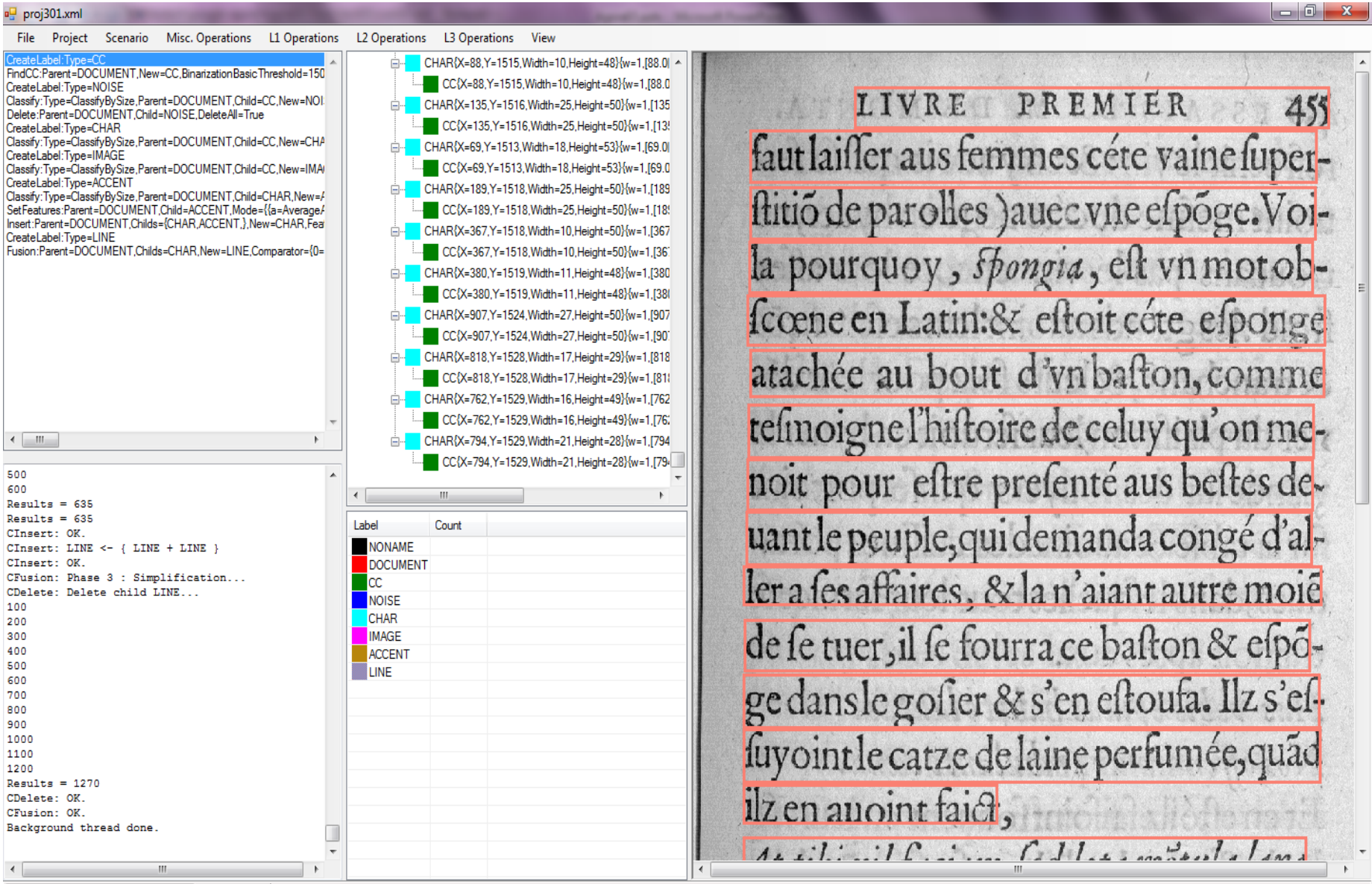
UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

AGORA

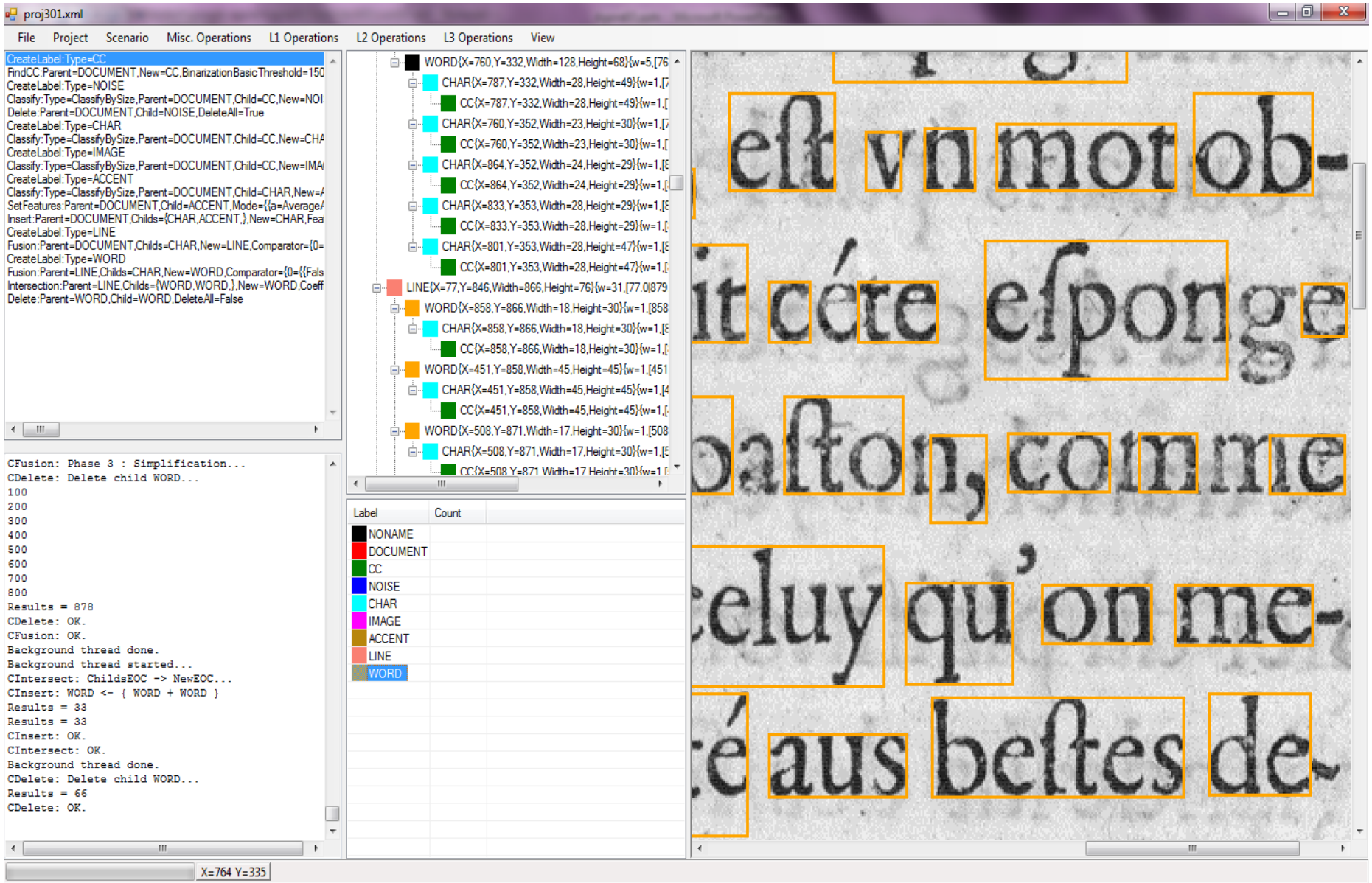


UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

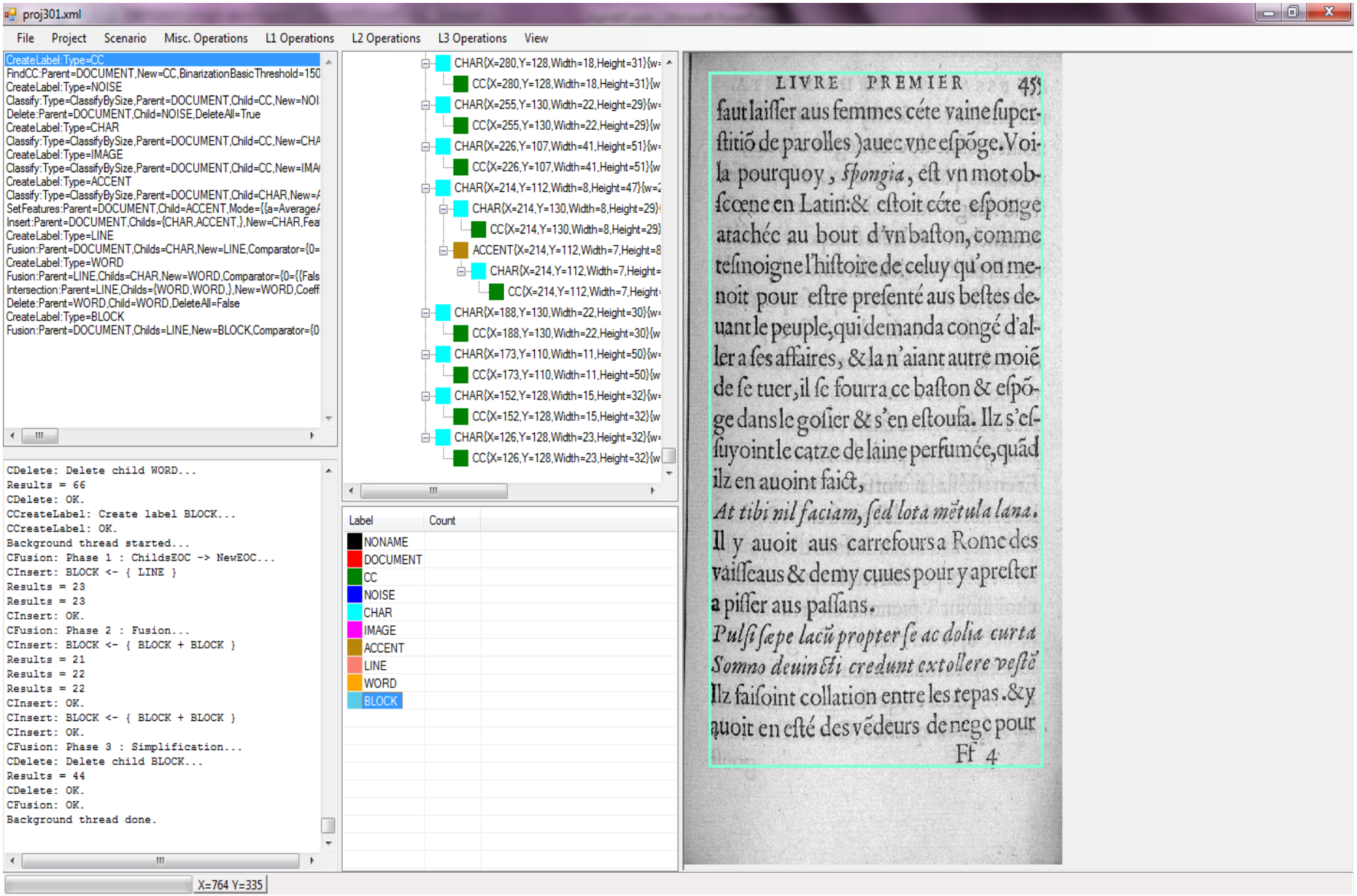
AGORA



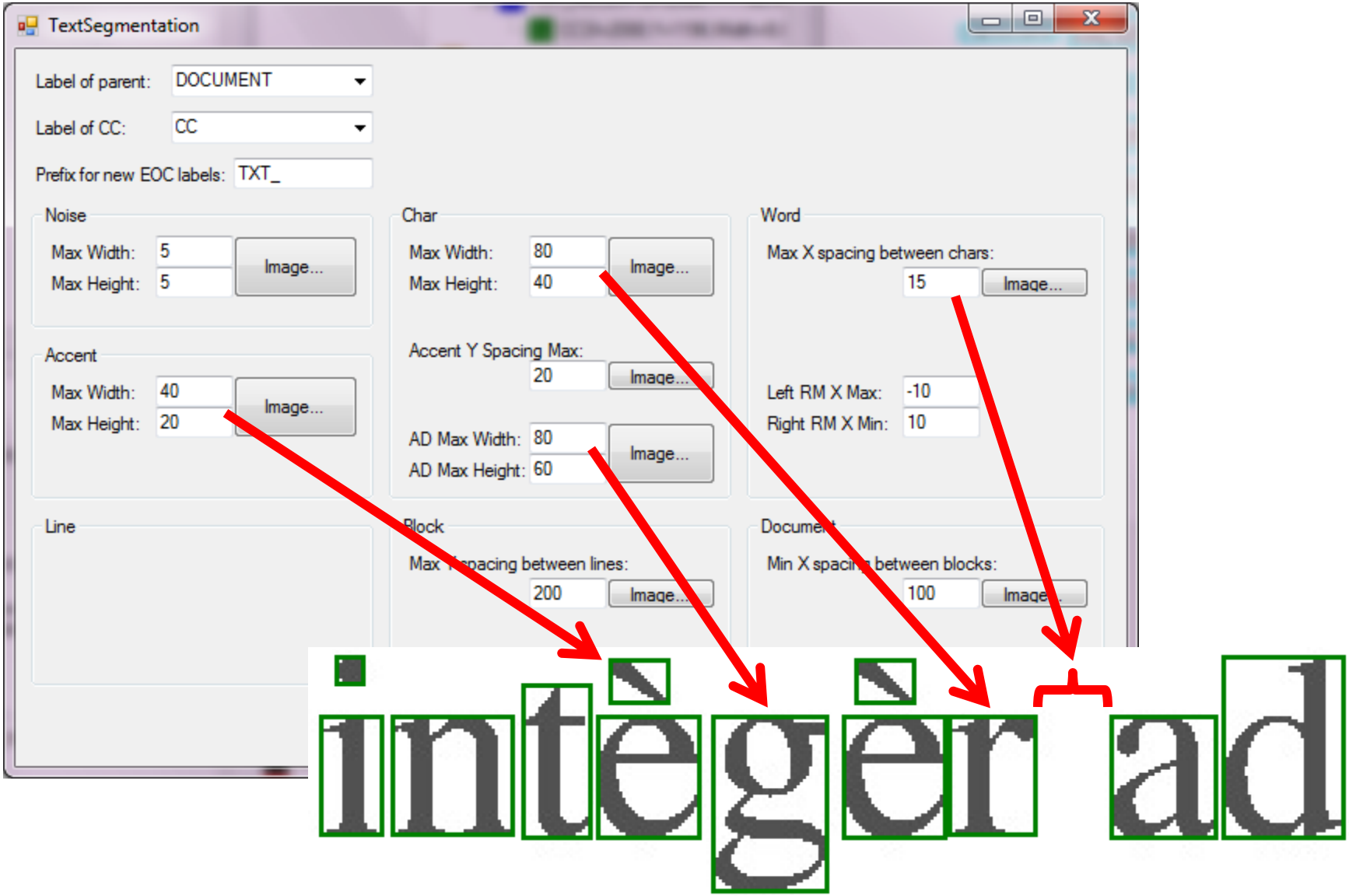
AGORA


proj301.xml
File
Project
Scenario
Misc. Operations
L1 Operations
L2 Operations
L3 Operations
View
CreateLabel:Type=CC
FindCC:Parent=DOCUMENT,New=CC,BinarizationBasicThreshold=150
CreateLabel:Type=NOISE
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=NOI
Delete:Parent=DOCUMENT,Child=NOISE,DeleteAll=True
CreateLabel:Type=CHAR
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=CHA
CreateLabel:Type=IMAGE
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=IMA
CreateLabel:Type=ACCENT
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CHAR,New=A
SetFeatures:Parent=DOCUMENT,Child=ACCENT,Mode={{a=AverageA
Insert:Parent=DOCUMENT,Childs={CHAR,ACCENT,},New=CHAR,Fea
CreateLabel:Type=LINE
Fusion:Parent=DOCUMENT,Childs=CHAR,New=LINE,Comparator={0=
CreateLabel:Type=WORD
Fusion:Parent=LINE,Childs=CHAR,New=WORD,Comparator={0={{Fals
Intersection:Parent=LINE,Childs={WORD,WORD,},New=WORD,Coeff
Delete:Parent=WORD,Child=WORD,DeleteAll=False
WORD{X=760,Y=332,Width=128,Height=68}{w=5,[76
CHAR{X=787,Y=332,Width=28,Height=49}{w=1,[7
CC{X=787,Y=332,Width=28,Height=49}{w=1,[
CHAR{X=760,Y=352,Width=23,Height=30}{w=1,[7
CC{X=760,Y=352,Width=23,Height=30}{w=1,[
CHAR{X=864,Y=352,Width=24,Height=29}{w=1,[8
CC{X=864,Y=352,Width=24,Height=29}{w=1,[
CHAR{X=833,Y=353,Width=28,Height=29}{w=1,[8
CC{X=833,Y=353,Width=28,Height=29}{w=1,[
CHAR{X=801,Y=353,Width=28,Height=47}{w=1,[8
CC{X=801,Y=353,Width=28,Height=47}{w=1,[
LINE{X=77,Y=846,Width=866,Height=76}{w=31,[77.0|879
WORD{X=858,Y=866,Width=18,Height=30}{w=1,[858
CHAR{X=858,Y=866,Width=18,Height=30}{w=1,[8
CC{X=858,Y=866,Width=18,Height=30}{w=1,[
WORD{X=451,Y=858,Width=45,Height=45}{w=1,[451
CHAR{X=451,Y=858,Width=45,Height=45}{w=1,[4
CC{X=451,Y=858,Width=45,Height=45}{w=1,[
WORD{X=508,Y=871,Width=17,Height=30}{w=1,[508
CHAR{X=508,Y=871,Width=17,Height=30}{w=1,[5
Label
Count
NONAME
DOCUMENT
CC
NOISE
CHAR
IMAGE
ACCENT
LINE
WORD
CFusion: Phase 3 : Simplification...
CDelete: Delete child WORD...
100
200
300
400
500
600
700
800
Results = 878
CDelete: OK.
CFusion: OK.
Background thread done.
Background thread started...
CIntersect: ChildsEOC -> NewEOC...
CInsert: WORD <- { WORD + WORD }
Results = 33
Results = 33
CInsert: OK.
CIntersect: OK.
Background thread done.
CDelete: Delete child WORD...
Results = 66
CDelete: OK.
X=764 Y=335

AGORA


proj301.xml
File
Project
Scenario
Misc. Operations
L1 Operations
L2 Operations
L3 Operations
View
CreateLabel:Type=CC
FindCC:Parent=DOCUMENT,New=CC,BinarizationBasicThreshold=150
CreateLabel:Type=NOISE
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=NOI
Delete:Parent=DOCUMENT,Child=NOISE,DeleteAll=True
CreateLabel:Type=CHAR
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=CHA
CreateLabel:Type=IMAGE
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CC,New=IMA
CreateLabel:Type=ACCENT
Classify:Type=ClassifyBySize,Parent=DOCUMENT,Child=CHAR,New=A
SetFeatures:Parent=DOCUMENT,Child=ACCENT,Mode={{a=Average
Insert:Parent=DOCUMENT,Childs={CHAR,ACCENT,},New=CHAR,Fea
CreateLabel:Type=LINE
Fusion:Parent=DOCUMENT,Childs=CHAR,New=LINE,Comparator={0=
CreateLabel:Type=WORD
Fusion:Parent=LINE,Childs=CHAR,New=WORD,Comparator={0={{Fals
Intersection:Parent=LINE,Childs={WORD,WORD,},New=WORD,Coeff
Delete:Parent=WORD,Child=WORD,DeleteAll=False
CreateLabel:Type=BLOCK
Fusion:Parent=DOCUMENT,Childs=LINE,New=BLOCK,Comparator={0
CDelete: Delete child WORD...
Results = 66
CDelete: OK.
CCreateLabel: Create label BLOCK...
CCreateLabel: OK.
Background thread started...
CFusion: Phase 1 : ChildsEOC -> NewEOC...
CInsert: BLOCK <- { LINE }
Results = 23
Results = 23
CInsert: OK.
CFusion: Phase 2 : Fusion...
CInsert: BLOCK <- { BLOCK + BLOCK }
Results = 21
Results = 22
Results = 22
CInsert: OK.
CInsert: BLOCK <- { BLOCK + BLOCK }
CInsert: OK.
CFusion: Phase 3 : Simplification...
CDelete: Delete child BLOCK...
Results = 44
CDelete: OK.
CFusion: OK.
Background thread done.
CHAR{X=280,Y=128,Width=18,Height=31}{w
CC{X=280,Y=128,Width=18,Height=31}{w
CHAR{X=255,Y=130,Width=22,Height=29}{w
CC{X=255,Y=130,Width=22,Height=29}{w
CHAR{X=226,Y=107,Width=41,Height=51}{w
CC{X=226,Y=107,Width=41,Height=51}{w
CHAR{X=214,Y=112,Width=8,Height=47}{w
CHAR{X=214,Y=130,Width=8,Height=29}
CC{X=214,Y=130,Width=8,Height=29}
ACCENT{X=214,Y=112,Width=7,Height=8
CHAR{X=214,Y=112,Width=7,Height=
CC{X=214,Y=112,Width=7,Height
CHAR{X=188,Y=130,Width=22,Height=30}{w
CC{X=188,Y=130,Width=22,Height=30}{w
CHAR{X=173,Y=110,Width=11,Height=50}{w
CC{X=173,Y=110,Width=11,Height=50}{w
CHAR{X=152,Y=128,Width=15,Height=32}{w
CC{X=152,Y=128,Width=15,Height=32}{w
CHAR{X=126,Y=128,Width=23,Height=32}{w
CC{X=126,Y=128,Width=23,Height=32}{w
Label
Count
NONAME
DOCUMENT
CC
NOISE
CHAR
IMAGE
ACCENT
LINE
WORD
BLOCK
LIVRE PREMIER 45
faut laiſſer aus femmes céte vaine ſuper-
ſtitiõ de parolles)auec vne eſpõge. Voi-
la pourquoy, Spongia, eſt vn mot ob-
ſcœne en Latin: & eſtoit céte eſponge
atachée au bout d'vn baſton, comme
teſmoigne l'hiſtoire de celuy qu'on me-
noit pour eſtre preſenté aus beſtes de-
uant le peuple, qui demanda congé d'al-
ler a ſes affaires, & la n'aiant autre moiẽ
de ſe tuer, il ſe fourra ce baſton & eſpõ-
ge dans le goſier & s'en eſtouſa. Ilz s'eſ-
ſuyoint le catze de laine perfumée, quãd
ilz en auoint faict,
At tibi nil faciam, ſed lota mẽtula lana.
Il y auoit aus carrefours a Rome des
vaiſſeaus & demy cuues pour y apreſter
a piſſer aus paſſans.
Pulſi ſæpe lacũ propter ſe ac dolia curta
Somno deuincti credunt extollere veſtẽ
Ilz faiſoint collation entre les repas. & y
auoit en eſté des vẽdeurs de nege pour
Ff 4
X=764 Y=335

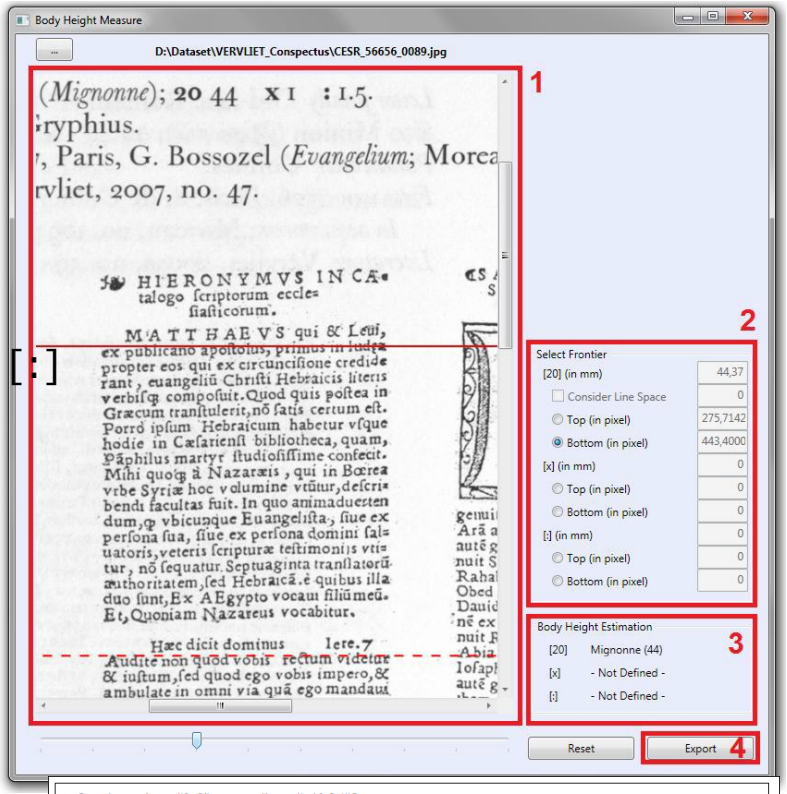
AGORA

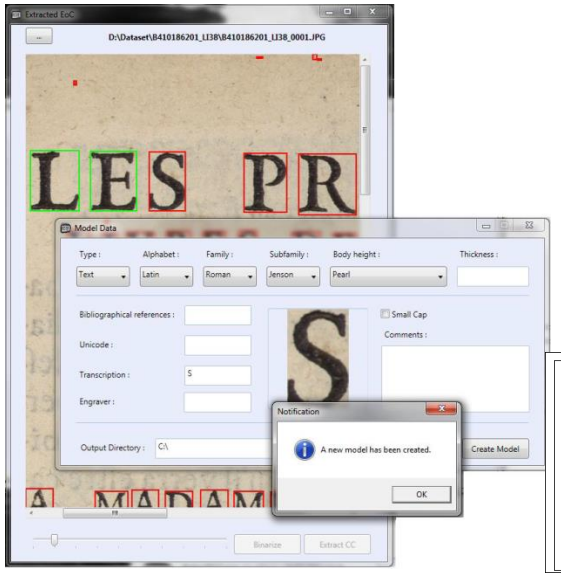

TextSegmentation
Label of parent:
DOCUMENT
Label of CC:
CC
Prefix for new EOC labels:
TXT_
Noise
Max Width:
5
Max Height:
5
Image...
Char
Max Width:
80
Max Height:
40
Image...
Word
Max X spacing between chars:
15
Image...
Accent
Max Width:
40
Max Height:
20
Image...
Accent Y Spacing Max:
20
Image...
Left RM X Max:
-10
Right RM X Min:
10
AD Max Width:
80
AD Max Height:
60
Image...
Line
Block
200
Image...
100
Image...
intègrad

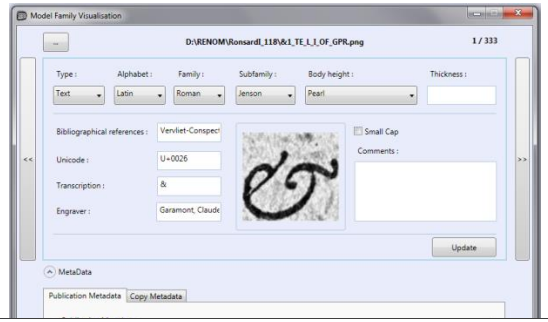
A walkthrough around interactive systems
Retro: Typographic analysis

Specific GUI

- Font models creation
- Model visualization
- Fonts Measurement and identification
  1. Select a page
  2. Selection of the frontier (top/bottom for [20], [x], [:] height), and display of measured values
  3. Estimated Body height designation
  4. Possibility to export computed information in xml







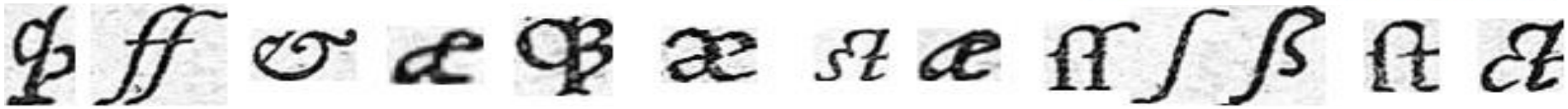

```
<?xml version="1.0" encoding="utf-16"?>
<!--RETRO Body Height file-->
<BodyHeight>
  <Filename Path="D:\Dataset\VERVLIET_Conspectus\CESR_56656_0089.jpg"
    <[20] Value="44,37" />
    <[x] Value="0" />
    <[:] Value="0" />
  </Typography>
</BodyHeight>
```

```
<?xml version="1.0" encoding="utf-16"?>
<!--RETRO Model file-->
<Model>
 <Metadata>
  <Book Author="" Title="" Place="" PrinterOrPublisher="" Date="" Format="" />
  <Copy Library="" CallNumber="" Digitalization="" Copyright="" CataloguerName="" />
 </Metadata>
 <Transcription Character="" Unicode="" />
 <Image Filename="" Folder="" Page="" Resolution="" />
 <Thumbnail Name="" Width="" Height="" PositionX="" PositionY="" />
 <Typography IsSmallCap="" Type="" Alphabet="" Family="" SubFamily="" BodyHeight="
          Thickness="" />
 <Description References="" Engraver="" Comments="" />
</Model>
```

A walkthrough around interactive systems

Retro: Typographic analysis

- Data generation
  - Which formats (XML + binary, color images)
  - Which meta-data for characters, glyphs, patterns (unicode, …)
- Training data generation
  - Which fonts ? Which models ? How many ?
  - Clean ? Degraded ?
- Early modern font classification
- Estheatic study of fonts
- Font Base Batyr du CESR
- Available data on Paradiit web site
  - The 'Gering' Pica-Roman [R80]
  - Cicéro (1478) (cf. [Vervliet2010] N°55)
  - Garamont's Great Primer Roman [R118]
  - Gros-romain (1549) (cf. [Vervliet2010] N°119)
  - Vérard Gothic Bâtarde Great Primer Roman

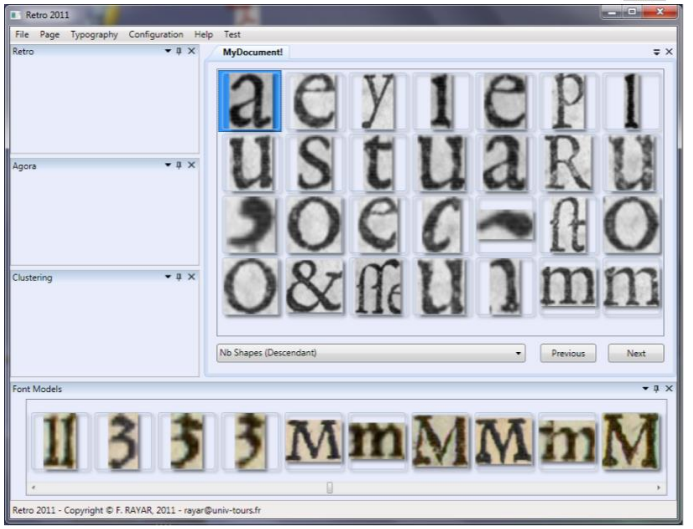


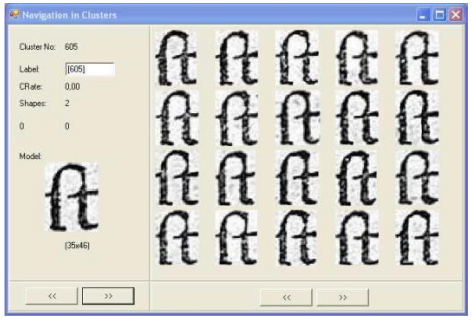
A walkthrough around interactive systems

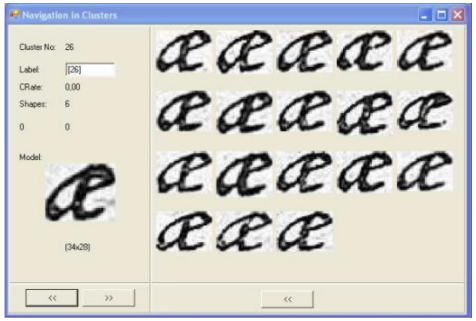
Retro: Assisted Transcription

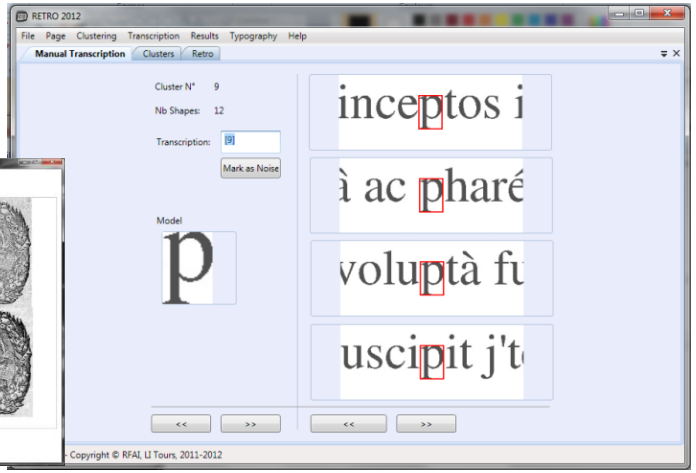
EoC Clustering ➔ Assisted transcription / indexing

- Feature selection
  - Pixels (NdG ? B&W ?)
  - Visual Features
- Comparison algorithms and metrics
  - Millions of matching
  - Cluster representatives
- Time complexity
  - 1 books = several days ➔ distribution
  - M. de files ➔ Mass of data







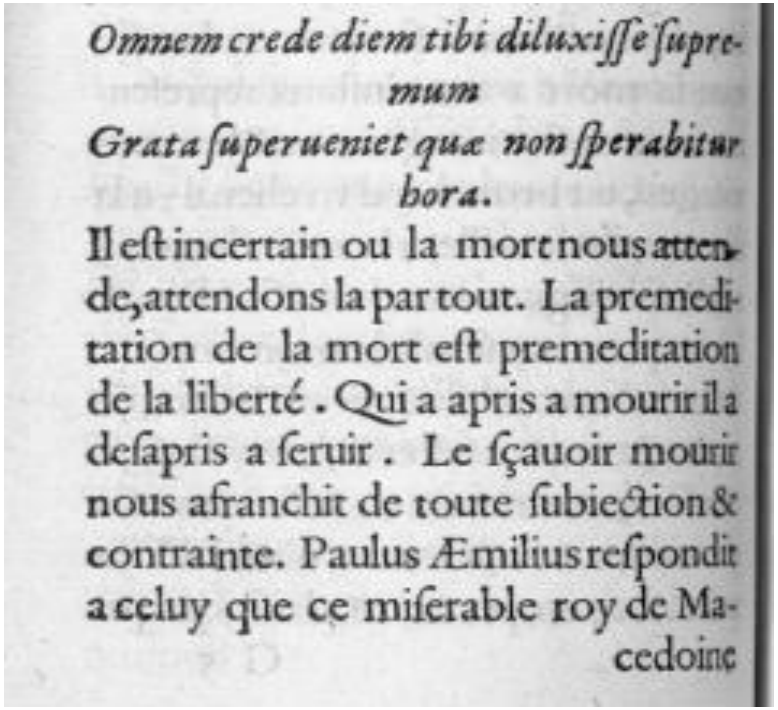


These tools are still to improve...

Extraction & Clustering of EoC

- ***Montaigne - 1557***
- 119 pages, 3260 blocs de texte centraux
- 125 744 composantes connexes (pseudo caracteres)
- 29 943 clusters
- 25 classes = 25% du texte
- 136 classes = 50%
- 1500 classes = 70%
- 20 000 classes = 90%
- 79% des classes comportent un seul élément
- Classe la plus grosse = 3%

- 1% des formes sont mal classées

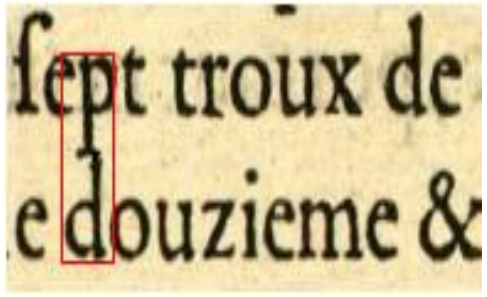

Omnem crede diem tibi diluxiſſe ſupremum
Grata ſuperueniet quæ non ſperabitur hora.
Il eſt incertain ou la mort nous attende, attendons la par tout. La premeditation de la mort eſt premeditation de la liberté. Qui a apris a mourir il a deſapris a ſeruir. Le ſçauoir mourir nous afranchit de toute ſubiection & contrainte. Paulus Æmilius reſpondit a celuy que ce miſerable roy de Macedoine

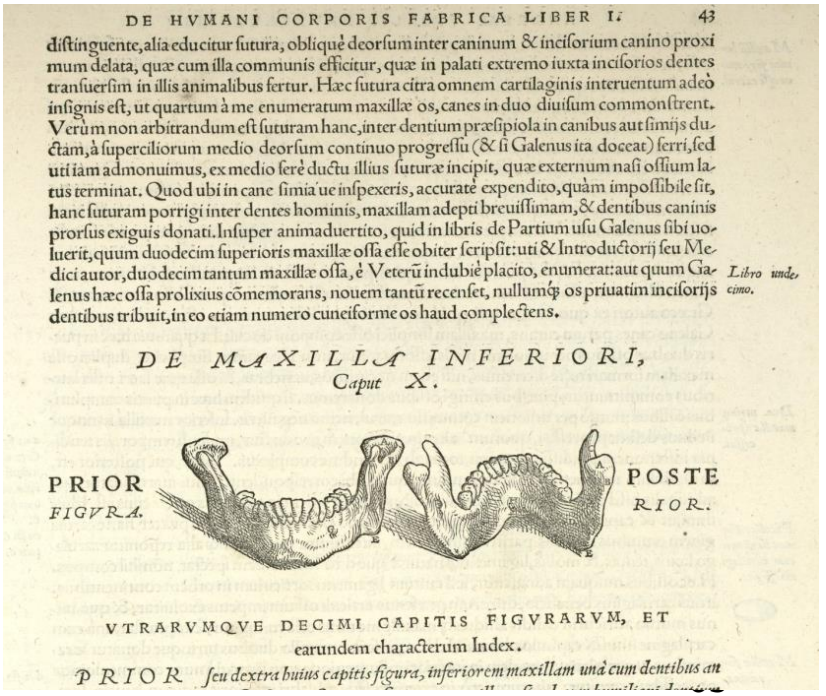
Extraction & Clustering of EoC

- ***Vésale – Latin - 1543***
- 150 pages avec notes en marges
- 1.062.081 composantes connexes (pseudo caractères)
- 40.000 classes (clusters)
- 57% des classes comportent un seul élément
- 90% des classes comportent moins de 10 occurrences
- Les ignorer durant la transcription correspond à un caractères manquants sur 14 → intolérable !!!

Pourquoi ?

- Bruits, dégradations
- Caractères connectés
- Caractères cassés
- …




DE HVMANI CORPORIS FABRICA LIBER I. 43

distinguente, alia educitur futura, obliquè deorsum inter caninum & incisorium canino proximum delata, quæ cum illa communis efficitur, quæ in palati extremo iuxta incisorios dentes transuersim in illis animalibus fertur. Hæc futura citra omnem cartilaginis interuentum adeò insignis est, ut quartum à me enumeratum maxillæ os, canes in duo diuisum commonstrent. Verùm non arbitrandum est futuram hanc, inter dentium præsipiola in canibus aut simijs ductam, à superciliorum medio deorsum continuo progressu (& si Galenus ita doceat) ferri, sed uti iam admonuimus, ex medio ferè ductu illius futuræ incipit, quæ externum nasi ossium latus terminat. Quod ubi in cane simiaue inspexeris, accuratè expendito, quàm impossibile sit, hanc futuram porrigi inter dentes hominis, maxillam adepti breuissimam, & dentibus caninis prorsus exiguis donati. Insuper animaduertito, quid in libris de Partium usu Galenus sibi uoluerit, quum duodecim superioris maxillæ ossa esse obiter scripsit: uti & Introductorij seu Medici autor, duodecim tantum maxillæ ossa, è Veterũ indubiè placito, enumerat: aut quum Galenus hæc ossa prolixius cõmemorans, nouem tantũ recenset, nullumq; os priuatim incisorijs dentibus tribuit, in eo etiam numero cuneiforme os haud complectens. *Libro undecimo.*

DE MAXILLA INFERIORI,
Caput X.

PRIOR FIGVRA.

POSTERIOR.

VTRARVMQVE DECIMI CAPITIS FIGVRARVM, ET earundem characterum Index.

PRIOR *seu dextra huius capitis figura, inferiorem maxillam unà cum dentibus an*

Try Agora and Retro (Practical works) →
http://rfai.li.univ-tours.fr/PagesPerso/jyramel/fr/cours_PEEN.html

Project web site ➔

https://sites.google.com/site/paradiitproject/