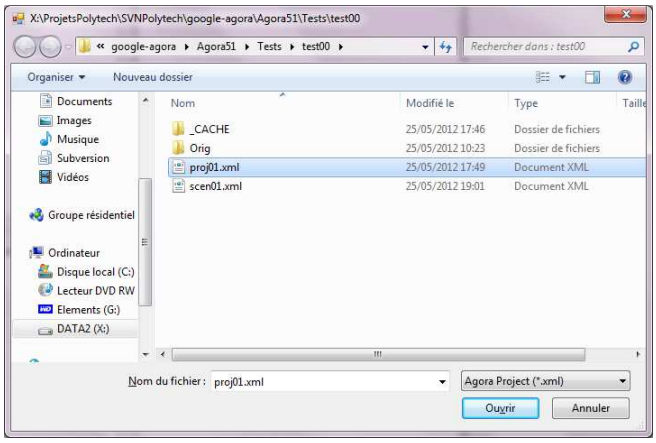
# First steps with Agora

## Introduction

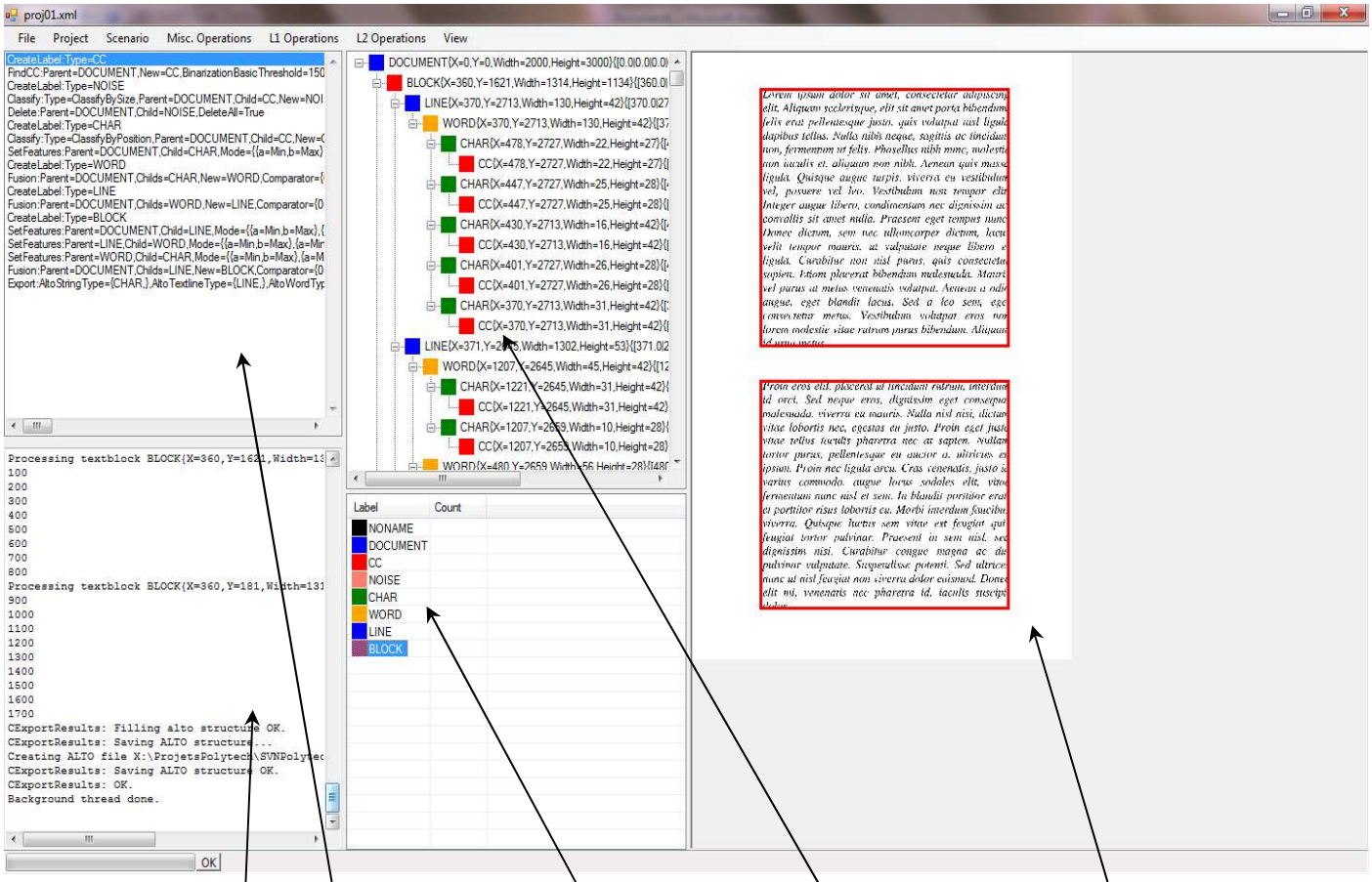
This tutorial will show how to analyze a very simple page with only 2 blocks of text. The final structure of the document will contains the following elements of content (EOC): CHAR, WORD, LINE and BLOCK. In this tutorial, we will only use the provided scenario and look at the Agora GUI and some details of each scenario step.

## Launch AGORA and open existing project "proj01.xml"

File → Open…




The project loads, and the associated scenario runs until the end. In the end, you probably obtain the following screen:




Technical feedback

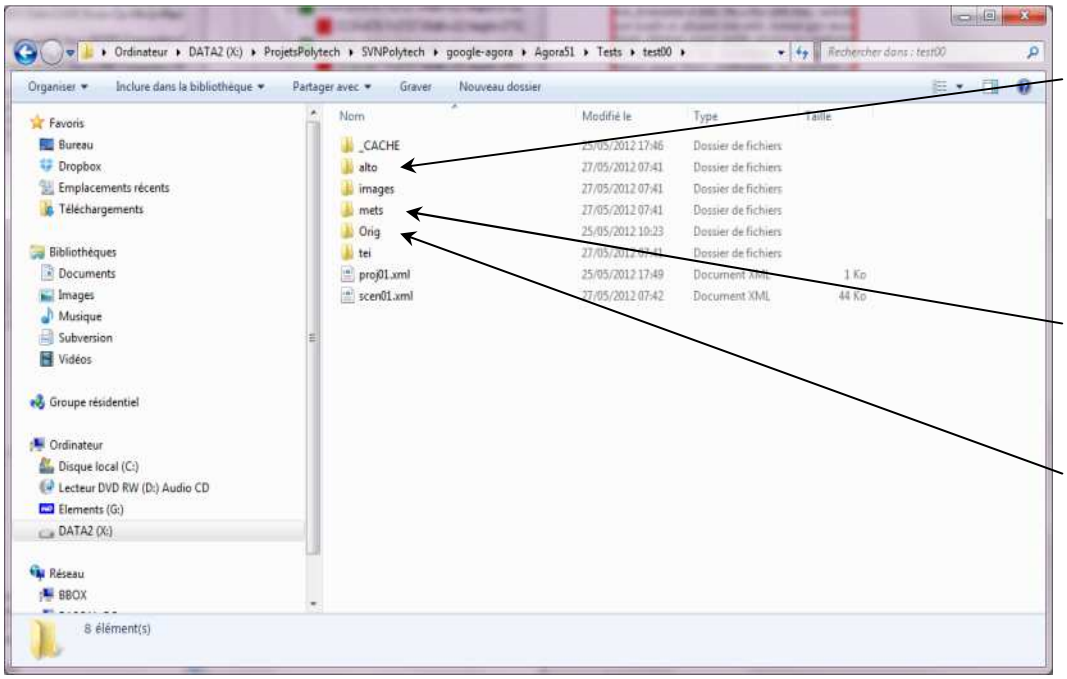
Scenario

Liste of EOC labels

Graph (tree) of EOC

Document

This scenario exports its results in ALTO format. In the project directory, directories were as follow:




Alto directory contains:
- ALTO generated files
- thumbnails of each EOC (png format) present in each ALTO file

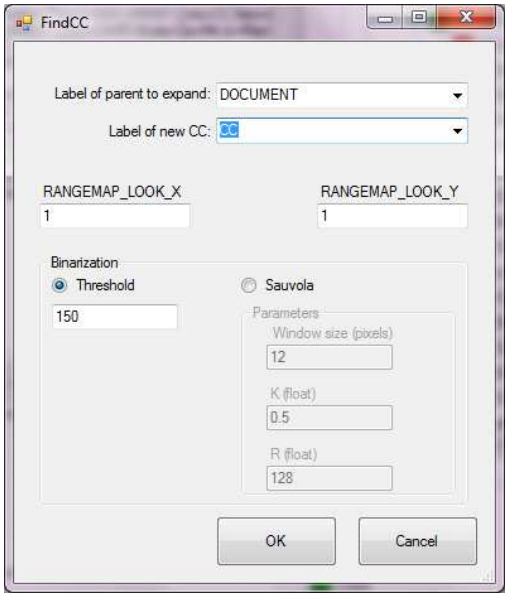
METS files. Their structure is the same as the ALTO files.

Original images (directory's name is stored in project file)

## Details of the scenario

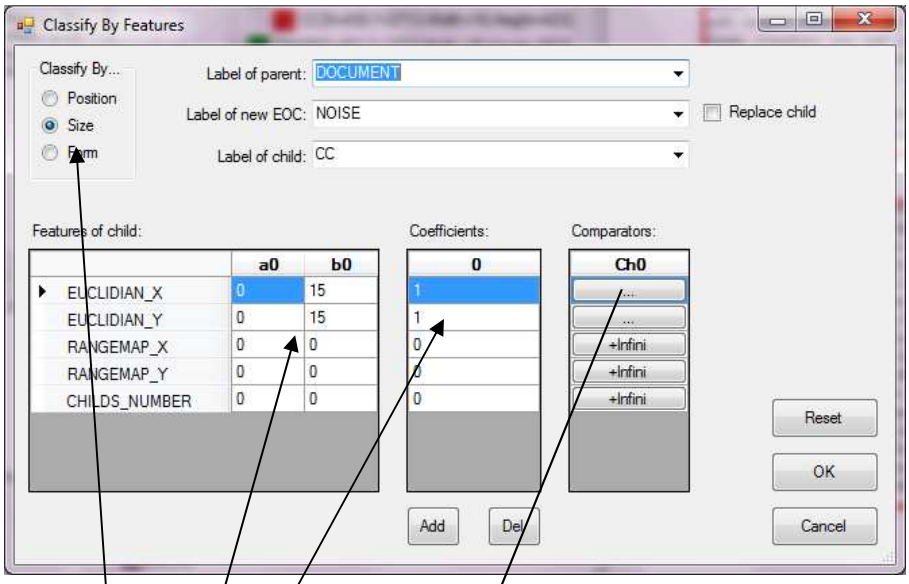
By double-clicking on a scenario step, you can visualize its parameters. This double-click is only active for "complex" steps !

### Expand an EOC : FindCC




This operator will expand any parent EOC by adding childs EOC to it. These childs EOC will be connected component identified in the parent's XY bounding box. 2 binarization pre-treatment algorithms are possible. RANGEMAP_XXX parameters will be useful for complex layouts with many blocks in the document; in this tutorial, default values are used.

## Insert new EOCs: classify by features

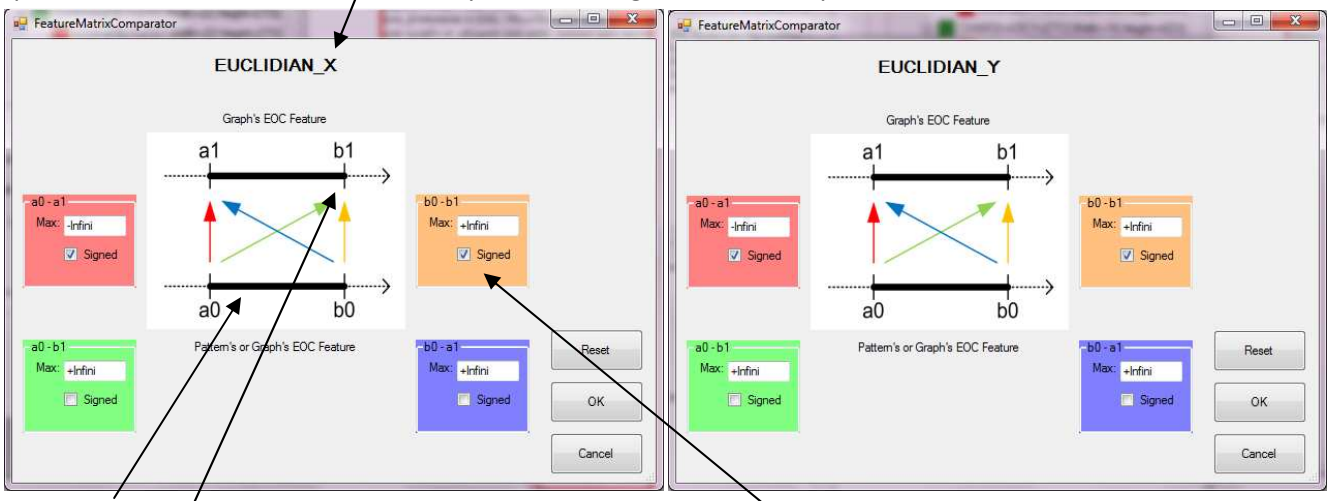



This means: a new EOC "NOISE" will be inserted between EOC "DOCUMENT" and one of its child EOC "CC" if rules are satisfied.

For this operator, rules are:

- Size of the child EOC is considered (i.e. EOC can be everywhere)
- Coefficients will be applied to each feature before comparison with comparators. Here, the "0" means simply that this corresponding feature doesn't care.
- between 0 and 15 pixels in X, and between 0 and 15 pixels in Y

Specifications of "between" are done by the following interface/comparators:




"a0" and "b0" are representing the features of the model of child pattern to extract in the image/graph. Here the child pattern is a "CC" with a0 = 0 and b0 = 15 with a "DOCUMENT" as Parent.

"a1" and "b1" are the features of the child "CC" found in the actual list of extracted CC (under "DOCUMENT") in the tree/current image. For specifying "between", we specify "a1 > a0" and "b1 < b0" thanks to the showed settings.