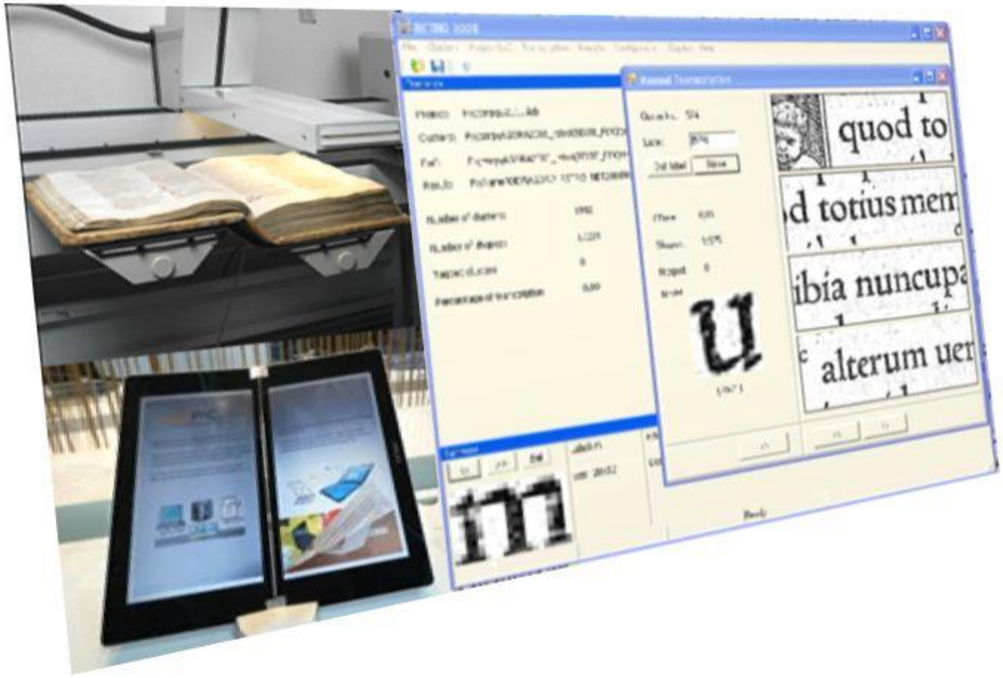
POLYTECH TOURS

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

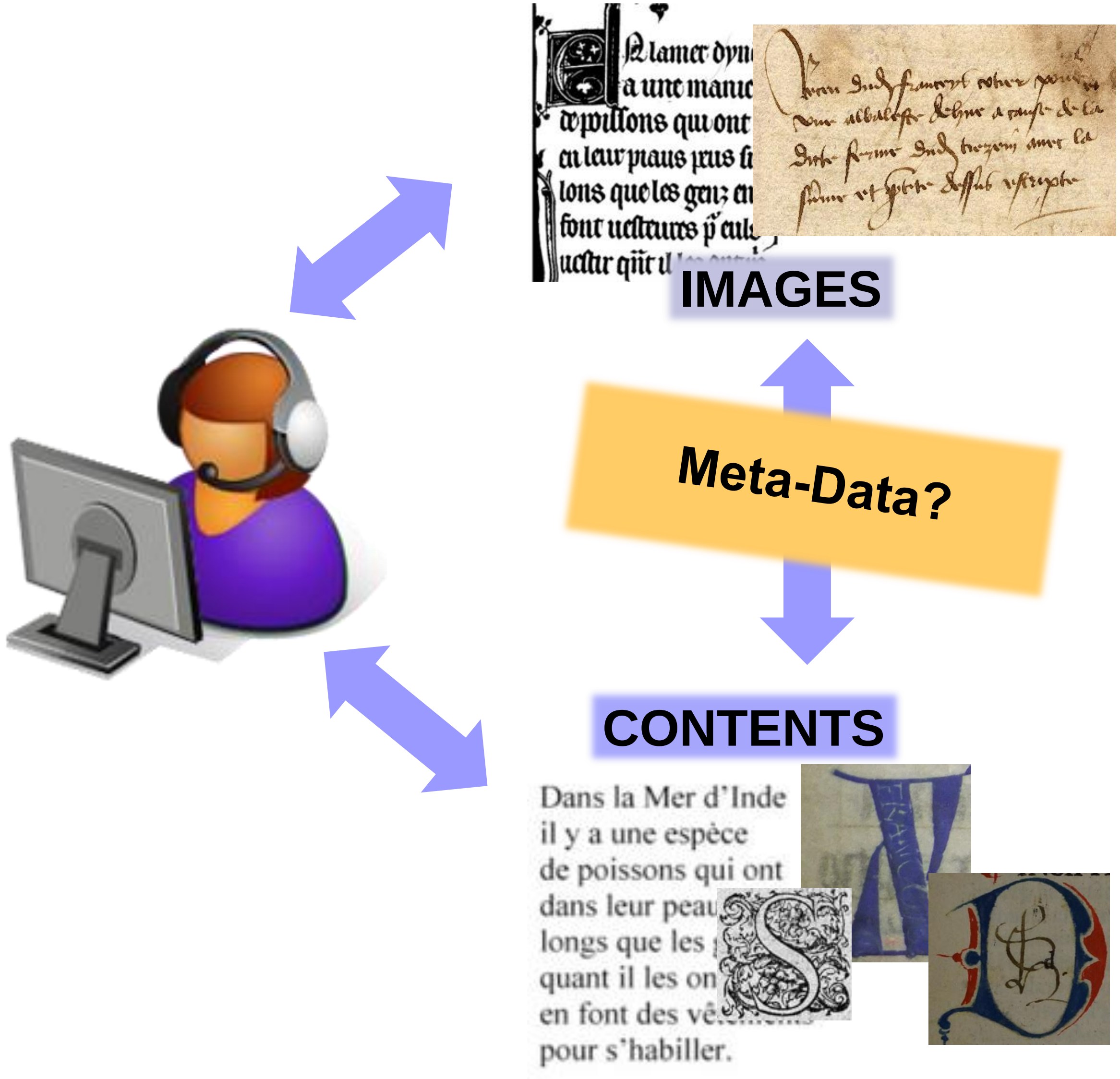
From pixels to content: An overview of the main techniques used in DIA

Jean-Yves RAMEL



LI
Laboratoire d'Informatique
EA 6300

From Pixels to Contents
Introduction



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

Document Image Analysis
Machine Learning
Content based images Retrieval

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

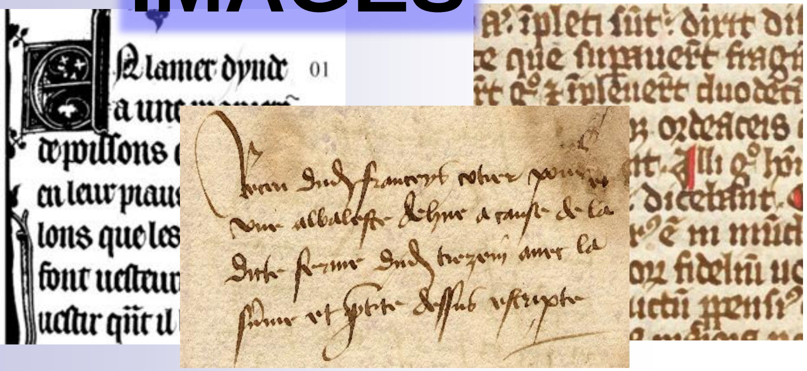
Generation and utilization of meta data

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…

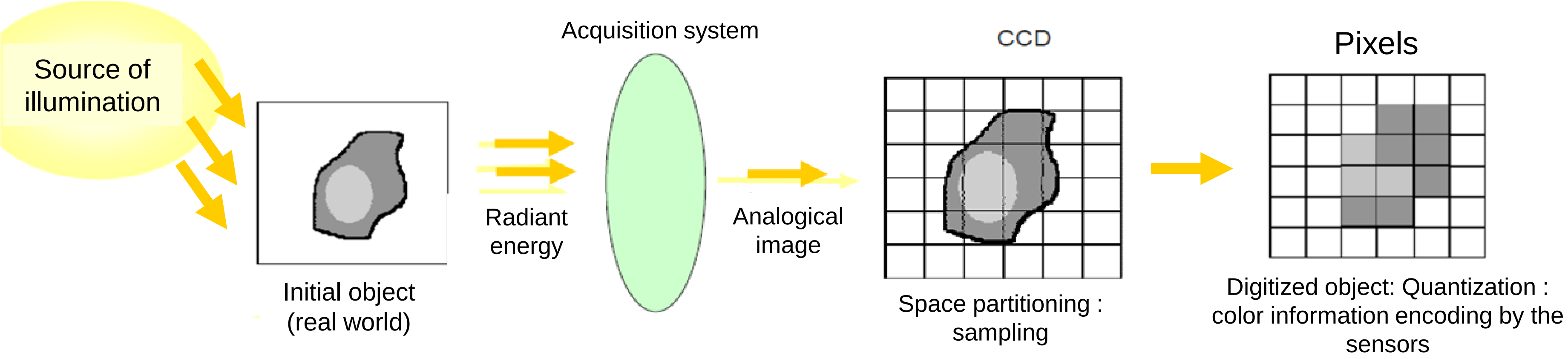
IMAGES



From Pixels…

Digitization ➔ Set of pixels ?

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



Source of illumination

Initial object (real world)

Radiant energy

Acquisition system

Analogical image

CCD

Space partitioning : sampling

Pixels

Digitized object: Quantization : color information encoding by the sensors

- **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

From Pixels…

What is an image?

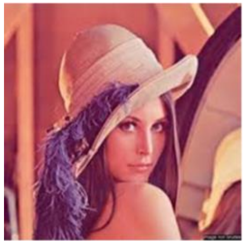
Image Quantization

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.

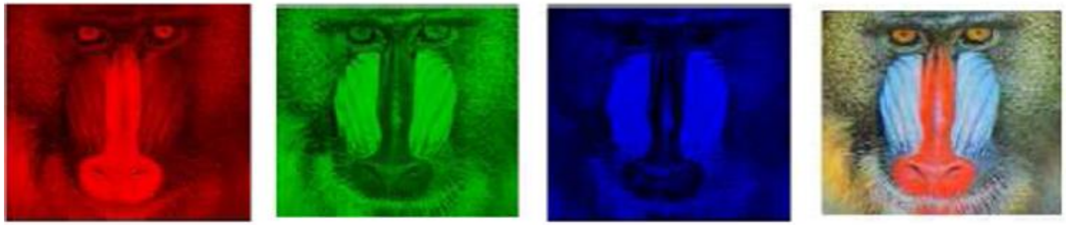
Color images (24 bits/pixel):
3 values of lighting intensity Red, Green, Blue

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


32 bits

6 bits

1 bit



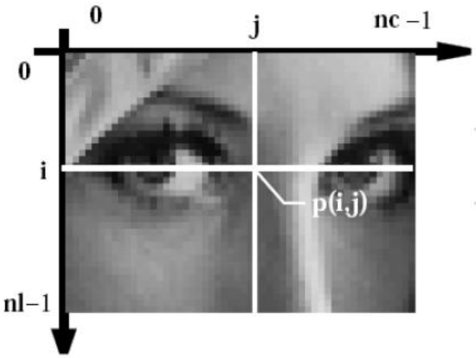
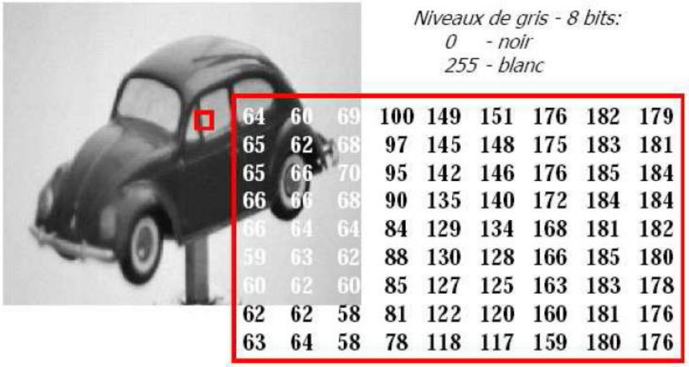


Image Representation & Processing

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



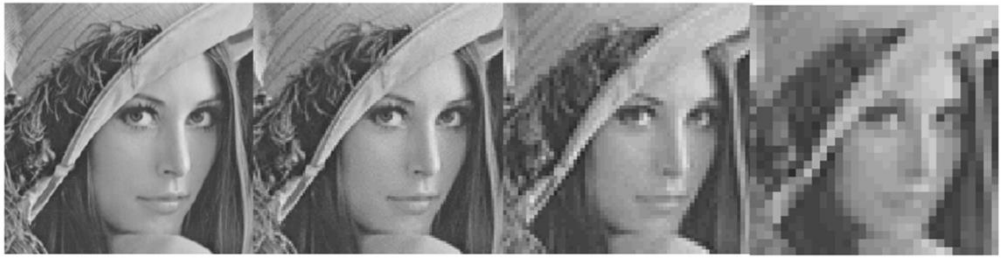
Niveaux de gris - 8 bits:
0 - noir
255 - blanc

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



256x256 128x128 64x64 32x32

- **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?

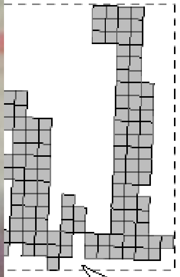
lated patterns can correspond
pixels) !

on the boundaries of the shapes
cess

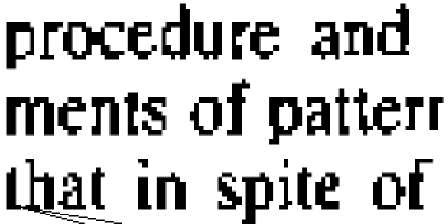
arance of characters until the

acters or touching characters

e resolution



Touching



procedure and
ments of patterr
that in spite of

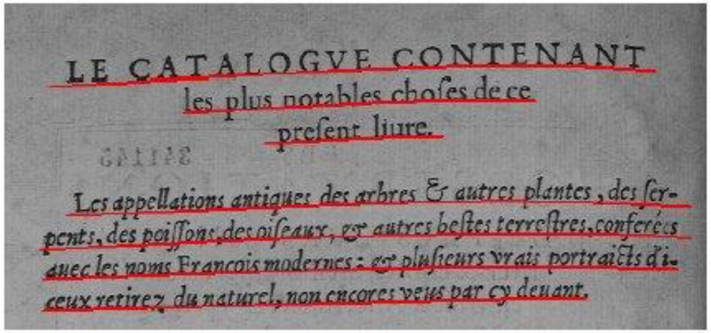
OCR errors

From Pixels…

What is Image (pre-)processing?

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

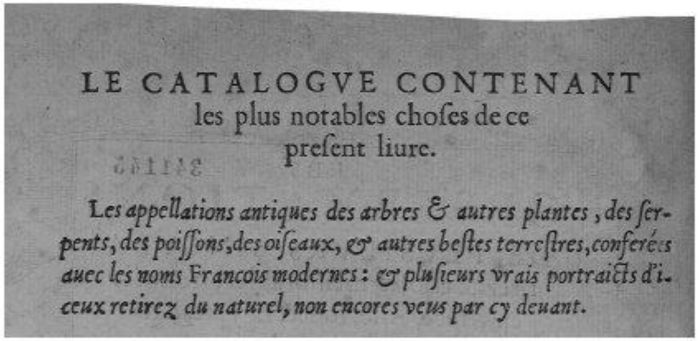
- Curvature and skew due to scanning
- Noise on boundaries, dots, blur, …



LE CATALOGVE CONTENANT
les plus notables choſes de ce
preſent liure.

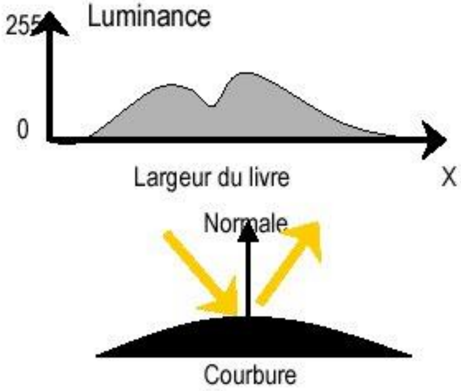
Les appellations antiques des arbres & autres plantes, des ſerpents, des poiſſons, des oiſeaux, & autres beſtes terreſtres, conferées auec les noms Francois modernes: & pluſieurs vrais portraicts d'iceux retirez du naturel, non encores veus par cy deuant.

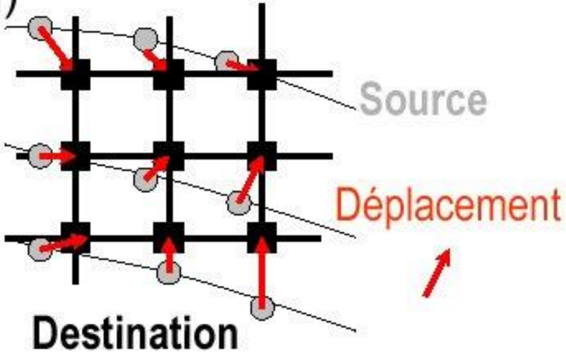
Corriger l'image (rotation, wrapping)

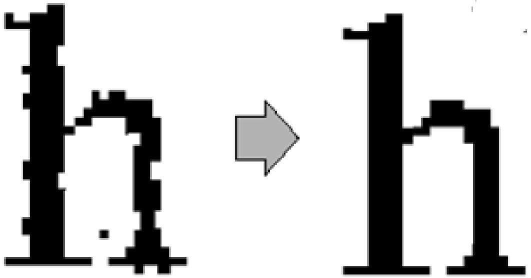


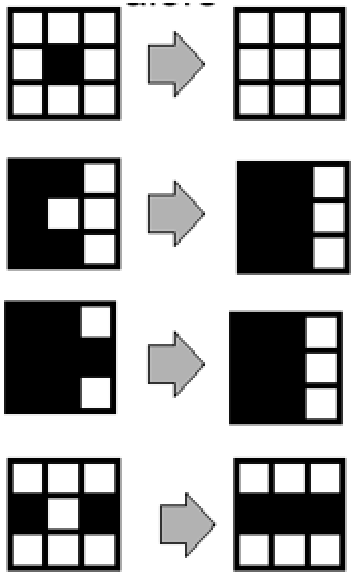
LE CATALOGVE CONTENANT
les plus notables choſes de ce
preſent liure.

Les appellations antiques des arbres & autres plantes, des ſerpents, des poiſſons, des oiſeaux, & autres beſtes terreſtres, conferées auec les noms Francois modernes: & pluſieurs vrais portraicts d'iceux retirez du naturel, non encores veus par cy deuant.





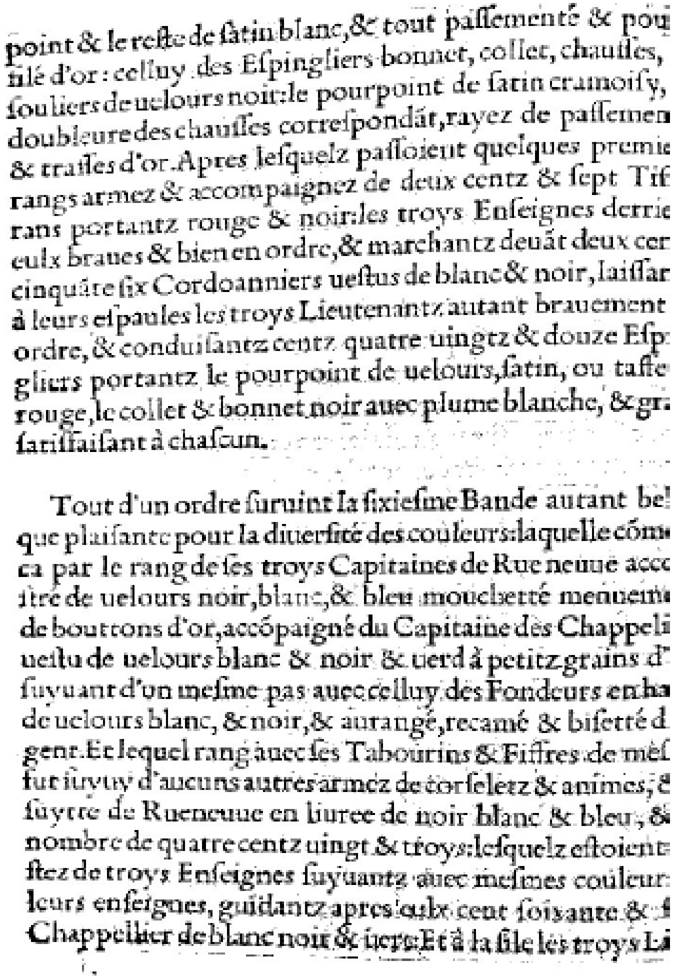




From Pixels…

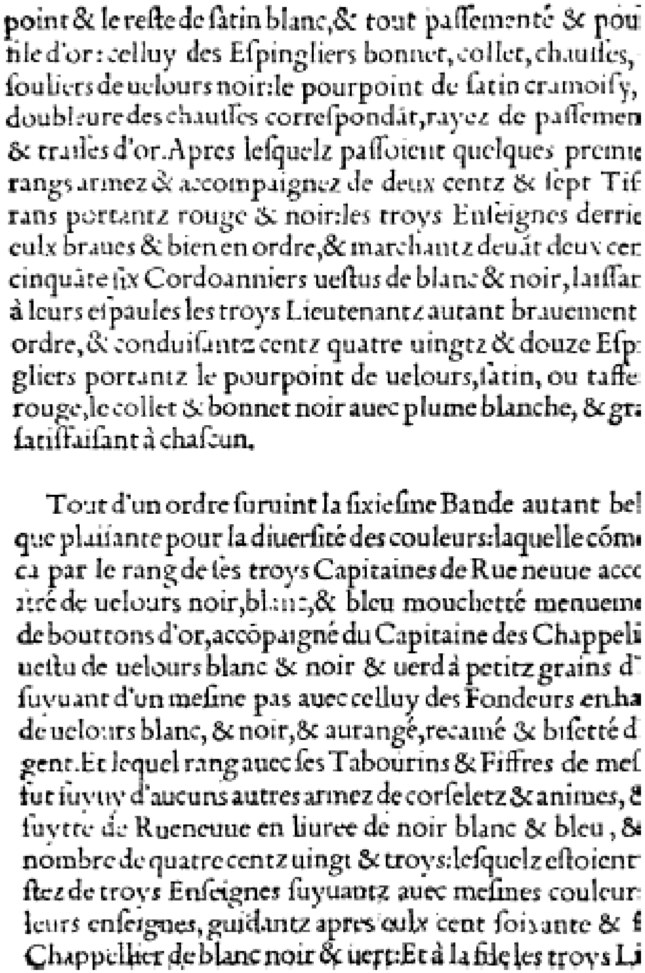
Image (pre-)processing

Curvature and skew correction is possible on text images



point & le reſte de ſatin blanc,& tout paſſementé & pou
tilé d'or: celluy des Eſpingliers bonnet, collet, chauſſes,
ſouliers de uelours noir:le pourpoint de ſatin cramoiſy,
doubleure des chauſſes correſpondāt,rayez de paſſemen
& traiſſes d'or.Apres leſquelz paſſoient quelques premie
rangs armez & accompaignez de deux centz & ſept Tiſ
rans portantz rouge & noir:les troys Enſeignes derrie
eulx braues & bien en ordre,& marchantz deuāt deux cen
cinquāte 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 ſuruint la ſixieſme Bande autant bel
que plaiſante pour la diuerſité des couleurs:laquelle cōm
ca par le rang de ſes troys Capitaines de Rue neuue acc
ſtré de uelours noir,blanc,& bleu mouchetté menuem
de bouttons d'or,accōpaigné 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, &
ſuytte 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 & ſ
Chappellier de blanc noir & uert:Et à la file les troys Li



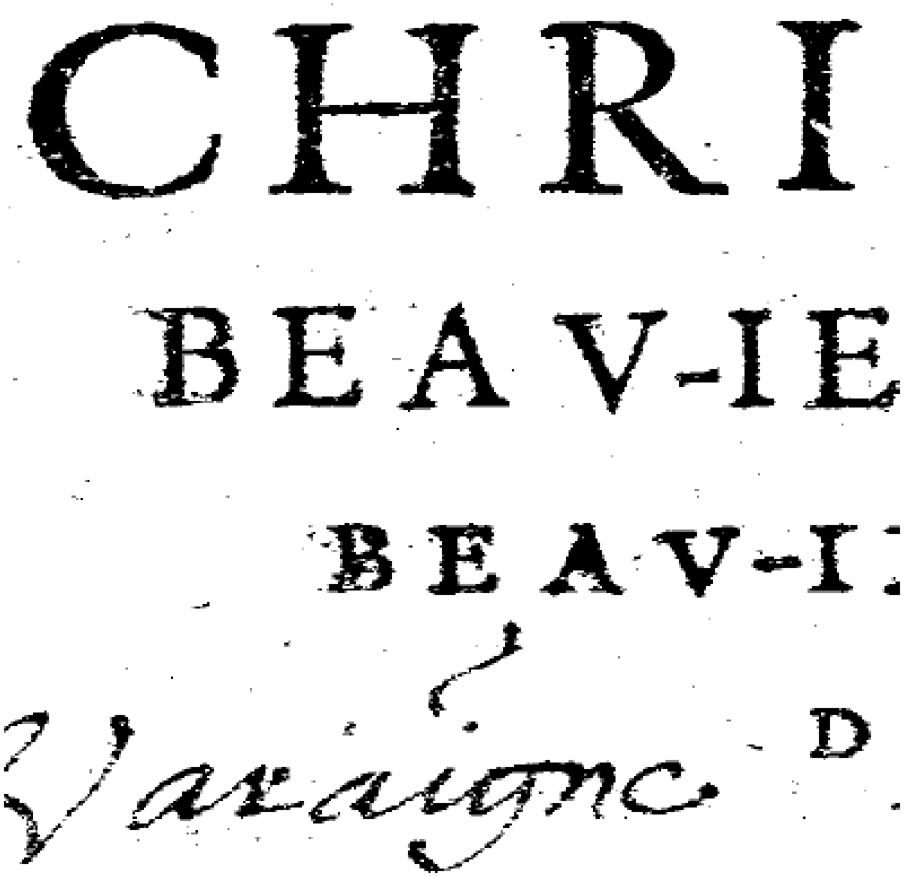
point & le reſte de ſatin blanc,& tout paſſementé & pou
tilé d'or: celluy des Eſpingliers bonnet, collet, chauſſes,
ſouliers de uelours noir:le pourpoint de ſatin cramoiſy,
doubleure des chauſſes correſpondāt,rayez de paſſemen
& traiſſes d'or.Apres leſquelz paſſoient quelques premie
rangs armez & accompaignez de deux centz & ſept Tiſ
rans portantz rouge & noir:les troys Enſeignes derrie
eulx braues & bien en ordre,& marchantz deuāt deux cen
cinquāte 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 ſuruint la ſixieſme Bande autant bel
que plaiſante pour la diuerſité des couleurs:laquelle cōm
ca par le rang de ſes troys Capitaines de Rue neuue acc
ſtré de uelours noir,blanc,& bleu mouchetté menuem
de bouttons d'or,accōpaigné 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, &
ſuytte 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 & ſ
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

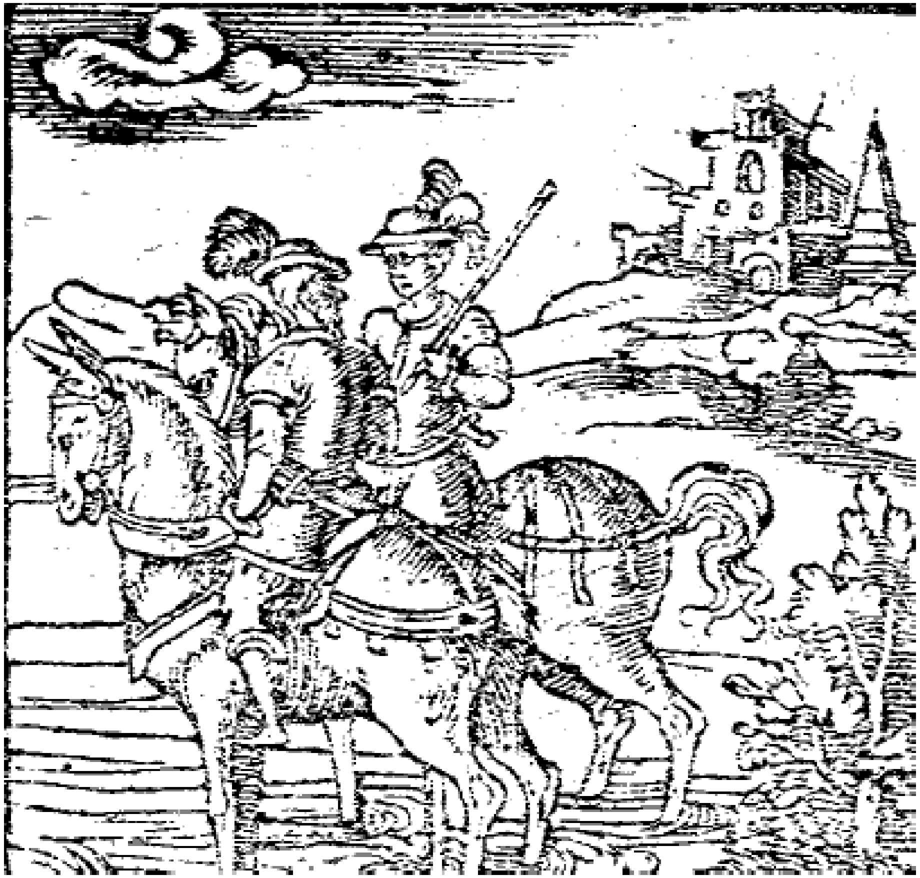


CHRI

BEAV-IE

BEAV-I

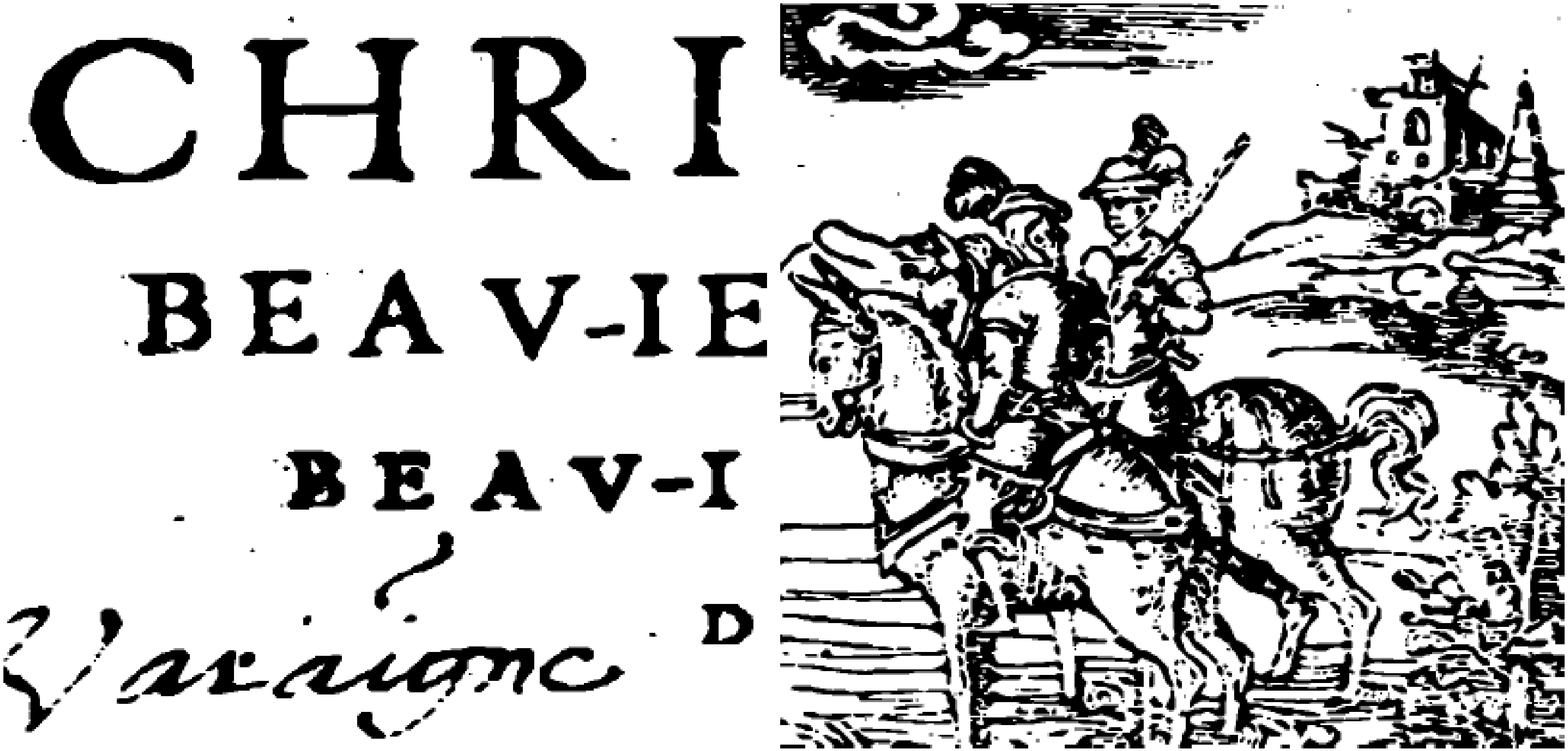
Varaigne D



From Pixels…

Image (pre-)processing

The problem is more complicated in case of heterogeneous content



CHRI

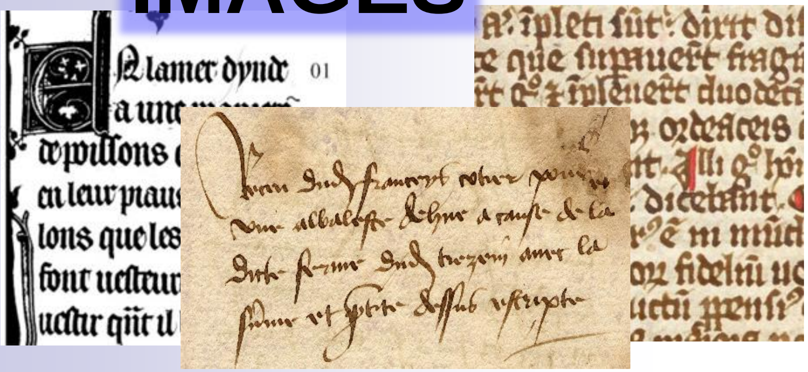
BEAV-IE

BEAV-I

Varaigne D

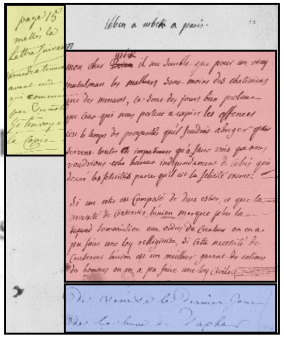
… to text and layout

IMAGES



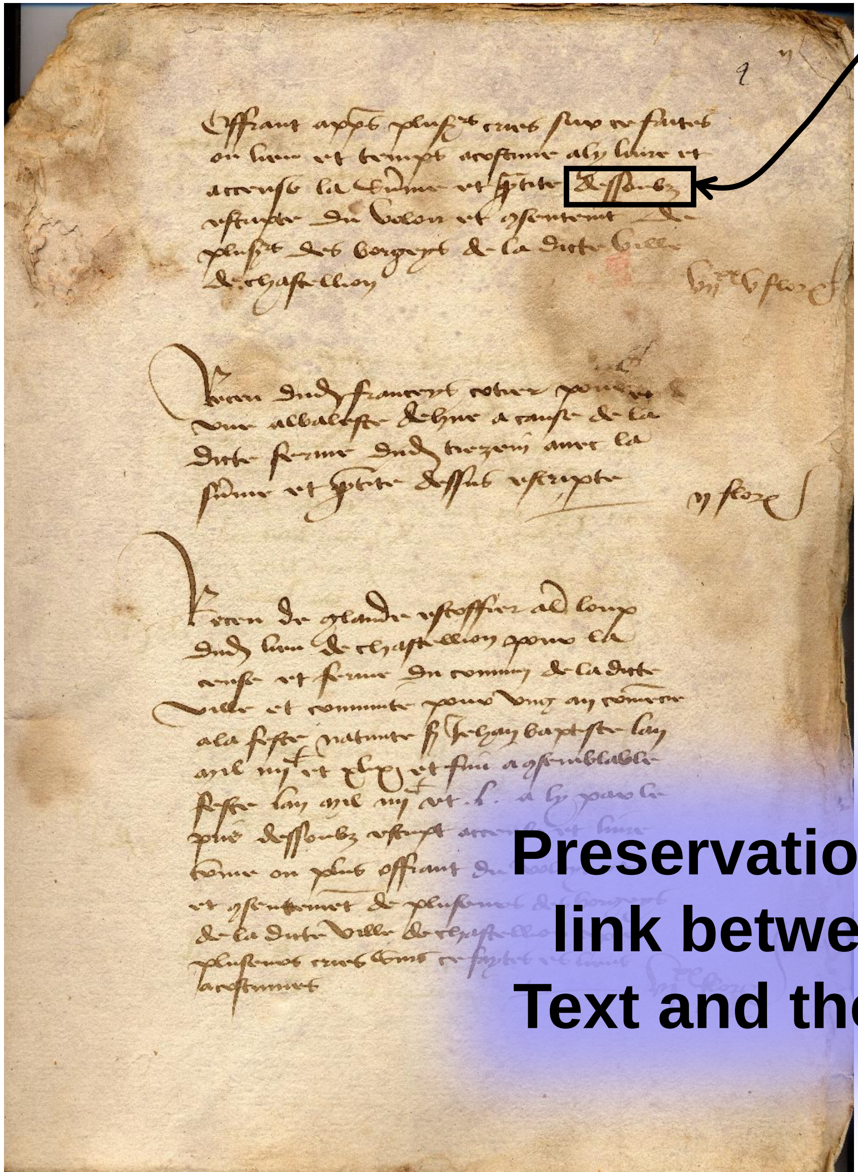
CONTENTS

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



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-alafeste nativite fi jehan baptiste lan
17-mil quatre cent et cinquante? A ly par le
18-pris dessouls esempt accense et hure ?
19-donnee 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 ?

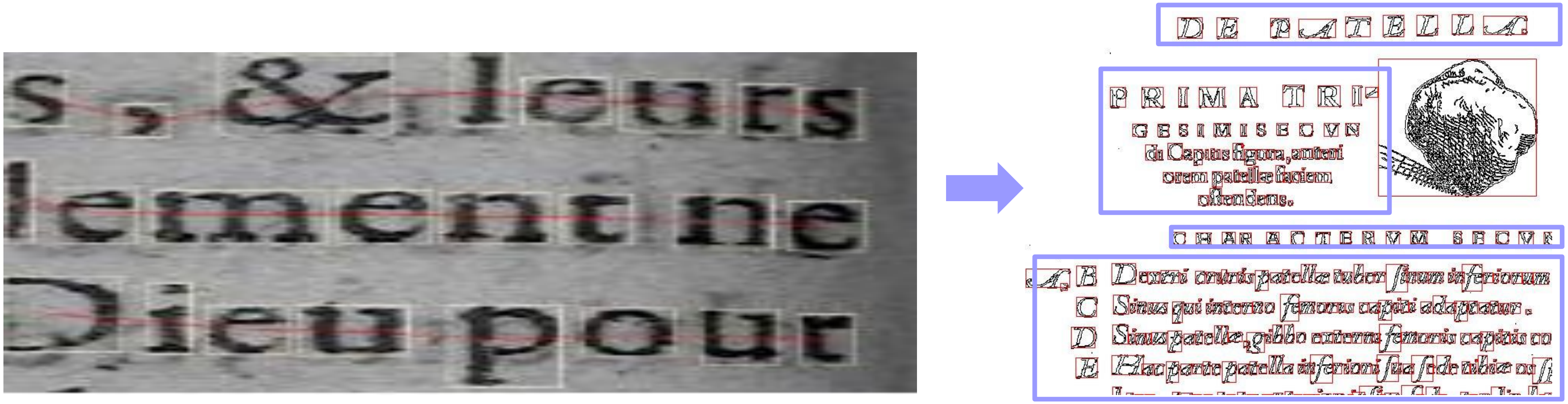
Preservation of the link between the Text and the Image

From pixels … to text

An overview of OCR mechanisms

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**

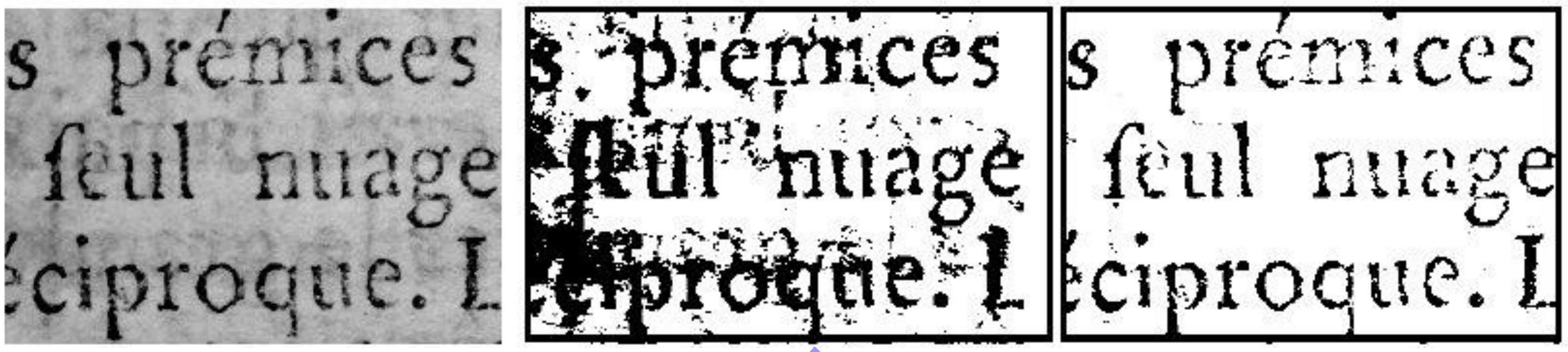


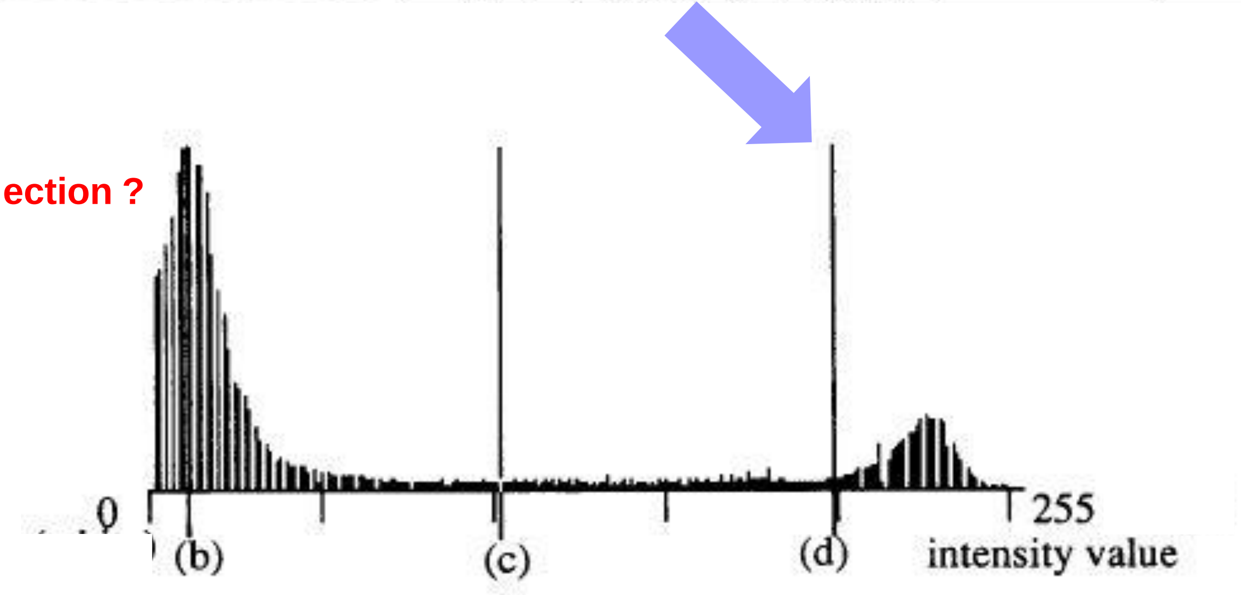
EoC

From pixels … to text
An overview of OCR mechanisms

Just to illustrate the difficulties…

- Most of the segmentation methods need a binarisation





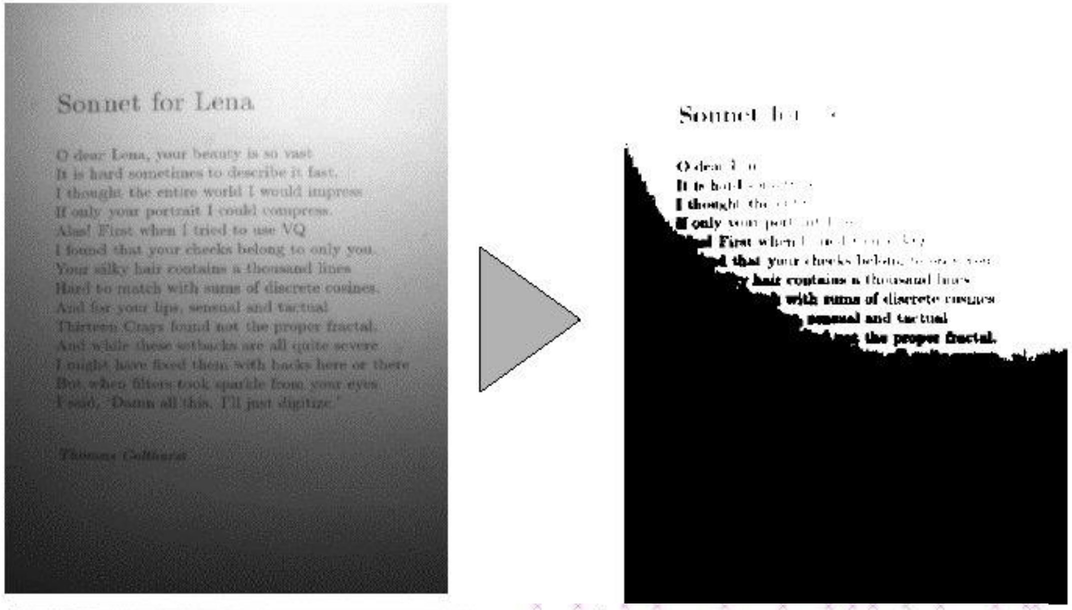
Threshold selection ?

From pixels … to text

An overview of OCR mechanisms

Just to illustrate the difficulties…

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



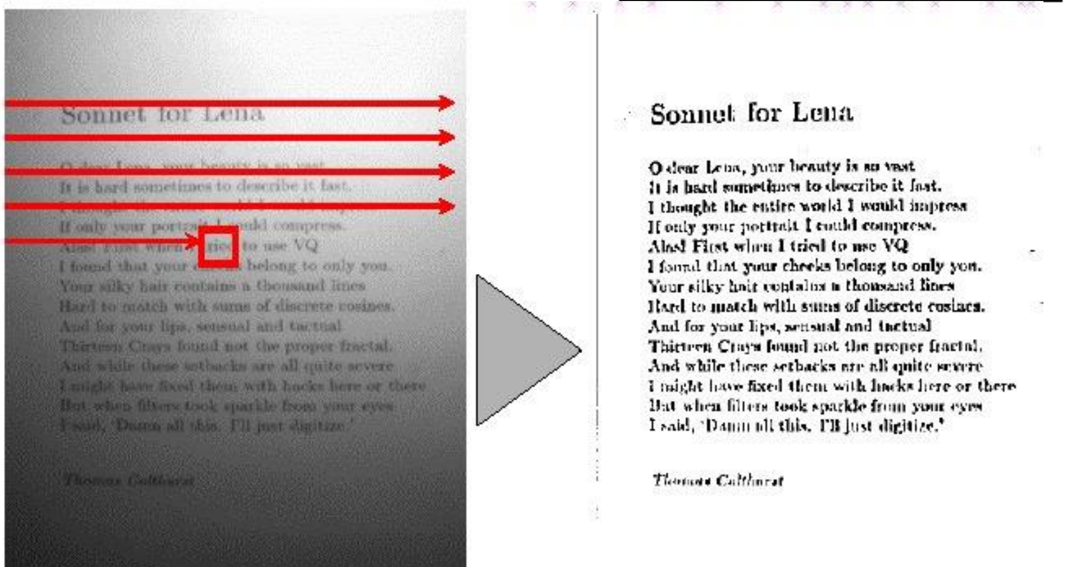
Sonnet for Lena

O dear Lena, your beauty is so vast
It is hard sometimes to describe it fast.
I thought the entire world I would impress
If only your portrait I could compress.
Alas! First when I tried to use VQ
I found that your cheeks belong to only you.
Your silky hair contains a thousand lines
Hard to match with sums of discrete cosines.
And for your lips, sensual and tactual
Thirteen Crays found not the proper fractal.
And while these setbacks are all quite severe
I might have fixed them with hacks here or there
But when filters took sparkle from your eyes
I said, 'Damn all this. I'll just digitize.'

Thomas Colthurst

- Local thresholds ➔

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



Sonnet for Lena

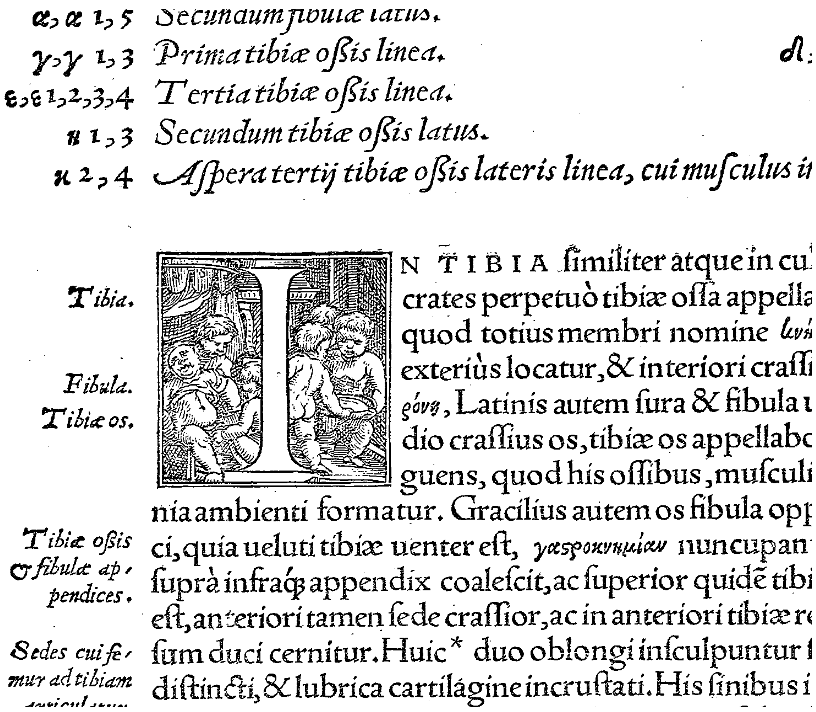
O dear Lena, your beauty is so vast
It is hard sometimes to describe it fast.
I thought the entire world I would impress
If only your portrait I could compress.
Alas! First when I tried to use VQ
I found that your cheeks belong to only you.
Your silky hair contains a thousand lines
Hard to match with sums of discrete cosines.
And for your lips, sensual and tactual
Thirteen Crays found not the proper fractal.
And while these setbacks are all quite severe
I might have fixed them with hacks here or there
But when filters took sparkle from your eyes
I said, 'Damn all this. I'll just digitize.'

Thomas Colthurst

From pixels … to text
An overview of OCR mechanisms

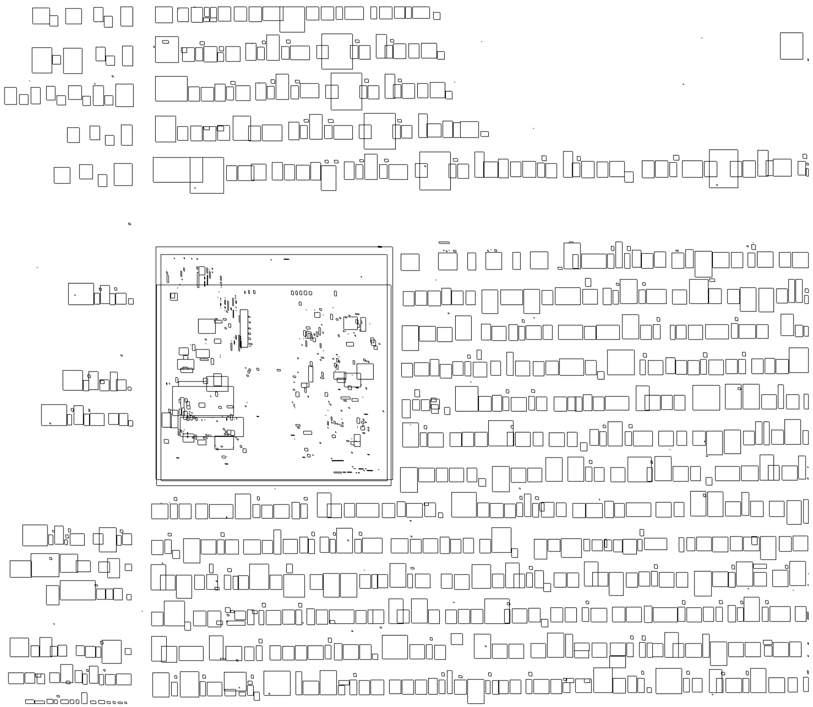
First step: Image segmentation / Connected components

- 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.* δ:
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. N TIBIA ſimiliter atque in cu
crates perpetuò tibiæ oſſa appella
quod totius membri nomine κνή
Fibula. exteriùs locatur, & interiori craſſi
Tibiæ os. ρόνη, Latinis autem ſura & fibula u
dio craſſius os, tibiæ os appellabo
guens, quod his oſſibus, muſculi
nia ambienti formatur. Gracilius autem os fibula opp
Tibiæ oßis ci, quia ueluti tibiæ uenter eſt, γαστροκνημίαν nuncupan
& fibulæ ap- ſuprà infraq; appendix coaleſcit, ac ſuperior quidẽ tibi
pendices. eſt, anteriori tamen ſede craſſior, ac in anteriori tibiæ re
Sedes cui fe- ſum duci cernitur. Huic* duo oblongi inſculpuntur ſ
mur ad tibiam diſtincti, & lubrica cartilágine incruſtati. His ſinibus i

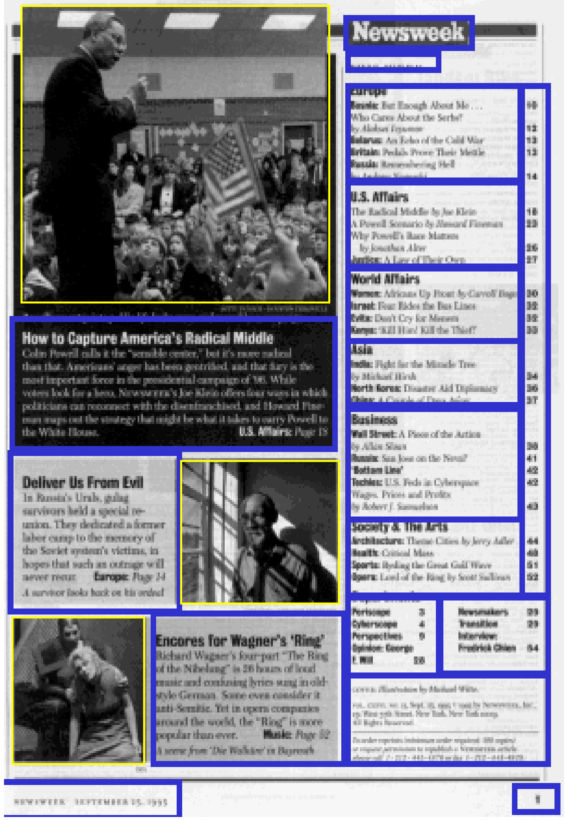


From pixels … to text

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…



Newsweek

Europe
Bosnia: But Enough About Me . . . 10
Who Cares About the Serbs?
by Aleksei Izyumov 12
Belarus: An Echo of the Cold War 13
Britain: Pedals Prove Their Mettle 13
Russia: Remembering Hell
by Andrew Nagorski 14

U.S. Affairs
The Radical Middle by Joe Klein 18
A Powell Scenario by Howard Fineman 23
Why Powell's Race Matters
by Jonathan Alter 26
Justice: A Law of Their Own 27

World Affairs
Women: Africans Up Front by Carroll Bogert 30
Israel: Fear Rides the Bus Lines 32
Evita: Don't Cry for Menem 33
Kenya: 'Kill Him! Kill the Thief!' 33

Asia
India: Fight for the Miracle Tree
by Michael Hirsh 34
North Korea: Disaster Aid Diplomacy 36
China: [illegible] 37

Business
Wall Street: A Piece of the Action
by Allan Sloan 38
Russia: San Jose on the Neva? 41
'Bottom Line' 42
Techies: U.S. Feds in Cyberspace 42
Wages, Prices and Profits
by Robert J. Samuelson 43

Society & The Arts
Architecture: Theme Cities by Jerry Adler 44
Health: Critical Mass 48
Sports: Ryding the Great Golf Wave 51
Opera: Lord of the Ring by Scott Sullivan 52

Periscope 3
Cyberscope 4
Perspectives 9
Opinion: George F. Will 58

Newsmakers 29
Transition 29
Interview: Fredrick Chien 54

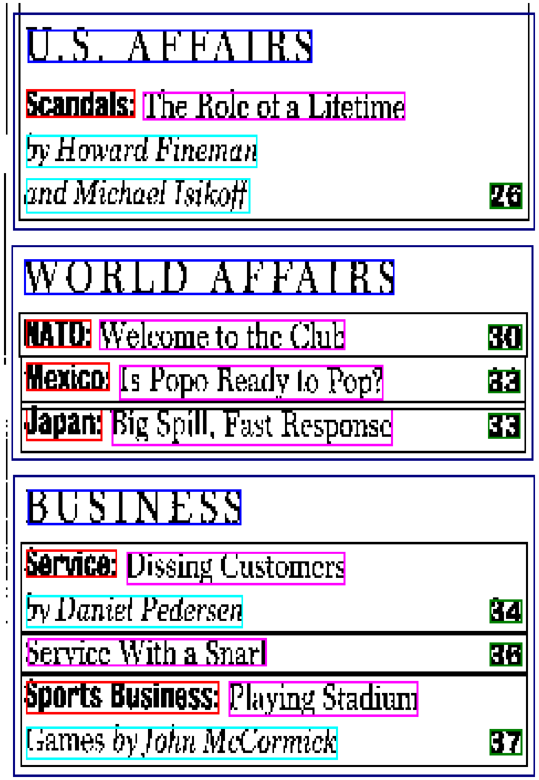
Cover: Illustration by Michael Witte.

How to Capture America's Radical Middle
Colin Powell calls it the "sensible center," but it's more radical than that. Americans' anger has been gentrified, and that fury is the most important force in the presidential campaign of '96. While voters look for a hero, Newsweek's Joe Klein offers four ways in which politicians can reconnect with the disenfranchised, and Howard Fineman maps out the strategy that might be what it takes to carry Powell to the White House. **U.S. Affairs:** Page 18

Deliver Us From Evil
In Russia's Urals, gulag survivors held a special reunion. They dedicated a former labor camp to the memory of the Soviet system's victims, in hopes that such an outrage will never recur. **Europe:** Page 14
A survivor looks back on his ordeal

Encores for Wagner's 'Ring'
Richard Wagner's four-part "The Ring of the Nibelung" is 20 hours of loud music and confusing lyrics sung in old-style German. Some even consider it anti-Semitic. Yet in opera companies around the world, the "Ring" is more popular than ever. **Music:** Page 52
A scene from 'Die Walküre' in Bayreuth

NEWSWEEK SEPTEMBER 25, 1995 1



U.S. AFFAIRS

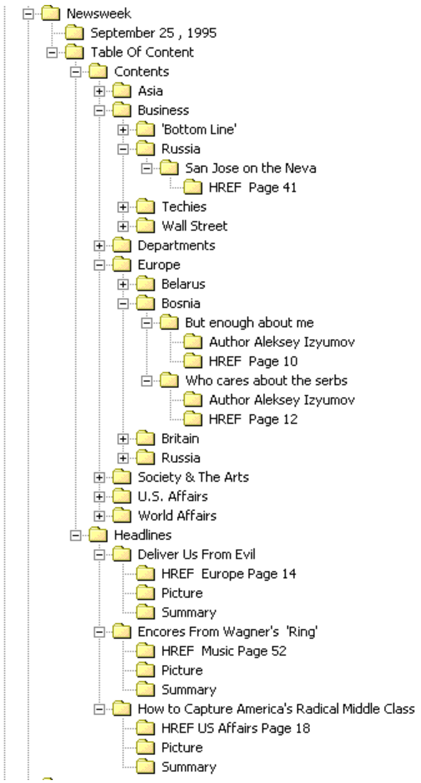
Scandals: The Role of a Lifetime
by Howard Fineman
and Michael Isikoff 26

WORLD AFFAIRS

NATO: Welcome to the Club 30
Mexico: Is Popo Ready to Pop? 32
Japan: Big Spill, Fast Response 33

BUSINESS

Service: Dissing Customers
by Daniel Pedersen 34
Service With a Snarl 36
Sports Business: Playing Stadium Games by John McCormick 37



```
Newsweek
  September 25 , 1995
  Table Of Content
    Contents
      Asia
      Business
        'Bottom Line'
        Russia
          San Jose on the Neva
            HREF  Page 41
        Techies
        Wall Street
      Departments
      Europe
        Belarus
        Bosnia
          But enough about me
            Author Aleksey Izyumov
            HREF  Page 10
          Who cares about the serbs
            Author Aleksey Izyumov
            HREF  Page 12
        Britain
        Russia
      Society & The Arts
      U.S. Affairs
      World Affairs
    Headlines
      Deliver Us From Evil
        HREF  Europe Page 14
        Picture
        Summary
      Encores From Wagner's  'Ring'
        HREF  Music Page 52
        Picture
        Summary
      How to Capture America's Radical Middle Class
        HREF US Affairs Page 18
        Picture
        Summary
```

From pixels … to text

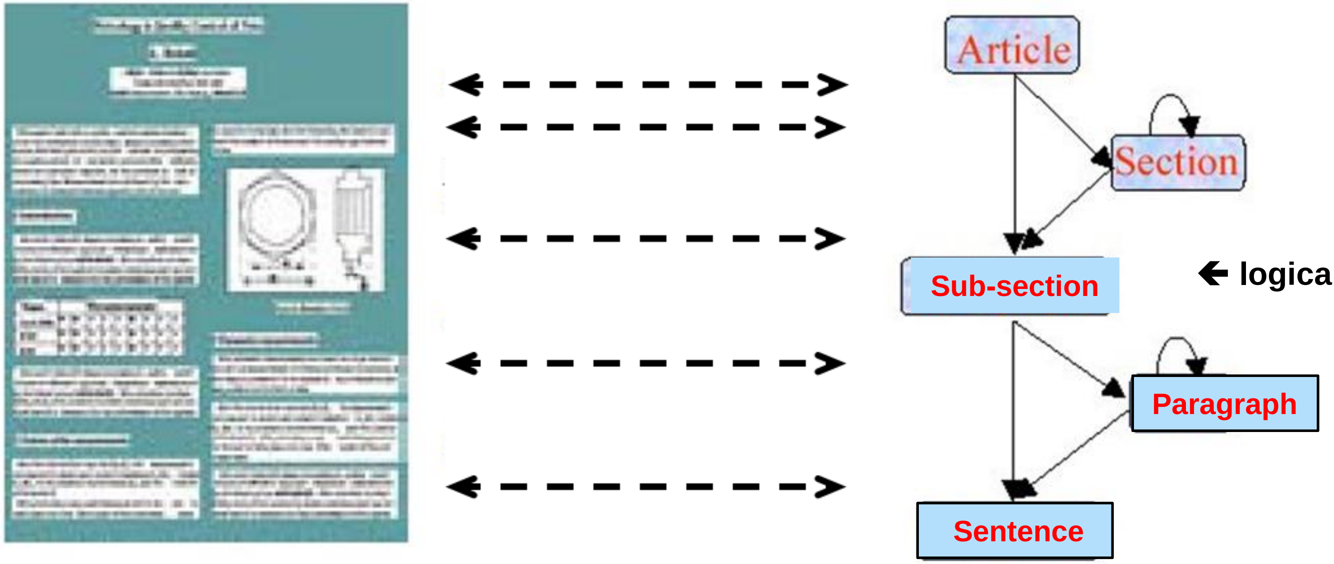
An overview of OCR mechanisms

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)

Physical structure ➔



Article

Section

Sub-section

Paragraph

Sentence

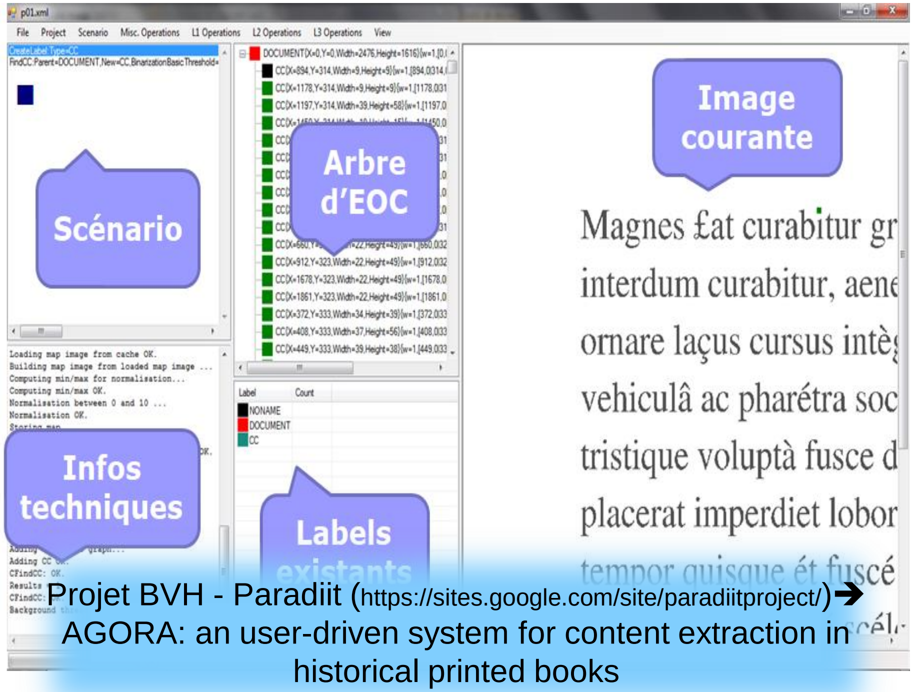
➔ logical structure

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



CreateLabel:Type=CC
FindCC:Parent=DOCUMENT,New=CC,BinarizationBasicThreshold=

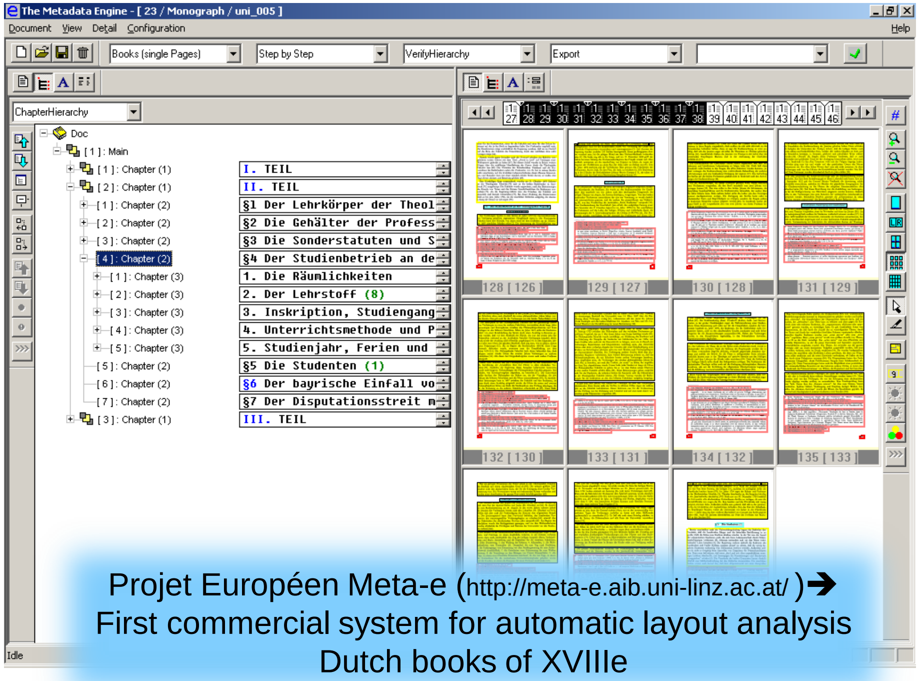
```
DOCUMENT{X=0,Y=0,Width=2476,Height=1616}{w=1,[0,(
  CC{X=894,Y=314,Width=9,Height=9}{w=1,[894,0314,(
  CC{X=1178,Y=314,Width=9,Height=9}{w=1,[1178,031
  CC{X=1197,Y=314,Width=39,Height=58}{w=1,[1197,0
  CC{X=660,Y=...,Width=22,Height=49}{w=1,[660,032
  CC{X=912,Y=323,Width=22,Height=49}{w=1,[912,032
  CC{X=1678,Y=323,Width=22,Height=49}{w=1,[1678,0
  CC{X=1861,Y=323,Width=22,Height=49}{w=1,[1861,0
  CC{X=372,Y=333,Width=34,Height=39}{w=1,[372,033
  CC{X=408,Y=333,Width=37,Height=56}{w=1,[408,033
  CC{X=449,Y=333,Width=39,Height=38}{w=1,[449,033
```

```
Loading map image from cache OK.
Building map image from loaded map image ...
Computing min/max for normalisation...
Computing min/max OK.
Normalisation between 0 and 10 ...
Normalisation OK.
Storing map
Adding CC OK.
CFindCC: OK.
Results
CFindCC: OK.
Background
```

Label	Count
■ NONAME	
🟥 DOCUMENT	
🟩 CC	

Magnes £at curabitur gr
interdum curabitur, aene
ornare laçus cursus intè
vehiculâ ac pharétra soc
tristique voluptà fusce d
placerat imperdiet lobor
tempor quisque ét fuscé

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



```
Doc
  [ 1 ] : Main
    [ 1 ] : Chapter (1)        I. TEIL
    [ 2 ] : Chapter (1)        II. TEIL
      [ 1 ] : Chapter (2)      §1 Der Lehrkörper der Theol
      [ 2 ] : Chapter (2)      §2 Die Gehälter der Profess
      [ 3 ] : Chapter (2)      §3 Die Sonderstatuten und S
      [ 4 ] : Chapter (2)      §4 Der Studienbetrieb an de
        [ 1 ] : Chapter (3)    1. Die Räumlichkeiten
        [ 2 ] : Chapter (3)    2. Der Lehrstoff (8)
        [ 3 ] : Chapter (3)    3. Inskription, Studiengang
        [ 4 ] : Chapter (3)    4. Unterrichtsmethode und P
        [ 5 ] : Chapter (3)    5. Studienjahr, Ferien und
      [ 5 ] : Chapter (2)      §5 Die Studenten (1)
      [ 6 ] : Chapter (2)      §6 Der bayrische Einfall vo
      [ 7 ] : Chapter (2)      §7 Der Disputationsstreit m
    [ 3 ] : Chapter (1)        III. TEIL
```

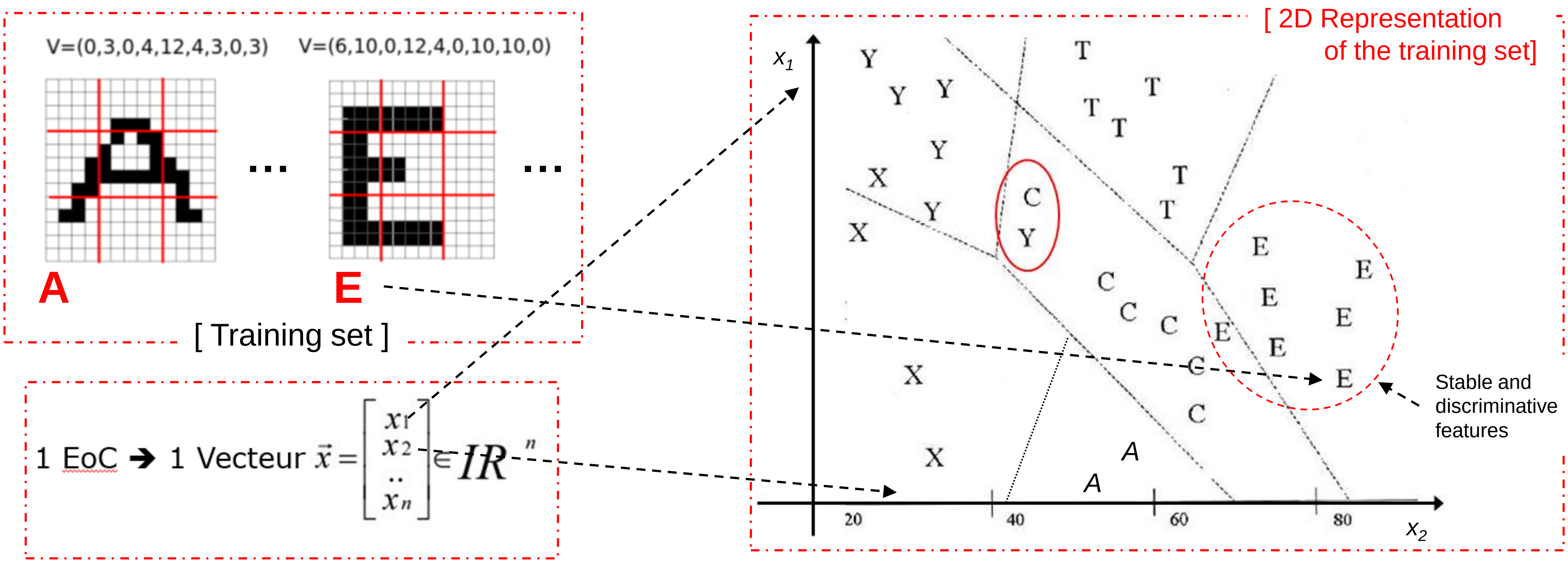
Projet Européen Meta-e (http://meta-e.aib.uni-linz.ac.at/)➔
First commercial system for automatic layout analysis
Dutch books of XVIIIe

From pixels … to text
An overview of OCR mechanisms

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)



[Training set]

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

[2D Representation of the training set]

Stable and discriminative features

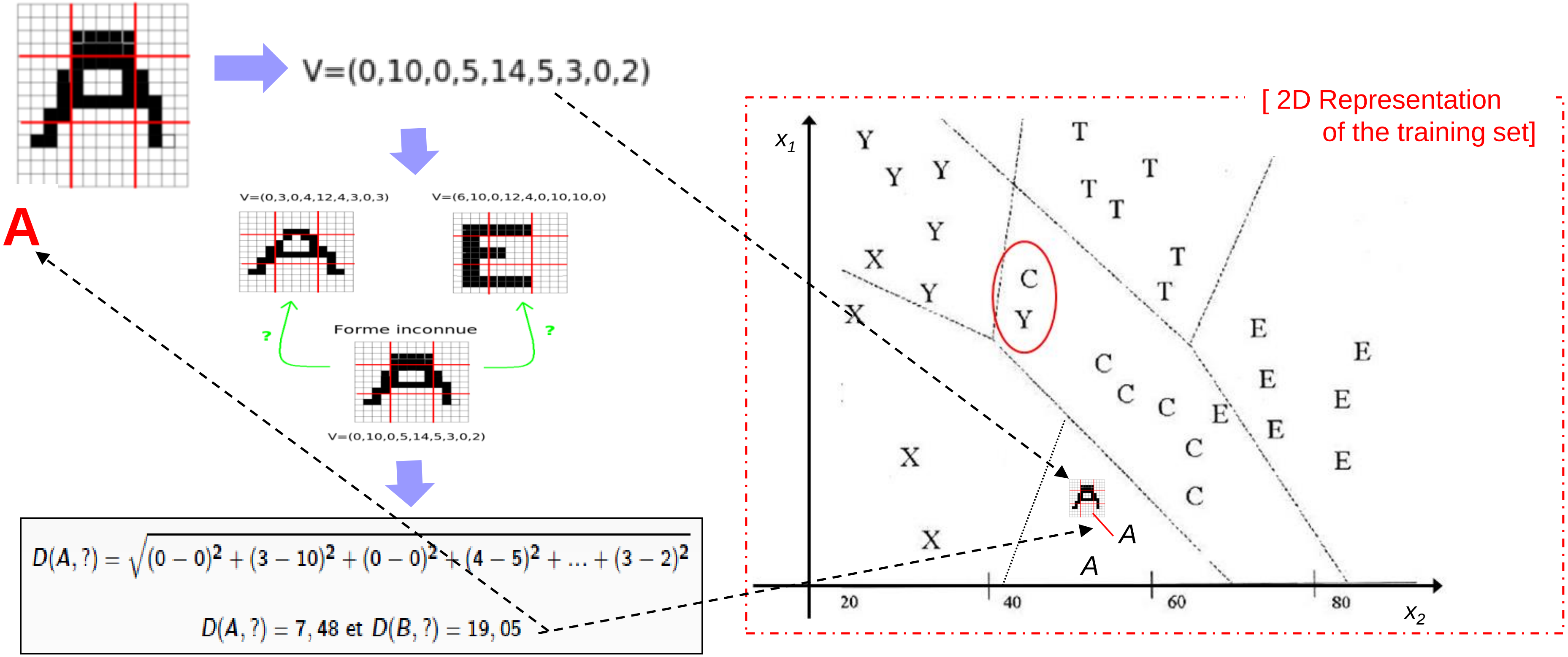
From pixels … to text

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)



[2D Representation of the training set]

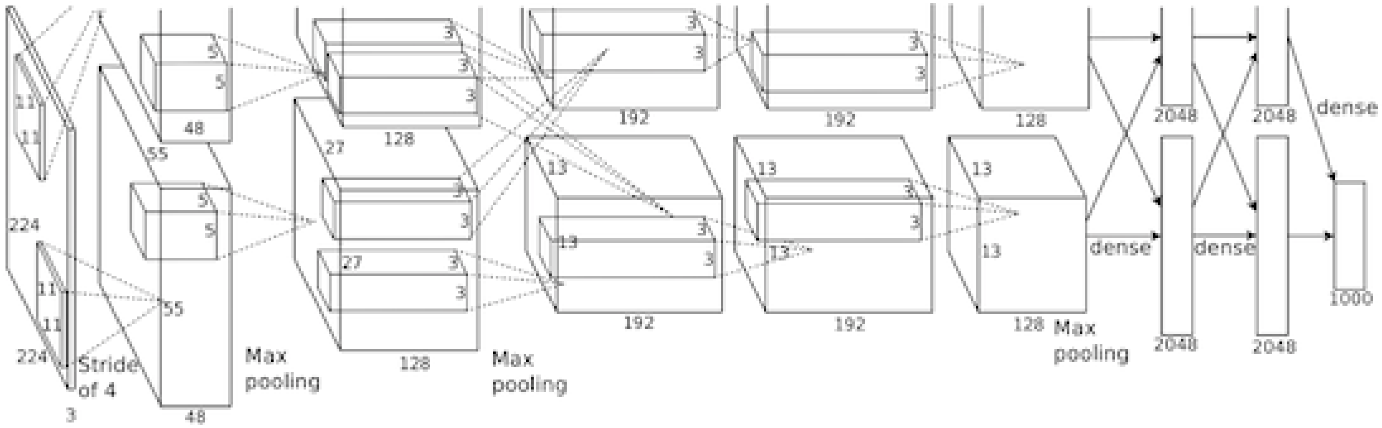
$$D(A,?) = \sqrt{(0-0)^2 + (3-10)^2 + (0-0)^2 + (4-5)^2 + \ldots + (3-2)^2}$$

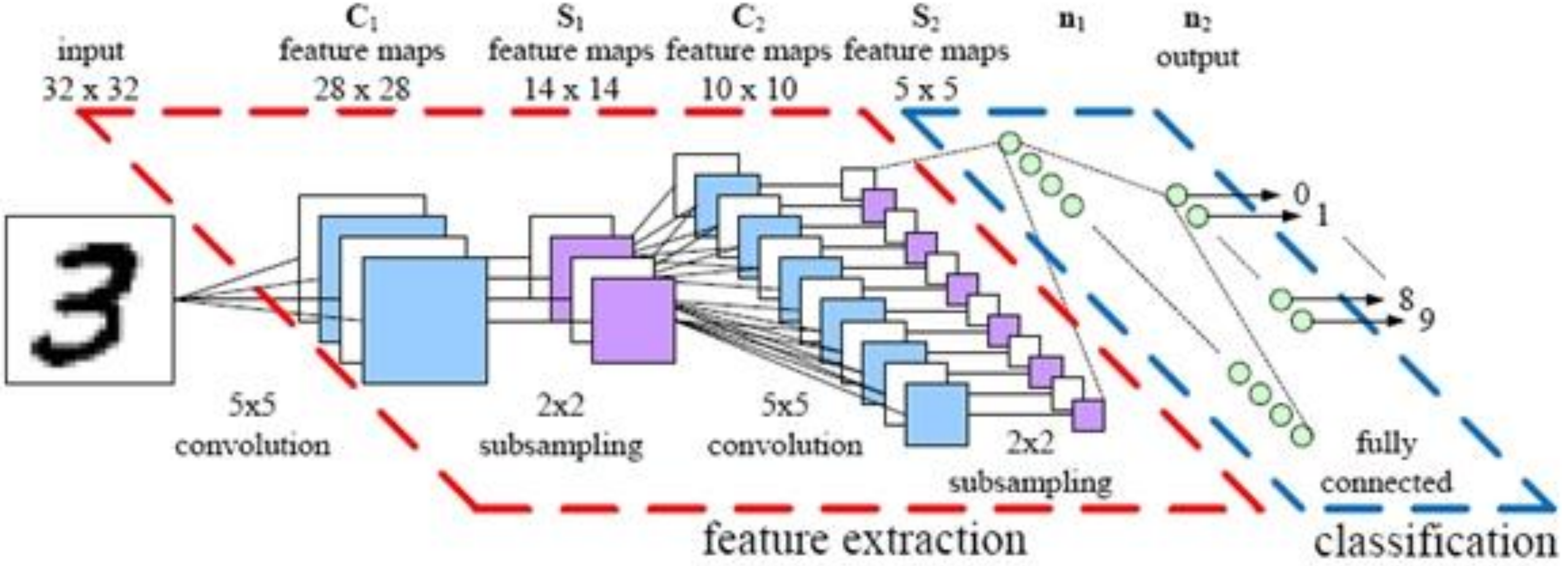
$$D(A,?) = 7,48 \text{ et } D(B,?) = 19,05$$

From pixels … to text

An overview of OCR mechanisms

Deep Learning (Conv. Neural Net)

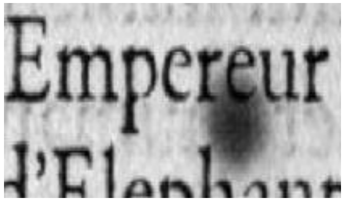




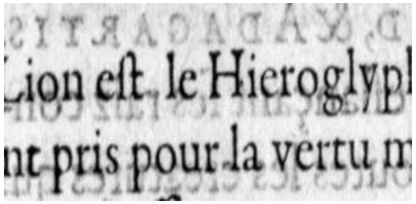
From pixels … to text

An overview of OCR mechanisms

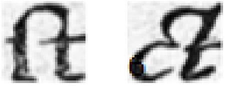
- **Why commercial OCR are not working well on historical documents?**
  - □ Noises and degradations
  - □ Unusual layout
  - □ Unsuited training set

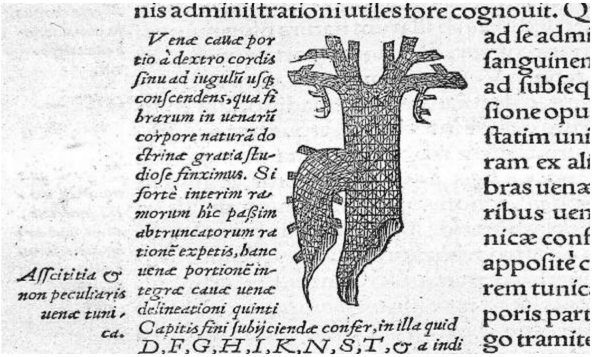


Empereur
d'Elephant



Lion eſt le Hieroglyp
nt pris pour la vertu n



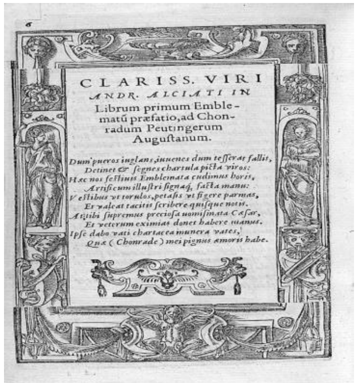


nis adminiſtrationi utiles fore cognouit. Q

Venæ cauæ portio à dextro cordis ſinu ad iugulũ uſq; conſcendens, qua fibrarum in uenarũ corpore naturã doctrinæ gratia ſtudioſe finximus. Si fortè interim ramorum hic paſsim abtruncatorum rationẽ expetis, hanc uenæ portionẽ integræ cauæ uenæ delineationi quinti Capitis fini ſubijciendæ confer, in illa quid D, F, G, H, I, K, N, S, T, & a indi

Aſſcititia & non peculiaris uenæ tunica.

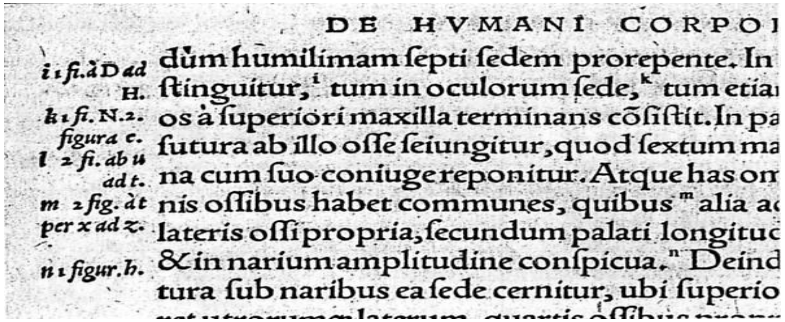
ad ſe admi
ſanguinen
ad ſubſeq
ſione opu
ſtatim uni
ram ex ali
bras uenæ
ribus uen
nicæ conf
appoſitè c
rem tunic
poris part
go tramite



CLARISS. VIRI
ANDR. ALCIATI IN
Librum primum Emblematũ præfatio, ad Chonradum Peutingerum Auguſtanum.

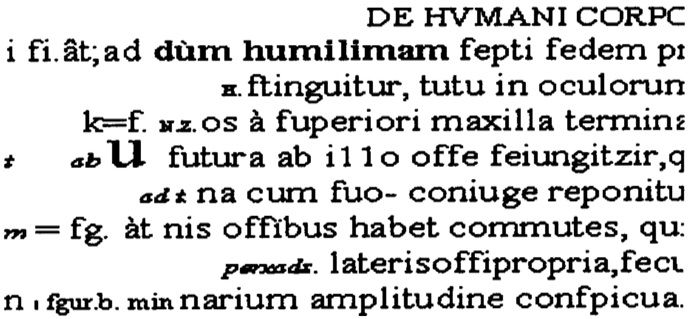
Dum pueros iuglans, iuuenes dum teſſera fallit,
Detinet & ſegnes chartula picta viros:
Hæc nos feſtiuis Emblemata cudimus horis,
Artificum illuſtri ſignaq; facta manu:
Veſtibus vt torulos, petaſis vt figere parmas,
Et valeat tacitis ſcribere quiſque notis.
At tibi ſupremus precioſa nomiſmata Cæſar,
Et veterum eximias donet habere manus.
Ipſe dabo vati chartacea munera vates,
Quæ (Chonrade) mei pignus amoris habe.

Fine Reader



DE HVMANI CORPO

i 1 fi. à D ad H. — dũm humilimam ſepti ſedem prorepente. In
ſtinguitur,[i] tum in oculorum ſede,[k] tum etia
k 1 fi. N. 2. — os à ſuperiori maxilla terminans cõſiſtit. In pa
figura c. — ſutura ab illo oſſe ſeiungitur, quod ſextum ma
l 2 fi. ab u ad t. — na cum ſuo coniuge reponitur. Atque has om
m 2 fig. à t per x ad z. — nis oſſibus habet communes, quibus[m] alia a
lateris oſſi propria, ſecundum palati longitu
n 1 figur. h. — & in narium amplitudine conſpicua.[n] Deind
tura ſub naribus ea ſede cernitur, ubi ſuperio



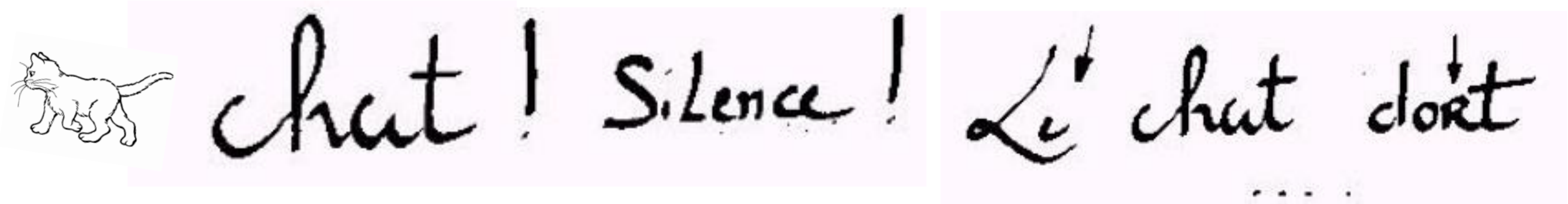
DE HVMANI CORPC
i fi.ât;ad **dùm humilimam** fepti fedem pı
ʜ. ftinguitur, tutu in oculorun
k=f. ɴ.z. os à fuperiori maxilla termina
t ab **u** futura ab i1 1o offe feiungitzir,q
ad t na cum fuo- coniuge reponitu
m = fg. àt nis offibus habet commutes, qu:
perxadz. laterisoffipropria,fecu
n ı fgur.b. min narium amplitudine confpicua.

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?

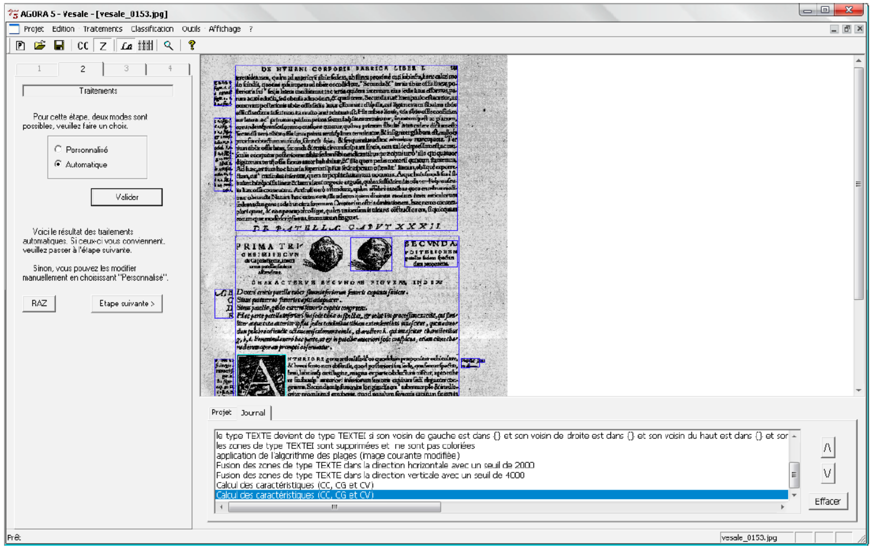
… to non-text

IMAGES

CONTENTS

From pixels … to non-textual contents

Pictorial Content is also of high interest



Pour cette étape, deux modes sont possibles, veuillez faire un choix.

○ Personnalisé
● Automatique

Valider

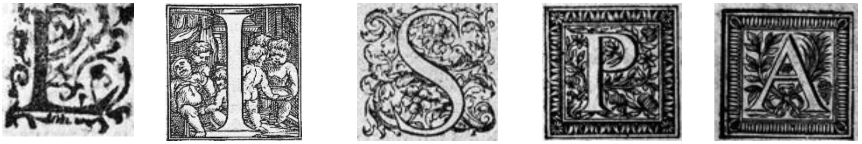
Voici le résultat des traitements automatiques. Si ceux-ci vous conviennent, veuillez passer à l'étape suivante.

Sinon, vous pouvez les modifier manuellement en choisissant "Personnalisé".

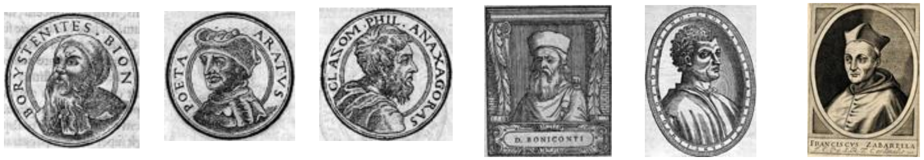
RAZ   Etape suivante >

```
le type TEXTE devient de type TEXTEI si son voisin de gauche est dans {} et son voisin de droite est dans {} et son voisin du haut est dans {} et sor
les zones de type TEXTEI sont supprimées et  ne sont pas coloriées
application de l'algorithme des plages (image courante modifiée)
Fusion des zones de type TEXTE dans la direction horizontale avec un seuil de 2000
Fusion des zones de type TEXTE dans la direction verticale avec un seuil de 4000
Calcul des caractéristiques (CC, CG et CV)
Calcul des caractéristiques (CC, CG et CV)
```

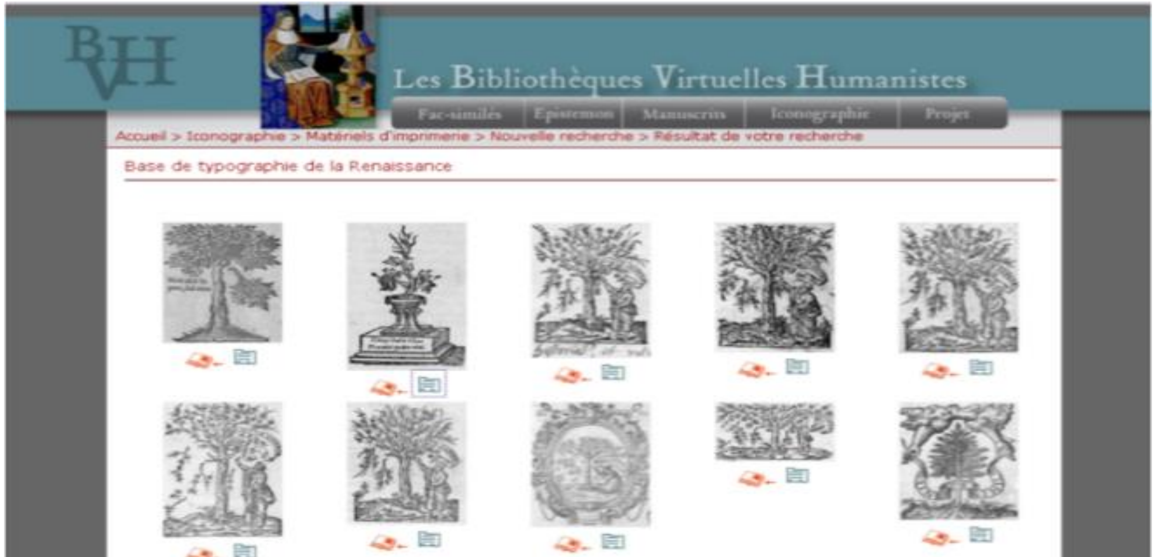
- Ornamental letters (+of 20000)



- Figures (+ de 1500)

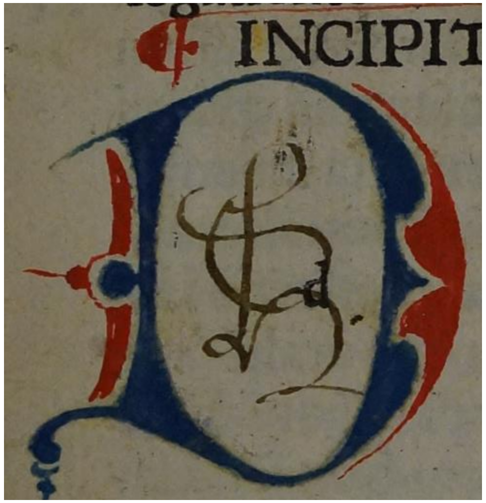


BATYR: http://www.bvh.univ-tours.fr



Accueil > Iconographie > Matériels d'imprimerie > Nouvelle recherche > Résultat de votre recherche

Base de typographie de la Renaissance



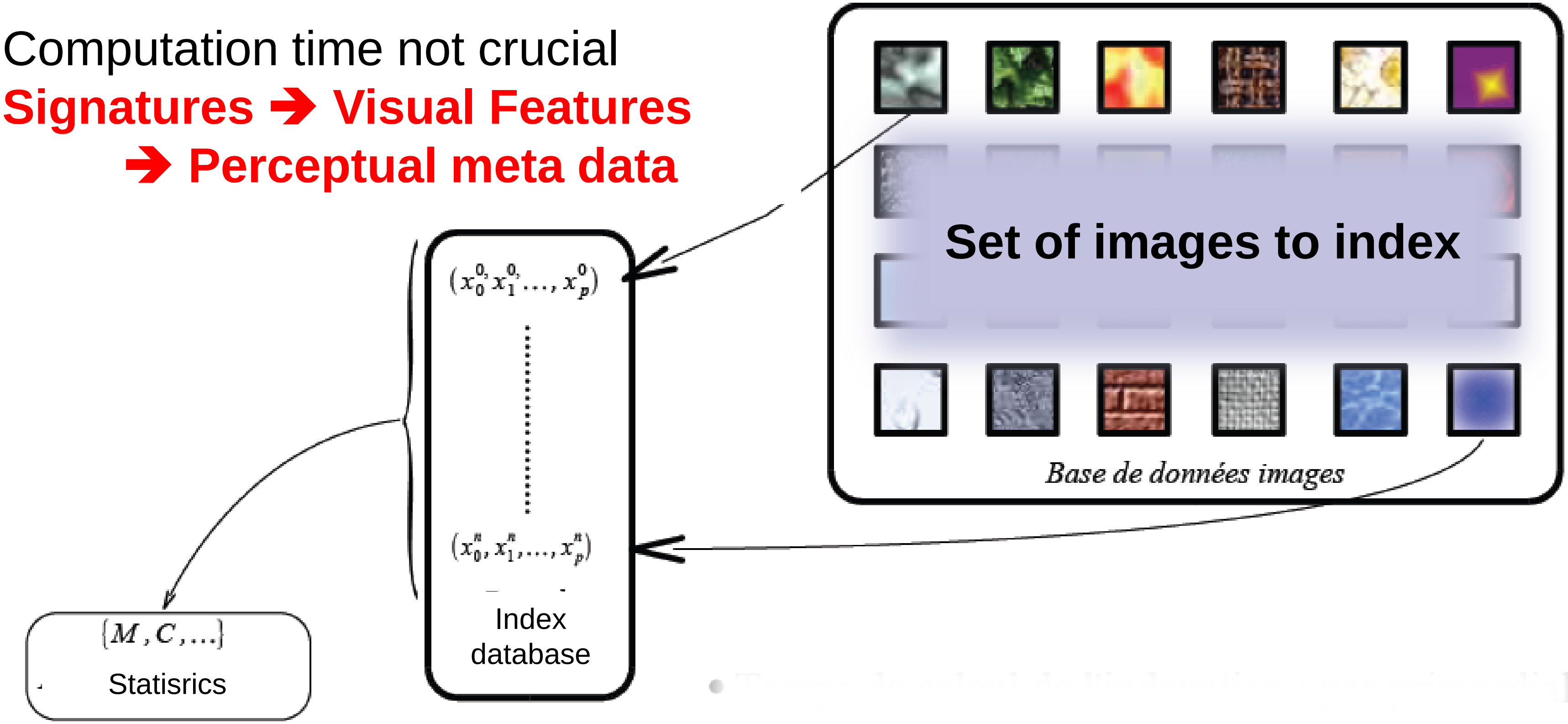
From pixels … to non-textual contents

An overview of Content Based Image Retrieval

Perceptual meta data instead of classical meta data

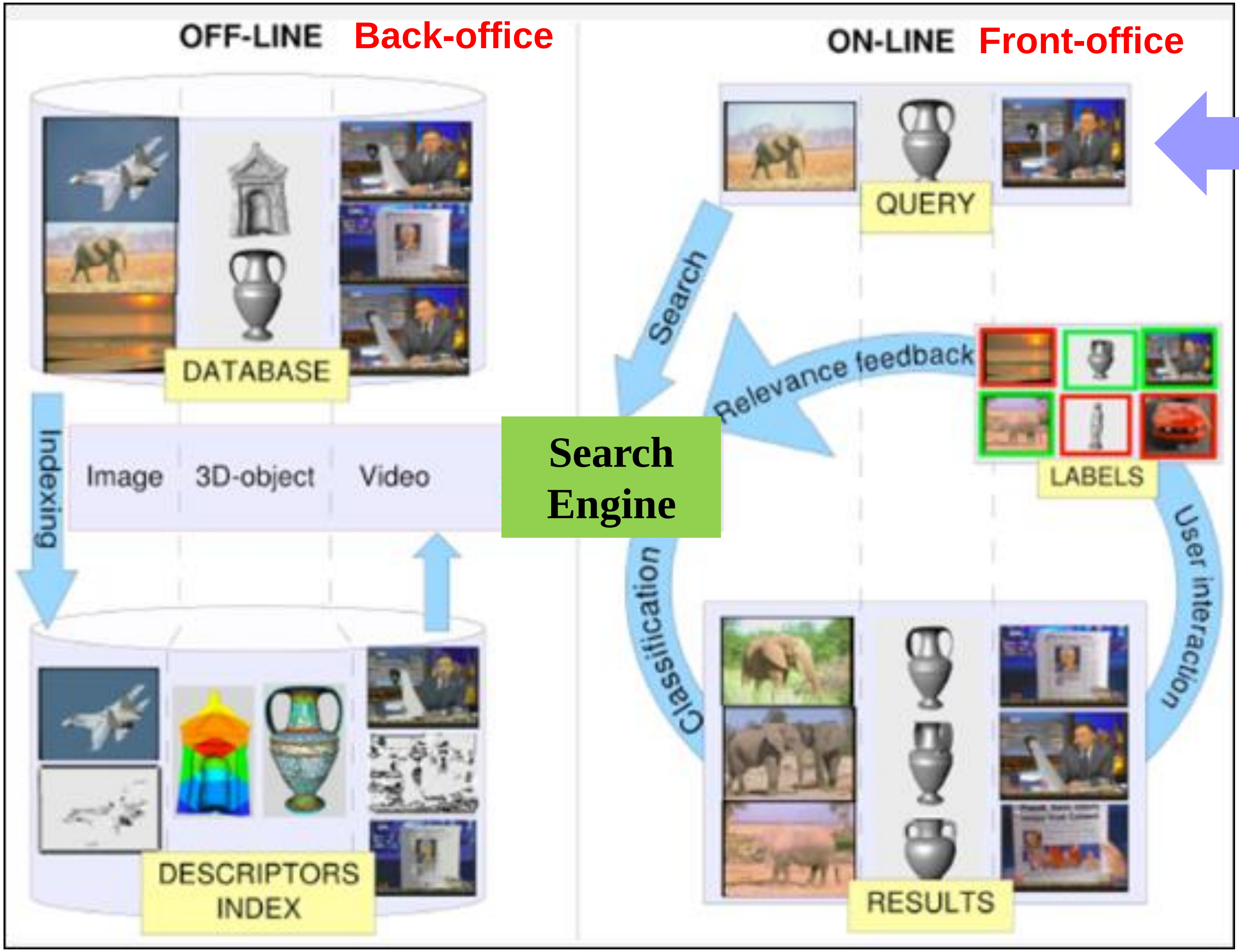
- Computation of signatures for all the images or even **sub-parts of the images (EoC)**

- Computation time not crucial
- **Signatures ➔ Visual Features**
  ➔ Perceptual meta data



From pixels … to non-textual contents

An overview of Content Based Image Retrieval



OFF-LINE Back-office

ON-LINE Front-office

Search Engine

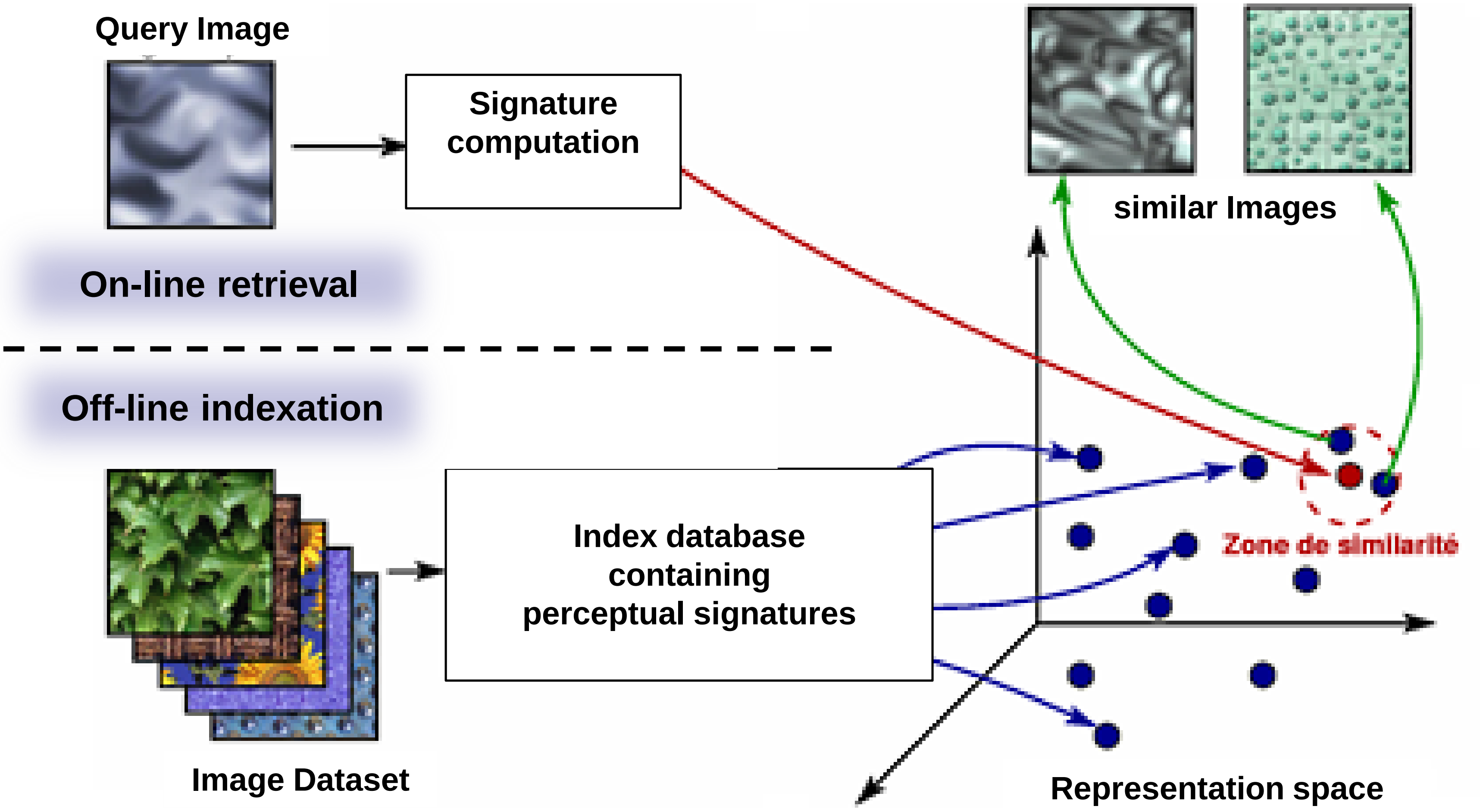
Using images as query instead of words

From pixels … to non-textual contents

An overview of Content Based Image Retrieval

It is again a question of features…

➔ We speak about signatures



Query Image

Signature computation

similar Images

On-line retrieval

Off-line indexation

Index database containing perceptual signatures

Zone de similarité

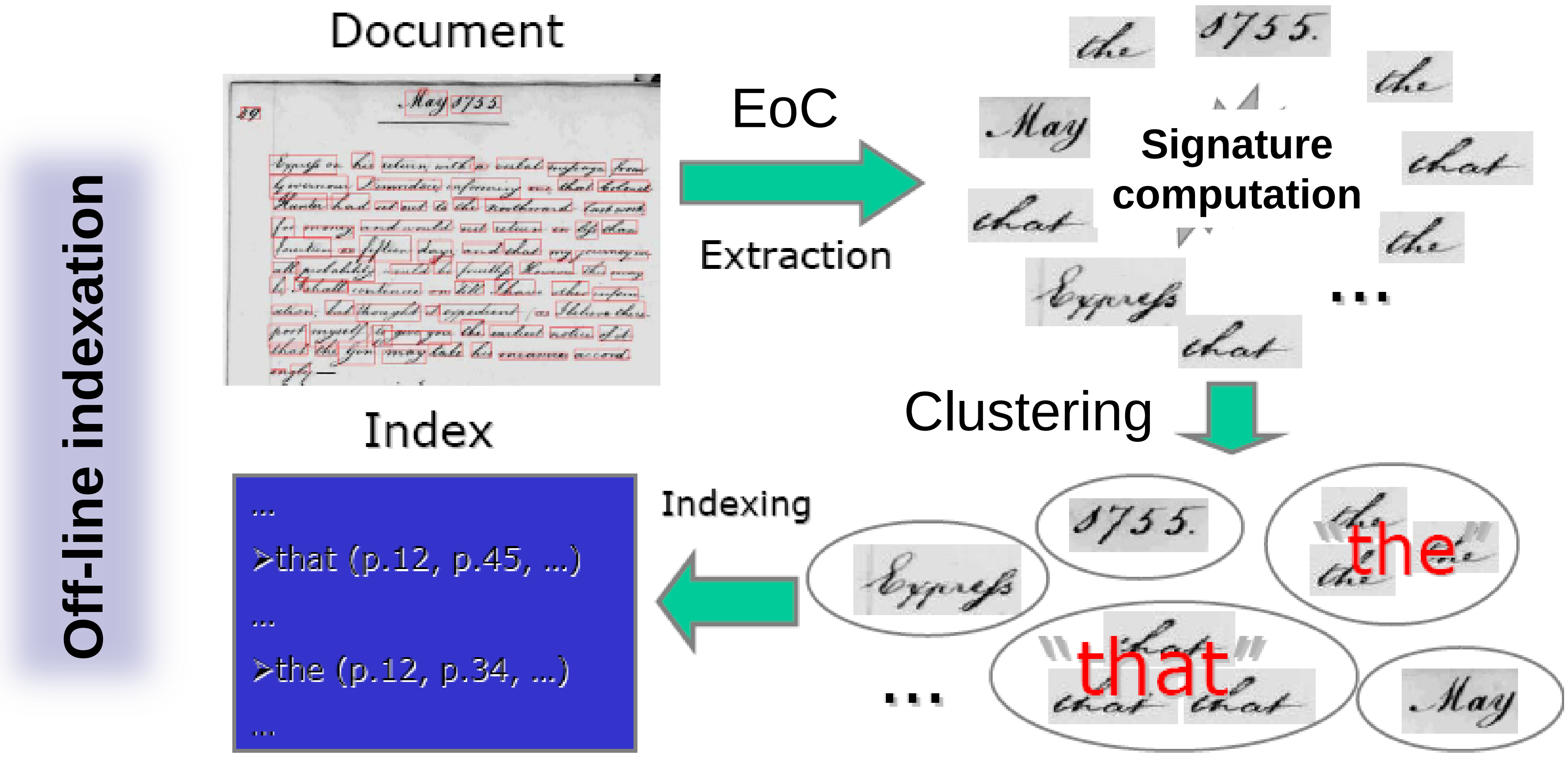
Image Dataset

Representation space

From pixels … to non-textual contents

An overview of Content Based Image Retrieval

- From CBIR to Word spotting



Off-line indexation

Document

EoC

Extraction

Signature computation

...

Clustering

Indexing

...

Index

...

>that (p.12, p.45, ...)

...

>the (p.12, p.34, ...)

...

From pixels … to non-textual contents

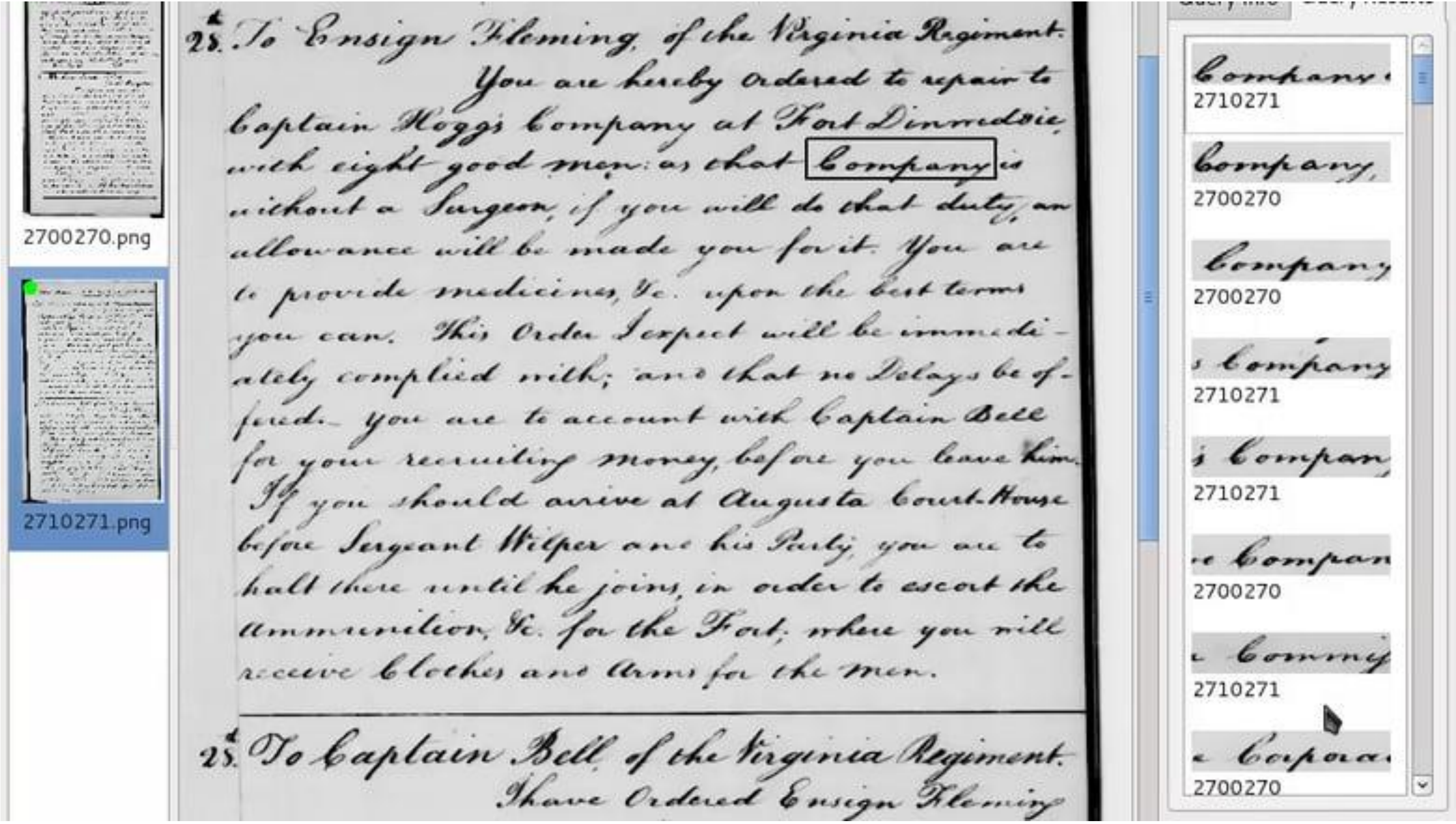
An overview of Content Based Image Retrieval

- From CBIR to Word spotting

On-line Retrieval

Query images

Results: retrieved images



2700270.png

2710271.png

25. To Ensign Fleming of the Virginia Regiment.
You are hereby Ordered to repair to
Captain Hoggs Company at Fort Dinwiddie,
with eight good men: as that Company is
without a Surgeon, if you will do that duty, an
allowance will be made you for it. You are
to provide medicines, &c. upon the best terms
you can. This Order I expect will be immedi-
ately complied with; and that no Delays be of-
fered. you are to account with Captain Bell
for your recruiting money, before you leave him.
If you should arrive at Augusta Court-House
before Serjeant Wilper and his Party, you are to
halt there until he joins, in order to escort the
ammunition, &c. for the Fort, where you will
receive Clothes and Arms for the men.

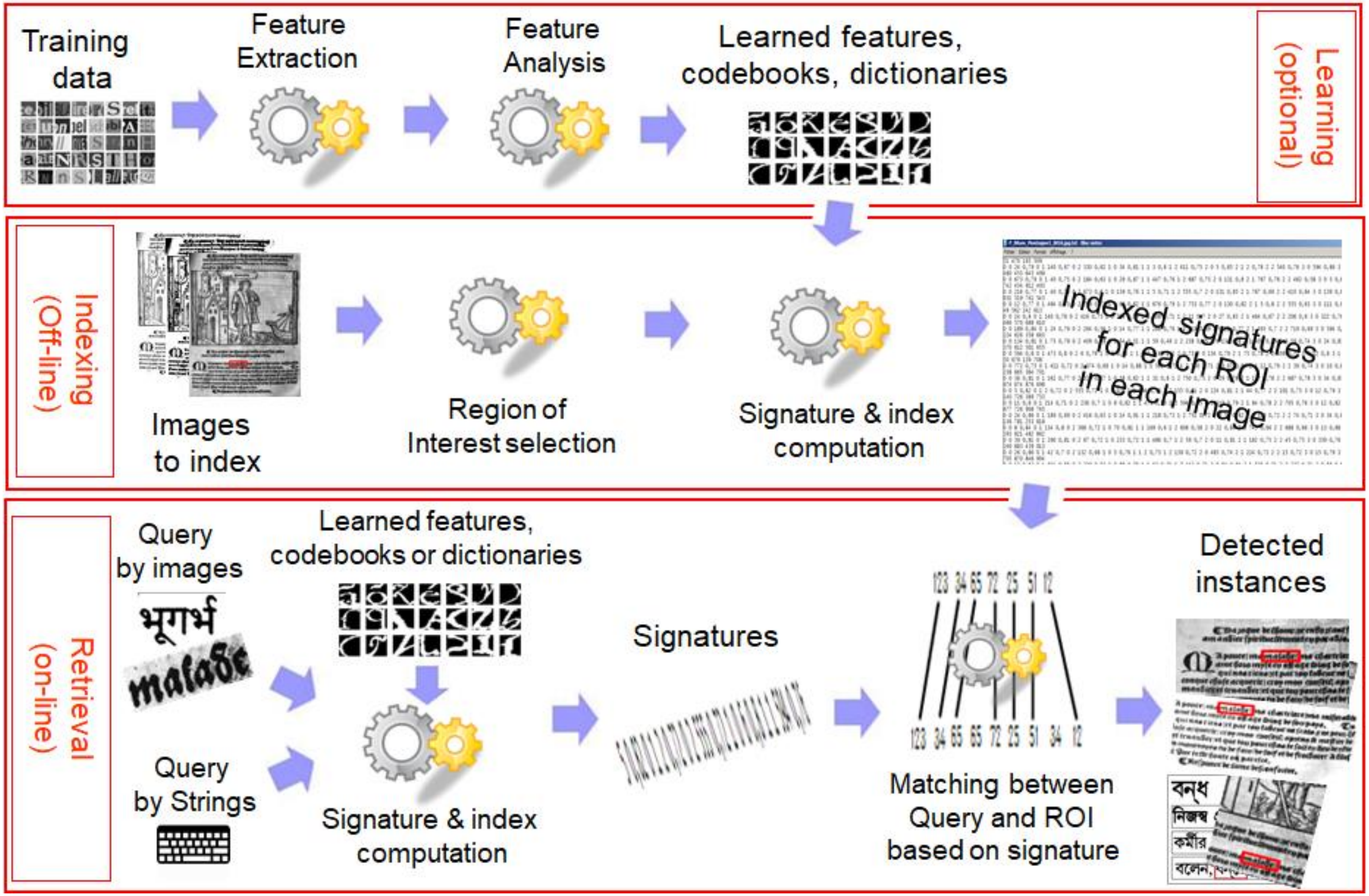
25. To Captain Bell of the Virginia Regiment.
I have Ordered Ensign Fleming

Company	2710271
Company,	2700270
Company	2700270
s Company	2710271
; Compan	2710271
e Compan	2700270
Compny	2710271
Copara	2700270

From pixels … to non-textual contents

An overview of Content Based Image Retrieval

- From CBIR to Word spotting



Training data

Feature Extraction

Feature Analysis

Learned features, codebooks, dictionaries

Learning (optional)

Indexing (Off-line)

Images to index

Region of Interest selection

Signature & index computation

Indexed signatures for each ROI in each image

Retrieval (on-line)

Query by images

भूगर्भ

malade

Query by Strings

Learned features, codebooks or dictionaries

Signature & index computation

Signatures

123 34 65 72 25 51 12

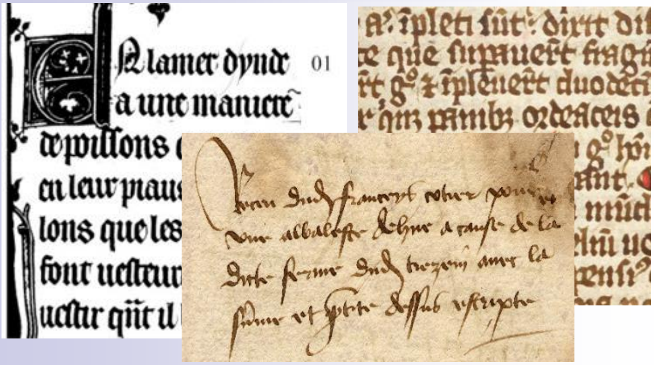
123 34 65 65 72 25 51 34 12

Matching between Query and ROI based on signature

Detected instances

Is it just a question of meta-data?

IMAGES



CONTENTS

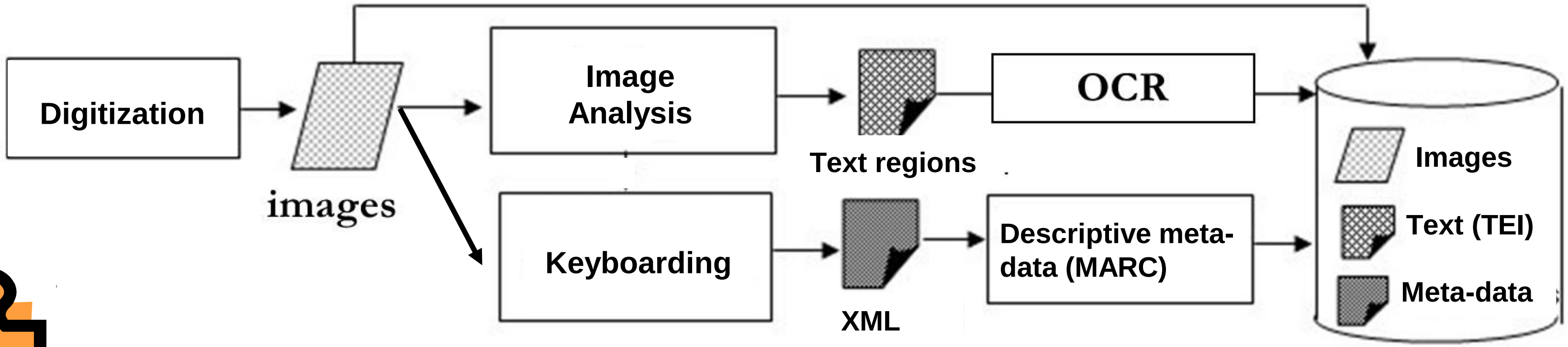
Dans la Mer d'Inde
il y a une espèce
de poissons qui ont
dans leur peau des poils si
longs que les ge
quant il les ont a
en font des vête
pour s'habiller.

The standard model

Descriptive meta-data + transcription

We have usually

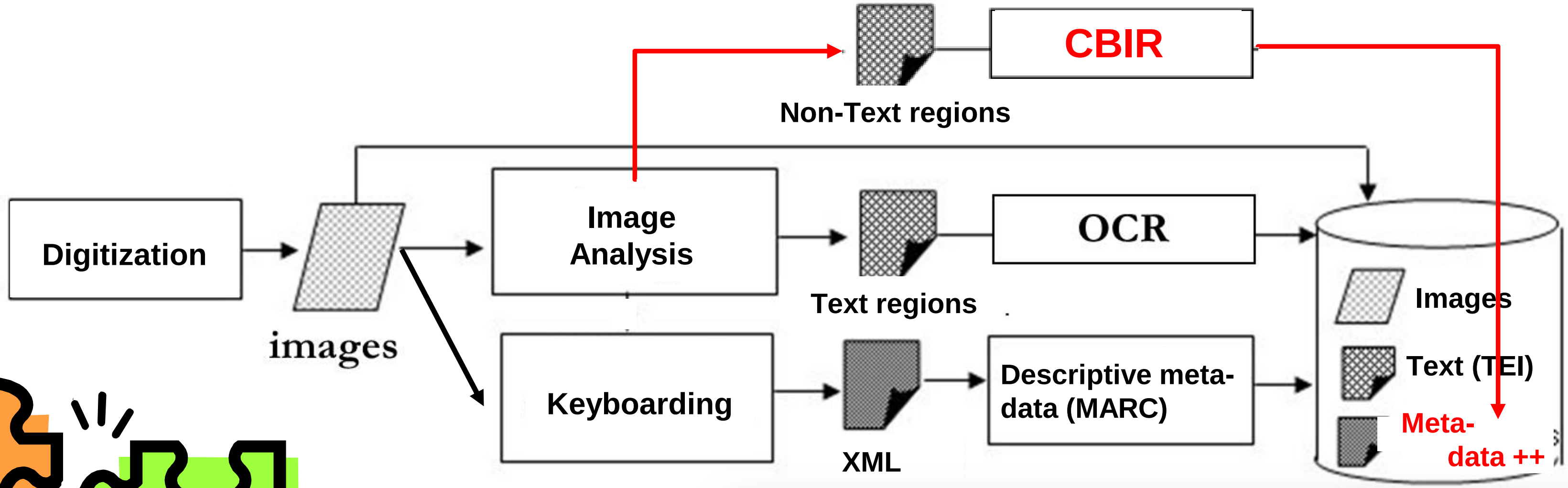
- Descriptive meta-data in standard formats **(MARC, EAD, Dublin Core, MODS, ...)**
  - Edited manually
  - “Semantical” information
- Text transcription associated to additional meta-data (TEI)
  - Semi-automatic transcription or manually edited
  - “Semantical” information

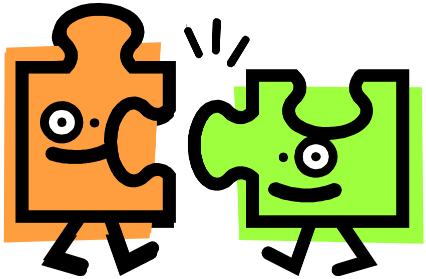


“Semantical” meta-data

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

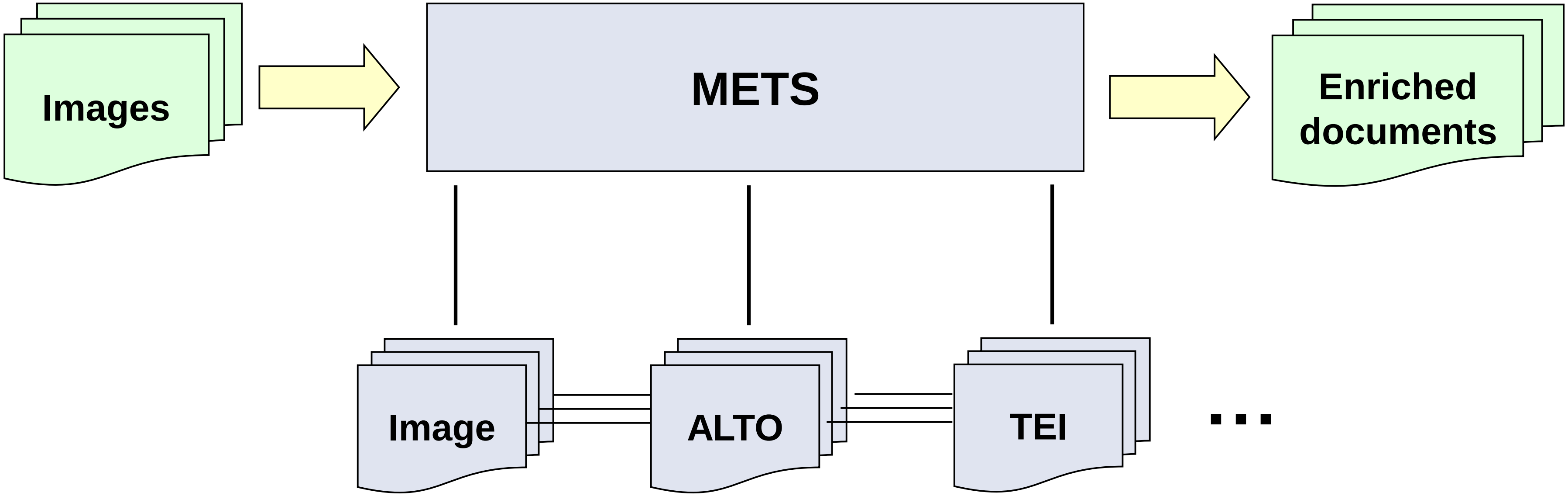
Is it just a problem of Meta-data?

What about the encoding formats?

Structuration of data and file formats is a difficult problem

- Data architects are needed

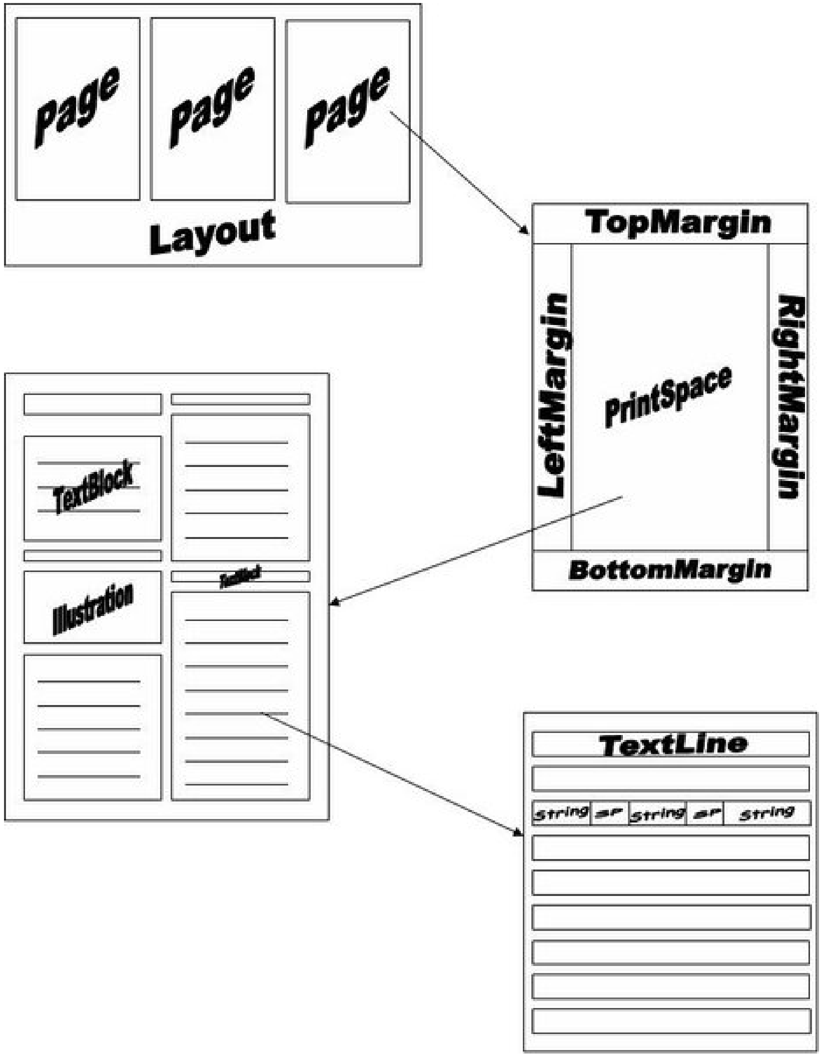
- Some interesting formats linked with previous discussions
  - **METS – Metadata Encoding and Transmission Standard**
  - ALTO – Analyzed Layout and Text Object
  - TEI – Text Encoding Initiative



Is it just a problem of Meta-data?
ALTO for OCR

ALTO = Analyzed Layout and Text Object

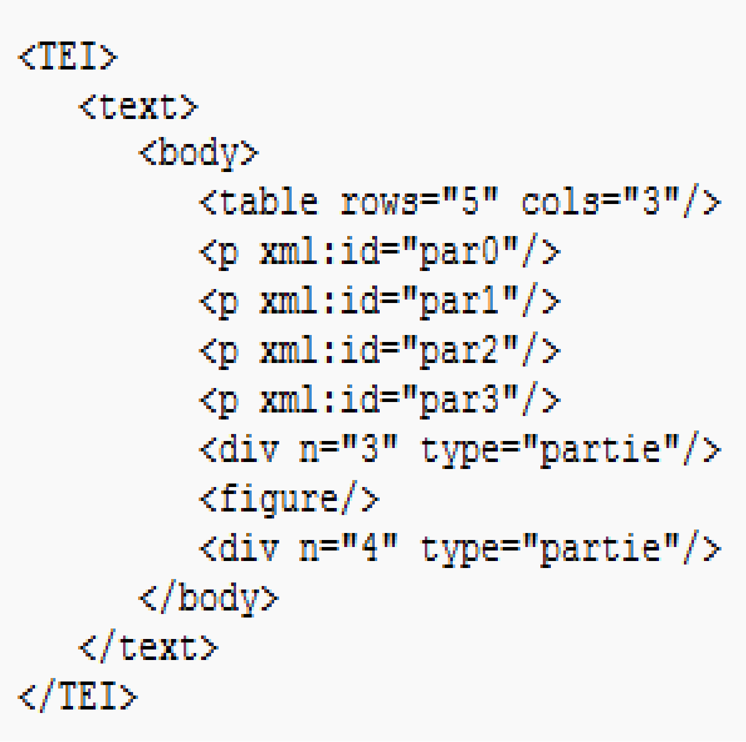
- Standard XML
- Created in 2003 during METAe project
- Developed by Graz, Linz, Innsbruck universities
- **Description of the content and the physical layout of one page**
- Used by several OCR software
- Adapted and used by the BNF and other libraries
- Drawbacks: huge / static



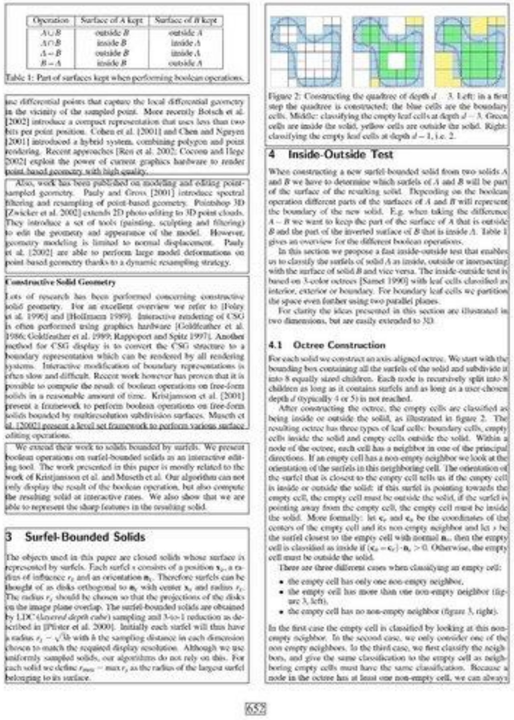
Is it just a problem of Meta-data?

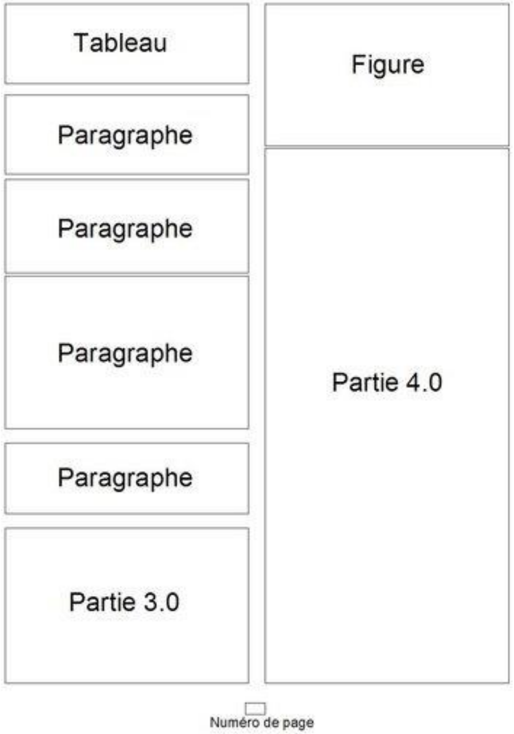
TEI for transcription and enriched contents

- **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>
```





Tableau

Figure

Paragraphe

Paragraphe

Paragraphe

Paragraphe

Partie 4.0

Partie 3.0

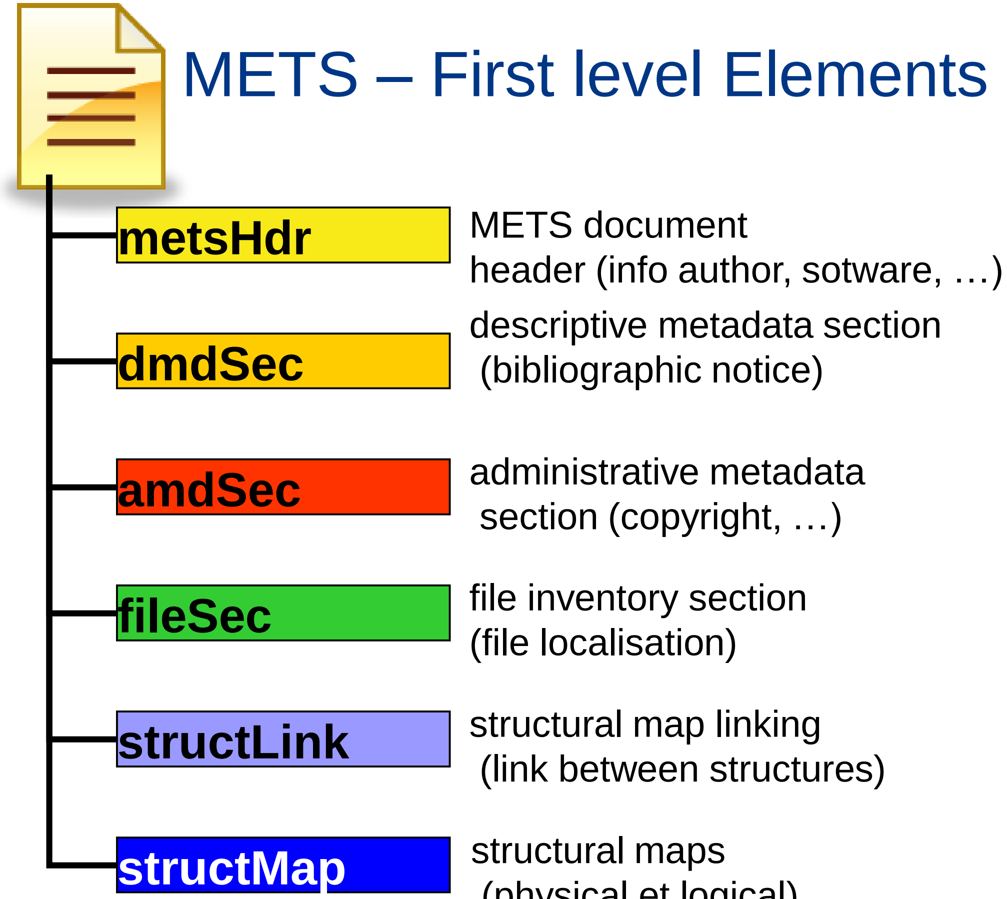
Numéro de page

Is it just a problem of Meta-data?

Link between meta data?

METS – Metadata Encoding and Transmission Standard

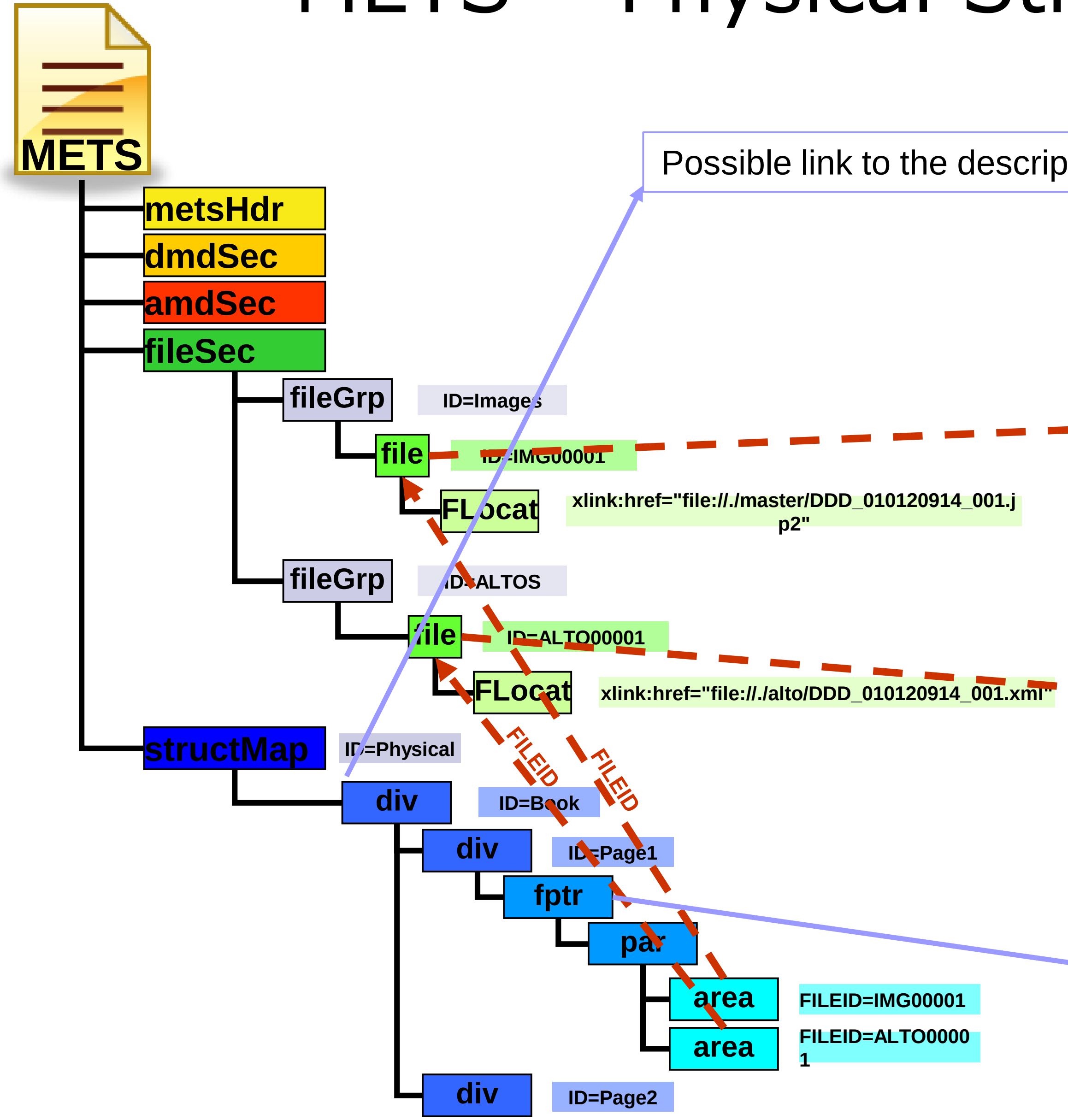
- Open XML Standard created in 2001 by the Digital Library Federation maintained by METS Editorial Board
- XSD-Schema
- **Linking between multimedia objects**
- Complete Description of digitized content (images, texts, audio, sculptures, …)
- Physical / logical structures
- Descriptive Meta data (DC,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



METS – First level Elements

metsHdr	METS document header (info author, sotware, …)
dmdSec	descriptive metadata section (bibliographic notice)
amdSec	administrative metadata section (copyright, …)
fileSec	file inventory section (file localisation)
structLink	structural map linking (link between structures)
structMap	structural maps (physical et logical)

METS – Physical Structure



METS

metsHdr

dmdSec

amdSec

fileSec

fileGrp ID=Images

file ID=IMG00001

FLocat xlink:href="file://./master/DDD_010120914_001.jp2"

fileGrp ID=ALTOS

file ID=ALTO00001

FLocat xlink:href="file://./alto/DDD_010120914_001.xml"

structMap ID=Physical

div ID=Book

div ID=Page1

fptr

par

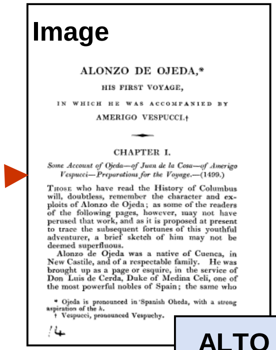
area FILEID=IMG00001

area FILEID=ALTO00001

div ID=Page2

Possible link to the descriptive meta data (dmd Sec)

Image



ALONZO DE OJEDA,*

HIS FIRST VOYAGE,

IN WHICH HE WAS ACCOMPANIED BY

AMERIGO VESPUCCI.†

CHAPTER I.

Some Account of Ojeda—of Juan de la Cosa—of Amerigo Vespucci—Preparations for the Voyage.—(1499.)

Those who have read the History of Columbus will, doubtless, remember the character and exploits of Alonzo de Ojeda; as some of the readers of the following pages, however, may not have perused that work, and as it is proposed at present to trace the subsequent fortunes of this youthful adventurer, a brief sketch of him may not be deemed superfluous.

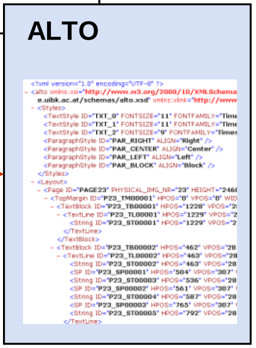
Alonzo de Ojeda was a native of Cuenca, in New Castile, and of a respectable family. He was brought up as a page or esquire, in the service of Don Luis de Cerda, Duke of Medina Celi, one of the most powerful nobles of Spain; the same who

* Ojeda is pronounced in Spanish Oheda, with a strong aspiration of the h.

† Vespucci, pronounced Vespuchy.

14

ALTO

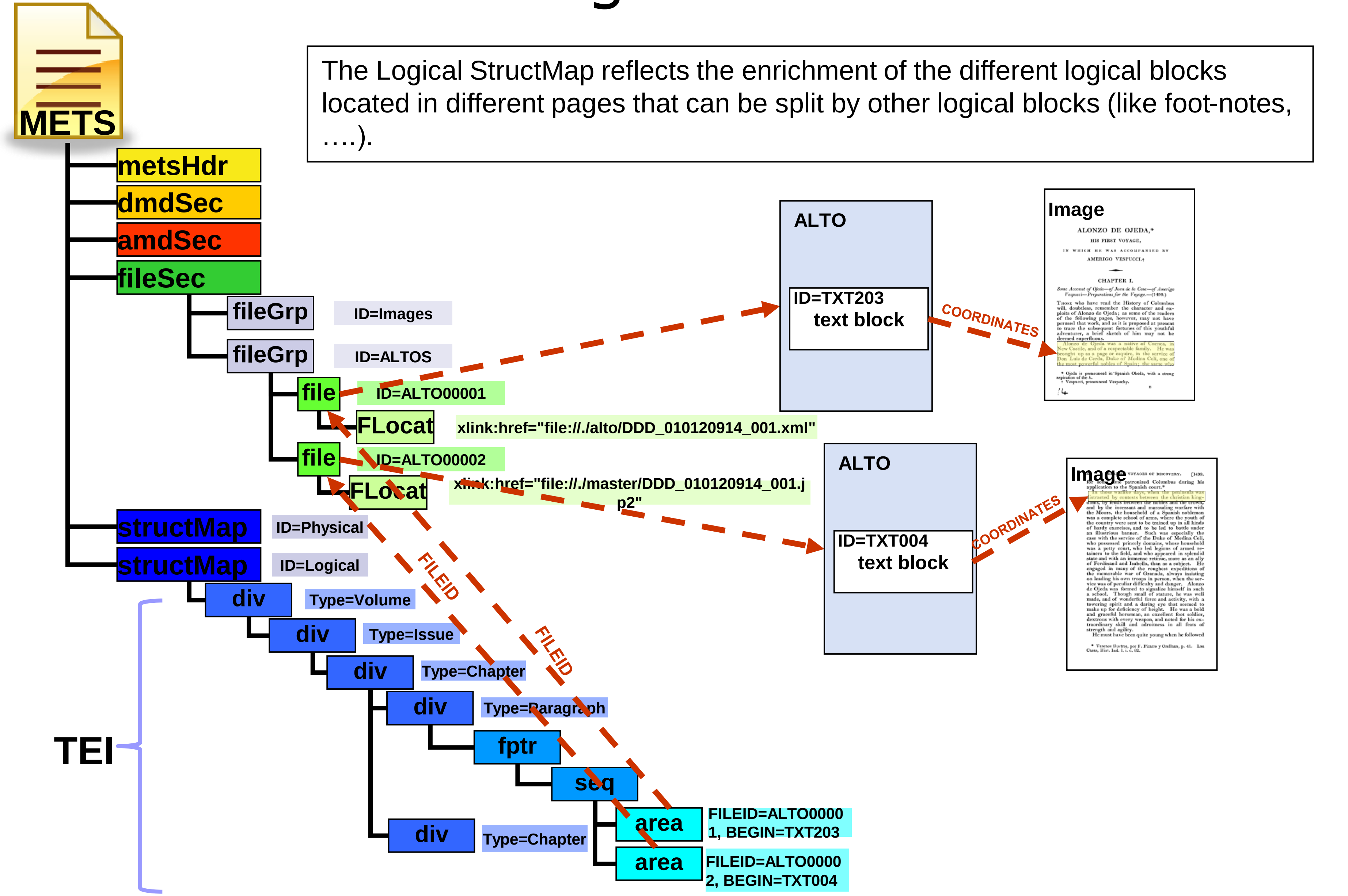


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <alto xmlns:xsi="http://www.w3.org/2000/10/XMLSchema
    e.uibk.ac.at/schemas/alto.xsd" xmlns:xlink="http://www
  - <Styles>
      <TextStyle ID="TXT_0" FONTSIZE="11" FONTFAMILY="Time
      <TextStyle ID="TXT_1" FONTSIZE="11" FONTFAMILY="Time
      <TextStyle ID="TXT_2" FONTSIZE="9" FONTFAMILY="Times
      <ParagraphStyle ID="PAR_RIGHT" ALIGN="Right" />
      <ParagraphStyle ID="PAR_CENTER" ALIGN="Center" />
      <ParagraphStyle ID="PAR_LEFT" ALIGN="Left" />
      <ParagraphStyle ID="PAR_BLOCK" ALIGN="Block" />
    </Styles>
  - <Layout>
    - <Page ID="PAGE23" PHYSICAL_IMG_NR="23" HEIGHT="2464
      - <TopMargin ID="P23_TM00001" HPOS="0" VPOS="0" WID
        - <TextBlock ID="P23_TB00001" HPOS="1228" VPOS="2
          - <TextLine ID="P23_TL00001" HPOS="1229" VPOS="2
              <String ID="P23_ST00001" HPOS="1229" VPOS="2
            </TextLine>
          </TextBlock>
        - <TextBlock ID="P23_TB00002" HPOS="462" VPOS="28
          - <TextLine ID="P23_TL00002" HPOS="463" VPOS="28
              <String ID="P23_ST00002" HPOS="463" VPOS="28
              <SP ID="P23_SP00001" HPOS="504" VPOS="307"
              <String ID="P23_ST00003" HPOS="536" VPOS="28
              <SP ID="P23_SP00002" HPOS="561" VPOS="307"
              <String ID="P23_ST00004" HPOS="587" VPOS="28
              <SP ID="P23_SP00003" HPOS="765" VPOS="307"
              <String ID="P23_ST00005" HPOS="792" VPOS="28
            </TextLine>
```

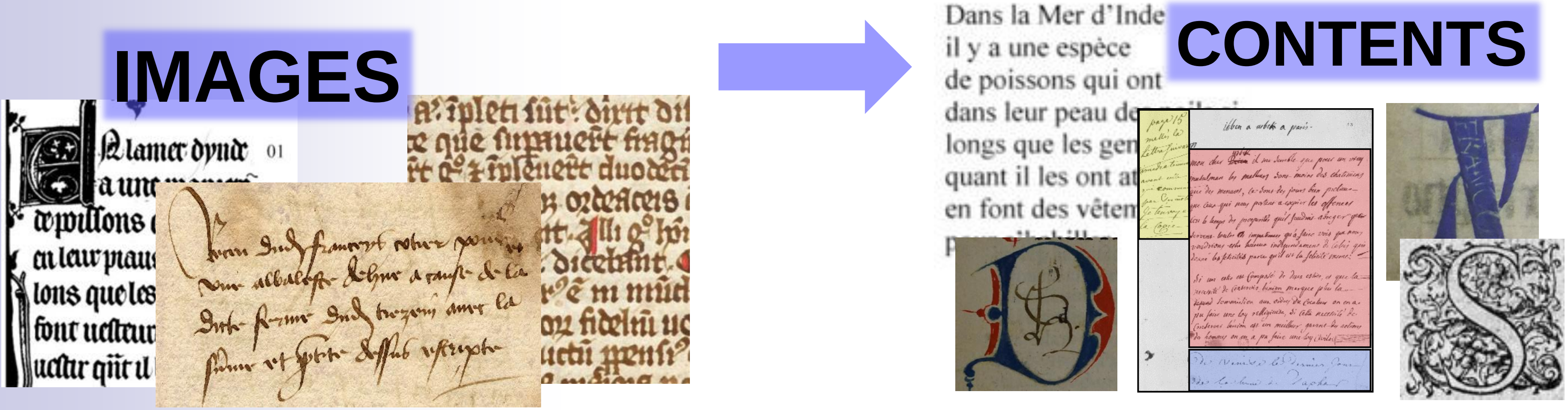
METS allows to specify the locations of resource files ➔ File pointers METS (fptr)

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 vêtem

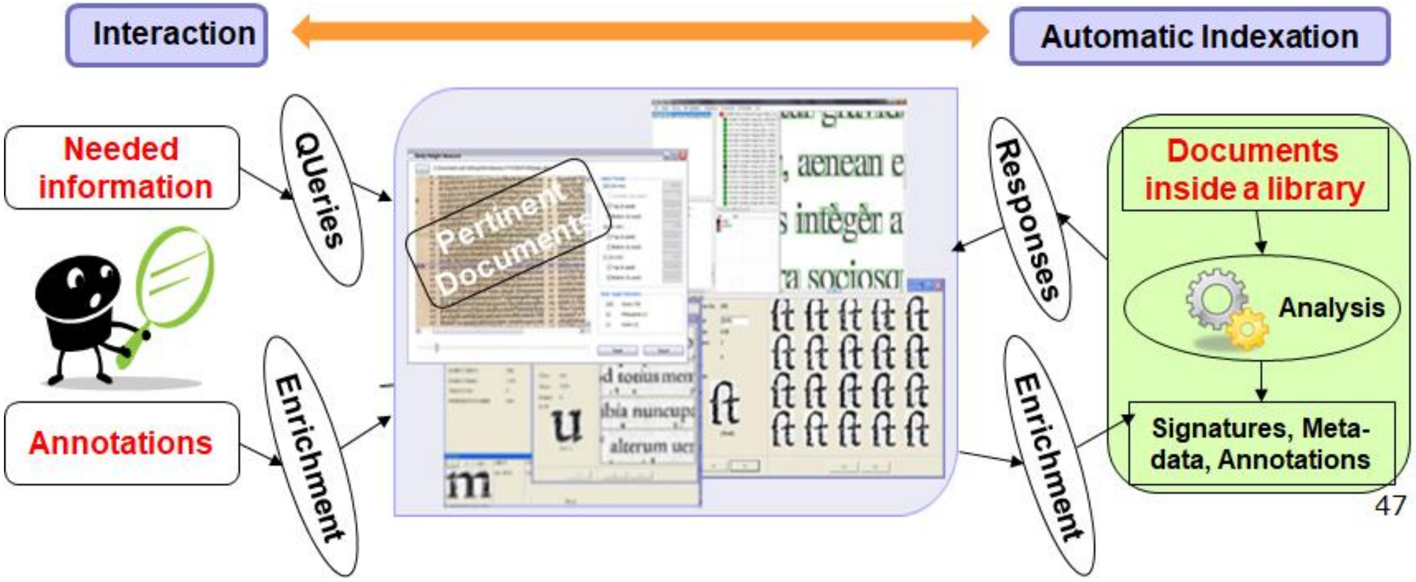
From Pixels to Contents
Conclusions

- **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



Interaction

Automatic Indexation

Needed information

QUeries

Pertinent Documents

Responses

Documents inside a library

Analysis

Signatures, Meta-data, Annotations

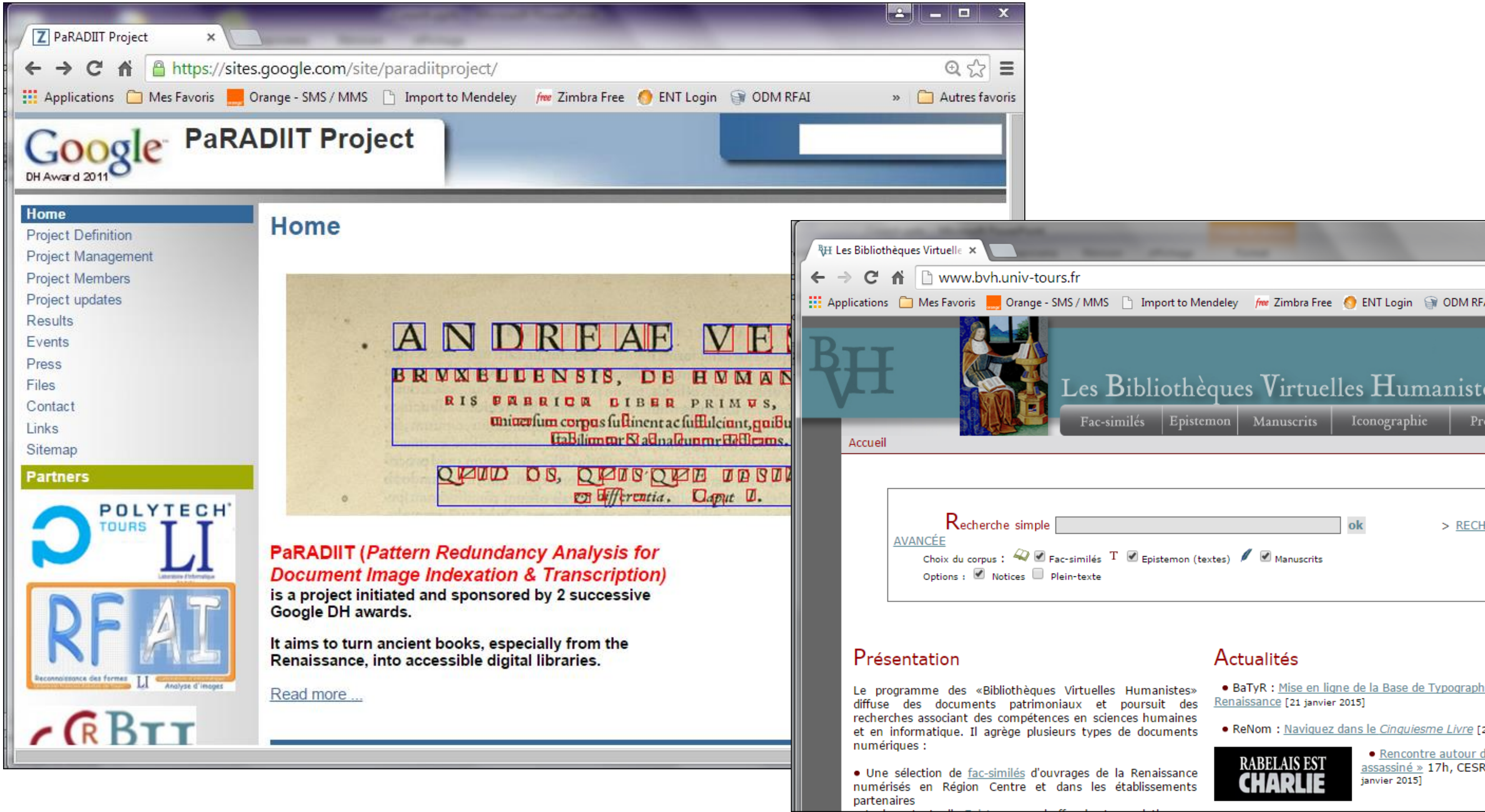
Annotations

Enrichment

Enrichment

Thanks…

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



PaRADIIT Project

Home

PaRADIIT (*Pattern Redundancy Analysis for Document Image Indexation & Transcription)* is a project initiated and sponsored by 2 successive Google DH awards.

It aims to turn ancient books, especially from the Renaissance, into accessible digital libraries.

Read more ...

Recherche simple

Choix du corpus : ☑ Fac-similés ☑ Epistemon (textes) ☑ Manuscrits

Options : ☑ Notices □ Plein-texte

Présentation

Le programme des «Bibliothèques Virtuelles Humanistes» diffuse des documents patrimoniaux et poursuit des recherches associant des compétences en sciences humaines et en informatique. Il agrège plusieurs types de documents numériques :

- Une sélection de fac-similés d'ouvrages de la Renaissance numérisés en Région Centre et dans les établissements partenaires

Actualités

- BaTyR : Mise en ligne de la Base de Typograph Renaissance [21 janvier 2015]
- ReNom : Naviguez dans le *Cinquiesme Livre* [2
- Rencontre autour d assassiné » 17h, CESR janvier 2015]