



Google Award Project

Pattern Redundancy Analysis for Document Image Indexation and Transcription

### **RETRO 2011 Progress**

RAYAR Frédéric 2012-01-06





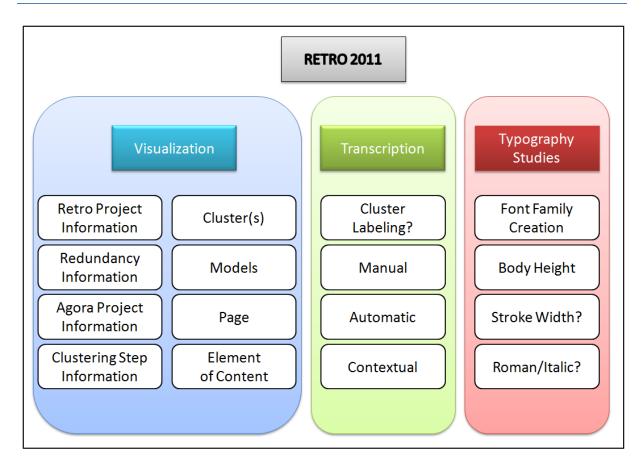
# Summary

Main windows	. 5
Navigation in Cluster	. 6
Navigation in page	. 6





## Introduction



Between April and July, several tasks have been done in the RETRO WP.

First, software conception related issues:

- Analysis of existing tools, libraries and source code (Retro2008, AForge.NET)
- Redaction of a Software Requirements Specification Document
- Validation of a enhanced RETRO2011 Design Draft

But also preliminary innovative work related to the *Transcription* Group:

- Research of OCROpus and Tesseract software as possible OCR Engine for RETRO2011
- Discussion about Super Resolution methods for text images regarding our application field

Then, we focused more on the Visualization Group, and made tools for typography studies regarding the need of end-users.

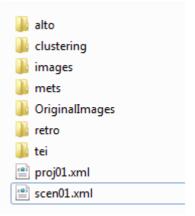




## Input files

A lot of effort has been made in order to have efficient and normalized input/output files for each WP of the project. Therefore one the first task was to gather the output files generated by the WPs AGORA and Clustering and be able to process them.

The following image presents the needed input for a RETRO project.



-	alto/	Alto description files generated by AGORA
		Images of all extracted components (block, line, letters)
_	clustering/	Clustering xml files Clustering algorithms description files
_	images/	Images of the AGORA project with normalized names
_	mets/	Mets description files generated by AGORA
_	OriginalImages/	Original set of images used for the AGORA project
_	tei/	Tei description files generated by AGORA
_	proj01.xml	AGORA project description file
_	scen01.xml	Description of the scenario used for the AGORA project





## Visualization

The following screenshots present the interface of the current demo version.

#### Main windows

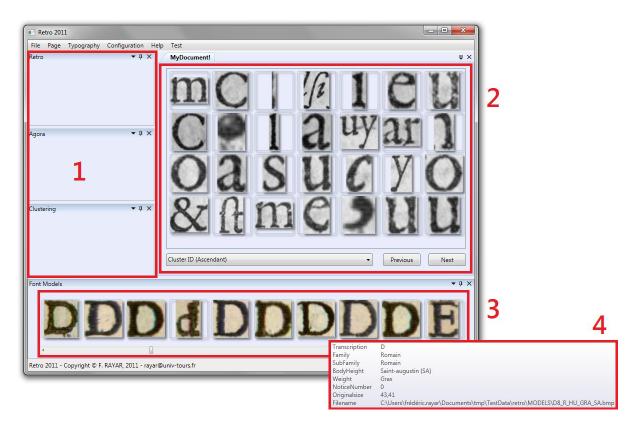


Figure 1. Main windows visualization

- **1** Panel where information summary regarding AGORA, Clustering and RETRO attributes of this project.
- Visualization of the Clusters, the image of a representative each cluster is displayed.
   A numbering is done, therefore both *Previous* and *Next* buttons are available.
   The possibility to sort the clusters regarding several criteria is also implemented.
- **3** All the shapes that are given as Font Model are displayed (currently 1500), possibility to navigate through the whole list quickly has been implemented.
- 4 Tooltips are available for each cluster and font model summering important information.





#### **Navigation in Cluster**

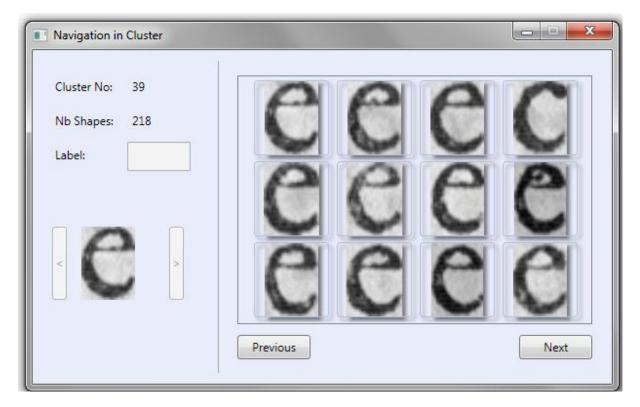


Figure 2. Cluster visualization

A click on a Cluster opens a new window to navigate inside the cluster.

Main information of this cluster are displayed and all his shapes.

As for the clusters, a numbering of the shapes is done.

#### Navigation in page

The current development involves the visualization of a page with augmented information extracted from AGORA, e.g. Extracted Element of content regarding a certain granularity (TextBlock, TextLine, and String).

Selection of an Element of Content of the page to view his attributes, and eventually its transcription will soon be possible.





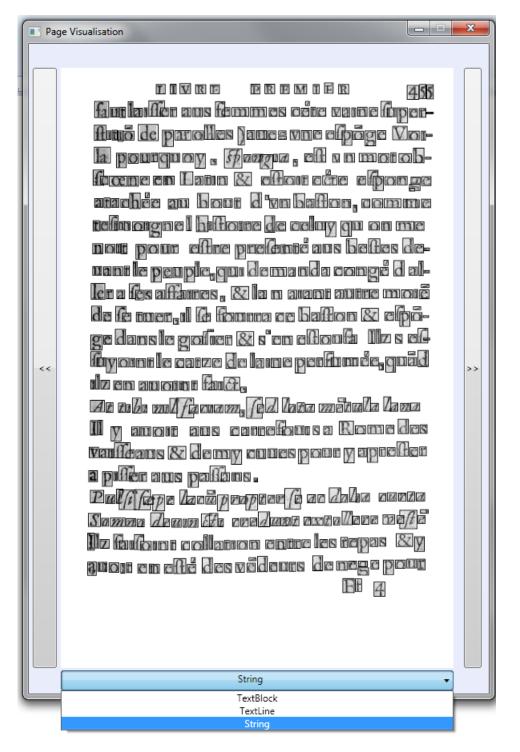


Figure 3. Visualization of a page (not fully implemented)





# Typography studies

A first tool for measuring body height of a font has been implemented.

Tests have been done with ground truth data (*"French Renaissance Printing Types: A Conspectus"* by Vervliet, 2010)

<ul> <li>chi &amp; addicorços vna, per Erdianum, que comme inis Grece, q Romane feribere ma lucria, mitrari definêt, fi ppêderint primă illis tipis Greci ilermonis mentis în eafdem &amp; nuncupatoria autoris epifiola.</li> <li>In epifiolă ad Romanos vnă.</li> <li>In epifiolă ad Romanos vnă.</li> <li>A d Corinthios duas, &amp; ci sa styrară côuicijs fit exagita- ta. Si gde apă Luciliă Sceuola ris des Atbuită, vt plus equo quâ a i deli tijs fuifie Grecanicas literas, vt ca res Satyrară côuicijs fit exagita- ta. Si gde apă Luciliă Sceuola ris des Atbuită, vt plus equo quâ a i deli de apă Luciliă Sceuola ris des Atbuită, vt plus equo quâ a i deli de apă Luciliă Sceuola ris des Atbuită, vt plus equo quâ a i de apă Luciliă Sceuola ris des Atbuită, vt plus equo quâ a a de fierbeneis, et a si de apă Luciliă sceuola ris de apă Luciliă sceuola ris de ca chere poffe Grecă vtbă.</li> <li>A d Thefialonicenfes duas</li> <li>A d Thiemonem vrăm</li> <li>In canonicas Petri duas</li> <li>In lude vram</li> <li>In ranonicas Petri duas</li> <li>In lude vram</li> <li>In vram lacobi, &amp; ci conto de con</li></ul>		Omnes Apoftoll PAVLI gets	LVS apoltor TEAU	Select Frontier	2
<ul> <li>mum Roterodamum, que commē tarij vice effe pofint, cum argu- mentis in eafdem &amp; nuncupatoria autoris epiftola.</li> <li>In epiftolā ad Romanos vnā.</li> <li>Mad Corinthios duas, &amp;</li> <li>Ad Corinthios duas, &amp;</li> <li>Ad Galatas vnam</li> <li>Ad Galatas vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Thiemonem vnām</li> <li>In canonicas P etti duas</li> <li>In ludę vnam</li> <li>In ludę vnam</li> <li>In runam lacobi, &amp;</li> <li>Consider Line Space</li> <li>Consider Line Space</li> <li>Top (in pixel)</li> <li>Bottom (in pixel)</li> </ul>		manas epiftolas, vna cum Canoni	nis Grece, G Romane feribere ma	[20] (in mm)	51,9
<ul> <li>mentis in eaſdem &amp; nuncupatoria autoris epífola.</li> <li>In epífolā ad Romanos vnā.</li> <li>Ad Corinthios duas, &amp; Ad Galatas vnam</li> <li>Ad Ephefios vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Theflalonicenfes duas</li> <li>Ad Timotheum duas</li> <li>Ad Thimotheum duas</li> <li>Ad Philemonem vnām</li> <li>In canonicas Petri duas</li> <li>In ludę vnam</li> <li>In vnam lacobi, &amp; Conta da conta da</li></ul>		mum Roterodamum, que comme	luerit, mirari definet, fi ppederint	Consider Line Space	
autoris epiftola. Nomant patebat i perita. Deinde Romanis iplis adeo quodă i deli tijs fuific Grecanicas literas, vi car res Satyrath couicijs fit exagita- ta. Sigdē apd Luciliŭ Seguola ri- dēs Albutiŭ, vi plus equo φιλιλ Ad Coloffenfes vnam Ad Timotheum duas Ad Timotheum duas Ad Timotheum duas Ad Thinotheum duas Ad Philemonem vnam In canonicas Petri duas In lude vnam In vnam lacobi, & Comant patebat i perita. Deinde Romanis iplis adeo quodă i deli tijs fuific Grecanicas literas, vi car tes Satyrath couicijs fit exagita- ta. Sigdē apd Luciliŭ Seguola ri- dēs Albutiŭ, vi plus equo φιλιλ λavœ, xœijs ingt Ptor Athenis, & apd Iuuenalē itomachā gipiā, negās fe ferre poffe Grecā vibē. Quāĝi ita Grece focur9 ē, vi a me te Grecis no latis queat itelligi, ob pafits admixtā Hebrei fermois prietatē. Qd fi maxie Greca. ac pure Greca fuifictoratis fupeffet le. Comant patebat iperitatis fupeffet le. (20) Deit tentto (51)		mentis in ealdem & nuncupatoria	vlum pene ta late patuiffe, g late	Top (in pixel)	292,919
<ul> <li>In epiftolä ad Romanos vnä.</li> <li>Ad Corinthios duas, &amp;</li> <li>Ad Galatas vnam</li> <li>Ad Galatas vnam</li> <li>Ad Galatas vnam</li> <li>Ad Bhelios vnam</li> <li>Ad Philippenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Timotheum duas</li> <li>Ad Timotheum duas</li> <li>Ad Timotheum duas</li> <li>Ad Thinotheum duas</li> <li>Ad Thinot</li></ul>		autoris epístola.	Romanii patebat i perifi. Deinde Romanis ipiis adeo quodă i deli	Bottom (in pixel)	489,335
<ul> <li>Ad Galatas vnam</li> <li>Ad Galatas vnam</li> <li>Ad Ephefios vnam</li> <li>Ad Philippenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Theffalonicenfes duas</li> <li>Ad Timotheum duas</li> <li>Ad Timotheum duas</li> <li>Ad Tinum vnam</li> <li>Ad Philemonem vnam</li> <li>In canonicas Petri duas</li> <li>In ludę vnam</li> <li>In nvam lacobi, &amp;</li> <li>Coloffenfes terre diffecter tudub?</li> <li>Ad Philemonem vnam</li> <li>Ad Coloffenfes vnam</li> <li>Ad Philemonem vnam</li> <li>Ad Philem</li></ul>		In epistola ad Romanos vna.	tijs fuisse Grecanicas literas, vt ea	[x] (in mm)	
Ad Ephefios vnam       des Albutin, vi plus equo φ1/λ/λ         Ad Philippenfes vnam       λκνα, καιμ, ingt Pior Athenis,         Ad Coloffenfes vnam       λκνα, καιμ, ingt Pior Athenis,         Ad Coloffenfes vnam       α and I ueunale itomachał g(pia, negás le ferre poffe Gręca vrbē,         Ad Timotheum duas       Quáği ita Gręce locul <sup>9</sup> č, vi a me re Gręcis nö fatis queat Itelligi, ob pafilja admixta Hebrei fermöis         Ad Philemonem vnam       ob pafilja admixta Hebrei fermöis         Prietatē. Qd fi maxie Gręca. ac pure Gręca fuiffet oratio, multu       Bottom (in pixel)         Body Height Estimation       Body Height Estimation		Ad Galatas vnam	ta. Sigde apd Lucilin Sceuolaria	🔘 Top (in pixel)	
Ad Coloffenfes vnam       & apd Iuuenalë itomachaž glpiž, negis fë ferre polfe Grçça vbë, Quă gi ia Grçce locur <sup>9</sup> ë, vt a me re Grçcis no fatis queat Itelligi, ob palits admixtă Hebrei fermóis prietatë.Qd fi maxie Grçca ac pure Grçca fuillet oratio , multă th adhue difficultatis fupeffet les       [:] (in mm)         Body Height Estimation       Imachaž glpiž, negis fë ferre polfe Grçça vbë, Quă gi ia Grçce locur <sup>9</sup> ë, vt a me re Grçcis no fatis queat Itelligi, ob palits admixtă Hebrei fermóis prietatë.Qd fi maxie Grçca ac pure Grçca fuillet oratio , multă th adhue difficultatis fupeffet les       Imachaž glpiž, Imachaž glpiž, Body Height Estimation		Ad Ephefios vnam	des Albutiu, vt plus equo 912 22	Bottom (in pixel)	
Ad Timotheum duas       Quã gi ita Grçce locut <sup>9</sup> č, vt a me te Grçcis no fatis queat Itelligi, ob pafiis admixtă Hebrei fermois       Image: Comparison of the state of the stat		Ad Coloffenfes vnam	& apd Iuuenale itomachat glpia,	[:] (in mm)	
Ad Philemonem vnam     Ob pafijs admixtā Hebrei fermóis     prietatē.Qd fi maxīe Gręca.ac     pure Gręca fuiffet oratito ,multū     in ude vnam     In vnam Iacobi,&     fori adhue difficultatis fugeffet les     fori adhue difficultatis fugeffet les     fori adhue difficultatis fugeffet les		Ad Timotheum duas	Quag ita Grece locut? e, vt a me	Top (in pixel)	
In canonicas Petti duas     prietatē.Qd fi maxie Gręca.ac     pure Gręca fuifictoratio _multů     in ude vnam     In vnam lacobi,&     fori domni a     fori domni a	-	Ad Titum vnam Ad Philemonem vnam	ob paffiz admixta Hebrei fermois	Bottom (in pixel)	
cauit veri? g explicuit videlicet [] - Not Defined -		<ul> <li>In canonicas Petri duas</li> <li>In Iudę vzam</li> <li>In vnam Iacobi, &amp;</li> <li>Treis Ioannis, &amp;</li> </ul>	pure Gręca fuisse oratio, muliů tň adhue difficultatis supesse i est thori, qđ cũ scriberet rudib <sup>9</sup> adhue & nup X po înstiatis, mysteria qa dã magis atrigit q tractauit & idi cauit veri <sup>9</sup> q explicuit, videlicet	[20] Petit-texte (51) [x] - Not Defined -	3

Figure 4. Body height measurement tool

- 1 Selected page
- 2 Selection of the, and associated values (pixels and mm)
- **3** Estimated Body height designation
- 4 Possibility to export computed information in xml for further use and studies.