

# RETRO 2014 User Guide

---

*Version 4.0*

- Clustering
- Visualization
- Modification
- Transcription
- Typography Study

Mai 25, 2014

---



---

## How to Contact Us:



Jean-Yves RAMEL  
Frédéric RAYAR

([ramel@univ-tours.fr](mailto:ramel@univ-tours.fr))  
([rayar@univ-tours.fr](mailto:rayar@univ-tours.fr))



Laboratoire d'Informatique  
Equipe Reconnaissance des formes et analyse d'images  
64, avenue Jean Portalis  
37200 – Tours  
France



For more contact information, please refer to the PaRADIIT project website  
<https://sites.google.com/site/paradiitproject/>

## Licence

RETRO 2012

Copyright © RFAI, LI Tours, 2011-2012










This program is free software: you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation, either version 3 of the License.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

## Partners

The authors would like to thank all the members of the BVH-CESR for their collaboration. This work has been supported by the Google Digital Humanities research Awards given to the Computer Science Laboratory of Tours (RFAI team).

	<a href="http://international.univ-tours.fr/">http://international.univ-tours.fr/</a>
	<a href="http://polytech.univ-tours.fr/">http://polytech.univ-tours.fr/</a>
	<a href="http://www.li.univ-tours.fr/">http://www.li.univ-tours.fr/</a>
	<a href="http://www.rfai.li.univ-tours.fr/">http://www.rfai.li.univ-tours.fr/</a>
	<a href="http://www.google.fr/intl/en/about/">http://www.google.fr/intl/en/about/</a>
	<a href="http://cesr.univ-tours.fr/">http://cesr.univ-tours.fr/</a>
	<a href="http://www.bvh.univ-tours.fr/">http://www.bvh.univ-tours.fr/</a>

# Content

---

CONTENT .....	5
<b>PART I - PROJECT MANAGEMENT .....</b>	<b>6</b>
INPUT DATA .....	7
NEW PROJECT .....	8
<b>PART II - CLUSTERING.....</b>	<b>9</b>
CLUSTERING PROCESS.....	10
LOAD CLUSTER .....	12
GENERATE STATS .....	12
CLUSTER TO MODEL.....	13
<b>PART III - VISUALIZATION .....</b>	<b>14</b>
CLUSTERS VISUALIZATION .....	15
CLUSTER NAVIGATION .....	17
PAGE NAVIGATION.....	19
ILLUSTRATION NAVIGATION .....	22
<b>PART VI - MODIFICATION .....</b>	<b>23</b>
CLUSTERS MODIFICATION .....	24
<b>PART V - TRANSCRIPTION.....</b>	<b>29</b>
MANUAL TRANSCRIPTION (I) .....	31
MANUAL TRANSCRIPTION (II) .....	32
AUTOMATIC TRANSCRIPTION .....	34
RESULTS EXPORTATION .....	36
TRANSCRIPTION EXPORTATION .....	37
<b>PART VI - TYPOGRAPHY STUDY.....</b>	<b>38</b>
FONT MODEL CREATION TOOL .....	39
VIEW FONT FAMILY TOOL.....	45
BODY HEIGHT MEASUREMENT TOOL .....	46
<b>PART VII - FAQ.....</b>	<b>48</b>
<b>Part VIII - GLOSSARY .....</b>	<b>50</b>

---

# PART I

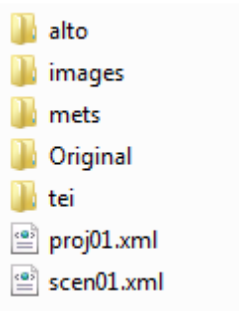
## PROJECT MANAGEMENT

---

# Input Data

---

The following image presents the needed input architecture for a RETRO project. All this architecture is normally build by using AGORA software.



- **alto/** Alto description files generated by AGORA  
Images of all extracted components (block, line, letters)
- **images/** Images of the AGORA project with normalized names
- **mets/** Mets description files generated by AGORA
- **Original/** 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

RETRO will then create *clustering/* and *retro/* folders.

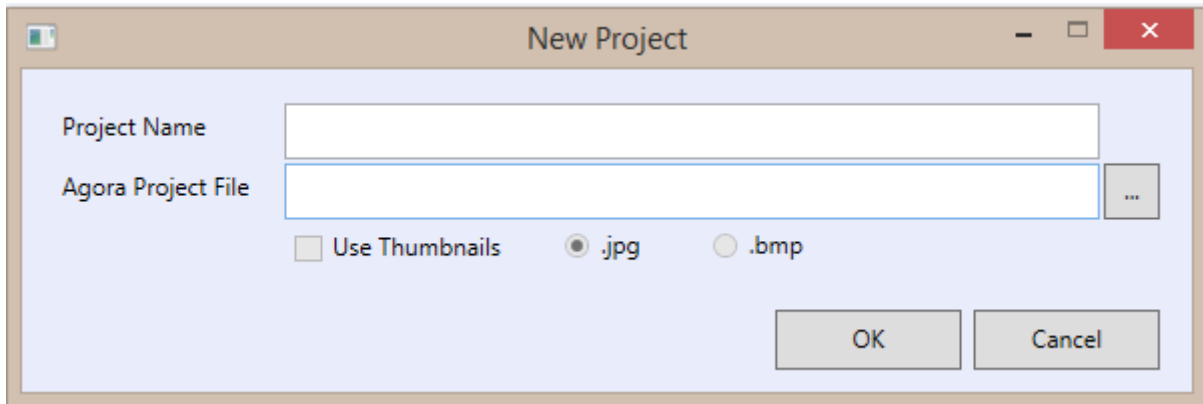
- **clustering/** Clustering xml files  
Clustering algorithms description file  
Clustering basic stats file
- **retro/** Project *\*.xml* and *\*.bin* files will be stored there
- **annotations/** Should be created manually.  
For Search and Annotations further purpose

# New Project

---

Create a new project by selecting *File/New project*.

A window will appear where you will specify the name of your project and the related AGORA project file.



*The “Use Thumbnails” Checkbox is made unavailable for now, and will be used later for future evolution of the software.*



*Note: The project is automatically saved after its creation. Therefore, \*.xml and \*.bin files will be created in the specified Project Directory.*



---

# PART II

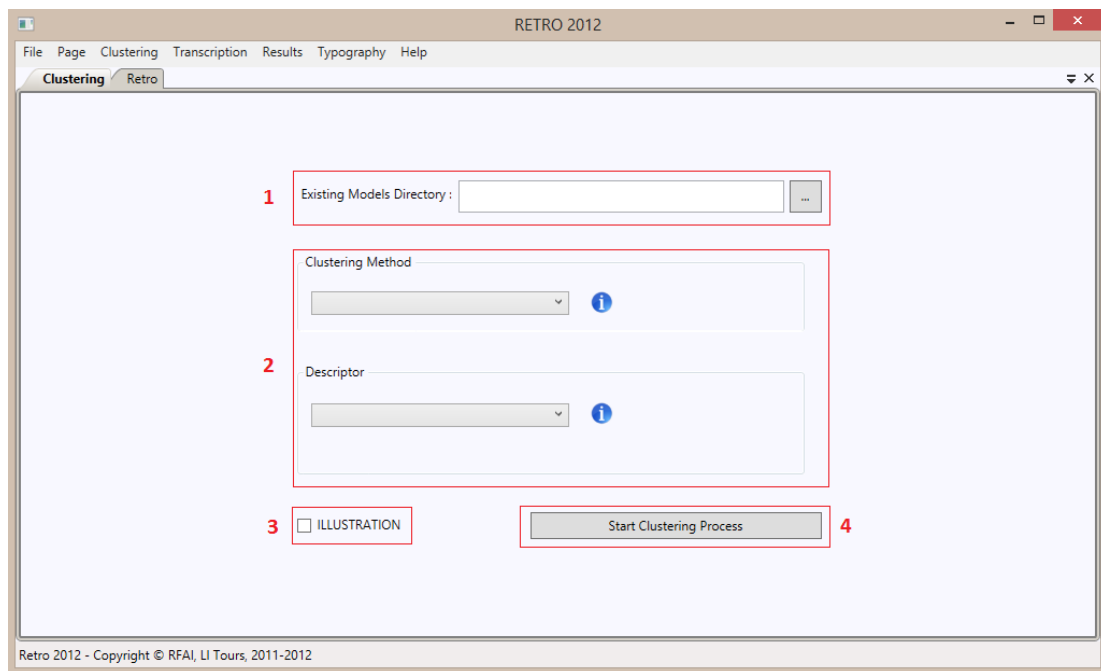
# CLUSTERING

---

# Clustering Process

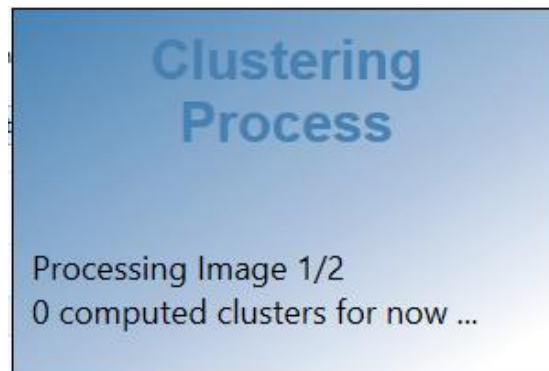
RETRO embedded the Clustering tool that will cluster the exported EoC from AGORA. This tool offers several options for a parameterized clustering.

Select *Clustering/Process Clustering* to open the clustering panel.

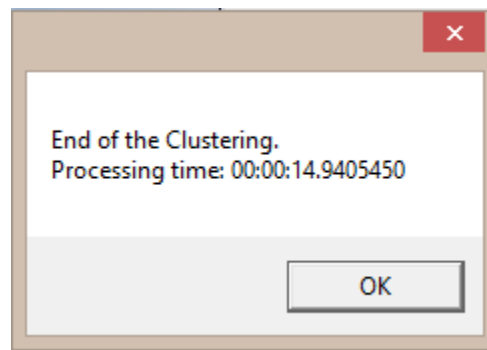


1. Allow the user to select a directory of Font Model that will be used as initial clusters during the clustering process. As Font Model have transcription, this allow to compute labeled clusters. A Font Model is a triplet {**png** grayscale image, **xml** file, **png** binary image}.
2. Allow the user to select the clustering method and the descriptor he wants to realize the Clustering.  
When the user select a clustering method or a descriptor, a window appear in order to configure the respective parameters.  
If the selected clustering method is "Template Matching", it's impossible to select a descriptor.
3. This Checkbox allow to precise if the user want to cluster Illustration EoC extracted by AGORA, instead of the Text EoC.
4. Start the Clustering process

A Dynamic Splash screen appears and will be present until the process is done.



Finally, a notification of the end of the Clustering will appear and give the time consumption of this process.



One xml file will be created for each clusters in the *clustering/* folder. An extract of one of the xml file is given below.

```
<?xml version="1.0" encoding="utf-8"?>
<cluster size="36" label="">
  <cc id="00000.0.0.0.1 path="\RETRO2014_Test\Book_Montaigne\alto\" "/>
  <cc id="00000.0.0.1.1 path="\RETRO2014_Test\Book_Montaigne\alto\" "/>
  ...
  <cc id="00000.0.9.2.6 path="\RETRO2014_Test\Book_Montaigne\alto\" "/>
</cluster>
```

## Load Clusters

---

The Clustering task directly exports its results in the cluster xml files.  
Therefore, to view and manipulate these clusters, the user should load them.

Select *Clustering/Load Clusters* to do so.



*Note: The loading of the clusters in the memory may be long, from a few seconds to several minutes depending on the number of letterform in the pages.*



*Note: The project is automatically saved after loading the clusters. The \*.bin file, that contains serialized clusters will be updated.*

## Generate Stats

---

Some basic statistics of the clustering process can be computed.

Select *Clustering/Generate stats* to do so.

An xml file, stats.xml will be created in the clustering/ folder.

An example of such a file is given below.

```
<?xml version="1.0" encoding="utf-8"?>
<stats>
  <clusters count="33" />
  <shapes count="447" />
  <histogram>
    <bin nbItems="1" nbClusters="10" />
    <bin nbItems="2" nbClusters="3" />
    <bin nbItems="3" nbClusters="0" />
    <bin nbItems="4" nbClusters="3" />
    <bin nbItems="5" nbClusters="1" />
    <bin nbItems="10" nbClusters="14" />
    <bin nbItems="50" nbClusters="2" />
    <bin nbItems="100" nbClusters="0" />
  </histogram>
</stats>
```

## Clusters To Models

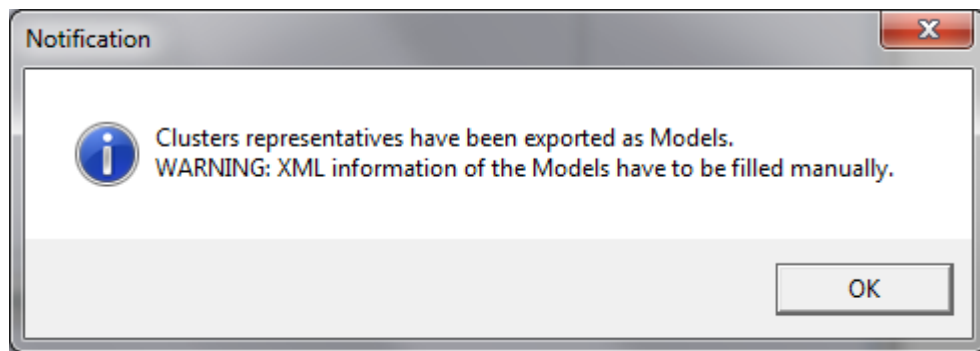
---

In order to exploit the previous clustering process and transcription, RETRO offers the possibility to create new Font Model family from labeled clusters.

Select *Clustering/Clusters To Models*.

A notification will indicate you that you have to select a directory where the Font Models will be exported, and display you a folder picker window.

After the exportation, a notification will appear.



*Note: A Font Model is a triplet {**png** grayscale image, **xml** file, **png** binary image}. We only have the transcription for each cluster that have been exported as Font Models. Therefore, the user has to manually fill, all the Font Model xml information manually.*

---

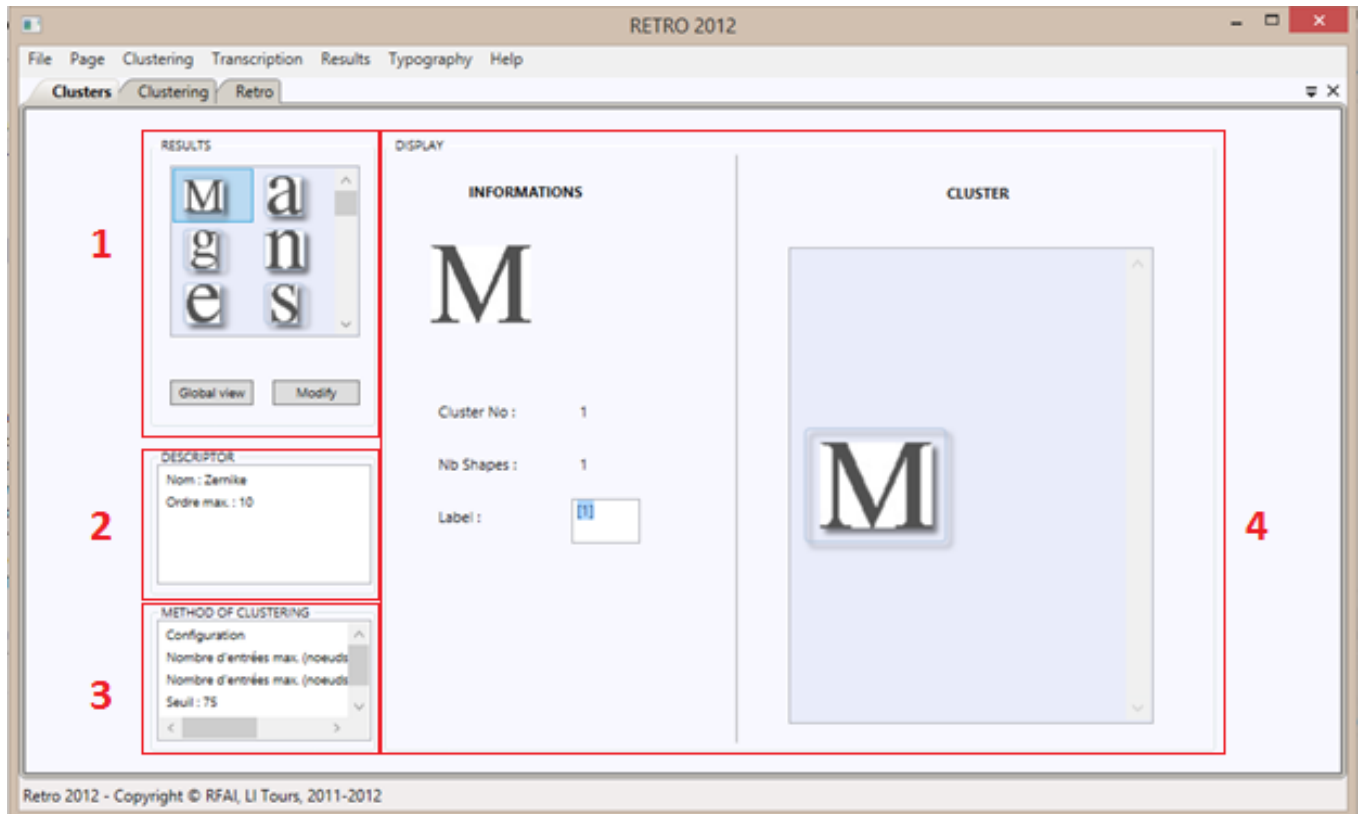
## PART III

# VISUALIZATION

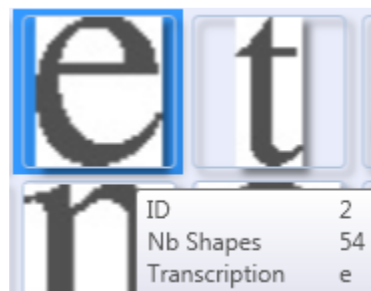
---

# Clusters Visualization

After loading the computed clusters, RETRO offers the possibility to view them. Select *Clustering/View Clusters* to open the clusters panel.



1. The image of a representative of each computed clusters is displayed. Basic information regarding the cluster appears in a tooltip when the mouse is passing over.



Two buttons are available. The “Global View button permit to display a larger view of the clusters.



You can filter the clusters regarding their labeling state.

A numbering is done, therefore both *Previous* and *Next* buttons are available.

The possibility to sort the clusters regarding several criteria is also implemented.

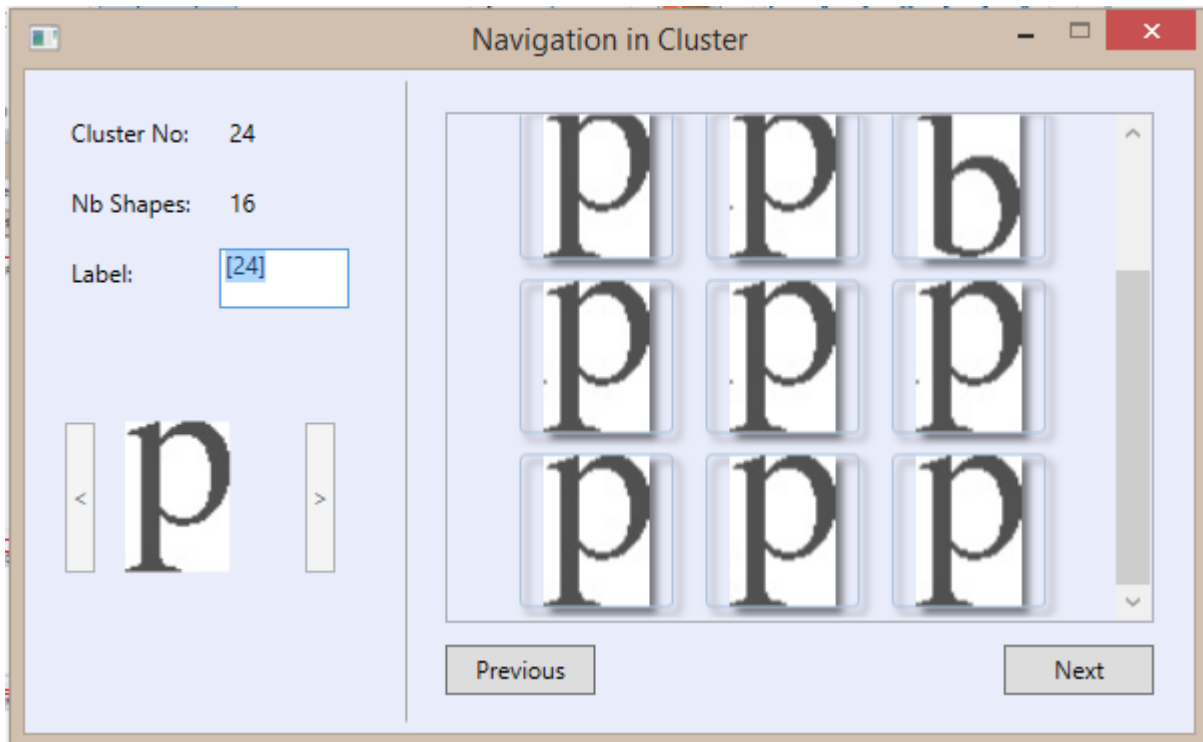
The “Modify” button permit to access to a window in order to modify the clusters.

2. Allow the user to see selected clustering method informations.
3. Allow the user to see selected descriptor informations.
4. Display the informations of the cluster which has been clicked in the panel 1.



## Cluster Navigation

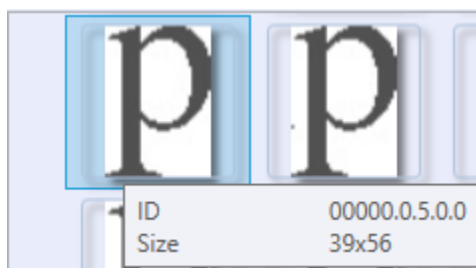
A click on a Cluster in the Global View opens a new window to navigate inside the cluster and visualize the shapes assigned to this cluster.



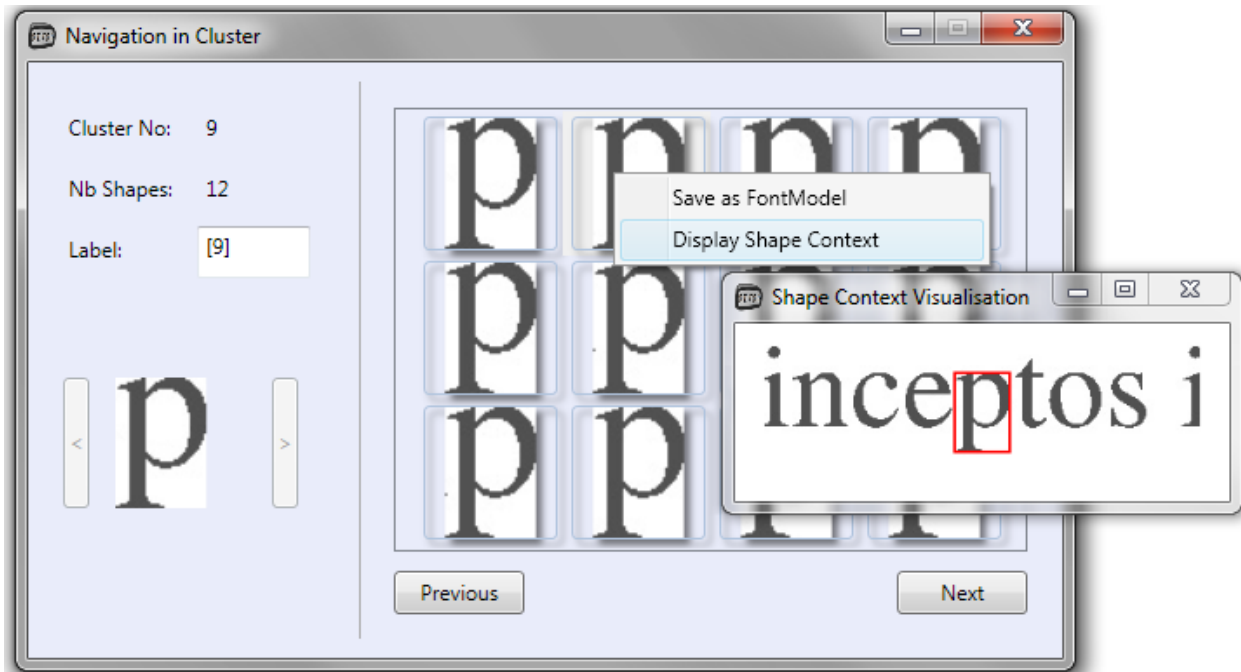
Main information and the cluster representative are displayed on the left of the window. Possibility to label the shape is also given through the “*Label*” textbox.

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

Basic information on a shape is available in a tooltip by passing the mouse over the shape.



Furthermore, as the understanding of some isolated shape in cluster may be ambiguous, the user has the possibility to display the Shape Context of a cluster element.

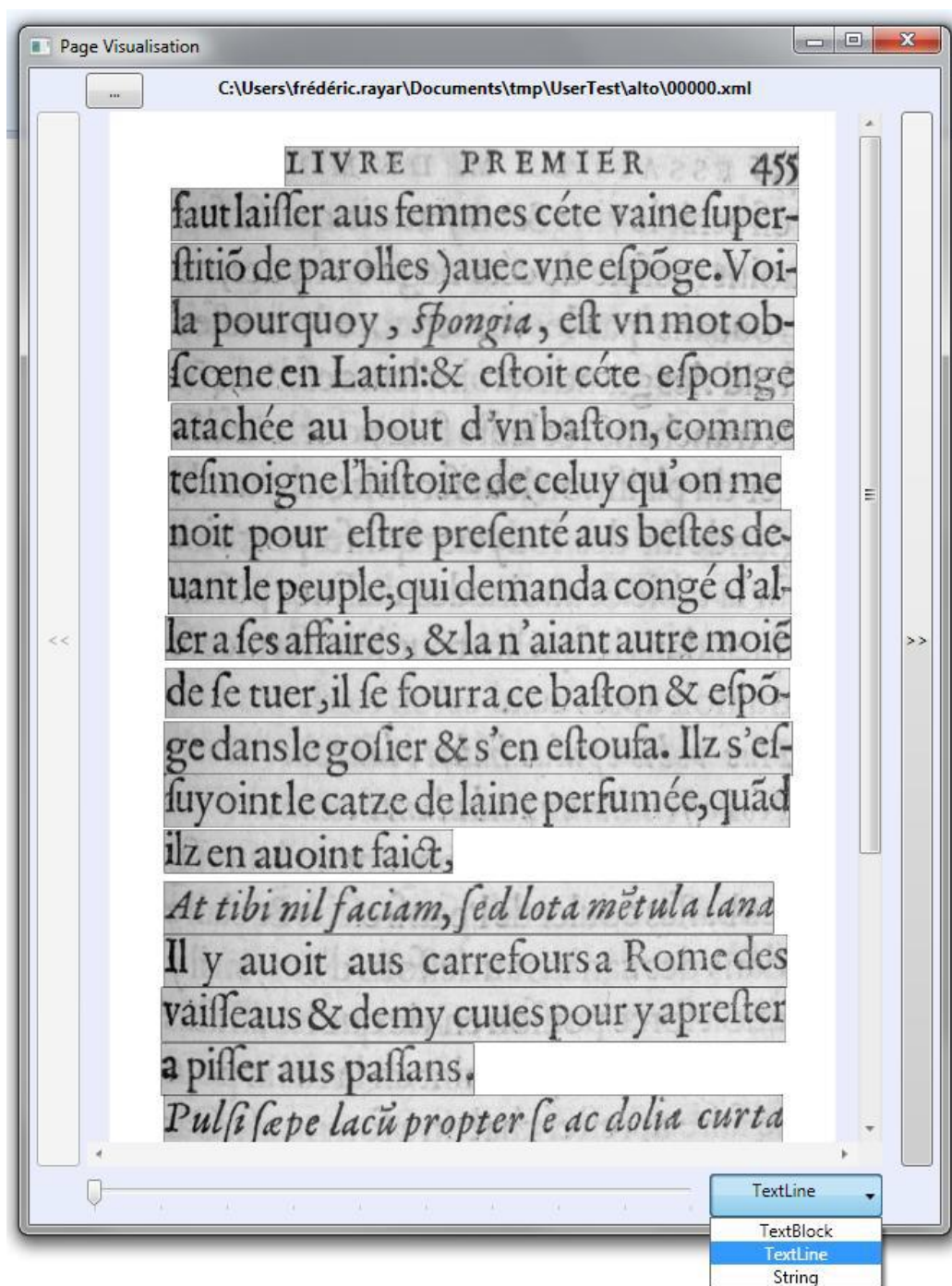


The window is accessible through a right-click, after selection of a shape. The window is always visible (until closed), and update automatically when a right-click is done on another shape.

## Page Navigation

In order to have a feedback on block extraction process of AGORA, RETRO gives the possibility to visualize the segmented pages.

Select *Page/Open a Page* to open the page navigation window.



Only the page processed by AGORA regarding the current project can be viewed (i.e. must be xml alto file in <dir>/alto).

Previous, Next buttons are available to navigate in the page of the book.  
An Open [...] button is also available to open any desired page of the book.

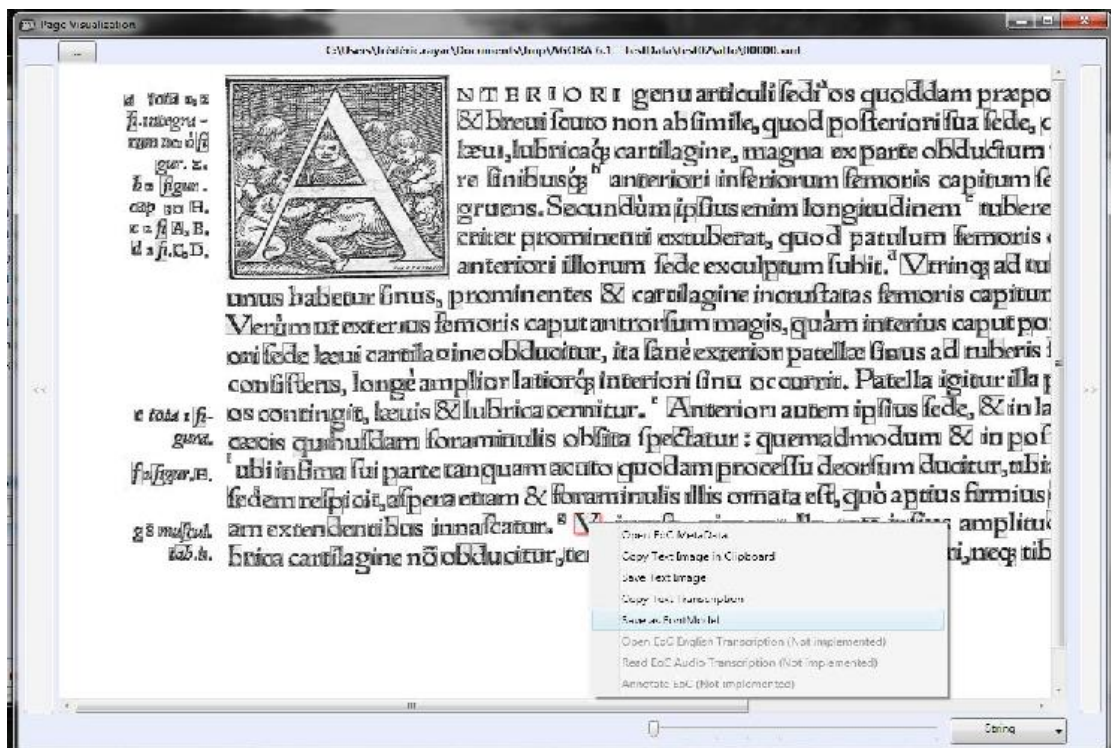
A zoom slider, and scrollbars are available to navigate within the page.

In accordance with the alto format, 3 granularities can be selected: TextBlock, TextLine and String (letterform). We refer to the block as Elements of Content (EoC).

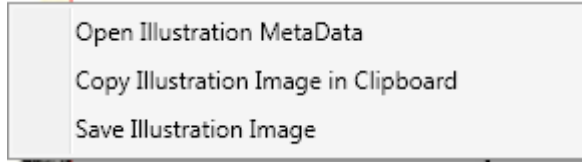


*Note: The display of the segmented page may take a few second to a minute for String depending on the number of letterform in the page.*

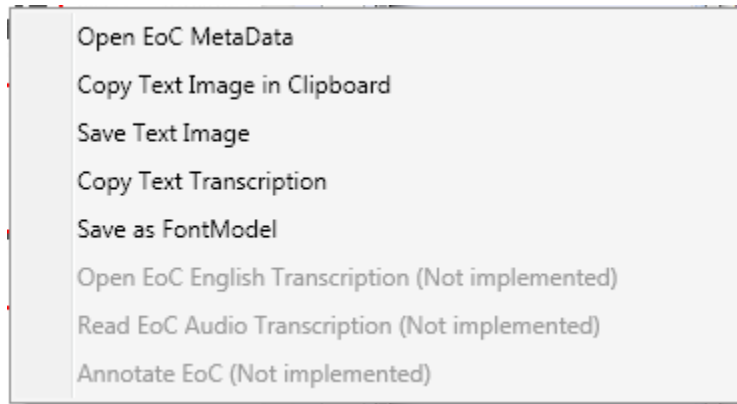
For each EoC an Contextual menu (accessible through a right click), is available, and present different functionalities regarding the type (*Illustration* or *Text*) and the Granularity of the selected EoC.



### Illustration EoC Context Menu:



### Text EoC Context Menu:



The “*Open EoC MetaData*” opens a new window, where a zoom, and available metadata (width, height ...) of the selected EoC are displayed. (Not fully implemented).



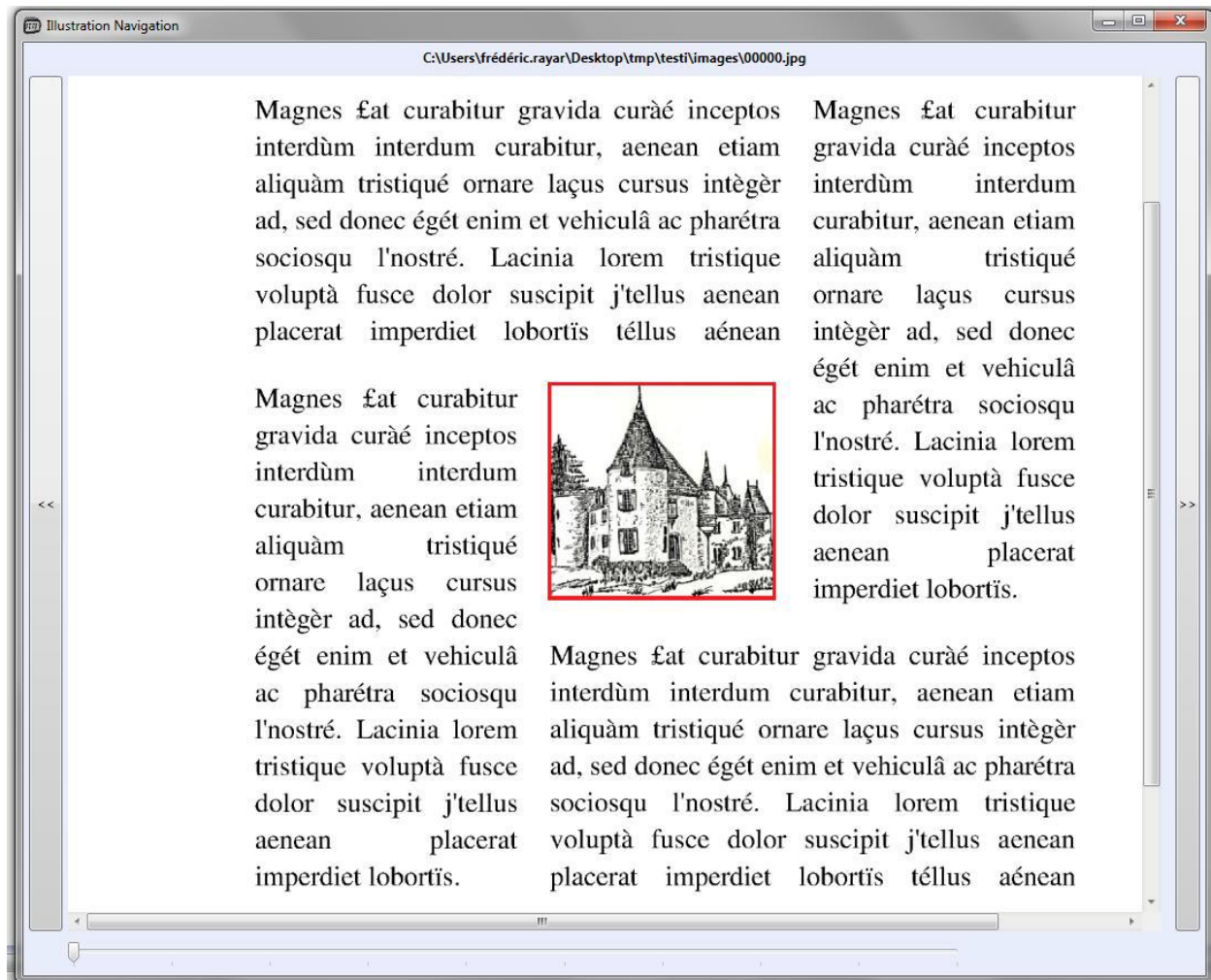
*Note: The “Save as Font Model” functionality is only Enable For Text EoC, and when String granularity is selected.*

# Illustration Navigation

For image dedicated purpose, an illustration navigation tool is available.

It allow the user to view only the page with illustration, where the illustration are highlighted with a red bounding box.

Select *Page/View illustration* to open the page navigation window.



---

# PART VI

## MODIFICATION

---

# Clusters Modification

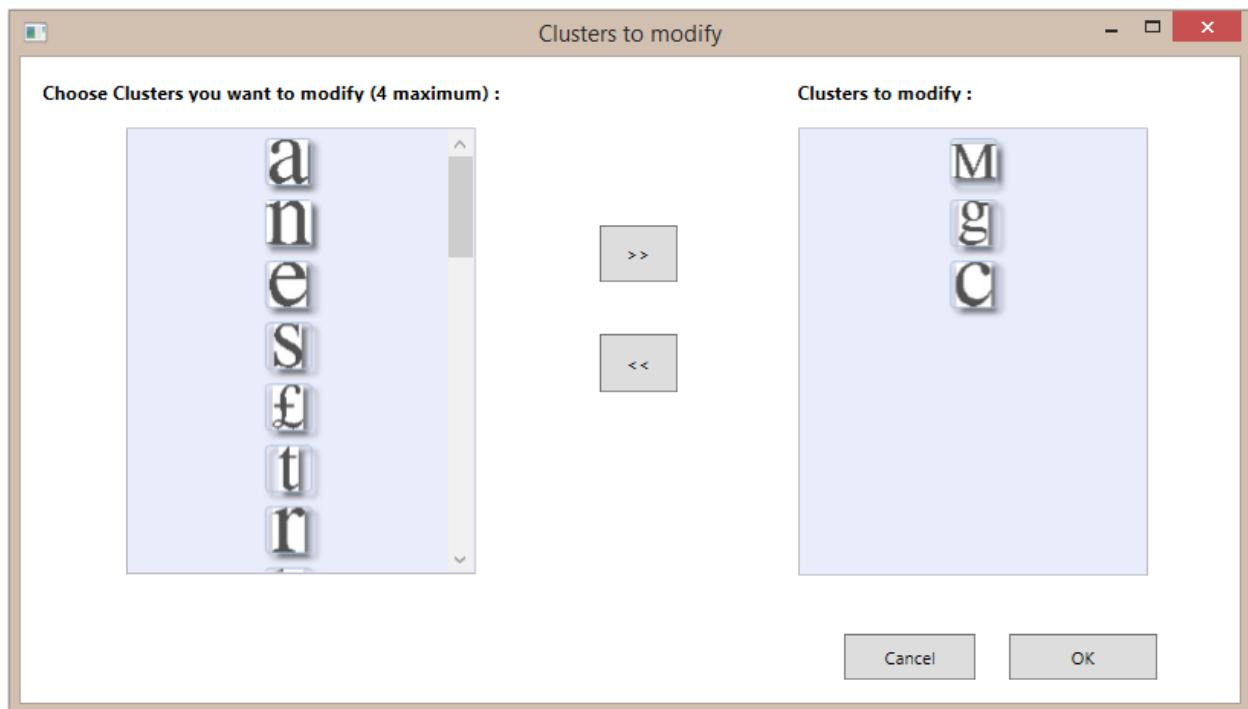
---

It's possible to change the shape affectation. To modify the Clusters

Click on the “Modify” button in Results Viewn to access to the clusters modification.



A new window appear in order to select the clusters that the user wants to modify

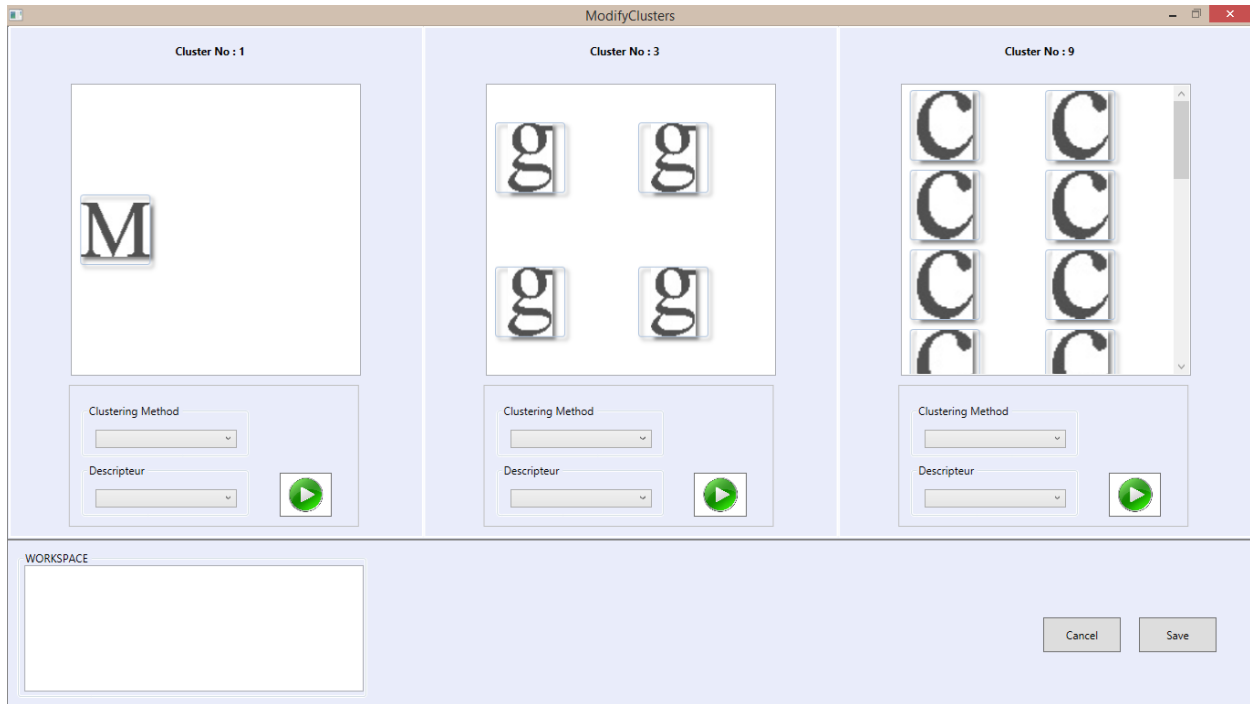


The user can select maximum four clusters to modify.

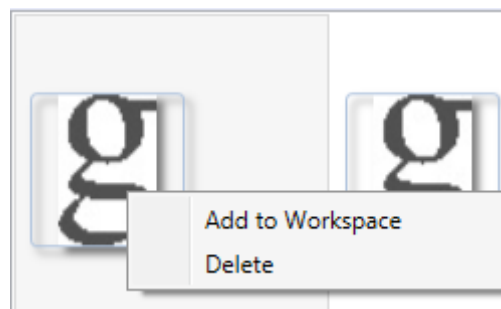


Then click Ok.

A window for the clusters modification appears and allow the user to change the affectation of each shape by drag and drop.

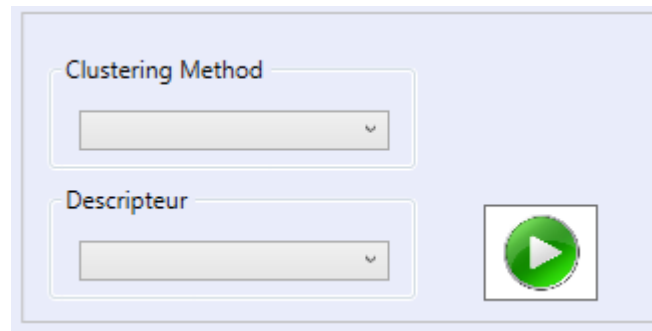


In addition, by right click on a shape, the user can delete definitively a shape or just put it in a workspace in order to reuse it later.



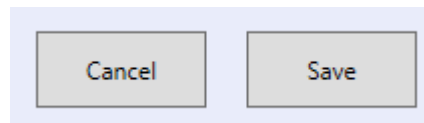
In a future version of RETRO, The user could realize a clustering process on a unique Cluster to create again new clusters from one clusters, if he think it would be better to divide a cluster.

That's why under each cluster, there is a panel for clustering process.

A light blue rectangular panel with a thin border. It contains two labels, "Clustering Method" and "Descripteur", each followed by a white rectangular dropdown menu with a small downward arrow on the right. To the right of these two dropdowns is a green circular button with a white right-pointing triangle in the center.

But this feature has not been implemented yet.

When the user has finished the modifications, if he wants to save what he has done, he has to click to the "Save" Button, else "Cancel Button".

A light blue rectangular panel with a thin border. It contains two gray rectangular buttons with black text. The button on the left is labeled "Cancel" and the button on the right is labeled "Save".

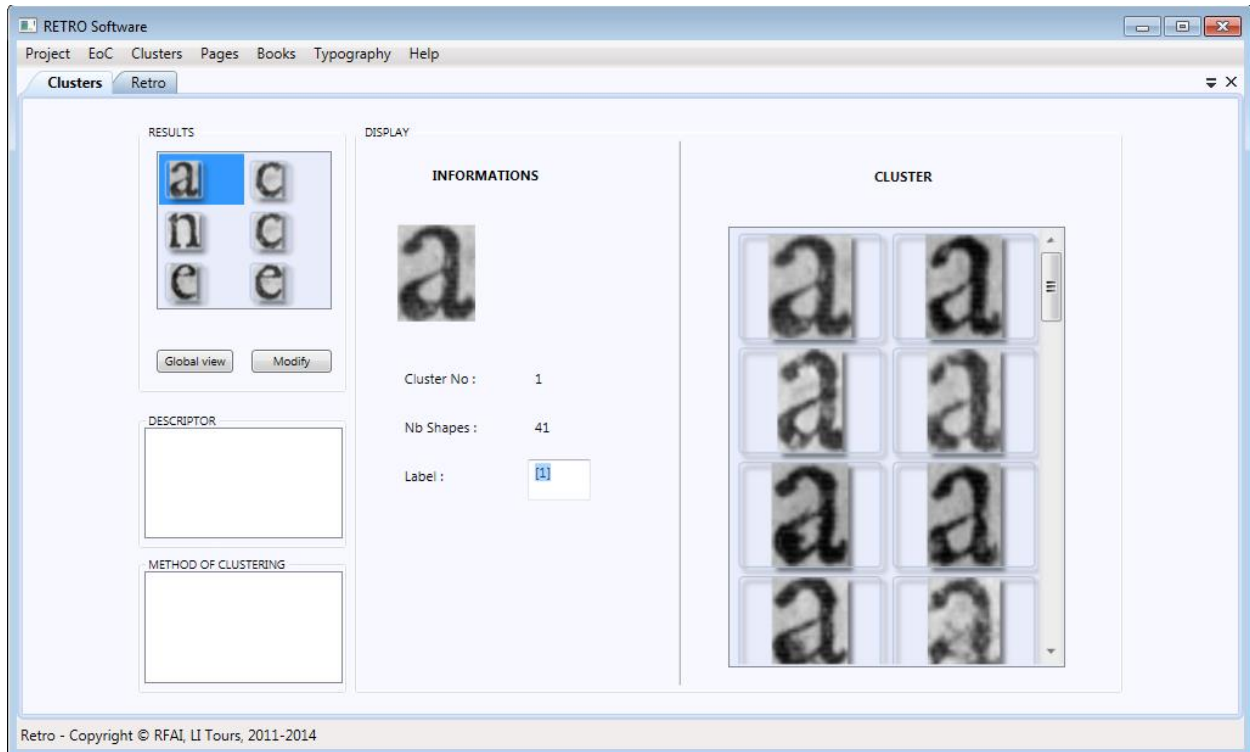
By clicking on the Save Button, it closes the clusters modification window and the changes are reflected in the Result View.

# Clusters Analyze

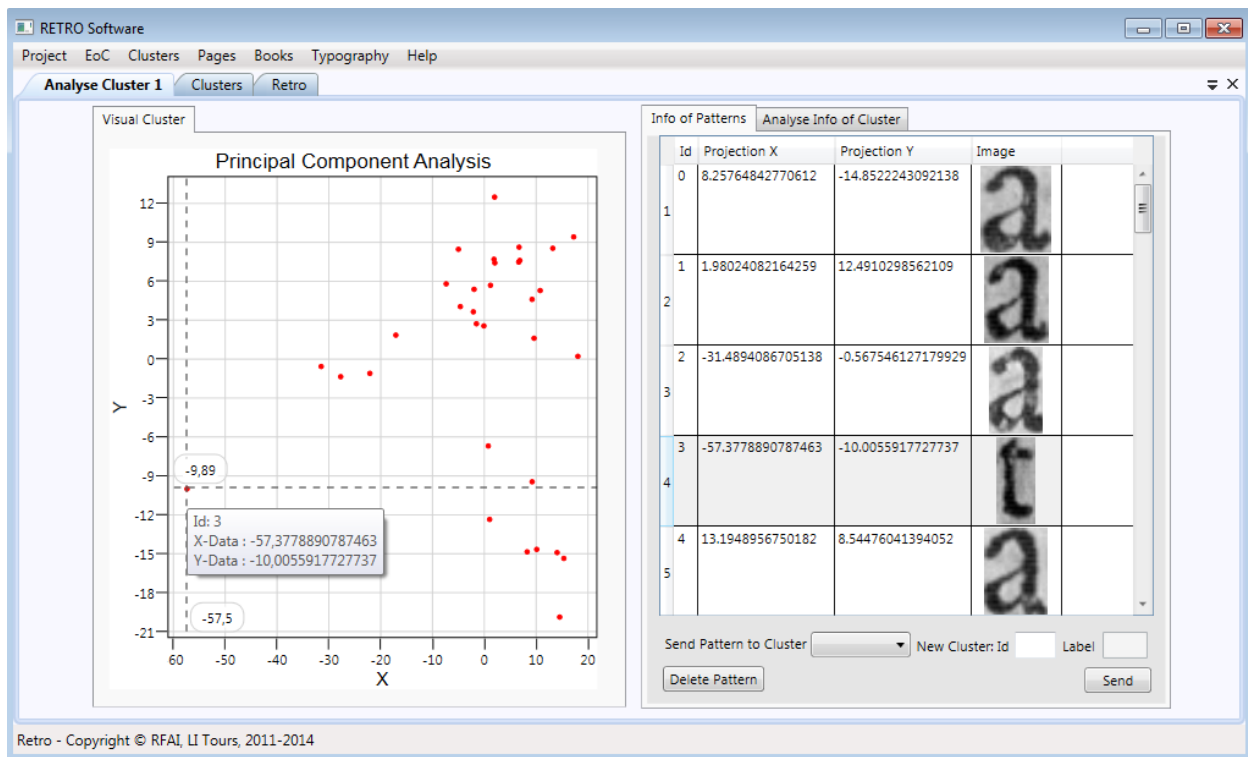
---

It's possible to analyze a Cluster and change the shape affectation, to modify the Clusters.

Choose a Cluster in in Results View and the click on the menu *Cluster/Analyse Clusters* to access the analyze Cluster.

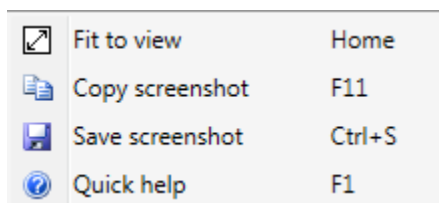


A new panel appears to show the Visualization of the cluster that the user wants to analyze.



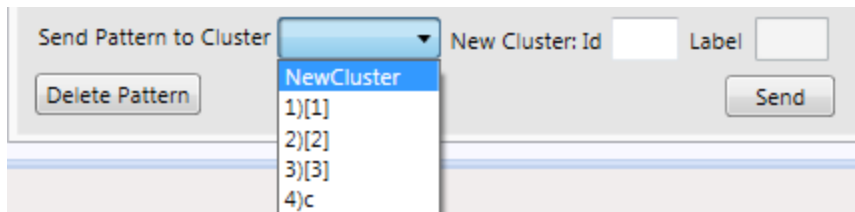
The projection of Principal Component Analysis of the Cluster shows on the 3D interactive dynamic coordinate axis on the left.

On this plan, user can zoom in, zoom out and drag the map to see those points of the Cluster. Each point is a pattern in the Cluster, when the mouse floats over a point, a tag floating will show the id and the position of pattern. When the user left clicks on the point, the pattern item in the list of “Info of Patterns” on the right side will be highlighted on grey. Clicking the right mouse button on the plan dynamic, a menu floating will be showed like this:



If user want to delete a pattern in the list right side, first choose a pattern in the list the click the button “Delete Pattern”.

If user want to send the pattern to another Cluster, Choose a Cluster in the comobox “NewCluster” then fill in the textbox with the Id and label of the new Cluster.



The comboBox is filled by the form: Id) Label , of a Cluster.

---

## PART V

# TRANSCRIPTION

---



## MANUAL TRANSCRIPTION (I)

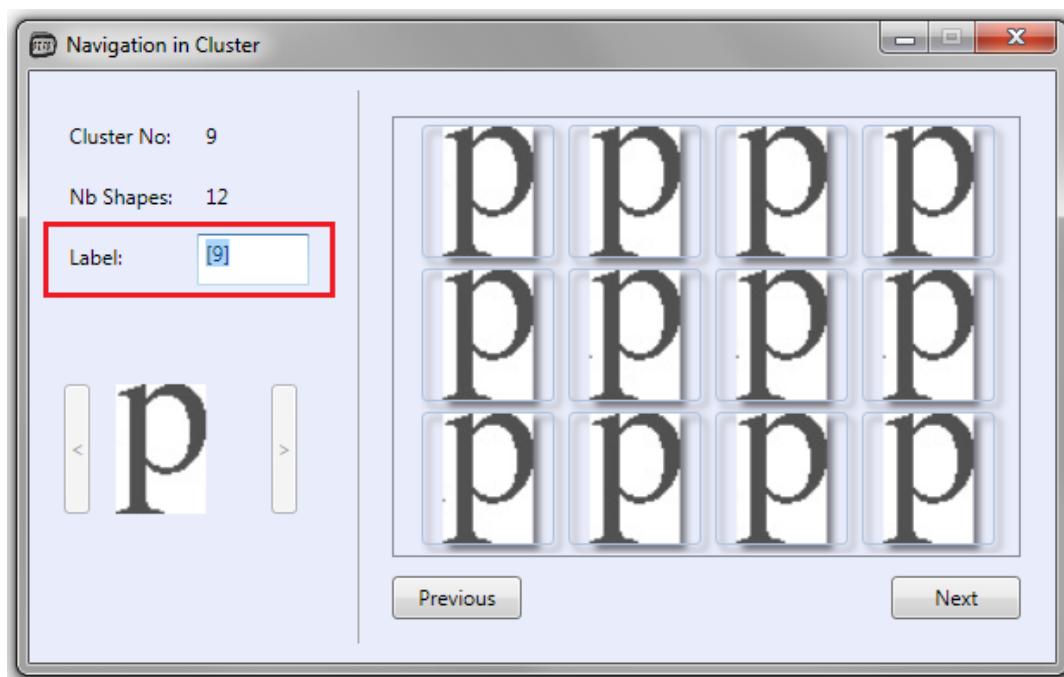
---

The first way to assign a label to a cluster is in the Cluster Navigation window (Global view) through the “*Label*” textbox.

The representative thumbnail is displayed, plus a quick glimpse on the shapes within the cluster is generally enough for the user to guess the transcription of the cluster.

A default transcription (the id of the cluster) is given by default.

No validation is required, closing this windows or clicking on another thumbnail in the Cluster panel of the main window automatically validate the labeling.

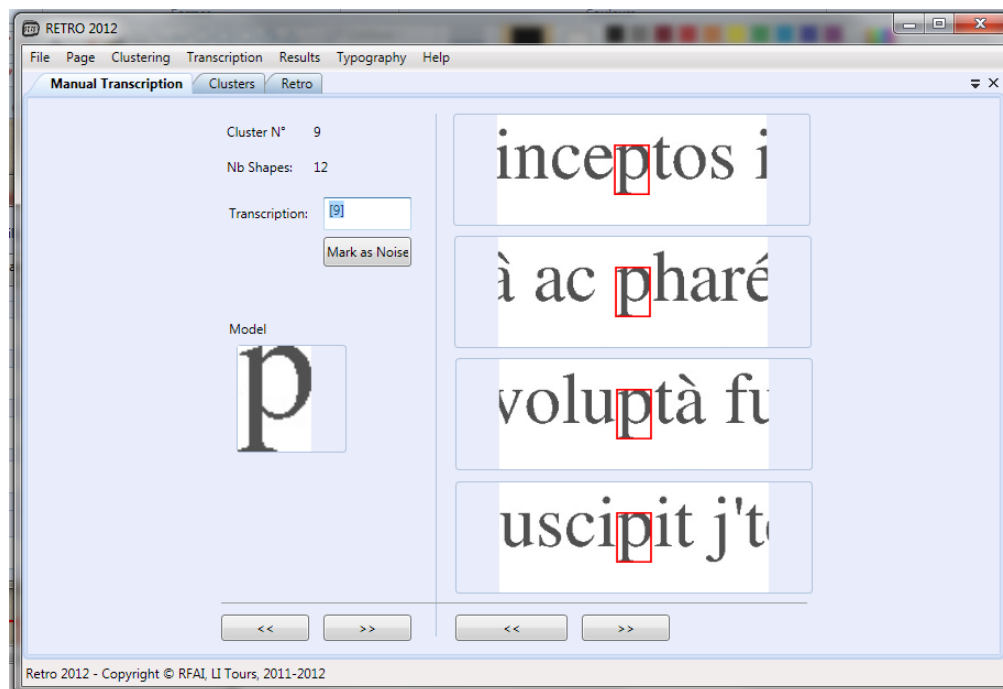


## MANUAL TRANSCRIPTION (II)

---

As the manual transcription may sometimes be quite a chore, a second interface, more optimized has been created for this task.

Select *Transcription/Manual Transcription* to display this interface.



Only non labeled clusters are displayed sorted regarding the sort selected in the Cluster Panel of the main window.

Information and cluster representative are displayed in the left of the window.

You can navigate through the non-labeled clusters with the Next/Previous buttons below the representative.

As sometimes, only isolated shapes may not be enough to determine the transcription, each shapes of the cluster is display within its context in the original images of the book, centered and surrounded by a red rectangle. Only 4 shapes in context are displayed, therefore Next/Previous buttons are available (below the 4 shapes in context).



Furthermore, to relieve the user in this manual transcription task and decrease this process time, optimization and shortcuts have been implemented.

The textbox is always focused, the user can enter the transcription and validate by pressing the **Enter** key. To assign the Noise label to a cluster, the user can either click on the Noise button or enter in the keyboard **Ctrl + N**.



*Note: As transcription is plain text, the user is free to assign costumed tag to describe special or rare font.*

# AUTOMATIC TRANSCRIPTION

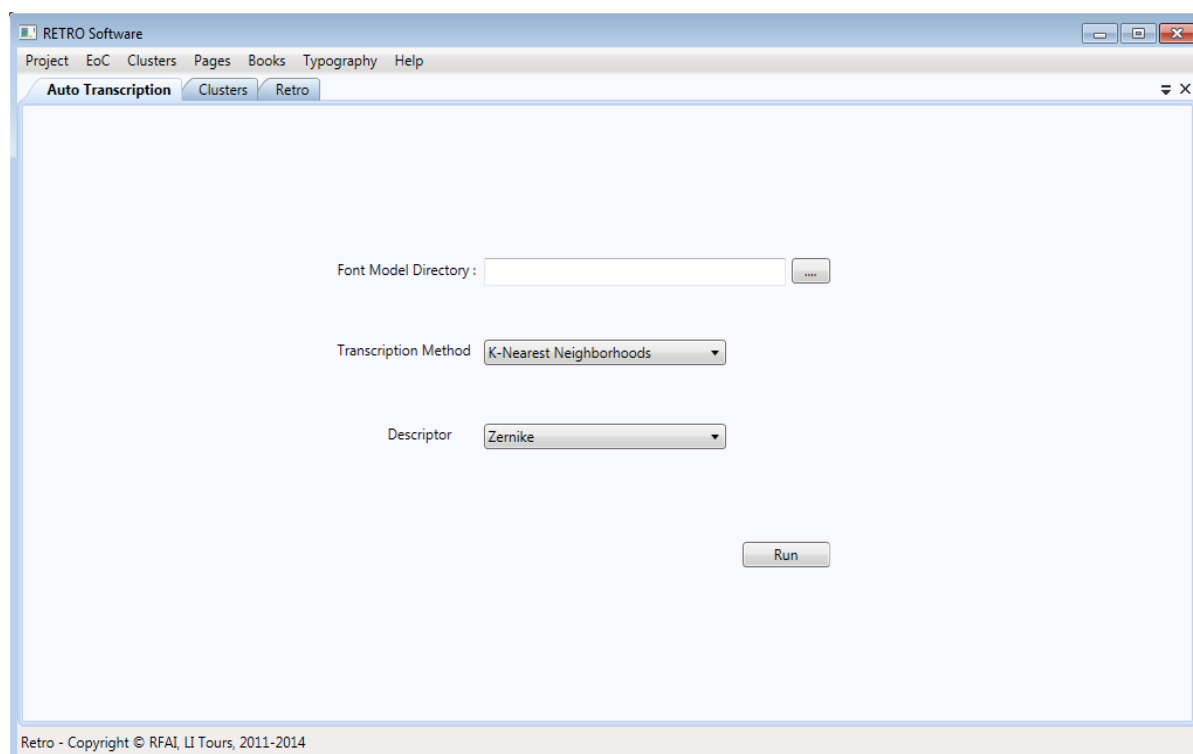
---

A simple OCR system has been implemented in the current prototype of RETRO.



*Note: The automatic transcription may take several minutes depending on the number of clusters to label and the number of font models.*

To run the automatic transcription, select *Cluster/Automatic Transcription*.



The recognition is based on the Font Models given as a learning database.

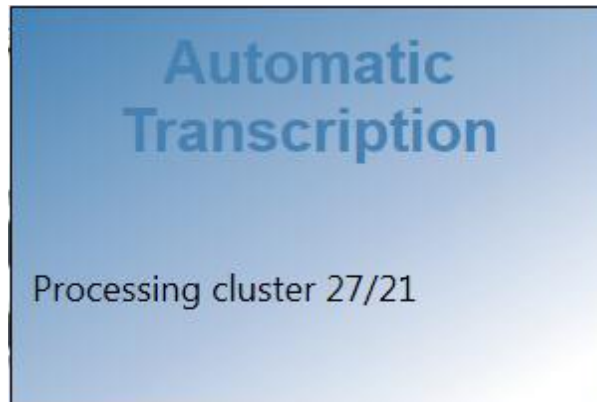
A FontModel is a triplet {**png** grayscale image, **xml** file, **png** binary image}.

The user has to indicate which learning database he wants to use.

Select a directory where the Font Models are stored in the output directory “*Font Model Directory*”. Then select a method and a descriptor for the transcription.

After you have set all the parameters, RETRO will analyze it, compute the number of found FontModel, and will ask the user if he want to proceed or not.

A Dynamic Splash screen appears and will be present until the process is done.



Then, a Message box will notify the user that the automatic transcription has been done.



*Note: At the end of the OCR process, clusters labeled only in the memory, you have to save the project to persist the labels.*

## Results Exportation

---

After a transcription task, manual or automatic, clusters are labeled only in the memory. The project must be saved to make the newly assigned labels persistent in the project (*File/Save Project*).

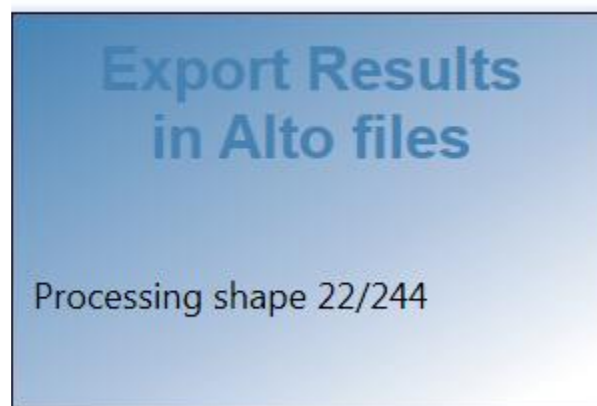
After the labeling of all clusters of a book, the user can export the results to the alto files. This exportation updated the *CONTENT* attributes of each *String* tag in the alto files.

To access this functionality, select *Results/Export As Alto*.



*Note: As the results exportation updated each String attributes in the alto files, it means the time is proportional to the number of shapes in the project. This means this task can take a considerable amount of time. Therefore, it is recommended to make one unique results exportation after fully labeling the clusters..*

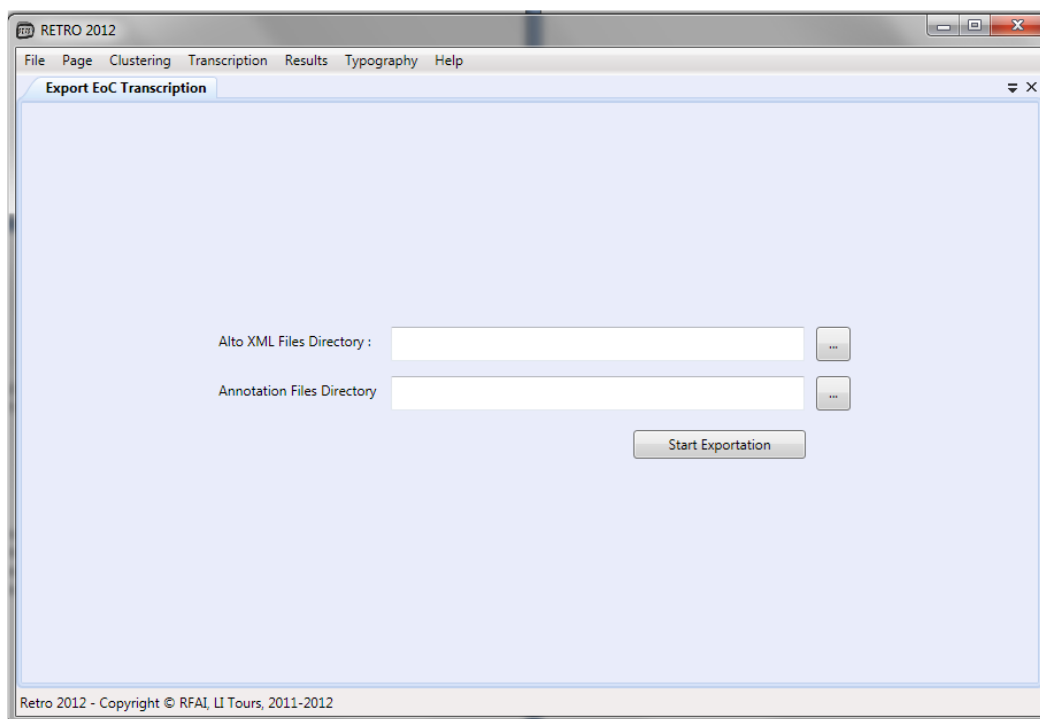
A Dynamic Slash screen appears and will be present until the process is done.



Then, a Message box will notify the user that the automatic transcription has been done.

# Transcription Exportation

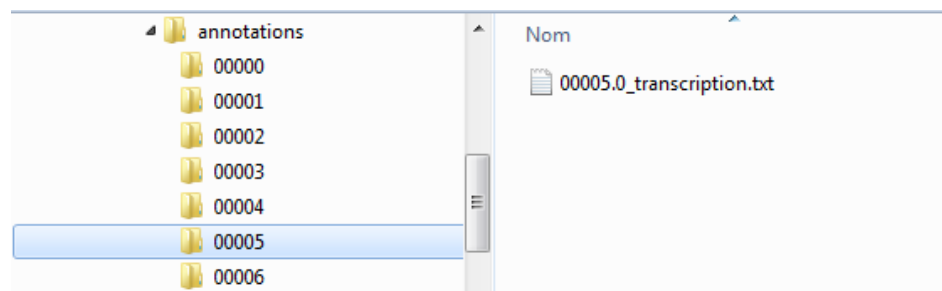
For an efficient and fast Search purpose in further studies, a functionality, available at *Results/Export EoC Annotation*, allow the user to export the transcription of each *TextBlock*.



*Note: The output directory, named “Annotation Files Directory”, must already exist!*

A Message box will notify the user that the automatic transcription has been done.

One subdirectory will be created for each page of the book and one \*.txt with a normalized name will be created for each existing TextBlock in one page.



---

## PART VI

# TYPOGRAPHY STUDY

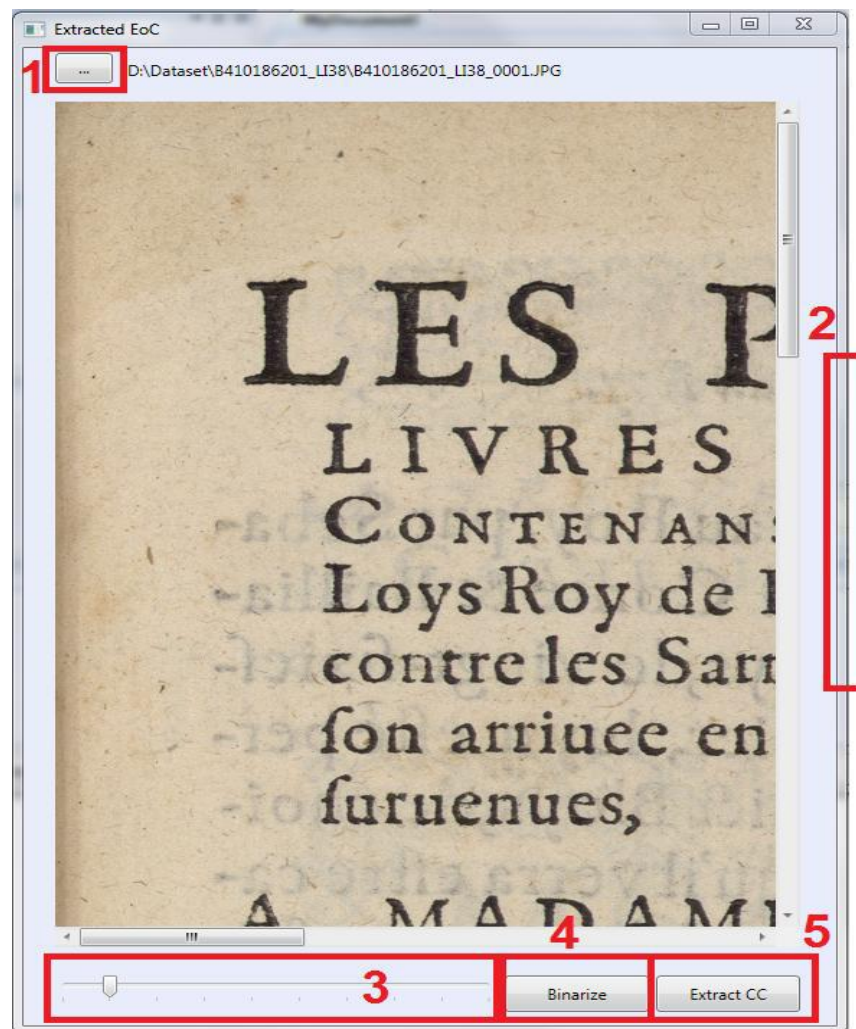
---

# Font Model Creation Tool

---

RETRO embedded a tool for creation of font model, and by extension font families. The tool is a standalone and can be used independently of any RETRO project.

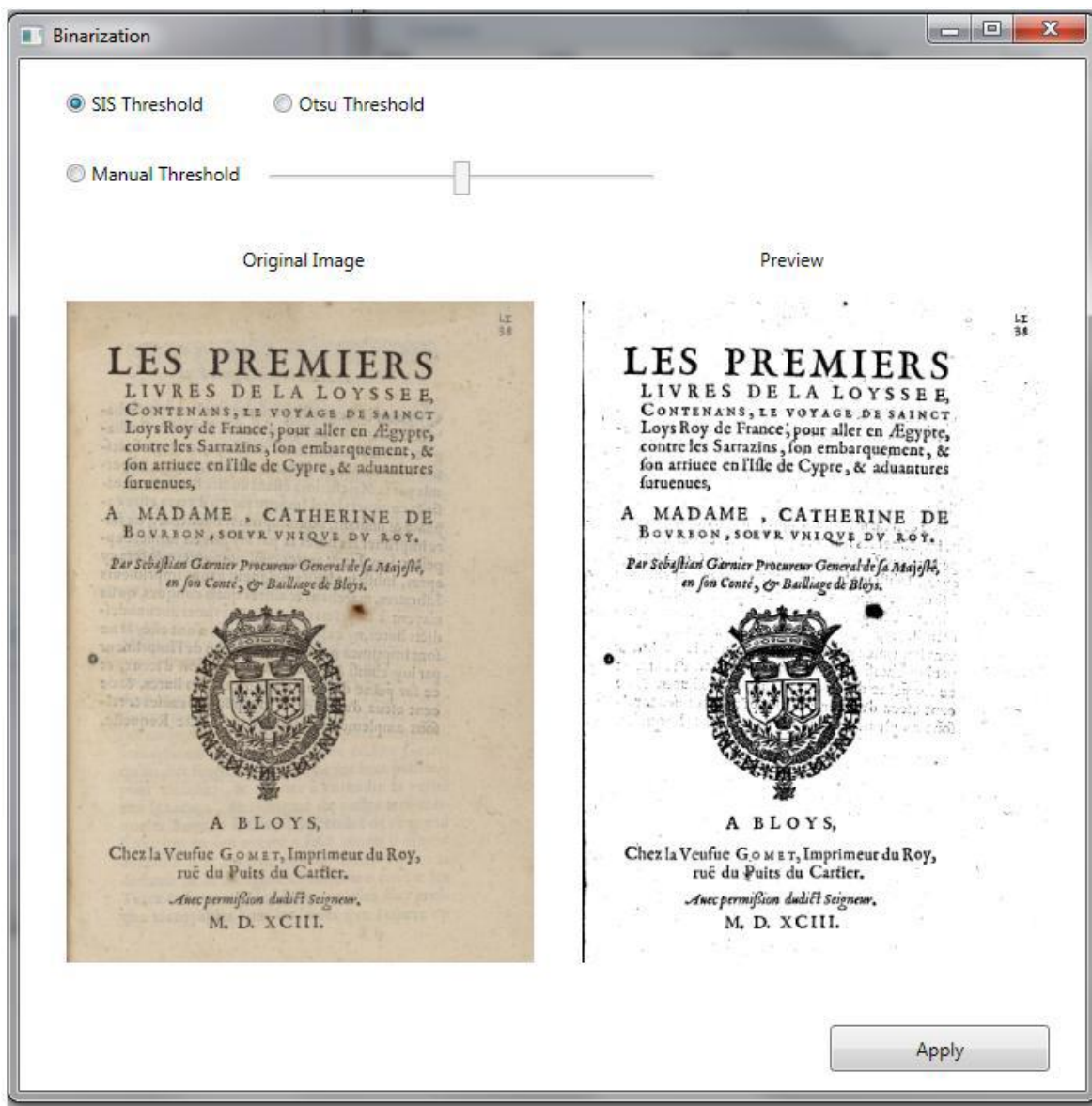
To open this tool, select *Typography/Add Model*, the window below will appear.



Here you have the possibility to:

1. Open the image of interest
2. Resize the windows regarding your screen size for more comfort
3. Zoom the image (up to x5)
4. Manage the binarization/thresholding process (optional)
5. Launch the extraction of Connected Component (CC)

If a customized binarization process is required, click on the *Binarize* button to open the interface below:



3 methods of binarization are proposed:

- SIS (Simple Image Statistics) method, used by default.
- Otsu binarization
- A manual threshold selection is possible (slider is only enable when the Manual option is checked)

Validate the process with the *Apply* button; else you can cancel by exiting with the red cross.



Computed CC will be highlighted with red bounding boxes, and a first form to fill the metadata of the image will appear.

You only need to fill once this form for each image, and then have the possibility to extract several models from this image.

The screenshot displays the 'Extracted EoC' application window. The main area shows a book cover with the title 'LES PREMIERS LIVRES DE LA LOYSSEE' in large, bold, black letters. Red bounding boxes are drawn around the text. Overlaid on the image is a 'Metadata' form with two sections: 'Book Metadata' and 'Copy Metadata'. The 'Book Metadata' section includes fields for Author, Title, Publication Site, Printer, Date, and Format. The 'Copy Metadata' section includes fields for Library, Pressmark, Digitalization, License, and Cataloguer. A 'Next' button is located at the bottom right of the form.

Extracted EoC

D:\Dataset\B410186201\_LI38\B410186201\_LI38\_0001.JPG

LES PREMIERS  
LIVRES DE LA LOYSSEE

Metadata

Book Metadata

Author :

Title :

Publication Site :

Printer :

Date :

Format :

Copy Metadata

Library :

Pressmark :

Digitalization :

License :

Cataloguer :

Next

A second window will then appear presenting a form to describe data related to the character itself.



Click on “Create model” to confirm the creation of a new model (datas are stored between two successive creation, except the Transcription and the Unicode fields).

You have the possibility to select the output directory, allowing the creation and organization of font models family. A notification will appear to inform the user that a new model has been created.

Furthermore, the bounding box of the newly created model will now be in green to allow the user track of his work.

The creation process produces 3 outputs for each model:

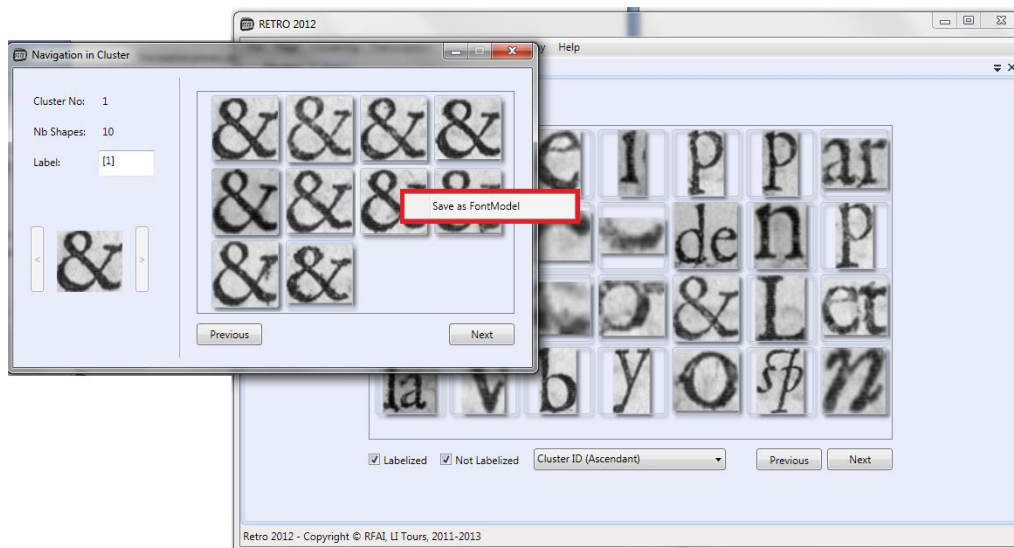
- a grayscale image (\*.bmp)
- a black and white image (\*.bmp)
- an XML description file

The schema of the xml output is presented below

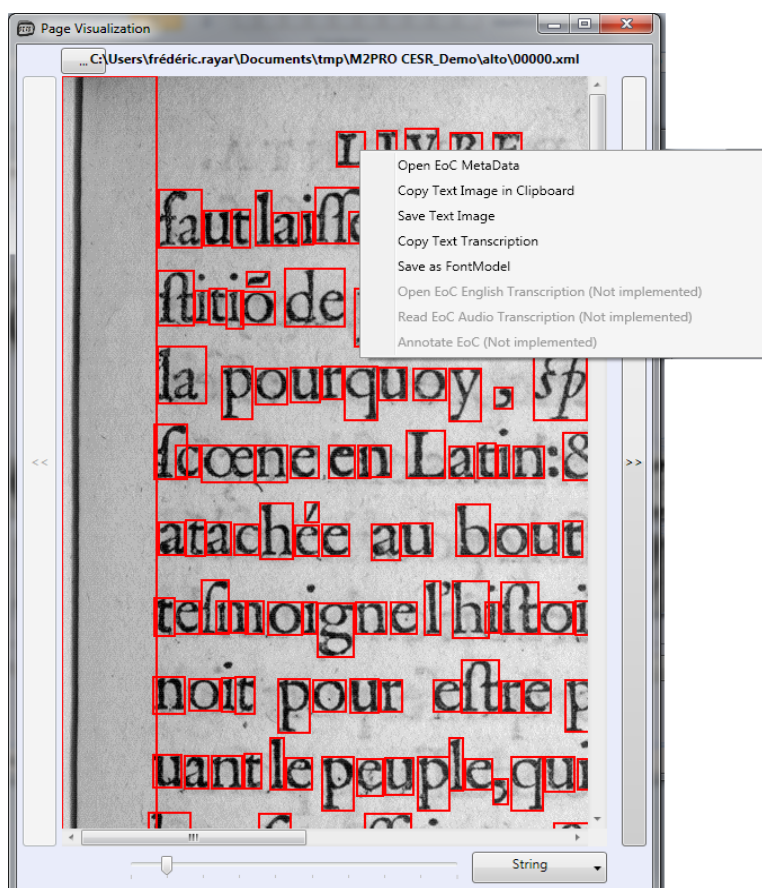
```
<?xml version="1.0" encoding="utf-16"?>
<!--RETRO Model file-->
<Model>
<Metadata>
  <Book Author="" Title="" Place="" PrinterOrPublisher="" Date="" Format="" />
  <Copy Library="" CallNumber="" Digitalization="" Copyright="" CataloguerName="" />
</Metadata>
<Transcription Character="" Unicode="" />
  <Image Filename="" Folder="" Page="" Resolution="" />
  <Thumbnail Name="" Width="" Height="" PositionX="" PositionY="" />
<Typography IsSmallCap="" Type="" Alphabet="" Family="" SubFamily="" BodyHeight=""
Thickness="" /> <Description References="" Engraver="" Comments="" />
</Model>
```

You also have the possibility to access this Font Model Creation Tool within the scope of an opened RETRO project:

1/ In the Custer Navigation window, with a right click on the desired model



2/ In the Page Navigation window, when the String granularity is selected, with a right click

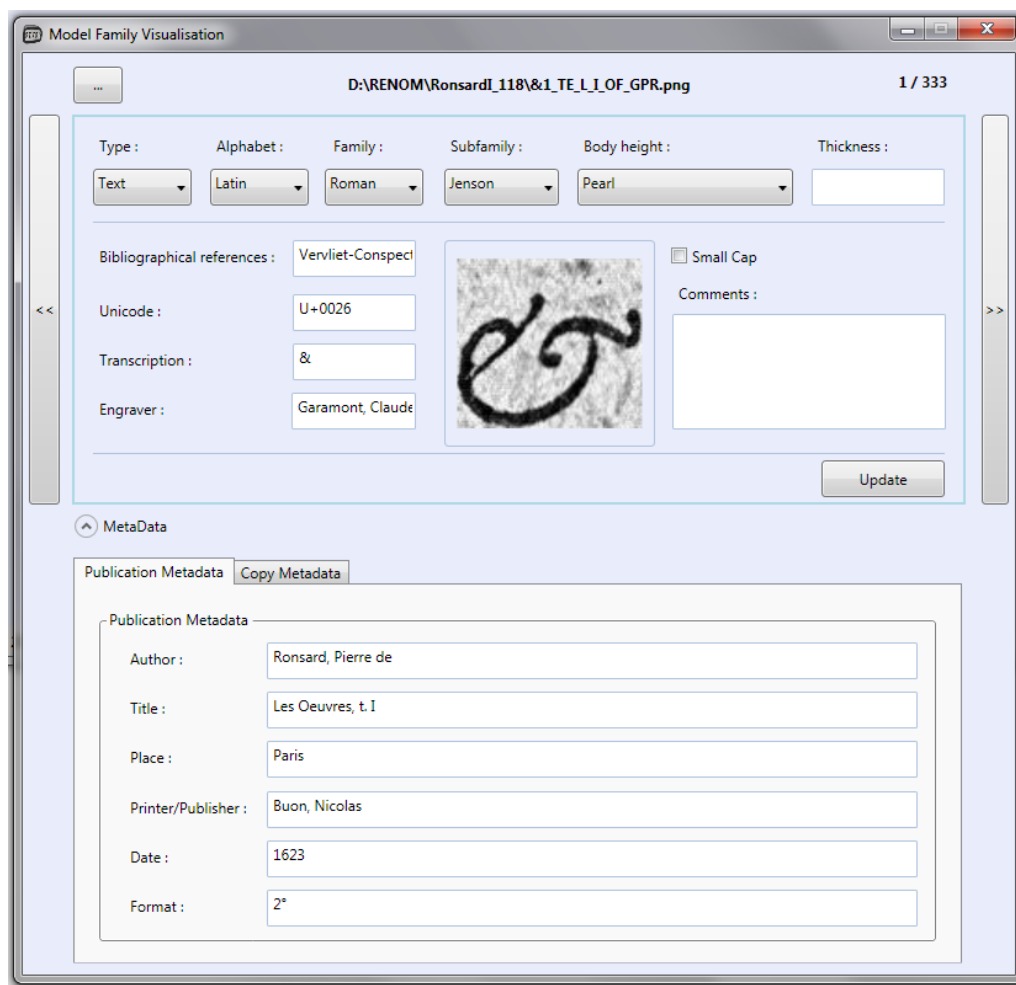


# View Font Family Tool

In order to check, and validate the indexed font family, RETRO embedded a View Font Family Tool.

The tool is a standalone and can be used independently of any RETRO project.

To open this tool, select *Typography/View ModelFamily* and select a directory of the font family extracted with RETRO.



You can navigate in the selected directory, with the Previos/Next Button, and the [...] to jump to a random model in the directory (note that you can select either the grayscale or the binary image).

You can also display/hide the panel with the Model metada.

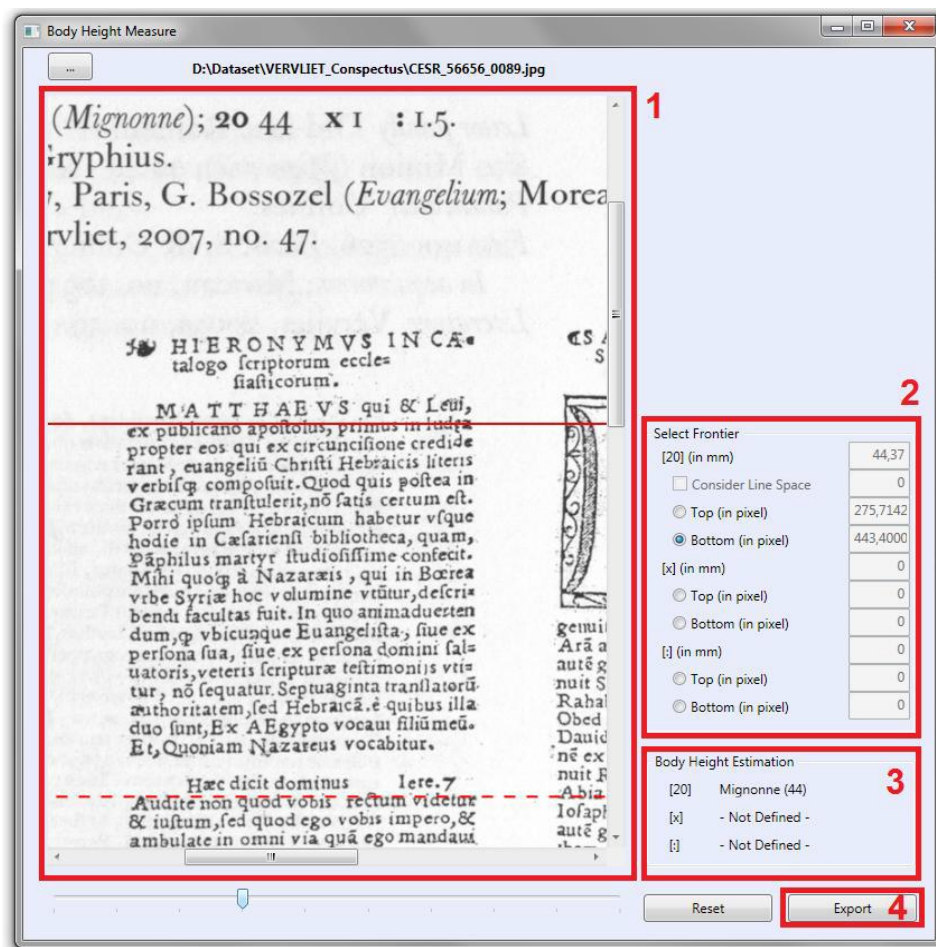
# Body Height Measurement Tool

RETRO embedded a tool for the body height measurement.

The tool is a standalone and can be used independently of any RETRO project.

To open this tool, select *Typography/BodyHeight* and select an image with the [...] button.

Zoom slider and scrollbar are available for a finer precision work.



1. Selected page
2. Selection of the frontier (top/bottom for [20], [x], [:] height), and display of measured values (in pixels and mm)
3. Estimated Body height designation
4. Possibility to export computed information in xml for further use and studies.

The output directory is *<appdir>/BODY HEIGHT/* where *<appdir>* is the directory of the executable Retro2011.exe

An example of an exported file is presented below (*CESR\_56656\_0089\_MI\_44.xml*):

```
<?xml version="1.0" encoding="utf-16"?> <!--RETRO Body Height file-->
<BodyHeight>
<Filename Path="D:\Dataset\VERVLiet_Conspetus\CESR_56656_0089.jpg" />
<Typography FrenchName="Mignonne" FrenchCode="MI" EnglishName="Minion">
    <[20] Value="44,37" />
    <[x] Value="0" />
    <[:] Value="0" />
</Typography>
</BodyHeight>
```

---

## PART VII

## FAQ

---



1. **What do I need to deploy RETRO on my computer?**

RETRO is only compatible with Microsoft Windows Operating System (XP and above).  
You will also need to install the .NET Framework 3.5.

---

## PART VIII

## GLOSSARY

---

**AGORA**

Software dedicated to digitized old documents. It achieves page layout analysis, text/graphics separation, pattern extraction and clustering of similar pattern. Current version: 6.3 (developed by Pascal Bourquin)

**ALTO**

Analyzed Layout and Text Object ALTO is an XML Schema that details technical metadata for describing the layout and content of physical text resources

**CC**

Connected Components

**Clustering**

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

**EoC**

Element of Content Blocks of text extracted by AGORA, it can be a paragraph, a line, a letter, *etc.*

**Font Model**

In RETRO, a Font Model is an input resource composed of a thumbnail of a letterform, and an xml file describing metadata and data (*e.g.* the transcription) of the letterform. Used during Automatic Transcription process.

**METS**

Metadata Encoding and Transmission Standard METS is an XML schema provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object, and is also used for expressing the complex links between these various forms of metadata present.

**Shape**

In RETRO, a Shape correspond to a letterform. It is linked to the alto String granularity. \*

**TEI**

Text Encoding Initiative TEI is an XML Schema, build from guidelines, and describing transcription of document images