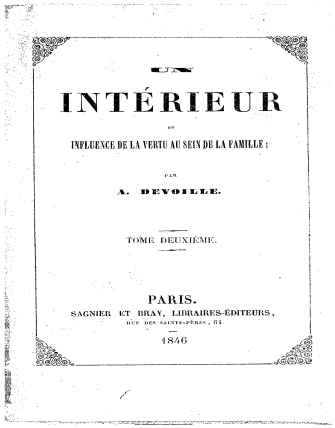
???: évaluation des résultats d'OCRisation

- Laboratoire LITIS, Université de Rouen
- Laboratoire d'Informatique de l'Université de Tours
- Bibliothèque nationale de France

Données

- Images de documents numérisés
  - Fichiers XML
    - Vérité-terrain (format ALTO)
      - Résultats d'OCRisation (ABBYY Fine Reader)



UN

INTÉRIEUR

DE

INFLUENCE DE LA VERTU AU SEIN DE LA FAMILLE ;

PAR

A. DEVOILLE.

TOME DEUXIÈME.

PARIS.

SAGNIER ET BRAY, LIBRAIRES-ÉDITEURS,

RUE DES SAINTS-PÈRES, 15.

1846

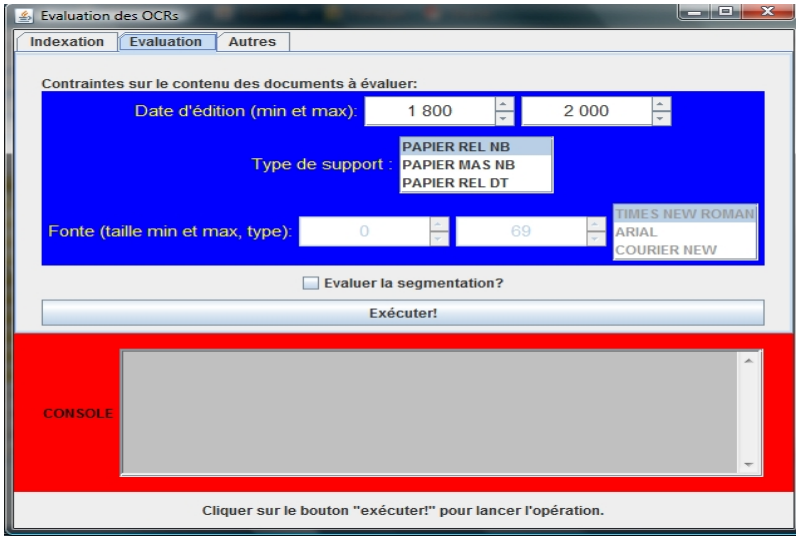

```
- <IMAGE Name="2753102616149 00002" PageIndex="1">
  <RAW_ANALYZE/>
  <BATCHER Date="26/01/2009 14:38:39" Version="1.0.0.0" EngineType="ABBYY" ReBatch="False">
  - <PLUGINS>
    - <PLUGIN Name="OriginalImage" Type="Imaging" UseBelow="False" EngineType="ABBYY">
    - <PLUGIN Name="OriginalBlocks" Type="Block" UseBelow="False" EngineType="ABBYY">
      <BLOCKS/>
      <EXTRA_BLOCKS/>
      </PLUGIN>
      <PLUGIN Name="Isako-Xml-Exporter" Type="Export" UseBelow="False" EngineType="ABBYY"/>
    </PLUGINS>
  </BATCHER>
  - <USER_PARAMS>
    - <VISIBLE>
      <PARAM Key="s1" Caption="---------- MARGES ----------" Value="" Default="False" Type="Separator"/>
      - <PARAM Key="Marges" Caption="Marges" Value="" Default="False" Type="List">
        <PARAM Key="topMargin" Caption="marge Haut " Value="alto.topmargin" Default="False" Type="List"/>
        <PARAM Key="leftMargin" Caption="marge Gauche " Value="alto.leftmargin" Default="False" Type="List"/>
        <PARAM Key="rightMargin" Caption="marge Droite " Value="alto.rightmargin" Default="False" Type="List"/>
        <PARAM Key="bottomMargin" Caption="marge Bas " Value="alto.bottommargin" Default="False" Type="List"/>
      </PARAM>
      <PARAM Key="printSpace" Caption="Corps de la page " Value="DocUnit.End" Default="False" Type="List"/>
      <PARAM Key="s2" Caption="---------- BLOC ----------" Value="" Default="False" Type="Separator"/>
      - <PARAM Key="Bloc" Caption="Bloc" Value="" Default="False" Type="List">
        <PARAM Key="table" Caption="Table " Value="alto.Table" Default="False" Type="List"/>
        <PARAM Key="illustration" Caption="Illustration, photo, carte, diagramme " Value="alto.illustration" Default="False" Type="List"/>
        <PARAM Key="graphicalElement" Caption="Decorations et autres elements graphiques " Value="alto.graphicalelement" Default="False" Type="List"/>
      </PARAM>
      <PARAM Key="s3" Caption="---------- NON-TEXTE ----------" Value="" Default="False" Type="Separator"/>
      - <PARAM Key="Non-texte" Caption="Non-texte" Value="" Default="False" Type="List">
        <PARAM Key="formula" Caption="Formule (mathématique, chimique) " Value="alto.formula" Default="False" Type="List"/>
        <PARAM Key="carte" Caption="Carte " Value="alto.carte" Default="False" Type="List"/>
        <PARAM Key="musicScore" Caption="Partition de musique " Value="alto.musicscore" Default="False" Type="List"/>
      </PARAM>
      <PARAM Key="s4" Caption="---------- ALERTES ----------" Value="" Default="False" Type="Separator"/>
      - <PARAM Key="Alertes" Caption="Alertes" Value="" Default="False" Type="List">
        <PARAM Key="illegible" Caption="Illisible " Value="alto.illegible" Default="False" Type="List"/>
        <PARAM Key="unworkable" Caption="Non-traité " Value="alto.unworkable" Default="False" Type="List"/>
      </PARAM>
    </VISIBLE>
    <HIDDEN/>
  </USER_PARAMS>
</IMAGE>
```

Outils utilisés

- Manipulation de fichiers XML :
  - Conversion de format
  - Extraction de contenu textuel
- Calcul de distance d’édition entre deux textes
  - OcrTk, institut ISRI, Université du Nevada
- Comparaison entre deux structures de documents
  - Outil Pink Panther

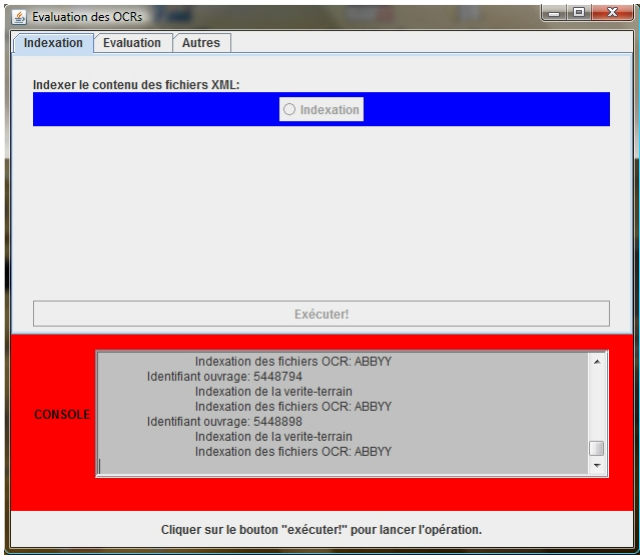
Interface graphique

- Gestion des outils d'évaluation
- Evaluation selon plusieurs critères
  - Date d'édition
  - Taille et type de fonte
  - Type de support



Utilisation (1)

- Indexation
  - Récupération de l'information utile dans les fichiers XML
    - Texte, coordonnées des blocs
    - Titre, année d'édition
  - Fichier index
    - Collection < livre < page < ligne
  - Effectuée une seule fois



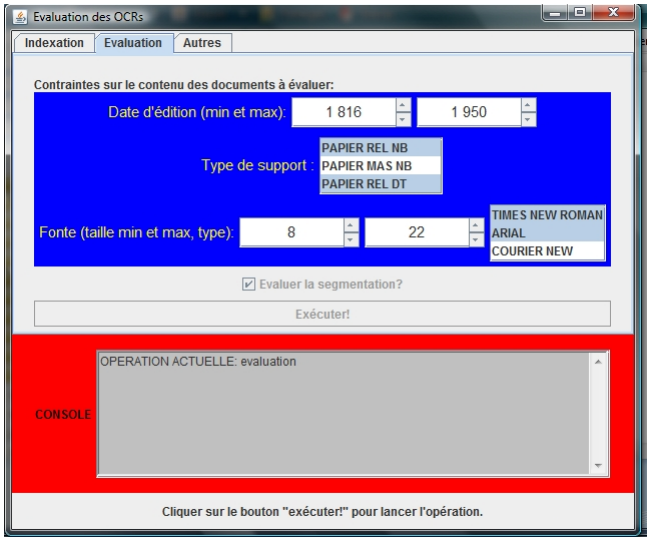
Index

- Fichiers XML
- Un fichier vérité-terrain
- Un fichier par système d'OCR

```
<?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>
```

Utilisation (2)

- Evaluation
  - Comparaison de 2 index (vérité-terrain et OCR)
  - Calcul de taux de reconnaissance
  - Deux modes :
    - Page
    - Ligne



Evaluation

- Alignement de 2 textes
- Calcul de distance d’édition
  - Mode page
    - Page vérité-terrain vs page OCR
  - Mode ligne
    - Comparer les lignes de texte (si correctement segmentées)
    - Calculer le taux de bonne segmentation
    - Pondérer

Résultats

- Rapport détaillé
  - Niveau mot et niveau caractère
  - Taux pour chaque caractère
  - Taux par type de caractère
  - Confusions
- Un rapport par OCR
- 2 Types de rapport :
  - Caractère
  - Mot

```
   886   Characters
     5   Errors
 99.44%  Accuracy

     0   Reject Characters
     0   Suspect Markers
     0   False Marks
  0.00%  Characters Marked
 99.44%  Accuracy After Correction

   Ins     Subst       Del   Errors
     0         0         0        0   Marked
     0         4         1        5   Unmarked
     0         4         1        5   Total

 Count   Missed   %Right
   137        0   100.00   ASCII Spacing Characters
    57        1    98.25   ASCII Special Symbols
     3        0   100.00   ASCII Digits
   112        1    99.11   ASCII Uppercase Letters
   557        2    99.64   ASCII Lowercase Letters
     6        0   100.00   Latin1 Control Codes
    14        0   100.00   Latin1 Lowercase Letters
   886        4    99.55   Total

Errors   Marked   Correct-Generated
     2        0   {nn}-{ti}
     1        0   {-}-{¬}
     1        0   {E}-{É}
     1        0   {}-{.}

 Count   Missed   %Right
    24        0   100.00   {<\n>}
   113        0   100.00   { }
     3        0   100.00   {'}
    21        0   100.00   {,}
     8        1    87.50   {-}
    24        0   100.00   {.}
     1        0   100.00   {2}
     1        0   100.00   {4}
     1        0   100.00   {5}
     1        0   100.00   {:}
```