

Les Bibliothèques Virtuelles Humanistes (The Humanistic Virtual Libraries, BVH) in Tours *10 years of digitization*

Graphical analysis and
graphematics for the BVH
corpus

Oslo, CERL Seminar, October 28, 2014



The Center for Renaissance Studies in Tours: a Research team and a Faculty



- *Graphical analysis* is the first step of OCRing historical documents
- *Graphematics* is the study of graphemes in a linguistic, cognitive and semiotic approach (M.-L. Demonet at the Institut Universitaire de France: history of semiotics in the Renaissance)
- The study of the written (typographic/ printed/ handwritten) document is one of the targets of the BVH



Public needs/ Research requirements

- On-line reading and downloading: binary pdf, pdf mrc, color pdf, xml/tei, epub,... for everyone
- Semantic web and RDF (in construction)
- Collaborative annotation, crowdsourcing (expected)
- Keyword access to images and texts
- **Full-text access (raw OCR/ enhanced OCR)**
- Choice among the transcription mode (Renaissance French language):
 - Quasi-diplomatic (a « facsimilar » restitution)
 - « Digital heritage » (with regularization)
 - Modernized

Tasks (1)



- Online publishing fac-similes of books, edited during the Renaissance, and manuscripts
- Extracting images from scanned pages
- Classifying and indexing graphic elements= AGORA, RETRO (PARADIIT project), and the BaTYR database
- Acquiring significant corpora of transcribed texts (mainly in Renaissance French)
 - By keyboarding
 - By OCRing (in construction)
- Processing texts with NLP software (PhiloLogic, TXM, Analog)
- Training in Digital Humanities (2 master courses)

Tasks (2)

- Re-building libraries of authors (Rabelais, Montaigne)
 - The Rablissime and ReNom projects (Rabelais, Biblissima)
 - The MONLOE (MONtaigne à L'Œuvre)
- Analyzing inks and *ductus* for handwritten documents
 - Inks: with a spectrometer (by fluorescence), IRHT, IRAMAT (University of Orleans)
 - Ductus: computing the specificities of « hands », IRHT, LIP6 (University Paris V)
- Cataloging incunabula: the CRII (Catalogues régionaux des Incunables Informatisé) project, and the reconstruction of provenances.

The BVH project and its partners

The OAI-PMH repository of the BVH is harvested by:

- The Bibliothèque nationale de France – Gallica
- Europeana – the BVH is a member of the Europeana Network
- Isidore: the platform for the Humanities launched in December 2010 by the CNRS « Huma-Num » Infrastructure

The BVH project is a member of :

- The TEI consortium
- The Centernet network
- The ADHO (digital humanities association)

The BVH project is:

- A resource center for the digital humanities of the Maison des Sciences de l'Homme Val de Loire (Orléans-Tours)
- A fac-simile provider and an aggregator for the libraries in Region Centre
- The head of the CAHIER consortium (Corpus d' Auteurs pour les Humanités: Informatisation, Edition, Recherche) since September 2011
- A partner of the Biblissima « Equipment of Excellence » since 2013
- A winner of the SUCCEED Award, may 2014



Classifying and indexing graphic elements:
PaRADIIT (2011-2013)
Pattern Redundancy Analysis for Document
Image Indexation and Transcription

Prof. Jean-Yves Ramel

Frédéric Rayar, Ph. D. Student (computing
science)

Remi Jimenes, Ph. D. Student (historian of
the book and typography), CESR-BVH



Testing OCRs (2011-2013)

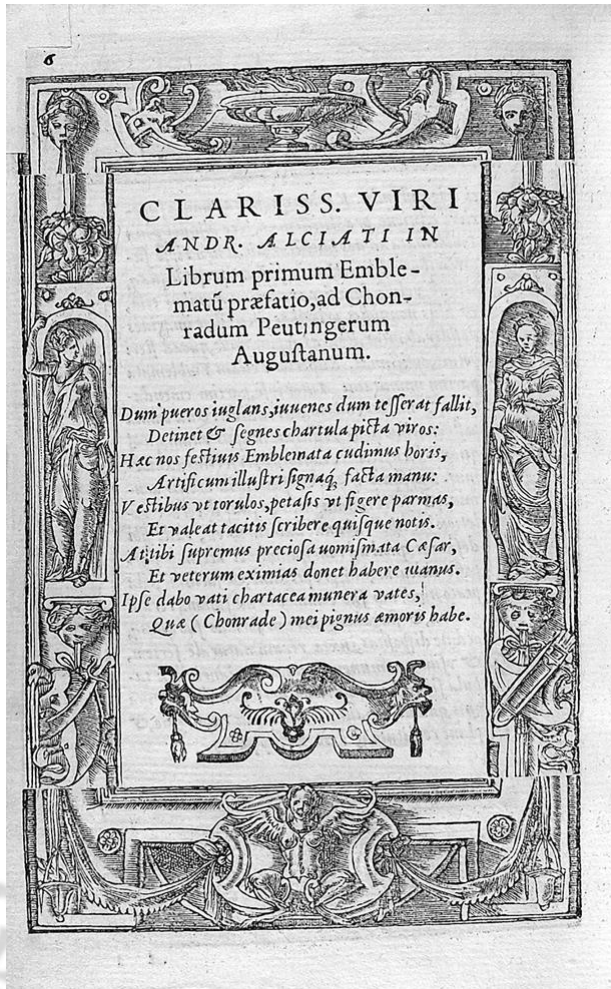
- FineReader 10 (ABBYY)
- IMPACT project: OCRopus-Tesseract, and the Rouen-BnF Ph. D. student
 - No improvement for Renaissance French
- Omnipage with/ without specific dictionaries
- BIT-Alpha (Tomasi company) in Wolfenbüttel and in Tours. Time consuming training – not competitive in comparison to keyboarding
- AGORA-RETRO (U. of Tours, JY Ramel), Google award for the PARADIIT project

Graphical analysis (Frederic Rayar, Computing Science Department, University of Tours)

Overview

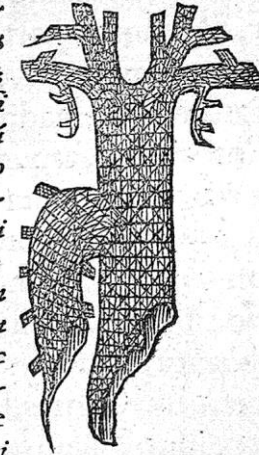
- Introduction
- Pattern Redundancy
- RETRO

Examples of Renaissance Books Issues



nis administrationi utiles fore cognouit. Q

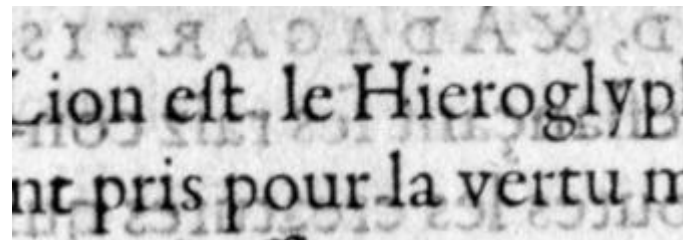
Vena caue portio à dextro cordis sinu ad iugulū usq; conscendens, qua fibrarum in uenarū corpore naturā doctrinae gratia studiose finximus. Si forte interim ramorum hic passim abtruncatorum ratione experis, hanc



Asscritia & non peculiaris uena tunica delineationi quinti ca.

Capitis fini subijcienda confer, in illa quid D, F, G, H, I, K, N, S, T, & a indi

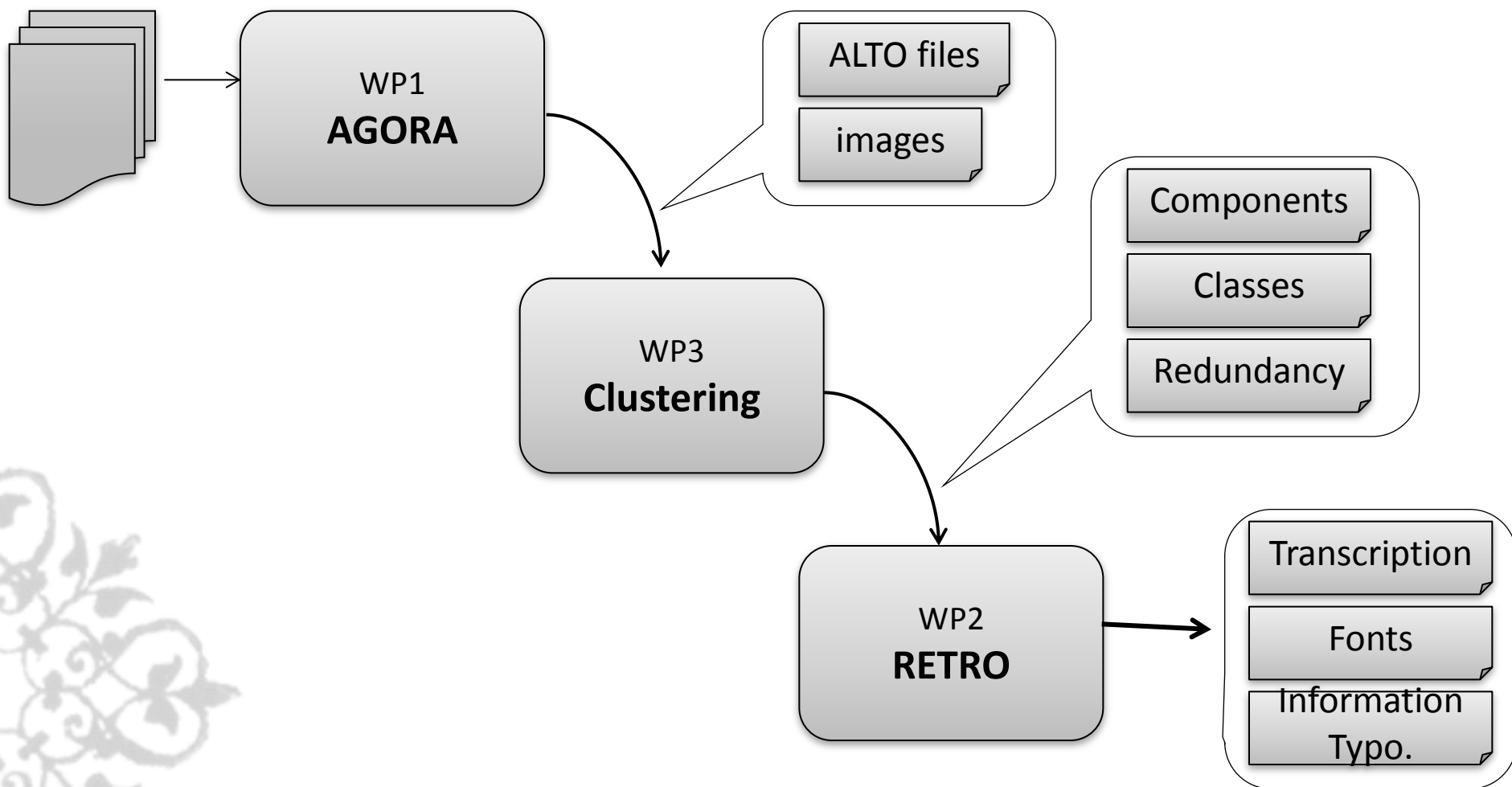
ad se admittit sanguinem ad subsequens operatione statim uniram ex aliis uenaribus uenica conf appositè c rem tunicae partis part go tramite



Empereur
d'Elephant

Projet PaRADIIT (funded by Google and IUF)

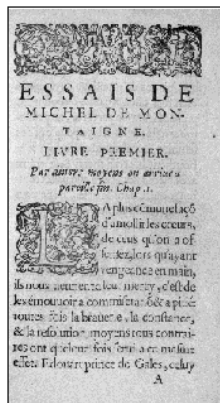
Pattern Redundancy Analysis for Document Image Indexation and Transcription



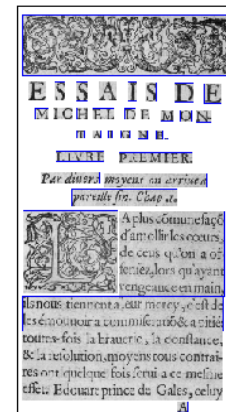
The PaRADIIT Workflow with AGORA and RETRO



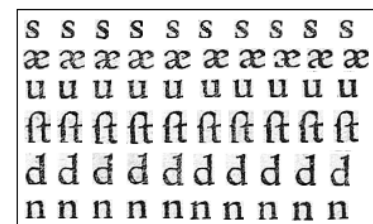
Livre



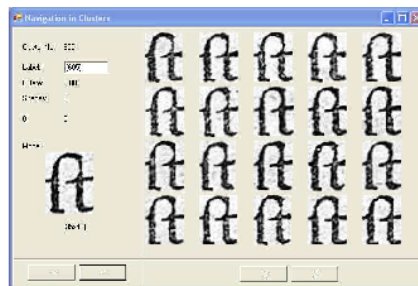
Page Numérisée
(Image)



Analyse de la
Mise en Page



Groupements
de Formes

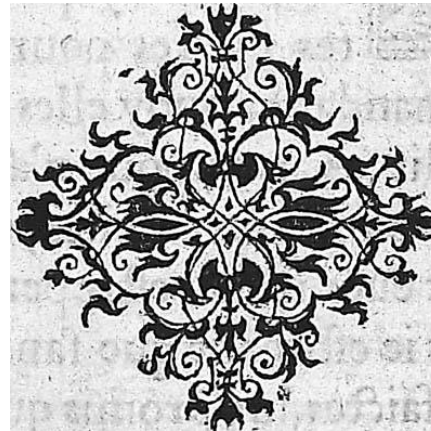
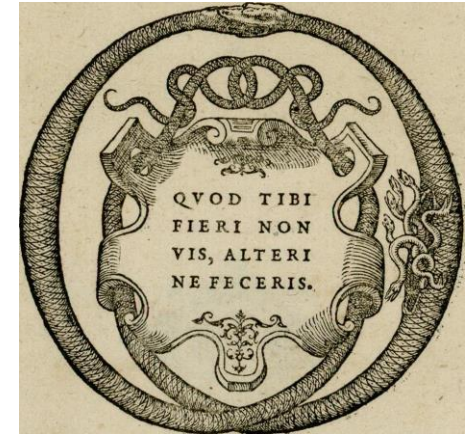


Transcription automatique
et/ou assistée



Texte

Ornaments extracted by AGORA. BaTYR : Base de Typographie de la Renaissance (Renaissance Typographic database)



The PaRADIIT objectives

Incremental and interactive analysis of graphics

User-driven Layout analysis or content extraction →
AGORA software

Computer assisted transcription → RETRO software

Getting information about typography to improve
transcription results by a redundancy analysis →
RETRO

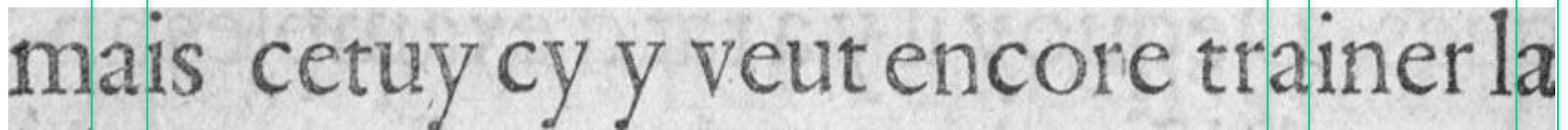
Expected Results → a software suite

An open source forge for AGORA and RETRO with

- An improved clustering method (pattern redundancy analysis)
- An interactive and collaborative transcription system
- New functionalities concerning typographical studies: creation of families of fonts in order to generate learning datasets

Pattern Redundancy Description

- Observation:
A text, ancient or not, is made up of sequences of similar patterns (e.g. letters, ...)
- Idea:
Analyzing **text redundancy** at image level



mais cetuy cy y veut encore trainer la

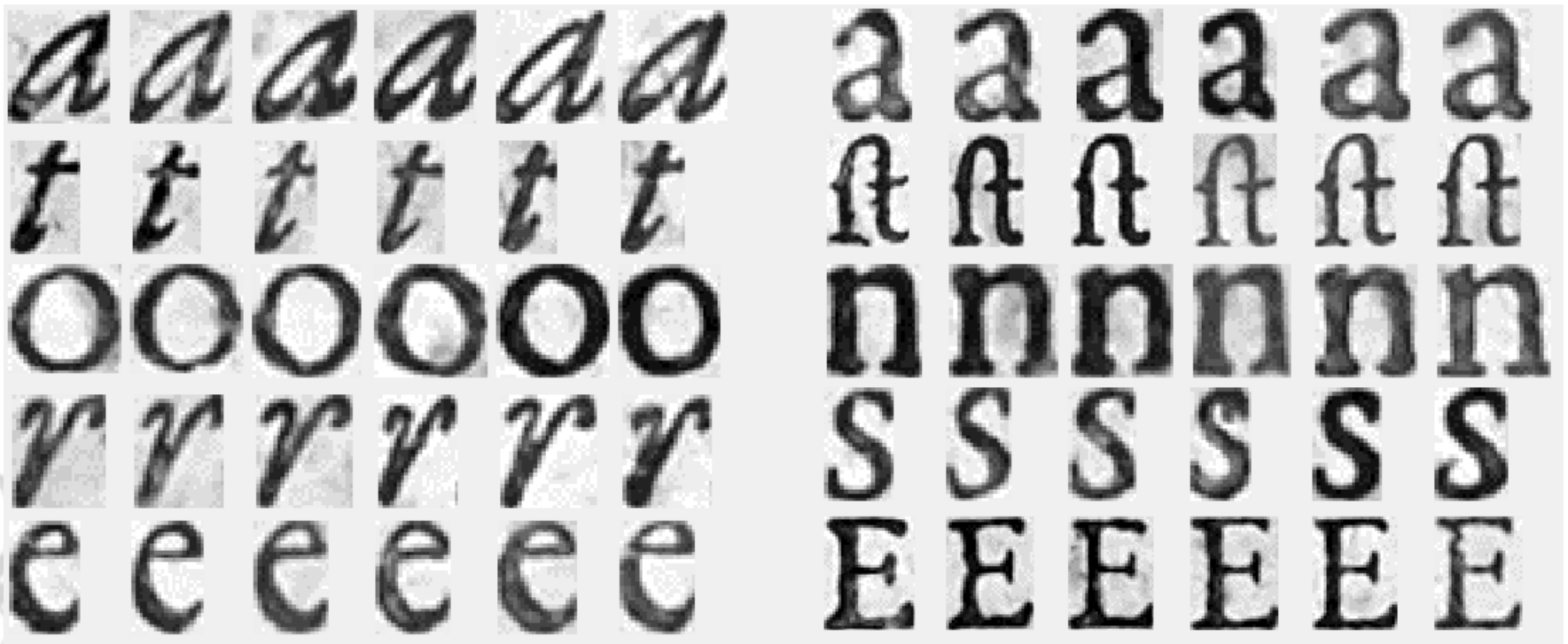
Pattern Redundancy Description

- Methods:
 - **Clustering** of extracted patterns to create classes of similar patterns
 - Comparison of patterns (matching techniques)
 - *N.B.* No prior knowledge of the meaning of these patterns
- Constraints:
 - Producing very homogeneous clusters
 - Producing a minimal number of clusters

Pattern Redundancy

Pattern clustering

Some (ideal) results on letters ...



RETRO

Visualization: Shape Context

The screenshot displays the RETRO 2012 software interface. The main window, titled "RETRO 2012", has a menu bar with "File", "Page", "Clustering", "Transcription", "Results", "Typography", and "Help". Below the menu bar, there are tabs for "Clusters" and "Retro". The main area shows a grid of handwritten characters. The first character in the top row, a lowercase 'e', is highlighted with a blue border. A dialog box titled "Navigation in Cluster" is open, showing the following information:

- Cluster No: 294
- Nb Shapes: 209
- Label: [294]


Below this information is a small image of the character 'e' with left and right navigation arrows. To the right of the dialog box is a "Shape Context Visualisation" window, which shows a grid of characters. The top row contains four 'e' characters. Below it, a larger image shows the word "vaiffeaus &" in a cursive script, with a red square highlighting the 'e' in "vaiffeaus". Below the grid are "Previous" and "Next" buttons. At the bottom of the main window, there are checkboxes for "Labelized" and "Not Labelized", both of which are checked. The footer of the window reads "Retro 2012 - Copyright © RFAI, LI Tours, 2011-2013".

RETRO

Transcription: manual labelling

The screenshot shows the RETRO 2012 software interface. The window title is "RETRO 2012". The menu bar includes "File", "Page", "Clustering", "Transcription", "Results", "Typography", and "Help". The "Transcription" menu is active, showing sub-tabs for "Manual Transcription", "Clusters", and "Retro".

On the left side, the following information is displayed:

- Cluster N°: 294
- Nb Shapes: 209
- Transcription: [294]
- Mark as Noise button
- Model: 

On the right side, four horizontal panels show snippets of handwritten text with a red square highlighting a specific character:

- Panel 1: "vaiffeaus &" with the 'e' highlighted.
- Panel 2: "ryaprefter" with the 'e' highlighted.
- Panel 3: "is & demy cu" with the 'e' highlighted.
- Panel 4: "s carrefours" with the 'e' highlighted.

At the bottom, there are navigation buttons: "<<" and ">>" for the left column, and "<<" and ">>" for the right column.

RETRO 2012 - Copyright © RFAI, LI Tours, 2011-2013

Typographic Studies Font Model Indexation Tool

Towards Automatic Transcription Systems

The screenshot shows a software window titled "Model Family Visualisation". The main content area displays the following information:

- File path: D:\RENOM\Ronsardi_118\&2_TE_L_I_OF_GPR.png
- Page number: 2 / 333
- Metadata fields:
 - Type: Text
 - Alphabet: Latin
 - Family: Italic
 - Subfamily: Old-face
 - Body height: Great Primer
 - Thickness: (empty field)
- Bibliographical references: Vervliet-Conspect
- Unicode: U+0026
- Transcription: &
- Engraver: Garamont, Claude
- Small Cap:
- Comments: calligraphique

A central image shows a calligraphic character, which is a stylized ampersand (&). The interface includes navigation arrows (<< and >>) and a "MetaData" button at the bottom left.

Automatic
Transcription

Processing cluster 5/326

Promotion

Open resources

- Available software :
 - AGORA 6.3 (Released on 2013/12/05)
→ OPEN SOURCE
 - RETRO 2012 - v2.5.1 (Released on 2012/11/25)
→ OPEN SOURCE
 - User Guide & Tutorials
 - Test data projects
- BaTyR (Base de Typographie de la Renaissance - A Database for Renaissance Typography, by Rémi Jimenes) :
 - 27 000 extracted ornaments
 - Database of fonts and sets of characters (*Work in progress*)

Promotion

Open resources

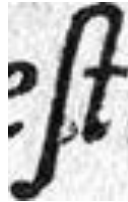
- Typographic font packages:
 - The 'Gering' Pica-Roman [R80] or Cicéro (1478) (cf. [Vervliet2010] N° 55)
 - Garamont's Great Primer Roman [R118] or Gros-romain (1549) (cf. [Vervliet2010] N° 119)
 - Vérard Gothic Bâtarde Great Primer Roman
 - Font family of 333 items from “*Les Oeuvres de Pierre de Ronsard*”
 - Used in ReNom Project
 - For Word Spotting purpose

→ <https://sites.google.com/site/paradiitproject/>

Some examples: towards the PICA project



U+FB06



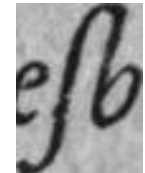
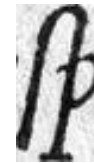
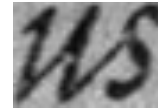
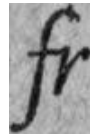
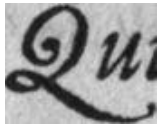
U+FB05

BUT



M+EEC5

The « ligatures » in Ronsard's *Œuvres* (Paris, Buon, 1623) lack in Unicode and MUFI :



Polysemy of glyphs:



= cum- (*cumulare*) , com- (*communion*), con- (*conqueror*)...



= ser- (*servus*), sen- (*sensus*)...

Online Historical Book Explorer

PARADIIT Project

An Online Historical Book Explorer

The PaRADIIIT (*Pattern Redundancy Analysis for Document Image Indexing and Transcription*) research project focused on layout analysis, text/graphics separation, Optical Character Recognition (OCR) and text transcription processes dedicated to old and precious books.

The originality of the work relies upon the analysis and exploitation of pattern redundancy in documents to enable the efficient indexing and quick transcription of books and the identification of typographic materials.

You can obtain more information about PaRADIIIT project at <https://sites.google.com/site/paradiitproject/>

This Online Book Explorer allows the users to easily interact with the digitized versions of historical books. Using a web interface, the users can navigate, visualize, search and get annotated copies of some specific elements of content extracted from historical books. They can also exploit and even enrich the available meta-data obtained with the help of PaRADIIIT indexing tools by adding their own annotations on all the parts of the books.

Please, consult the [Help](#), to quickly understand how to use the Book Explorer.

Partners



Requirements



Contacts:

[Frédéric RAYAR](#)
[Jean-Yves RAMÉL](#)





What format of output? METS/ ALTO and XML Files
What about the big data? The hundred books of the BVH
The Slow Flow...