# Interactive indexation and transcription of historical printed books

**Jean-Yves RAMEL, Nicolas SIDERE**

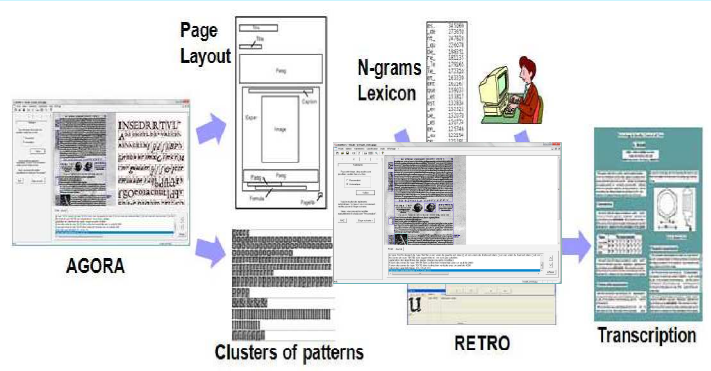
**Laboratoire d'Informatique (LI)**
Université François Rabelais, Tours
64 av. Jean Portalis 37200 TOURS, France
ramel@univ-tours.fr , nicolas.sidere@univ-tours.fr

## Context of the work

This work is part of a French Project dedicated to the preservation of cultural heritage realized in Tours city (Loire Valley). It aims to improve accessibility of historical books and to take away the barriers that stand in the way of the mass digitization. The difference with others projects comes from the very close pluri-disciplinary collaboration between specialists in Computer vision (**RFAI – LI Lab.**) and historians (specialists of the Renaissance period from **BVH team - CESR-Tours**).

## Proposed Architecture

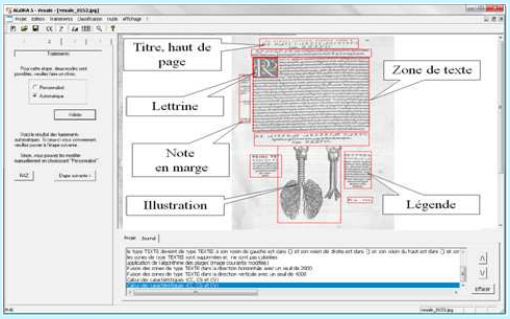
- Page Layout Analysis and content extraction with **Agora**
- **Pattern Redundancy Analysis** with clustering techniques
- Text transcription and Typography Analysis with **Retro**
- Learning from users and from available data




Genarl view of the proposed Framework of PaRADI-IT

## User-driven Analysis of images with Agora

- Used and improved since 2004 (CESR)
- Layout analysis (XML files describing the structure - Alto)
- Extraction of specific elements (textual or graphical)



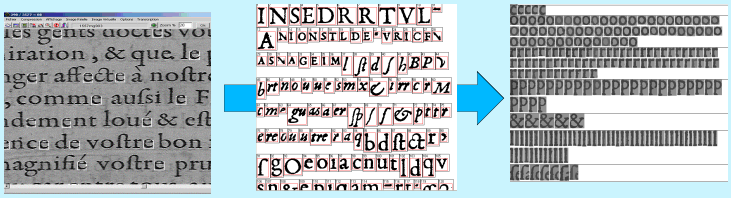

Agora software v2007: http://www.rfai.li.univ-tours.fr/pagesperso/ramel/fr/work1.html

- Extraction and indexation of ornamental materials

Image database available on: http://www.bvh.univ-tours.fr

## Pattern Redundancy Analysis

- Analyzing redundancy in images (text part) ➔ a text, ancient or not, is made up of sequences of similar patterns
- Clustering of similar patterns to create groups (classes)
- Comparison of patterns without prior knowledge about the meaning of these patterns
- Production of a minimal number of homogeneous clusters



From images to clusters of patterns (connected components or glyphs)

## Computer Assisted Transcription with Retro

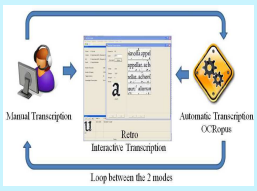
- For tagging the clusters using unicode
- Cluster visualization
- Characters (CCs) in context
- Creation (selection) of new templates

Retro software v2007 forText Transcription by tagging the clusters

- Deal with Early Modern fonts & allow **typography analysis**

## Inside a loop !

- Cooperation between manual, automatic (OCR) and contextual (dictionaries) contributions




This project is funded by a French Research Agency, University of Tours and Google Digital Humanities Award 2010