



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



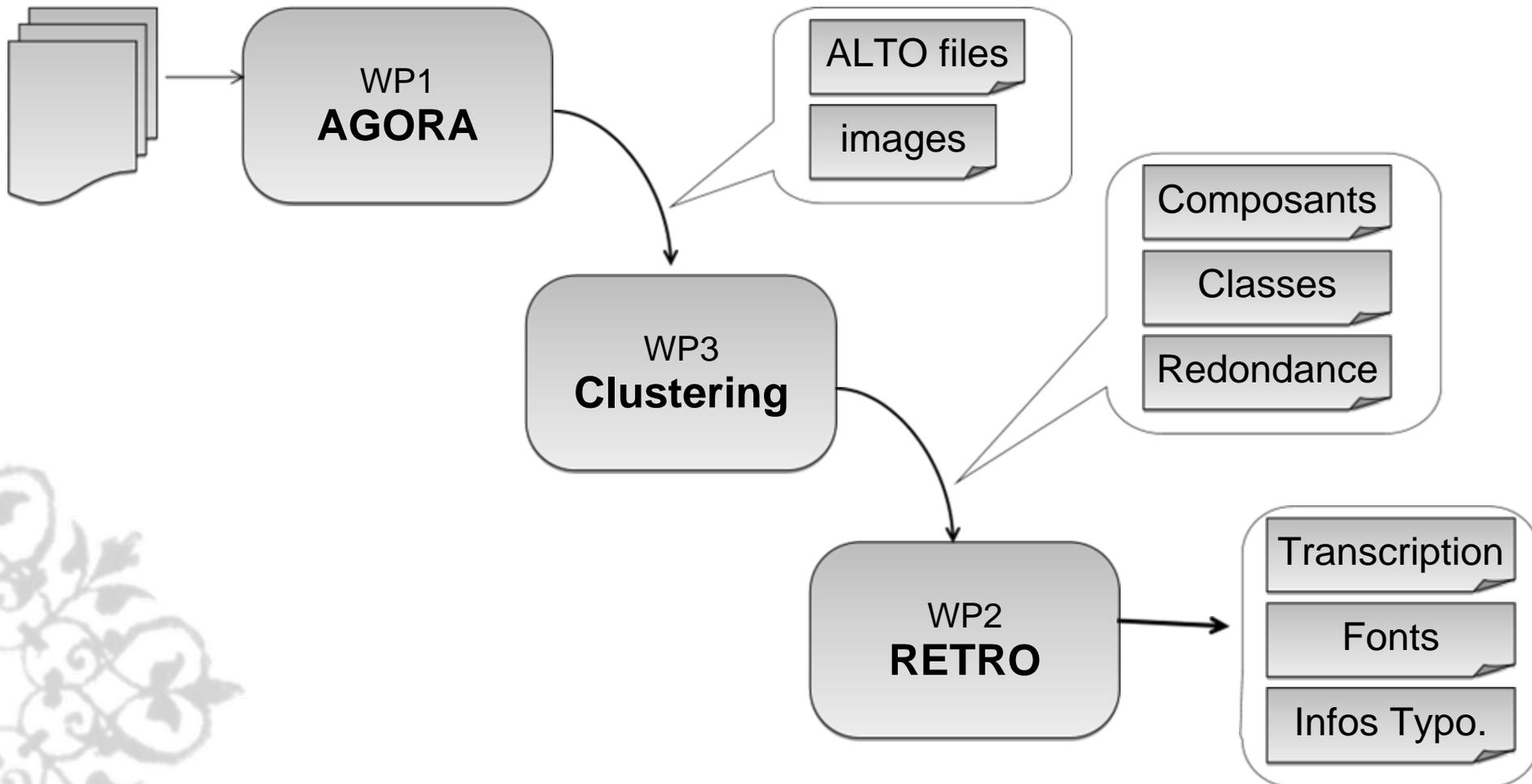
**Bourse Google 1, Laboratoire
d'Informatique : l'OCR RETRO et
le classement des caractères**



Frédéric Rayar (LI, Tours)

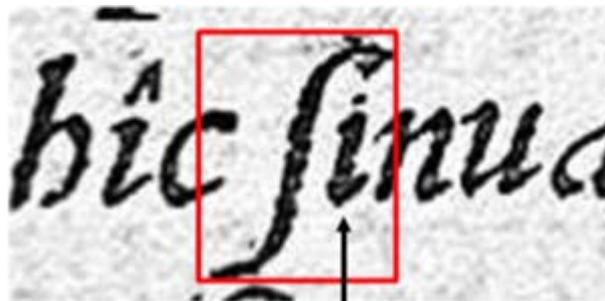
Projet PaRADIIT

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 Renaissance
- WP3 - Clustering:
 - Importante amélioration en cours de réalisation.
 - Travail réalisé par Thierry Brouard



These pixels must not
be taken into account
when computing features



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

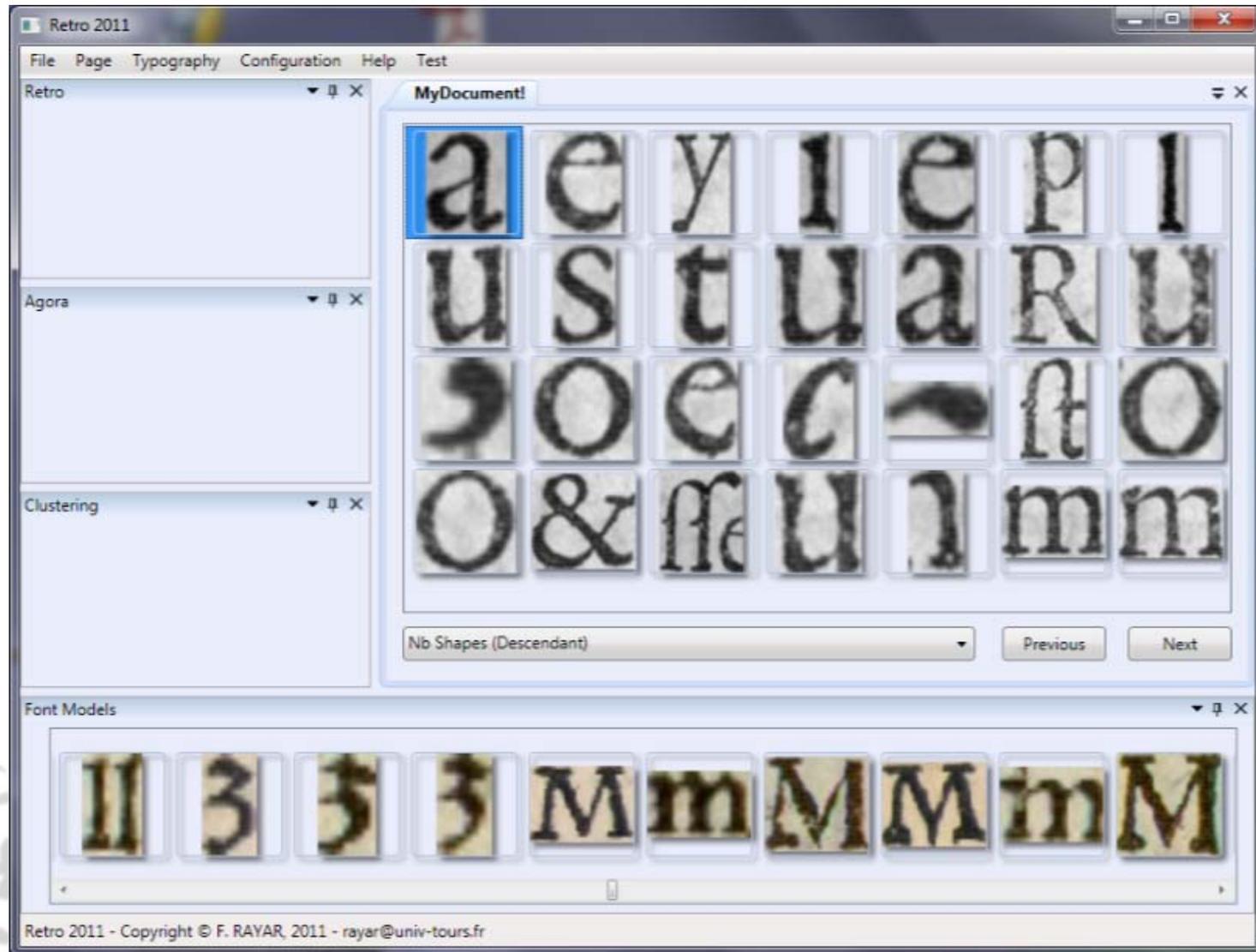
Font Family
Creation

Body Height

Stroke Width?

Roman/Italic?

RETRO



RETRO 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

The screenshot displays the RETRO 2008 software interface. The main window shows a document with red bounding boxes around characters, indicating the software's character-level indexing capabilities. The document text is in Latin, starting with a large decorative initial 'E' followed by a paragraph of text. The interface includes a menu bar (File, Clusters, EoC, Transcription, Results, Zoom, Configuration, Display, Help) and a sidebar with statistics (Project, Clusters, EoC, Result, Number of clusters, Number of shapes, Tagged clusters, Percentage of transcription). The bottom status bar shows the file path: C:\Documents and Settings\Renil\Bureau\Retro_Typo2\Patterns.

Project: ?
Clusters: ?
EoC: ?
Result: ?
Number of clusters:
Number of shapes:
Tagged clusters:
Percentage of transcription:

0410186201_1937_0003.JPG

Ruerendissimo in Christo Patri Eustathio
Croyo, Antrebatensium Episcopo,
Franciscus Syluius S.

Expectabam occasionem mihi dari, ut
aliquid laborum meorum tibi nuncuz
parem. Eam mihi nuper oblatam esse
vehementer laetor. Quum enim mente
superiore enarrarem discipulis meis
Orationem M. Ciceronis pro A. Licin-
nio Archia Poeta, visum est mihi com-
mode sumi argumētū, ut tibi dedicare
Comētarios, quos iēā orationē scriber-
remus. Ut enim puerum Ciceronem Archias docuit, ita tu
me preceptore puer usus es. Ut meritum munus preceptor-
ti Cicero hac defensione perfoluit, ita tibi discipulo munus
te non indignum offerre videor, quo tibi magis diuolida fiet
hac Ciceronis oratio, ex qua cognoscas quantam voluptatē
atq; utilitatem capere possis ex his literarum studiis, quae tu

Patterns
Label: C:\Documents and Settings\Renil\Bureau\Retro_Typo2\Patterns

RETRO : module d'indexation de caractères

Add Typographic Model

Famille
 Romain
 Italique

Sous-Famille
 Humane
 Garalde
 Capitale droite
 Capitale penchée

Graisse
 Gras
 Maigre

Corps du Texte
Saint-augustin (SA)

Transcription: Code Unicode:

Référence bibliographique :

Description :

 (39,42)

Ok Cancel

RETRO : module de mesure de corps de caractères

Body Height Measure

C:\Documents and Settings\Remi\Bureau\TYPOGRAPHIE\bases dimz

errantem p...
 4 da gnali noie vocari: specialivero Martē dici. Grēci au
 5 tē vīq; adeo oēs sī de aeluro senserunt: vt nobis merito li
 6 ceat gloriari q̄ p̄mi oīm tā vulgatu in lingua eorū erro
 7 rē dephēderim? Est igit̄ Felis: quē vulgo Catū noiam?
 8 nec meo quidē iudicio inepte. Veteres enī Catū astutū
 9 dicebāt: & qđ nos in p̄sentia Catum. A q̄ Catones p̄mo
 10 volūt appellatos. CIC. Sed hō cat? nunq̄ terminat nec
 11 magnitudinis: nec diuturnitat̄ modū. IDē. Quis igit̄
 12 prudētē (& vt ita dicā) Catū nō ex ip̄i? habitu: s; ex ali
 13 qua re externa iudicet. Astutissimū autē cēnā aial p̄clat
 14 ostendit. PLY. Feles ingt quo silētio: q̄ lembus vestigiis
 15 obrepūt aibus: q̄ multo speculatu in musculos exiliūt.
 16 Excremēta sua obruūt terra. itelligētes odorē illū idicē
 17 sui esse. Sed infinitis pene testimonis p̄bari pōt. Felem
 18 apud nos. Apud grācos Aelurō id aiale: qđ dicim?
 19 Catū. CI. Ipsi q̄rū vanitas ridef̄. Aegyptiū nullā beluā
 20 mfi ob aliquā vtilitatē p̄secrauerūt: velut Ibim qđ maxi
 21 mā vim serpentū cōficiat. Crocodilū qđ terrore arceat
 22 latrones. Felē qđ pelle ei? scuta cōtegāt. Ichneumones
 23 qđ nō sinat serpentes maximā ei? regionis pestē augeti
 24 IDē. Aegyptiorū morē q̄s ignoret: q̄s imbutā mētes
 25 prauitatū errorib; q̄uis carnificinā potū? subirēt q̄ Ibim
 26 aut aspidē: aut felē: aut canē: aut crocodilū violent. HE
 27 Rodot? ὄψορκαίη ἠσλαυίοντιη ἡσθεία πρραυμά

4 sagacitatē. VI
 5 Canis. Dicūf
 6 tiū filii. VIR.
 7 quēs. PLY. Si
 8 domos pepere
 9 frāmitu abiic
 10 ostētū faciūt. J
 11 sces etiā feras:
 12 ceri pilces: s; f
 13 serpētū filii C
 14 tiū Catellus:
 15 tulinā carnē el
 16 Catulos lactēt
 17 ad placida nu
 18 aliqñ p̄ cane at
 19 ticos atq; laga
 20 ga nō lōge abe
 21 cū rufā canes
 22 ritatē pducere
 23 tū maris appe
 24 tiliētē. LAbe
 25 Hinc Catuliti
 26 ra aperit̄ vt N
 27 cat̄ gestiētē na
 28 fctis. Eft̄ at̄

Select Frontier

[20] (in mm) 78,83

Consider Line Space 0

Top (in pixel) 67,90979

Bottom (in pixel) 365,8421

[x] (in mm) 1,59

Top (in pixel) 241,8421

Bottom (in pixel) 247,8421

[:] (in mm) 2,65

Top (in pixel) 314,8421

Bottom (in pixel) 324,8421

Body Height Estimation

[20] Cicero (78)

[x] Philosophie (1)

[:] Cicero (2)

Reset Export

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 PaRADIIIT 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 VARIOLOG**
 - *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'alumine 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 \$2 *

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

Add MARC Record

Show MARC tag documentation links

Enregistrer

Recherche Z39.50

Change framework: Incunables

0 1 2 3 4 5 6 7 8

200 ? - Titre et mention de responsabilité -

- a	titre propre *	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- b	type de document *	Livre ancien	<input type="checkbox"/>	<input type="checkbox"/>
- c	titre propre d'un auteur différent	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- d	titre parallèle	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- e	complément du titre	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- f	titre mention de resp.	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- g	mention de responsabilité suivante	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>

210 ? - Publication, production, diffusion, etc. -

- a	lieu de publication	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- c	nom de l'éditeur, du diffuseur	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- d	date de publication	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>

215 ? - Description matérielle -

- a	importance matérielle	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- c	autres caractéristiques matérielles	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>
- d	format	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>

KOHA

(*en test*)

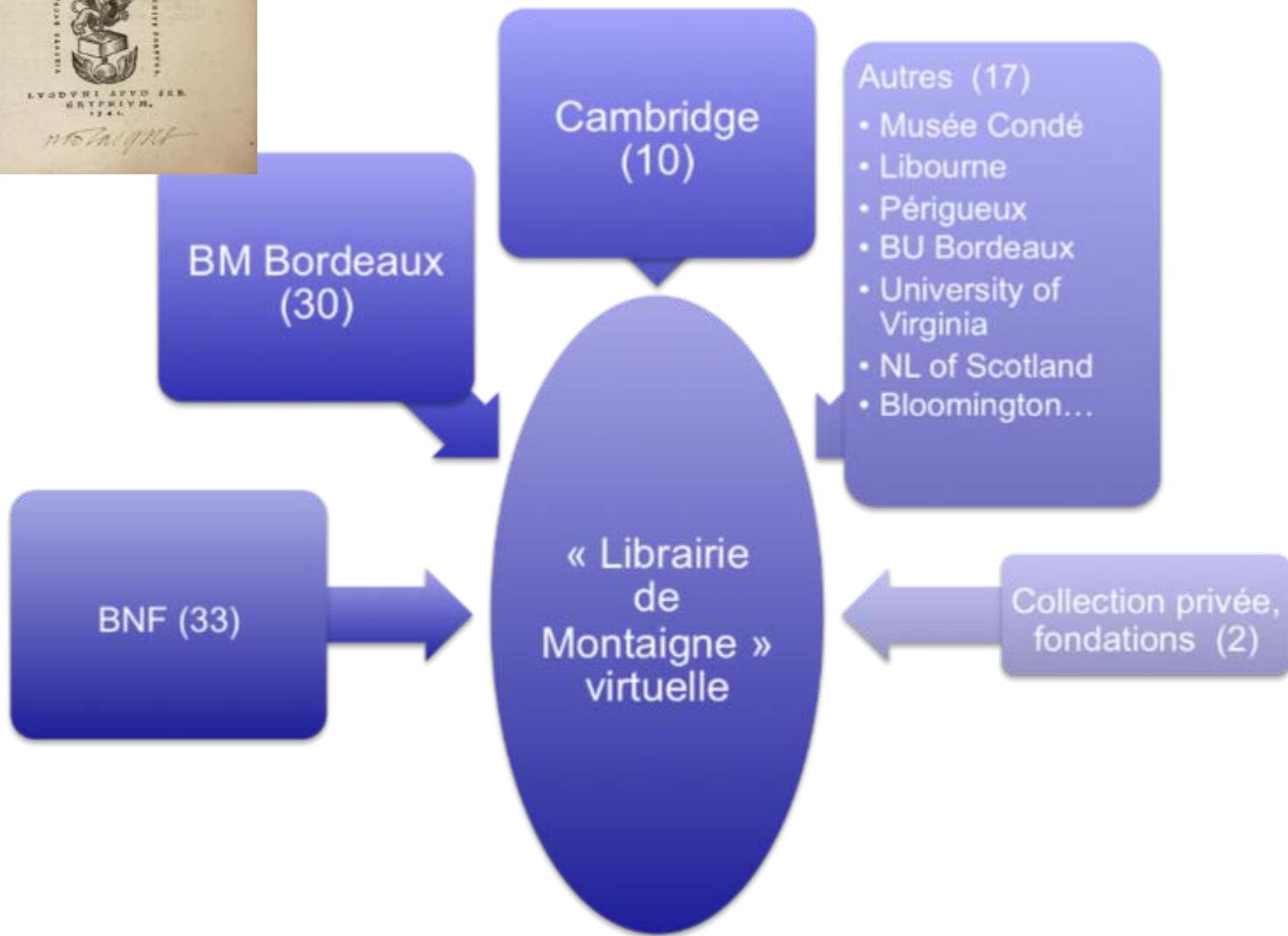
/	Ax	Ex	G
Coll	Sign		
Layout	Typo	Ill	
Music			
Bibl	Réf		
Inc	Exp	Colo	
Rubr	FRubr		
Rel	"Q"	Prix	
Arm	Supp	Fili	
Date	≤D≥	D≥	
≤D	¿D?		



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 ?



Nouvelles fonctionnalités : exemple du *Lucrece*

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



Bibliothèques Virtuelles Humanistes

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

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. ã⁴ e⁴ i² A-Z⁴ Aa-Zz⁴ AAa-ZZz⁴ AAA⁴ *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.



Tout contenu des [Bibliothèques Virtuelles Humanistes](#) (hors les originaux des images soumis à convention) est mis à disposition selon les termes de la [licence Creative Commons Paternité-Pas d'Utilisation Commerciale-Pas de Modification 2.0 France](#).



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



Denis Maurel (LI, Tours)



Technopôle ValCoNum :
problématiques de numérisation de
documents historiques en masse

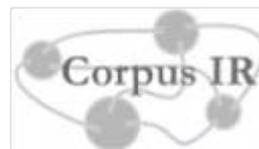


Mickaël Coustaty (L3i, La Rochelle)



Les Bibliothèques Virtuelles Humanistes

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



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