Google Award Project

Pattern Redundancy Analysis for Document Image Indexation and Transcription

---

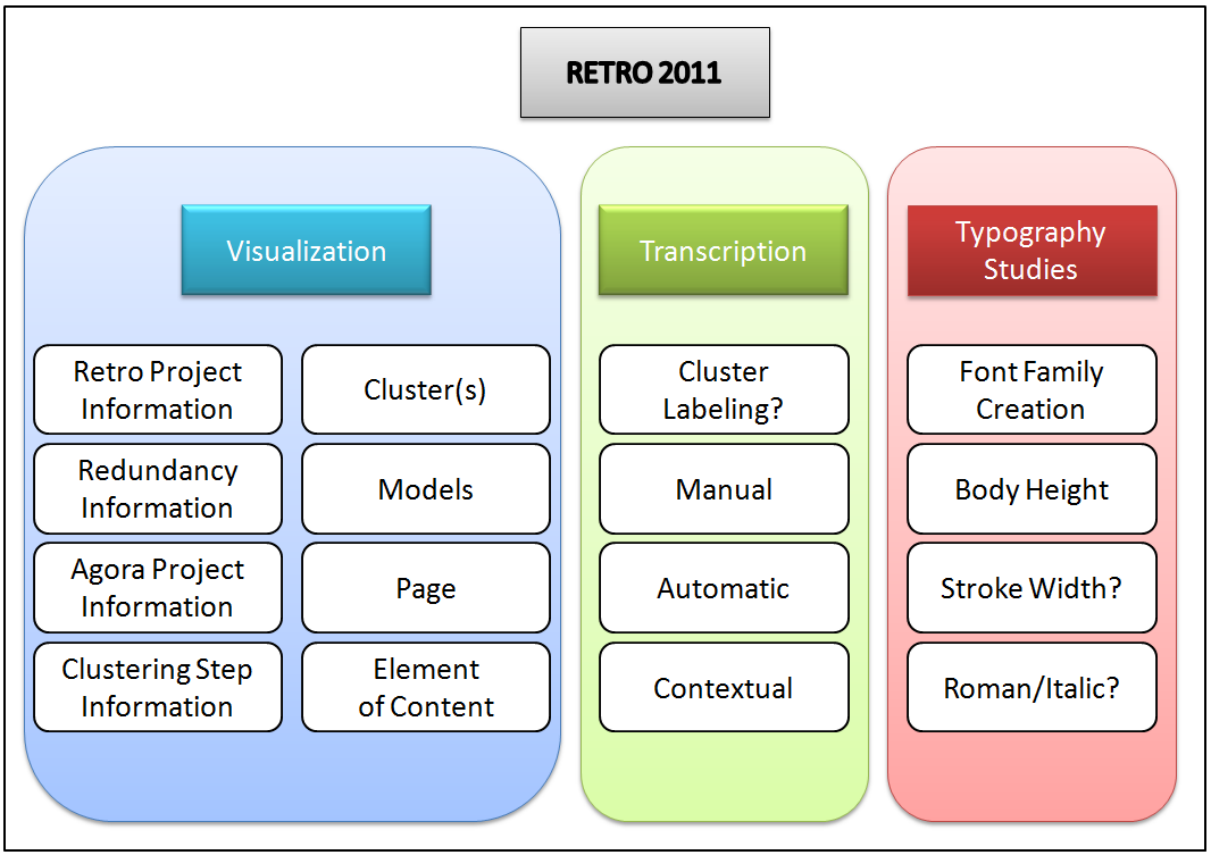
RETRO 2011 Progress

RAYAR Frédéric

2012-01-06

Summary

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:

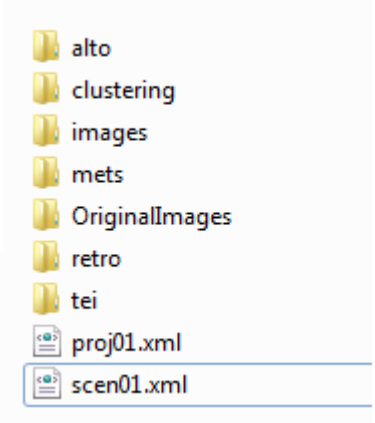
- Research of *OCROpus* and *Tesseract* software as possible OCR Engine for RETRO2011
- Discussion about Super Resolution methods for text images regarding our application field

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

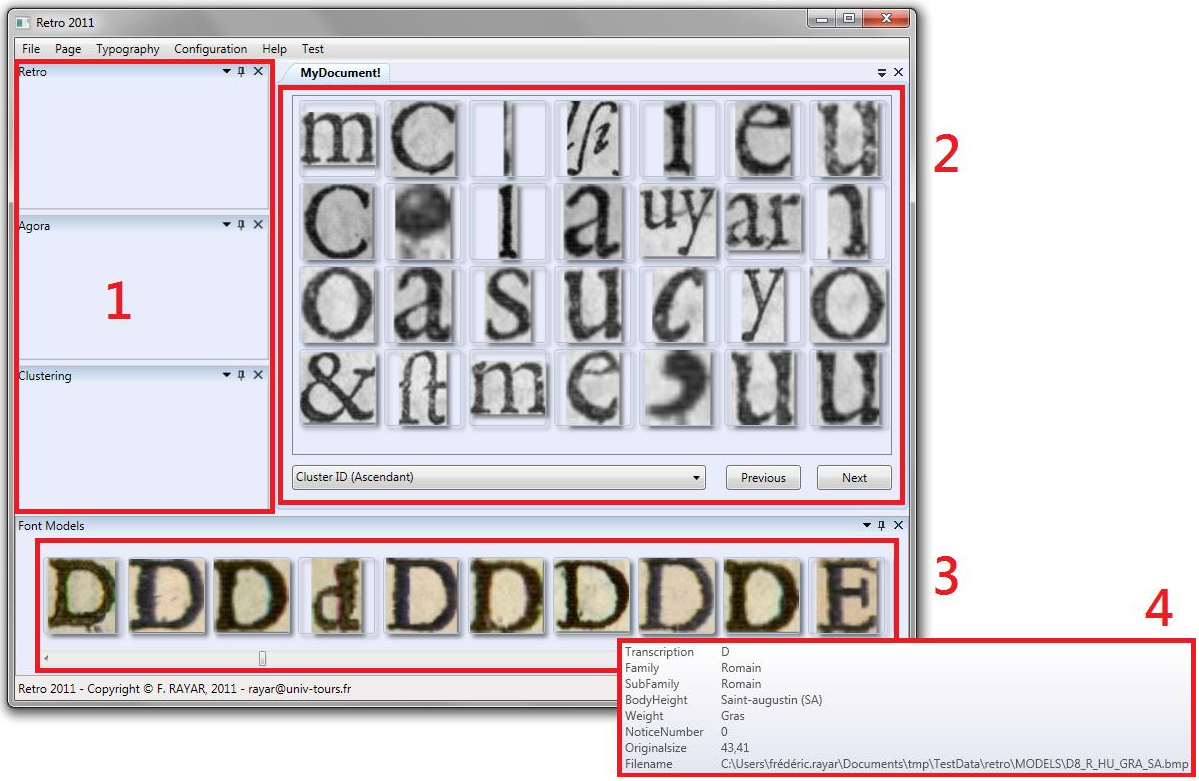


Figure 1. Main windows visualization

1 Panel where information summary regarding AGORA, Clustering and RETRO attributes of this project.

2 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

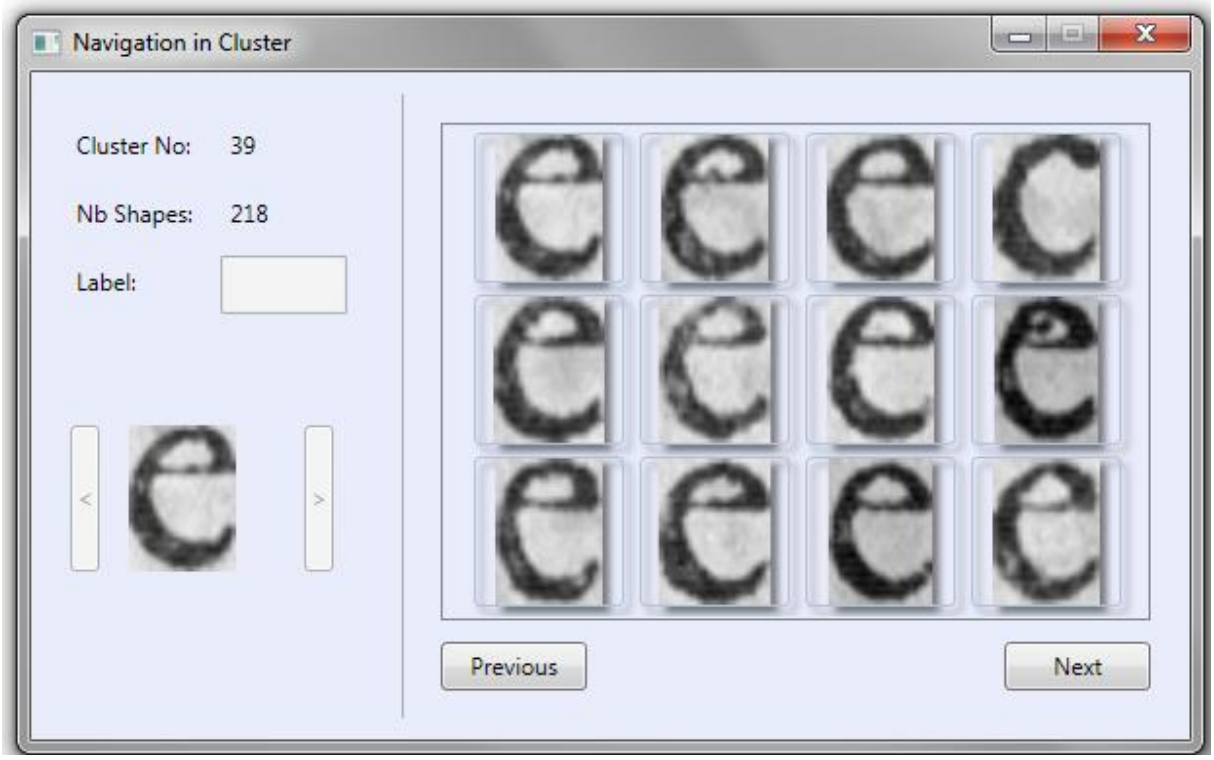


Figure 2. Cluster visualization

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.

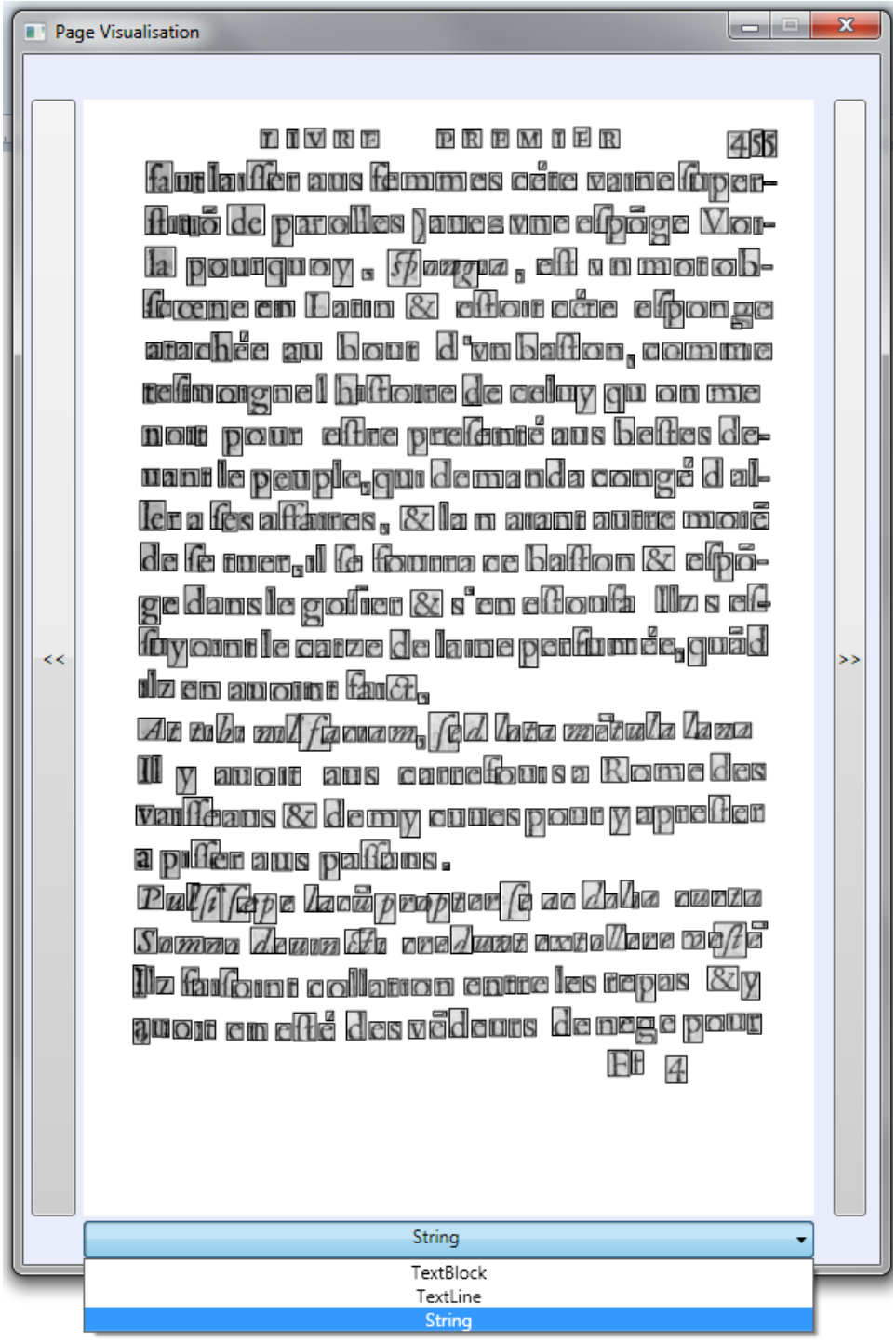


Figure 3. Visualization of a page (not fully implemented)

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)

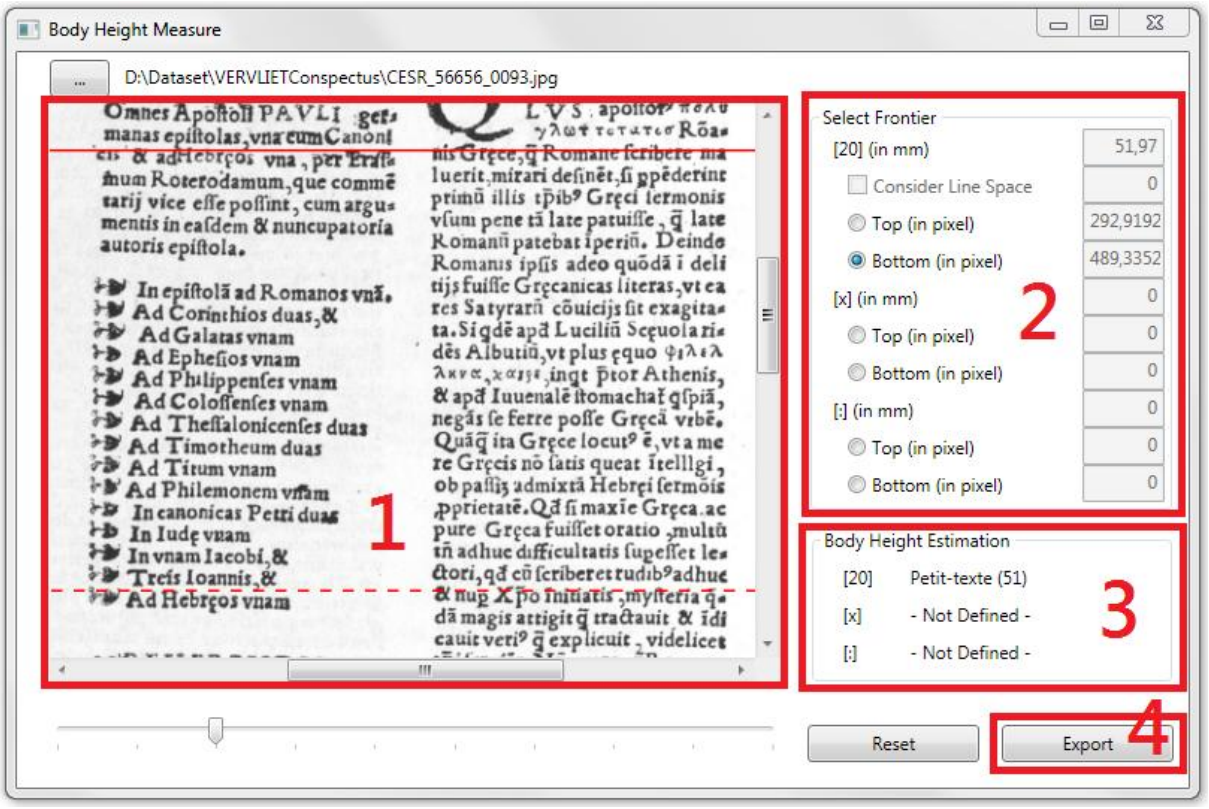


Figure 4. Body height measurement tool

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.