Contributions au projet BVH
Aide à l'indexation et la transcription d'ouvrages numérisés

- Laboratoire d'informatique , Université de Tours
  - JY Ramel, N. Ragot, T. Brouard, M. Delalendre, S. Barrat,

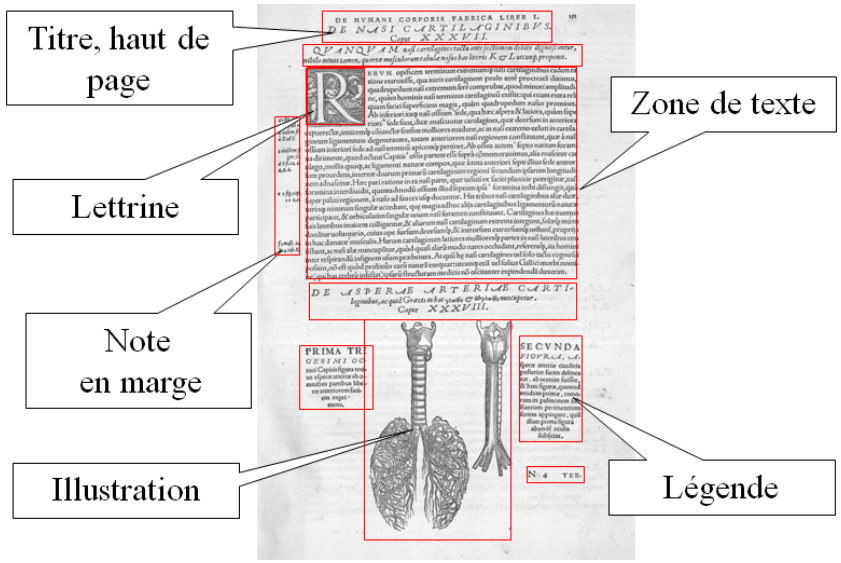
- **En partie en collaboration avec le LITIS Rouen et la BNF**

Plan

- Extraction d'éléments de contenus et Indexation visuelle
  - Segmentation des pages
  - Extraction d'éléments de contenus
- OCR et transcription assistée
  - Exploitations du logiciel Retro (transcription, analyse typographique)
  - OCR robuste
  - Evaluation des OCR
- Contribution à l'architecture Web BVH
- Travaux futurs

Extraction d'éléments de contenus et indexation visuelle

- **Logiciel AGORA**
- Analyse de structures guidée par scénario
- Scénario construit par l'utilisateur
- Analyse incrémentale
- Rien de prédéfini (adaptation aux besoins)
  - Interactivité
  - Assistant



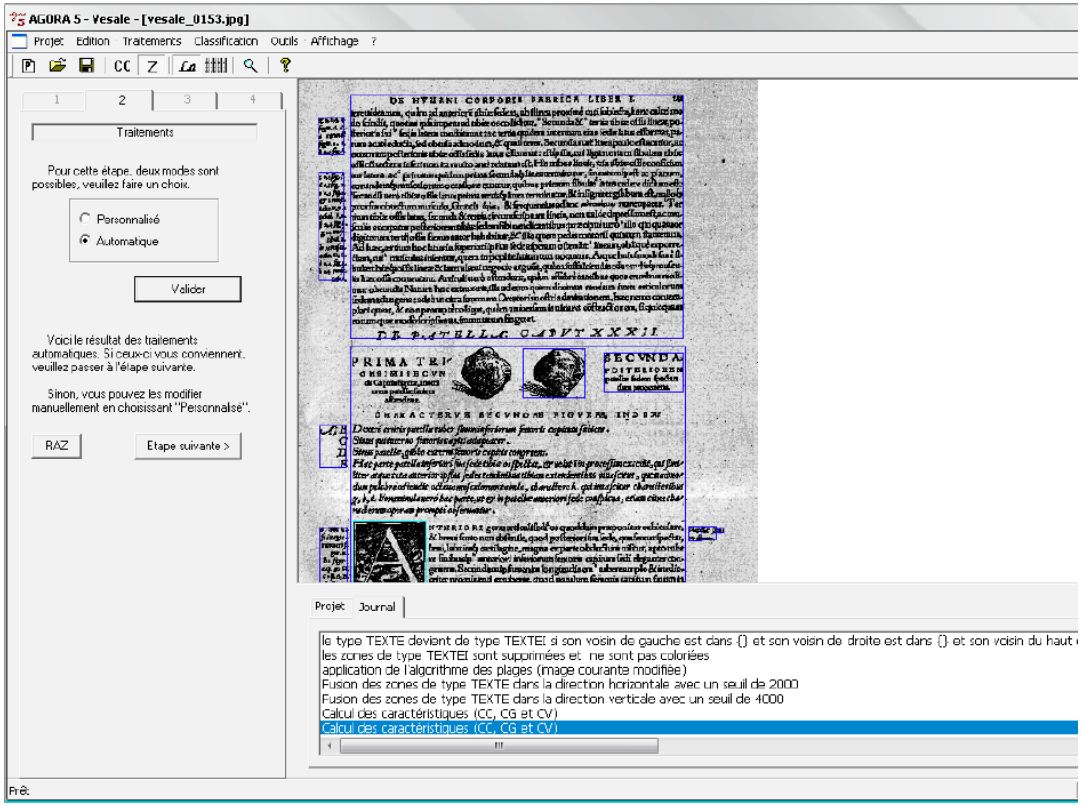
- **Bases de lettrines (+de 15000) et de marques typographiques**

- **Bases de portraits (+ de 1500)**

Voir sur http://www.bvh.univ-tours.fr

Extraction d'éléments de contenus et indexation visuelle

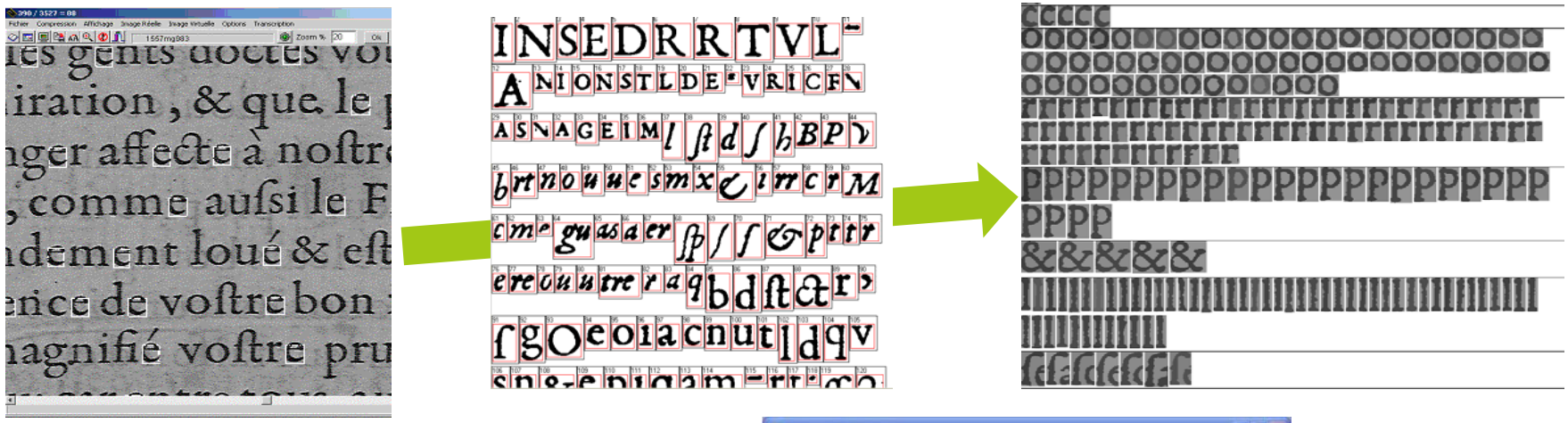
- **Logiciel AGORA**
- Utilisé depuis 2004 (CESR)
- Analyse de structures
- Extraction de contenus
- Etudié en M2Pro



- A améliorer encore…
  - La segmentation des paragraphes en lignes et des lignes en mots est une étape importante qui doit être améliorée
  - La gestion de la ponctuation et des accents doit être réalisée

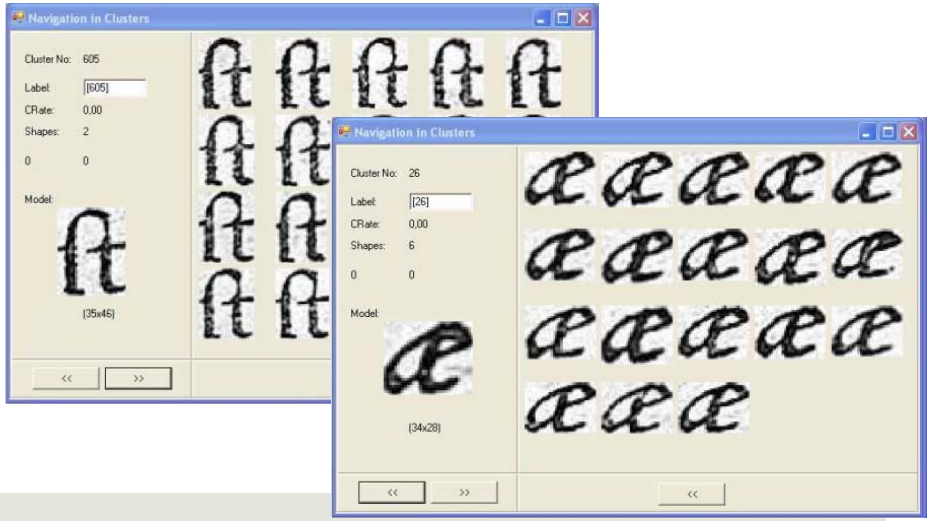
OCR et Transcription

RETRO : Analyse de redondance de formes



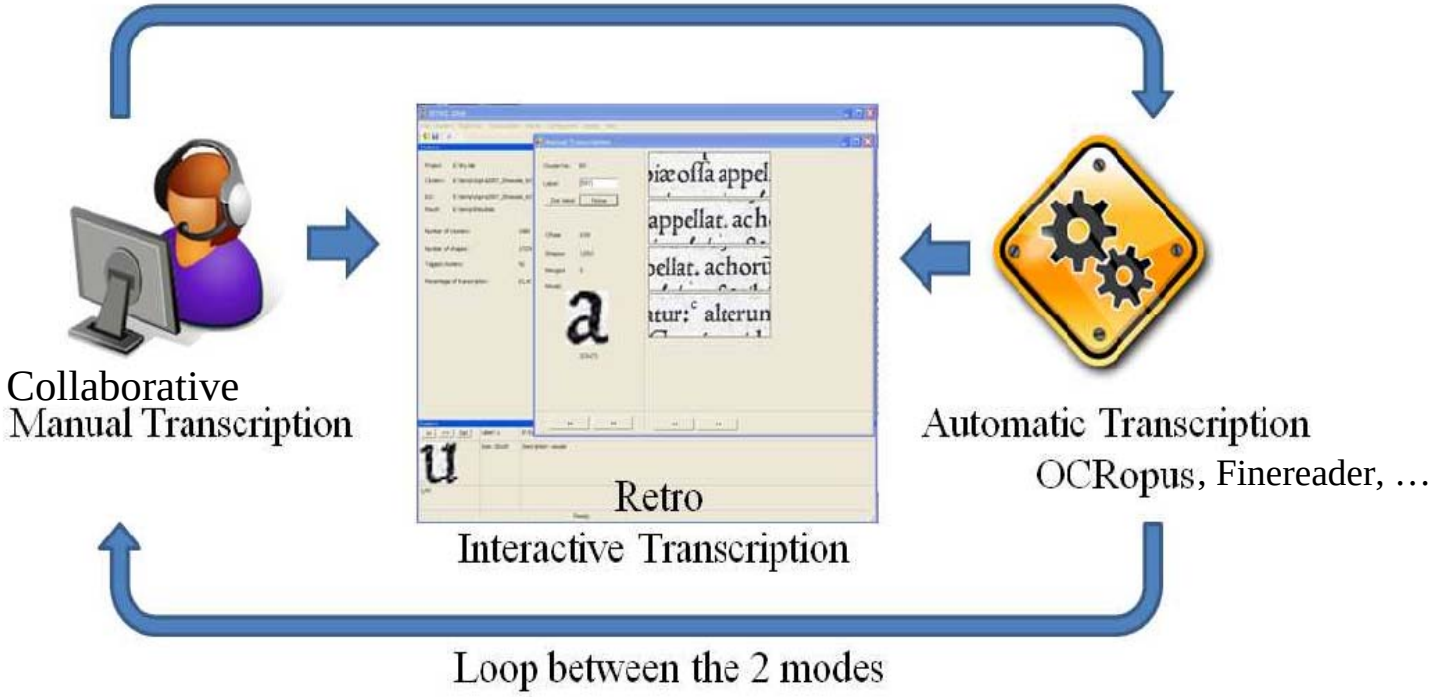
- Constitution d'une base de polices anciennes
- Reconnaissance automatique et adaptation des moteurs de transcription
- Exploitation sur les éléments graphiques

- Suivi de matériels typographiques ?
- ...



OCR et Transcription

RETRO : Analyse de redondance et transcription

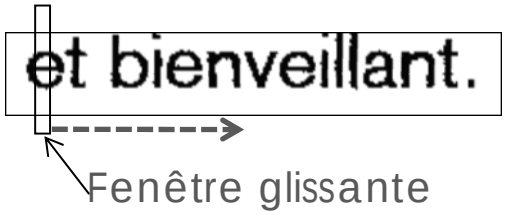


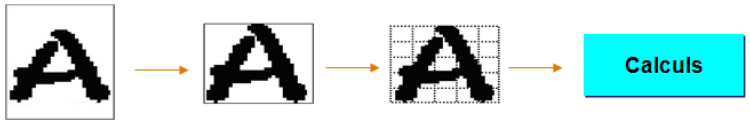
es_	345069
_de	273650
nt_	247820
_qu	226078
de_	198351
re_	181135
_le	179165
le_	172310
et_	163339
ent	162167
que	158033
_et	153817
est	132834
_en	132523
ue_	132070
_es	130774
en_	125746
_au	122194
on_	121291
ne_	115442
_la	115189

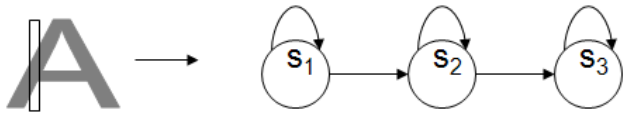
- Transcription interactive et collaborative
- Création et utilisation de dictionnaires, lexiques et modèles de langage adaptés

OCR robuste

- Caractéristiques utilisées :
  - Pas de segmentation de la ligne en mots
  - Fenêtre glissante
  - Grille + Gradient







- Apprentissage
  - Polyfontes + modèles de bruits

Famille	Base d'aprentissage		Base de test	
Humanes	Berkeley Old Style	*Berkeley Old Style Italic*	Centaur	*Centaur Italic*
Garaldes	Garamond	*Garamond Italic*	Caslon Old Face	*Caslon Old Face Italic*
Réales	Times New Roman	*Times New Roman Italic*	Perpetua	*Perpetua Italic*
Didones	Bodoni	*Bodoni Italic*	Walbaum	*Walbaum Italic*
Mécanes	Rockwell	*Rockwell Italic*	Lubalin	Lubalin Italic
Linéales	Futura	*Futura Italic*	Arial	*Arial Italic*
Incises	Optima	*Optima Italic*	Friz Quadrata	*Friz Quadrata Italic*
Scriptes	Mistral	*Shelley Allegro Script*	Monoline Script	Brush Script
Manuaires	Graphite MM	Broadway	OCR-A	Ondine
Gothiques	Old English	*Old English Italic*	Clairvaux	Fraktur

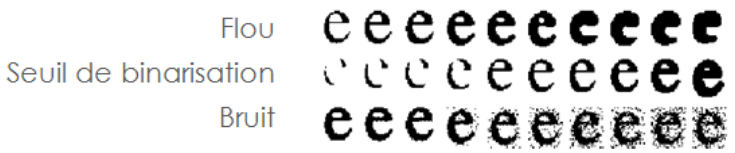
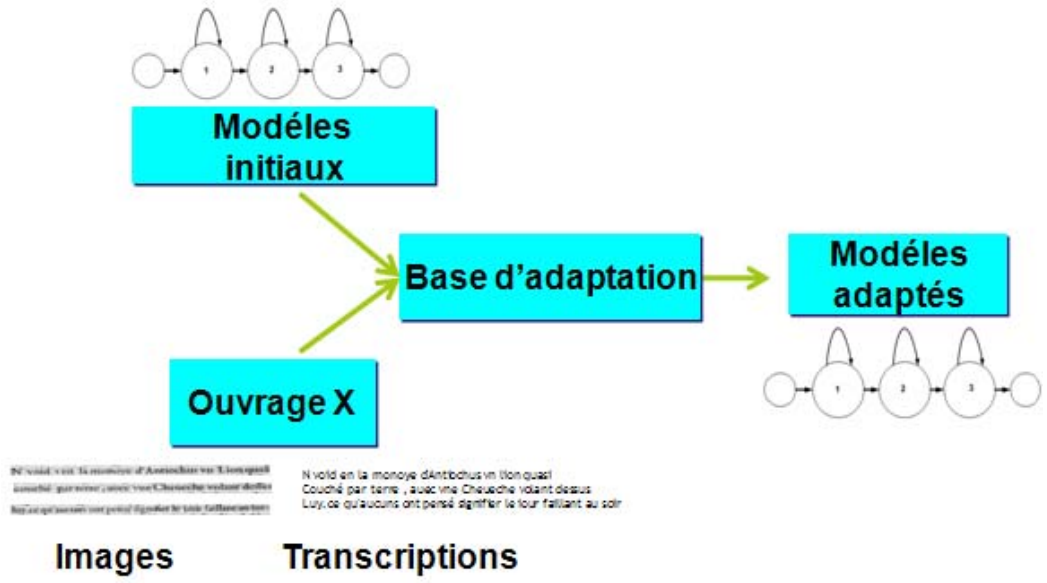


Illustration du modèle de Baird

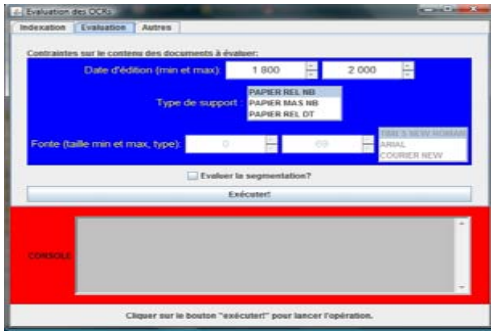
OCR robuste

- Reconnaissance par HMM
- Adaptation : modification des parameters des HMM pour mieux modéliser un nouveau type de données
- Adaptation supervisée



Outils pour l'évaluation d'OCR BNF

- Conversion de format
  - Manipulation de fichiers XML
  - Extraction de contenu textuel et formatage
- Calcul de distance d'édition entre 2 textes
  - Outil ISRI – Université du Nevada
- Comparaison entre 2 structures de documents
  - Evaluer la qualité de la segmentation
    - Outil Pink Panther – Université du Maryland
  - Rapport détaillé (1 rapport par OCR)
  - Niveau mot et niveau caractère
  - Taux pour chaque caractère ou par type de caractères
  - Confusions




```
<?xml version="1.0" encoding="UTF-8" ?>
- <Collection>
  - <Book>
      <BookID>5406202</BookID>
      <genre>MONOGRAPHIE</genre>
      <dateEdition>1879</dateEdition>
      <agentOperation>Groupement Safig</agentOperation>
    - <Page>
        <resolution>0300,0300</resolution>
        <supportOrigine>PAPIER REL NB</supportOrigine>
        <numeroPage>6</numeroPage>
      - <TextLine>
          <TextLineID>0</TextLineID>
          <TextLineFontFamily>TIMES NEW ROMAN</TextLineFontFamily>
          <TextLineFontSize>19</TextLineFontSize>
          <TextLineLeft>528</TextLineLeft>
          <TextLineRight>1017</TextLineRight>
          <TextLineTop>1140</TextLineTop>
          <TextLineBottom>1197</TextLineBottom>
          <TextLineContent>ARMORIAL</TextLineContent>
          <TextLineNbCharacters>8</TextLineNbCharacters>
        </TextLine>
      - <TextLine>
          <TextLineID>1</TextLineID>
          <TextLineFontFamily>TIMES NEW ROMAN</TextLineFontFamily>
          <TextLineFontSize>30</TextLineFontSize>
          <TextLineLeft>207</TextLineLeft>
          <TextLineRight>1332</TextLineRight>
          <TextLineTop>1327</TextLineTop>
          <TextLineBottom>1416</TextLineBottom>
          <TextLineContent>DU NIVERNAIS.</TextLineContent>
          <TextLineNbCharacters>13</TextLineNbCharacters>
        </TextLine>
      - <TextLine>
          <TextLineID>2</TextLineID>
```

Outils pour l'évaluation d'OCR BNF

- Ouvrage: Recueil des antiquités Gauloises et Françoises
- Lexique Renaissance: 150 mille mots

	Omnipage Pro 17	Fine Reader Pro
Sans lexique	89.53%	
Lexique renaissance	89.81%	
Lexique moderne	89.82%	89.36%

- Trés forte influance de la segmentation et du lexique (moderne)

Ouvrages	Omnipage segment. Normale + lexique	Omnipage segment. Ocropus + lexique	OCR **sans lexique**	OCR adapté **sans lexique**	OCR – lignes "propres" sans lexique	OCR **adapté** – lignes "propres" sans lexique
Antiquités Gauloises	89.82%	85.93%	86.08%	88.51%	91.78%	94.98%
Expédition chrestienne	86.48%	61.25%	67.82%	75.61%		
Les treselegantes annales	85.6%	73.92%	76.84%	80.7%		
Les histoires de Diodore	90.19%	83.82%	83.36%	85.32%		

Architecture Web BVH

- Encadrements de stages et projets étudiants (Polytech Tours)
- Choix des technologies utilisées
  - Moteurs d'indexation XTF
  - Technologies d'encodage (TEI, METS, ALTO)
  - Visualisation d'images (zoom, couplage texte-image)
- Mise en place sur les serveurs BVH

Travaux futurs

- Couplage RETRO (transcription interactive) et logiciel OCR (reconnaissance des clusters)
- Exploitation de connaissances linguistiques (lexiques, statistiques) - modèle de langages (n-gram).
- Amélioration de la segmentation de texte (Ligne, mot, ponctuation, accent, ...)
- Détection de régions d'interet dans les images de documents
- Google European Digital Humanities Award 2010-2011
- ANR Digidoc 2011-2014
  - Numérsiation cognitive (LIRIS Lyon)
  - Prédiction de résultats OCR (LITIS Rouen, BNF)

Questions ?