Les Bibliothèques Virtuelles Humanistes
http://www.bvh.univ-tours.fr

Assemblée générale 2011
Partie 6

Projets en cours et nouveaux programmes de recherche

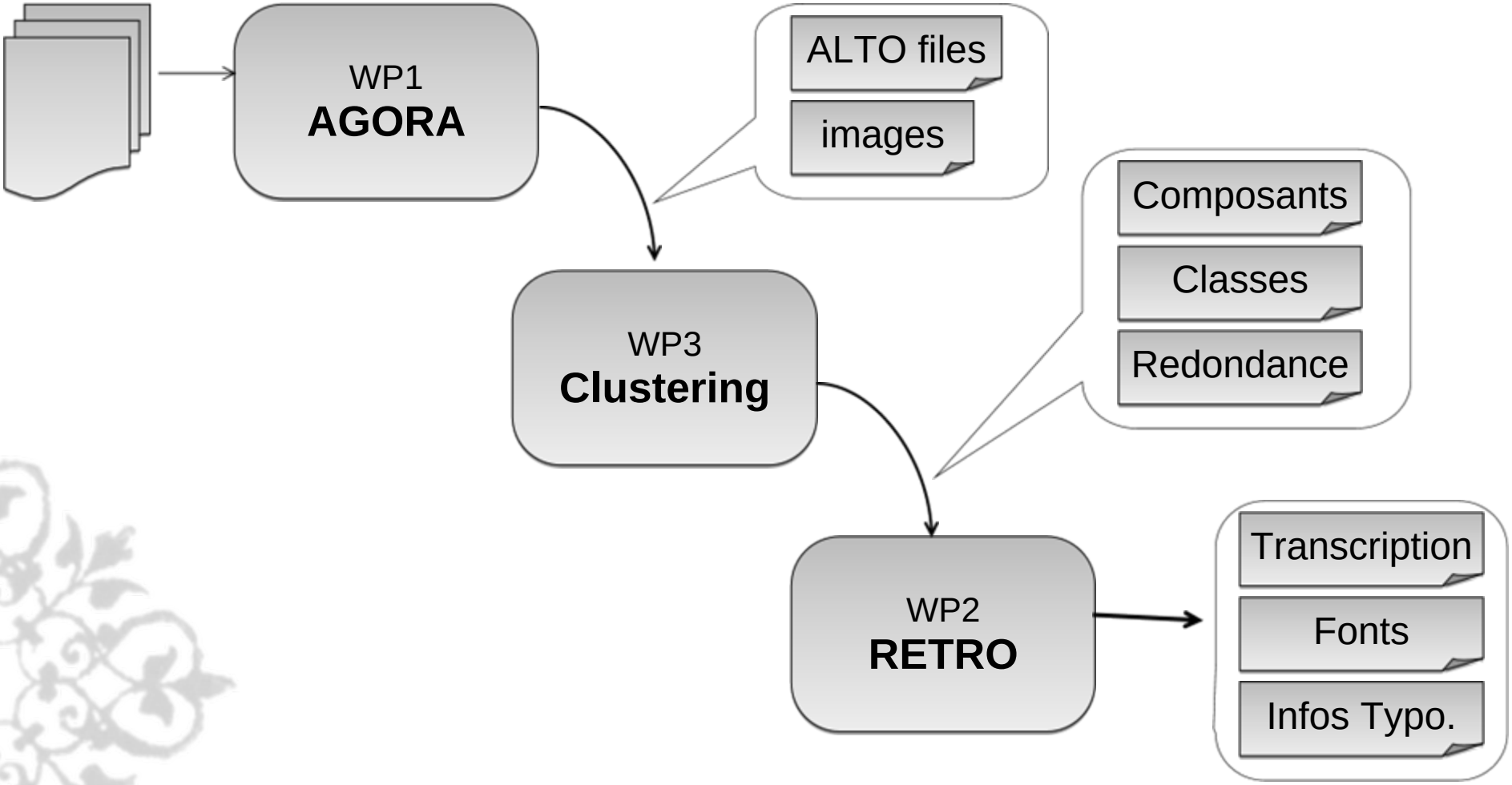
Marie-Luce Demonet et l'équipe des BVH
Lundi 5 décembre 2011
CESR, Tours

Bourse Google 1, Laboratoire d'Informatique : l'OCR RETRO et le classement des caractères

Frédéric Rayar (LI, Tours)

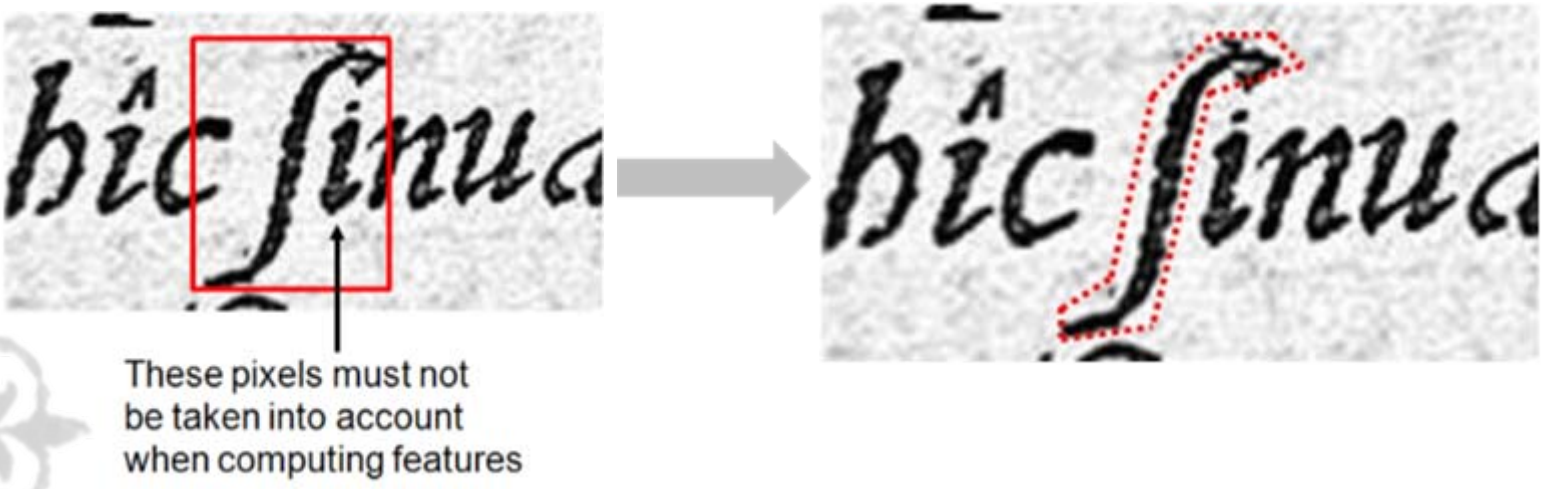
Projet PaRADIIT

Pattern Redundancy Analysis
for Document Image Indexation and Transcription

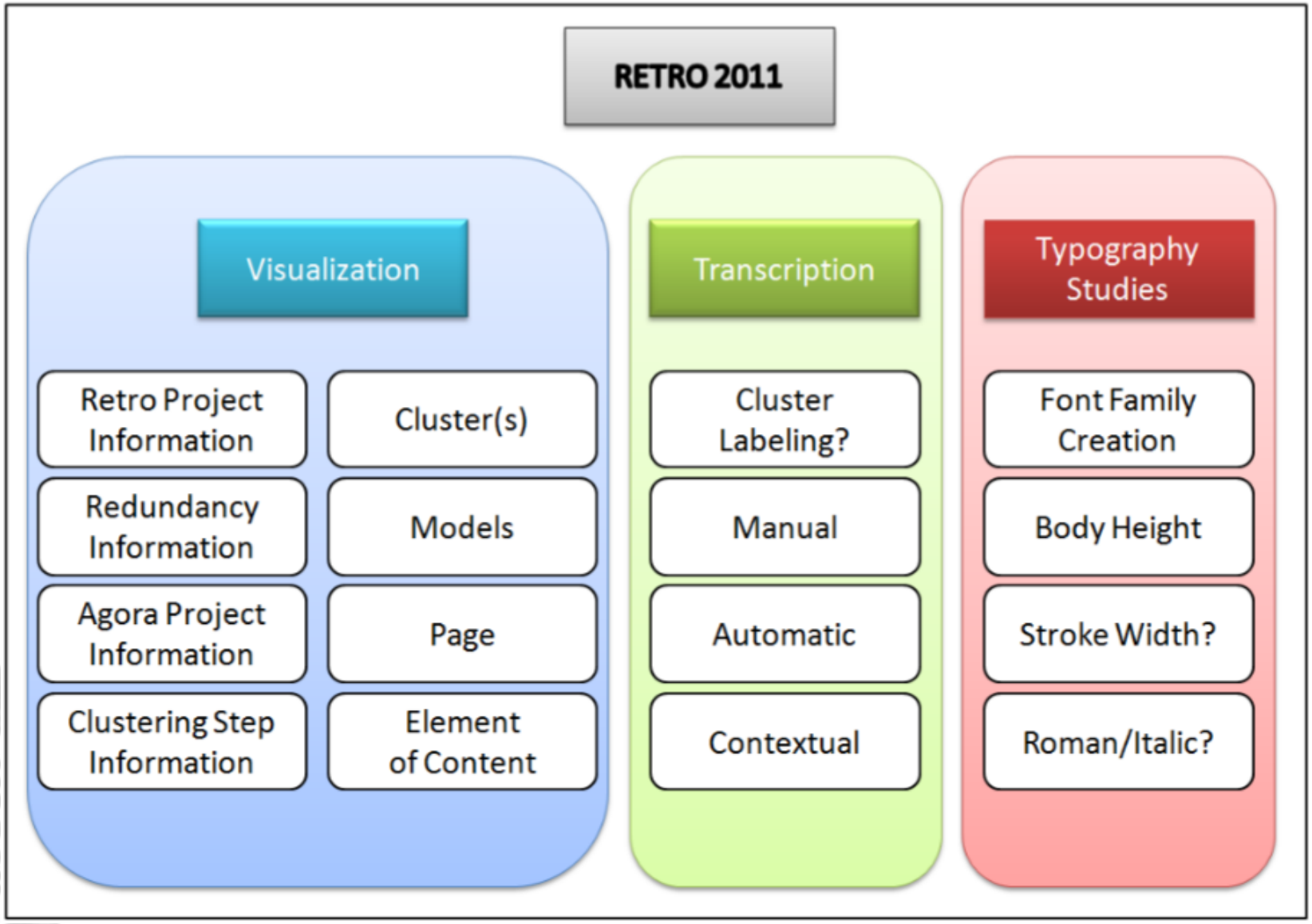


AGORA & Clustering

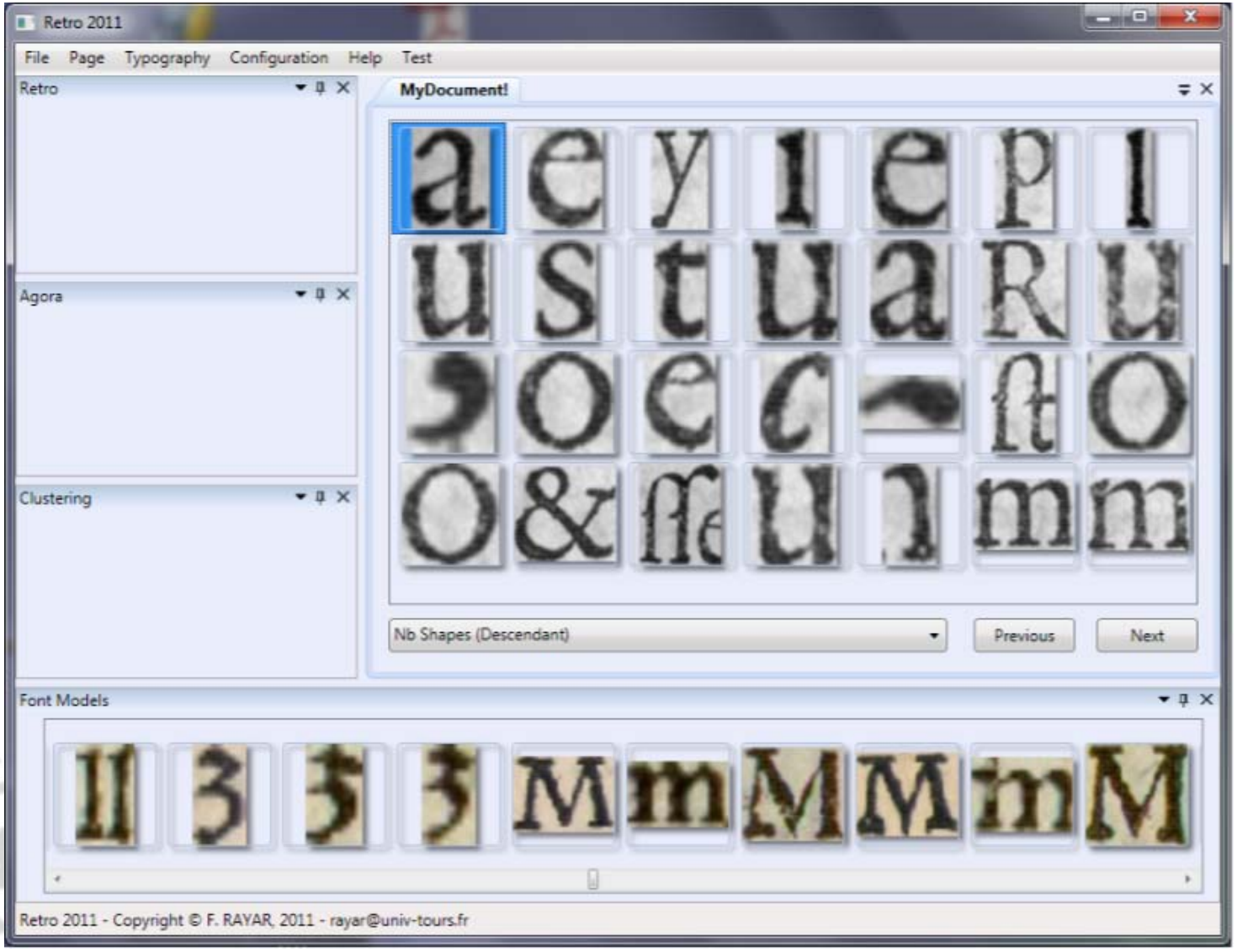
- WP1 – Segmentation:
  - Nouvelle version en cours de création par Pascal Bourquin.
  - Alignement des sorties sur les standards ALTO, METS, TEI Renaissanc
- WP3 - Clustering:
  - Importante amélioration en cours de réalisation.
  - Travail réalisé par Thierry Brouard



RETRO


RETRO 2011
Visualization
Retro Project Information
Cluster(s)
Redundancy Information
Models
Agora Project Information
Page
Clustering Step Information
Element of Content
Transcription
Cluster Labeling?
Manual
Automatic
Contextual
Typography Studies
Font Family Creation
Body Height
Stroke Width?
Roman/Italic?

RETRO


Retro 2011
File
Page
Typography
Configuration
Help
Test
Retro
Agora
Clustering
MyDocument!
Nb Shapes (Descendant)
Previous
Next
Font Models
Retro 2011 - Copyright © F. RAYAR, 2011 - rayar@univ-tours.fr

RETRO et la typographie ancienne

Constitution d'une base de caractères typographiques :

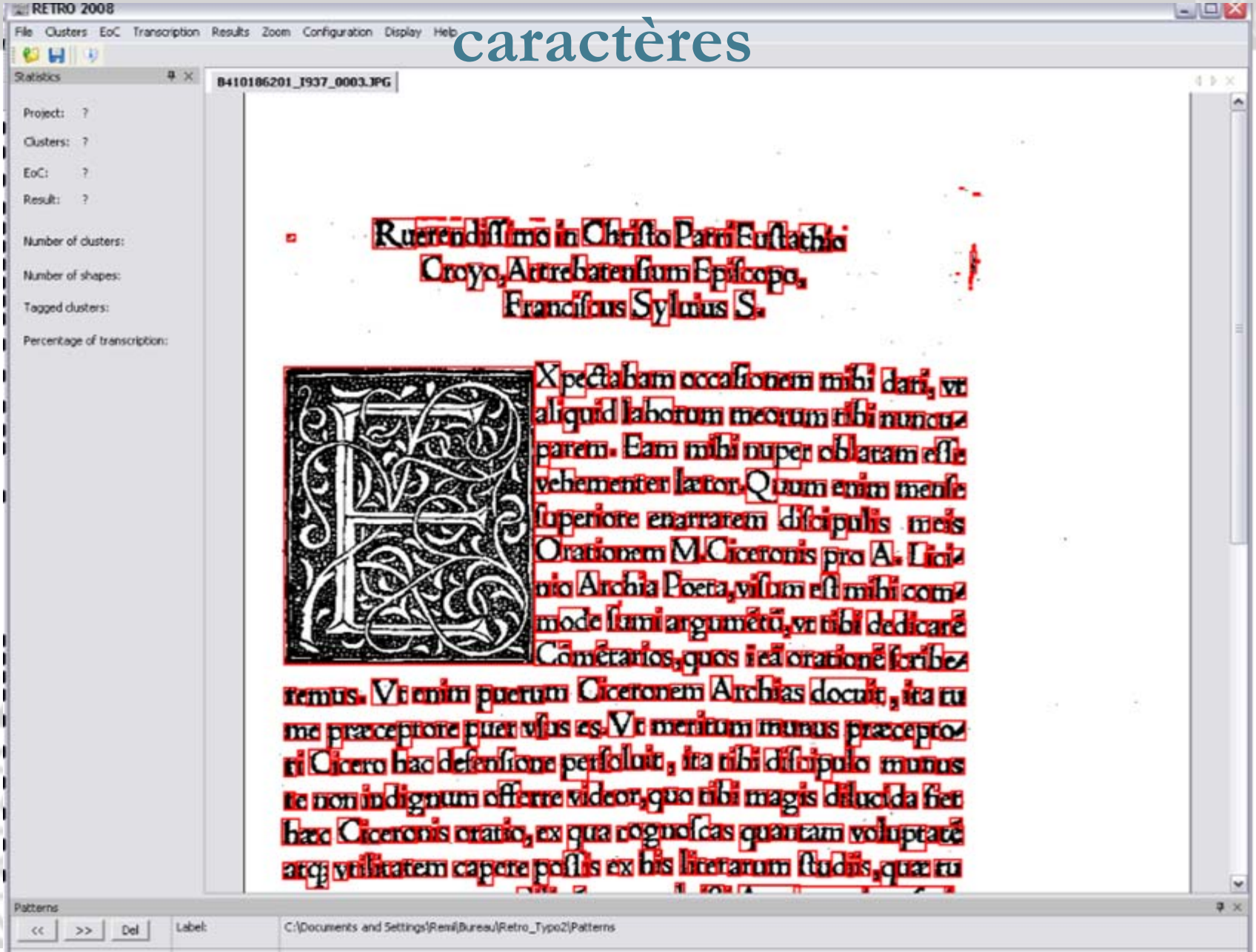
- **enjeux :**

1. améliorer l'OCR
2. (à terme) analyse de la circulation et de l'évolution des matériels typographiques.

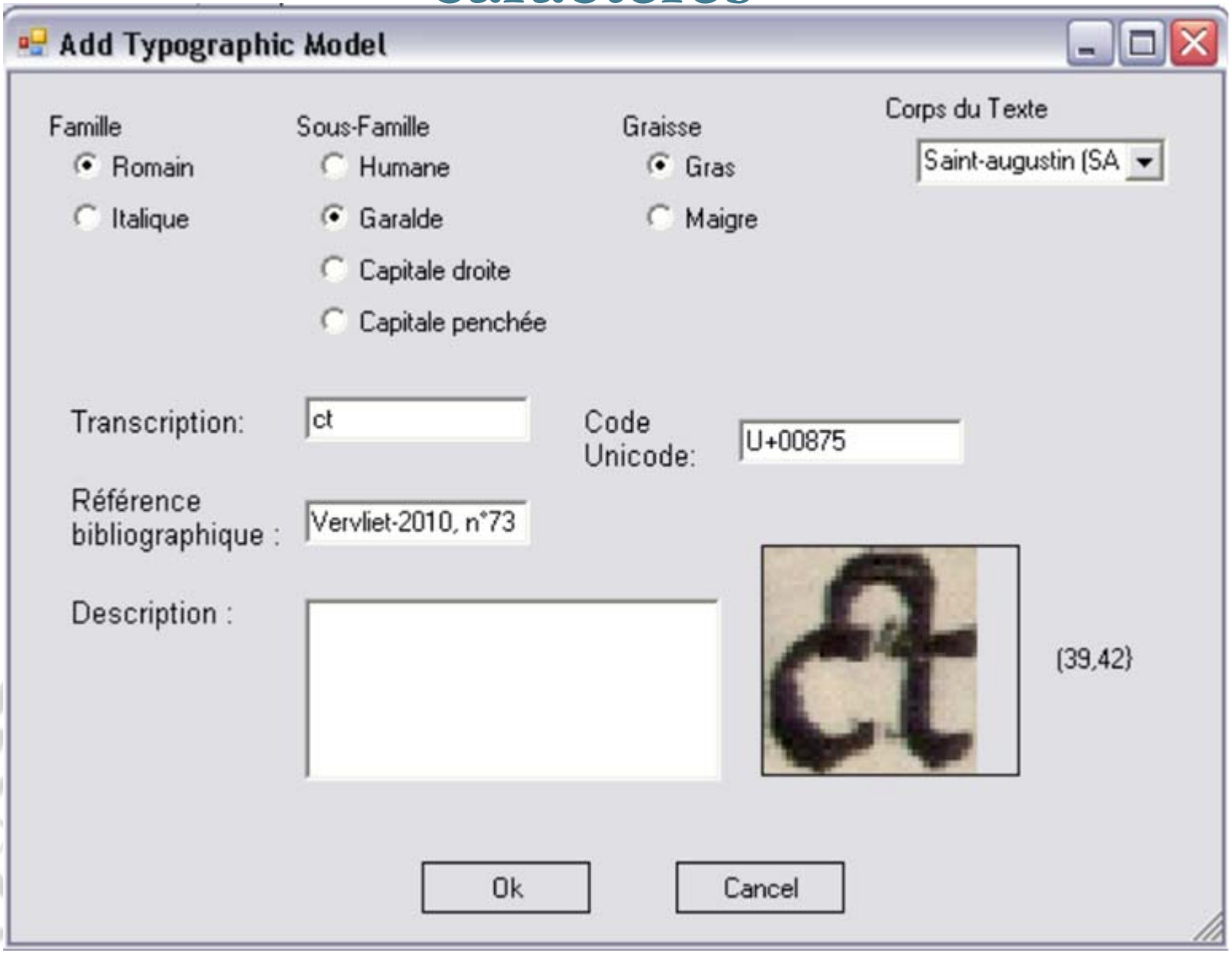
Problèmes :

- **le format de stockage des données (XML + vignettes binaires + vignettes couleur)**
- **les métadonnées : indexation des caractères**

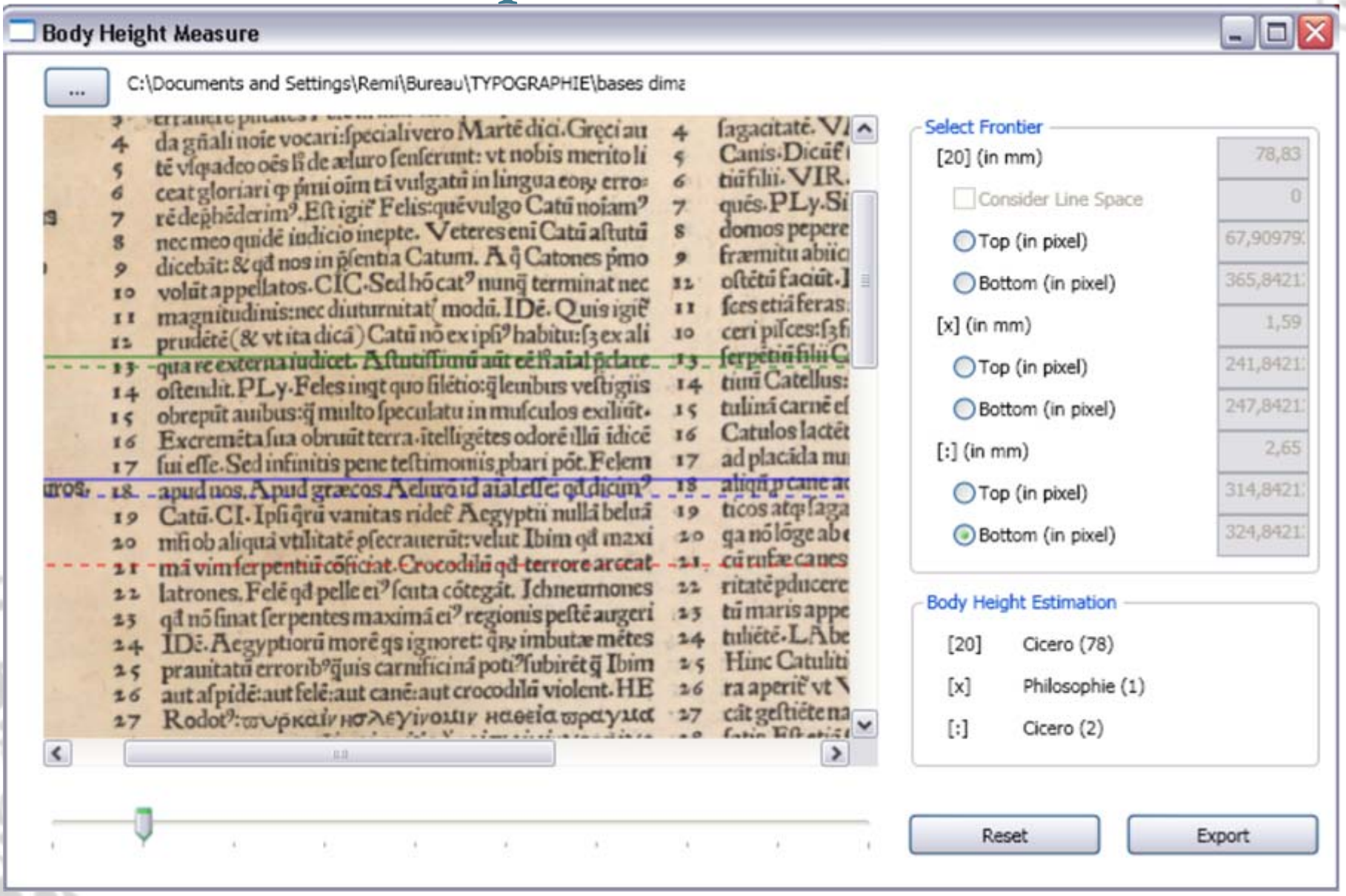
RETRO : module d'indexation de caractères



RETRO : module d'indexation de caractères



RETRO : module de mesure de corps de caractères



RETRO et la typographie ancienne

Difficultés rencontrées

- Segmentation des caractères (résolu)
- Gestion des accents et des éléments flottants (en cours)
- Codage des caractères : coïncidence entre polices numériques et polices anciennes (en réflexion)

Futurs Travaux

- **Le Projet PaRADIIT a réellement démarré en Avril 2011**
- **Fin 2011/ Début 2012**

  AGORA2011 Beta

  RETRO2011 Beta

  Clustering2011 Beta

- **Avril 2012**

  AGORA2011 Release v1

  RETRO2011 Release v1

  Clustering2011 Release v1

Google Award renouvelé pour une année supplémentaire!

Bourse Google 2, CESR :
Les outils *Franscriptor* et *Varialog*

Bourse Google 2

- Google 2a: Automatisation de la transformation en version patrimoniale, puis modernisée (par méthode n-grammes et dictionnaires), avec la société Digiscrib (Tours-La Riche)
  - *vifue > vifve > vive CRÉATION DE FRANSCRIPTOR*

- Google 2b: outil de requête à partir de la variation graphique (application de règles linguistiques + dictionnaires), avec le laboratoire FORELL (Poitiers, Marie-Hélène Lay) *CRÉATION DE VARIALOG*
  - *La requête <VIF> donne vifs, vifz, vive, viue, vifues, vifves, vifues, …*

Informatisation des *Catalogues régionaux des incunables.*

MCC / CESR

Informatisation des CRI

- L'objectif du projet est la diffusion en ligne, gratuite, d'un catalogue interopérable, rassemblant l'ensemble des données des CRI et enrichi de liens vers d'autres catalogues ou fac-similés d'incunables en ligne, grâce à nos collaborations européennes.
- Format choisi : UNIMARC-TEI
- Paramétrage d’un SIGB open-source pour la retroconversion des notices.
- Le mois dernier, les données de l’ISTC (Incunabula Short Title Catalogue) ont été envoyées par la British Library afin de récupérer automatiquement certaines métadonnées.

KOHA
(en test)

d - Note sur l'enluminure de l'exemplaire / 394 $a +

e - Note sur la reliure de l'exemplaire / 395 $a +

f - Note sur la reliure / URL - 395 $u +

g - Note sur la reliure / Texte de lien - 395 $z Version numérisée +

h - Note sur les mentions manuscrites / 396 $a +

i - Reproduit comme / Note - 456 $z +

j - Reproduit comme / URL - 456 $u +

k - Reproduit comme / Texte de lien - 456 $b

l - Relié à la suite de / Numéro notice - 482 $0

m - Relié à la suite de / auteur - 482 $a

n - Relié à la suite de / titre - 482 $t

o - Relié à la suite de / titre parallèle - 482 $l

p - Relié à la suite de / lieu publication - 482 $c

q - Relié à la suite de / nom éditeur - 482 $n

r - Relié à la suite de / date de publication - 482 $d

s - Relié à la suite de / première responsabilité - 482 $f

t - Relié à la suite de / responsabilité suivante - 482 $g

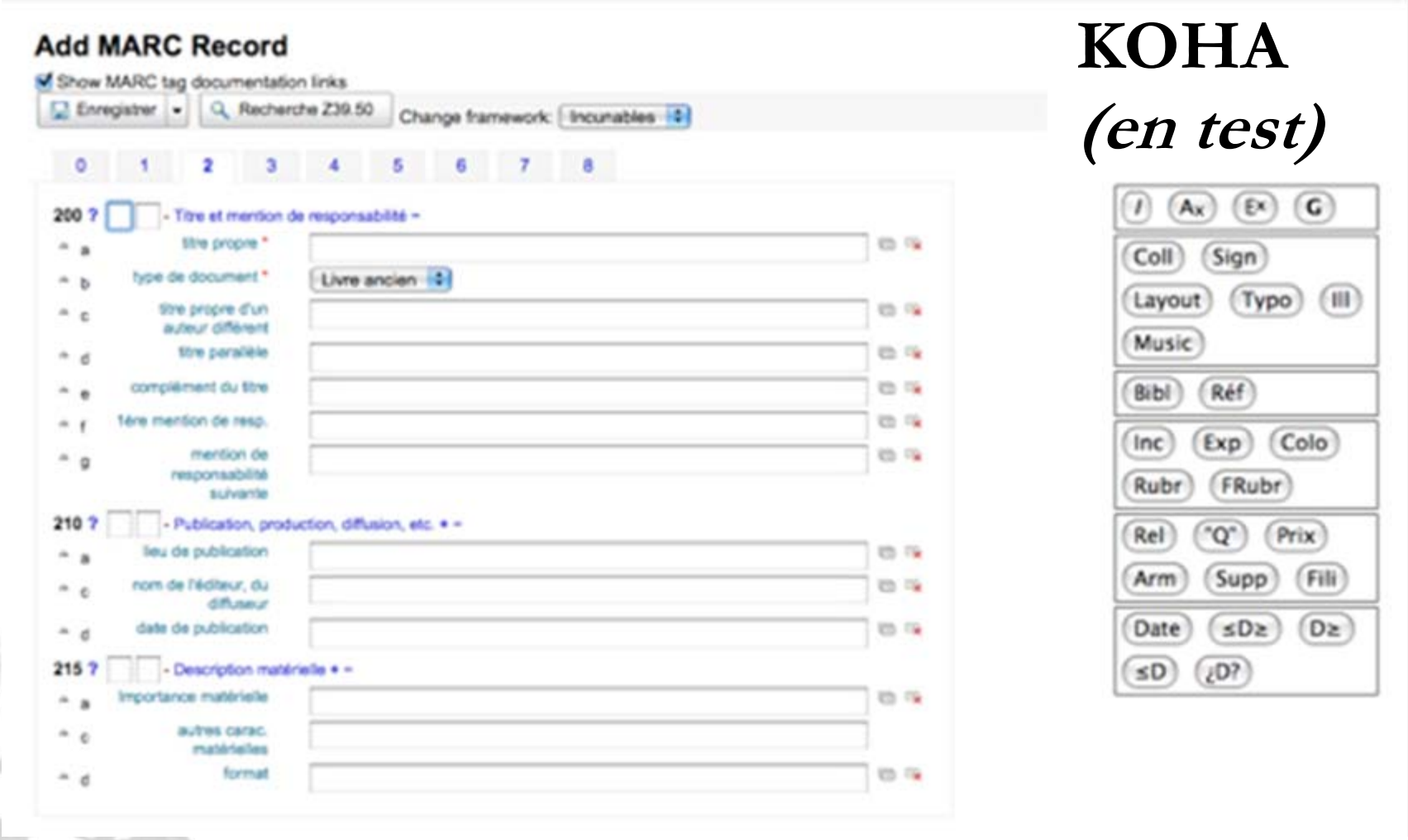
u - Relié à la suite de / URL - 482 $u

v - Note sur la préservation / Opération - 318 $a +

w - Note sur la préservation / Date - 318 $c +

x - Note sur la disparition - 939 $a

Informatisation des CRI



KOHA

(en test)

Nouveaux programmes de recherche 2012

MONLOE « Montaigne à l'œuvre »

- Exemple d'un corpus d'auteur: la « librairie » de Montaigne et le projet « Montaigne à l'œuvre »
  - Partenariat avec l'IRHT (reconnaissance des écritures)
  - Partenariat avec l'Ecole nationale des Chartes (exploitation sémantique)
  - Collaboration avec Archeovision Bordeaux (3D)

- Projet ANR Corpus déposé en octobre 2011
  - Accord de la BnF, de la BM de Bordeaux, de Libourne, de l'Université de Virginie, de Cambridge
  - Transcriptions expertes d'Alain Legros, selon 3 niveaux
- Résultats au printemps 2012 ?


HIERONYMI
VIDAE CREMO-
LVGDVNI APVD SEB.
GRYPHIVM,

Cambridge (10)

BM Bordeaux (30)

Autres (17)
- Musée Condé
- Libourne
- Périgueux
- BU Bordeaux
- University of Virginia
- NL of Scotland
- Bloomington…

« Librairie de Montaigne » virtuelle

BNF (33)

Collection privée, fondations (2)

Nouvelles fonctionnalités : exemple du Lucrèce

1. Pages en jpg avec zoom et téléchargement page à page

2. Nouvelle ergonomie de consultation et présentation des pdf. Trois types de téléchargement au lieu de 2:
   1. PDF couleur sans feuilletage ni OCR brut
   2. PDF binarisé sans OCR
   3. PDF compressé avec OCR brut en texte caché

Lucretius Carus, Titus
De rerum natura libri sex [exemplaire annoté par Montaigne], 1563

http://www.bvh.univ-tours.fr/Consult/index.asp?numfiche=764

Publication : Paris : Gaultier, Philippe
Lyon : Rouillé, Guillaume
Impression : Paris : Gaultier, Philippe
Format : 4°
Collation : [20], 1-559, [1 bl., 4, 4 bl.] p. (sig. $\tilde{a}^4$ e^4 i^2 A-Z^4 Aa-Zz^4 AAa-ZZz^4 AAA^4 $*^2$)

Localisation : Cambridge, Cambridge University Library, The Montaigne Library
Cote : Montaigne_Lucrece1563

Numérisation : Cambridge - Phase One P65+ 2011
Mise en ligne : 27/07/2011

Extrait de la convention établie avec les établissements partenaires pour la numérisation

- Ces établissements autorisent la numérisation des ouvrages dont ils sont dépositaires (fonds d'Etat ou autres) sous réserve du respect des conditions de conservation et de manipulation des documents anciens ou fragiles. Ils en conservent la propriété et le copyright, et les images résultant de la numérisation seront dûment référencées.
- Le travail effectué par les laboratoires étant considéré comme une «œuvre» (numérisation, traitement des images, description des ouvrages, constitution de la base de données, gestion technique et administrative du serveur), il relève aussi du droit de la propriété intellectuelle et toute utilisation ou reproduction est soumise à autorisation.
- Toute utilisation commerciale restera soumise à autorisation particulière demandée par l'éditeur aux établissements détenteurs des droits (que ce soit pour un ouvrage édité sur papier ou une autre base de données).
- Les bases de données sont déposées auprès des services juridiques compétents.

Projet régional (obtenu) : l'APR « ReNom »

Denis Maurel (LI, Tours)

Technopôle ValCoNum : problématiques de numérisation de documents historiques en masse

Mickaël Coustaty (L3i, La Rochelle)

Les Bibliothèques Virtuelles Humanistes

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

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