europeana
newspapers

ICT POLICY SUPPORT PROGRAMME

Europeana Newspapers

Contrôle de la qualité OCR

Christian Clausner, USAL

Traduction : jean-Philippe Moreux, Bibliothèque nationale de France

University of Salford MANCHESTER

PRImA
Pattern Recognition & Image Analysis Research Lab

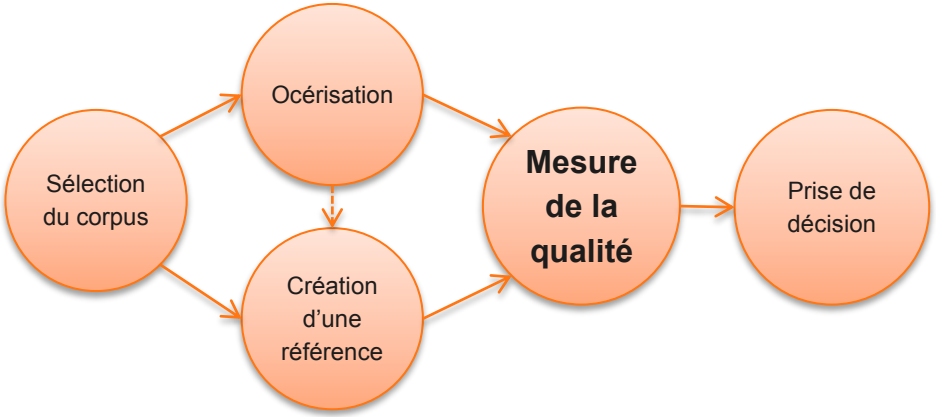
europeana
newspapers

Partie 1 : Introduction

Cas d'usage, workflow, outils, formats

Contrôle de la qualité de l'OCR – **Cas d'usage**

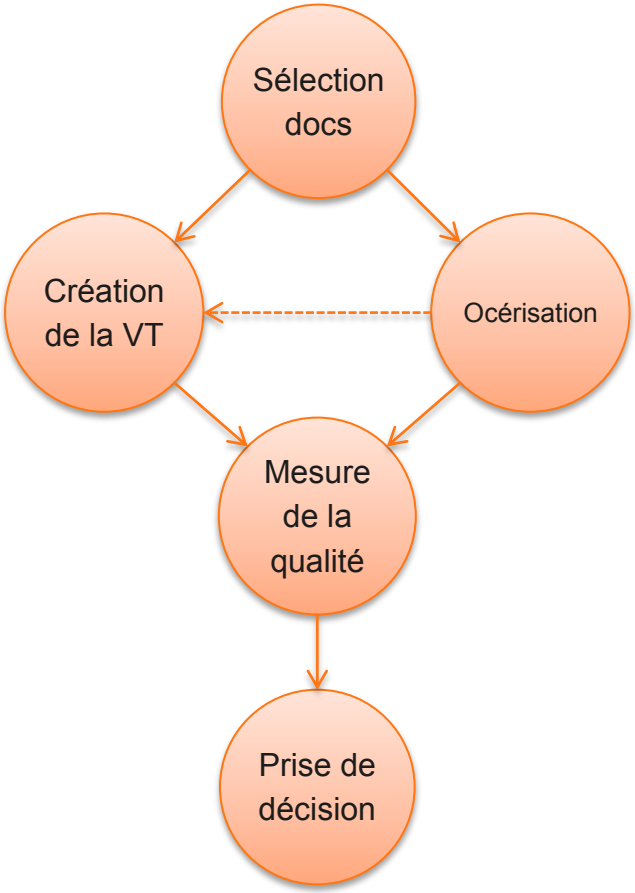
- Etudes avant-projet
  - Évaluer si une collection d'imprimés est apte à être océrisée pour un usage particulier
- Contrôle de la qualité
  - Contrôler le résultat de l'OCR pendant ou après la production



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contrôle de la qualité de l'OCR – **Workflow**

1. Sélectionner un (petit) ensemble de documents réprésentatifs du corpus.
2. Océriser les documents dans la chaîne de numérisation cible.
3. Créer les documents de référence (vérité terrain, VT)
4. Mesurer la qualité en sortie de la chaîne de numérisation.
5. Prendre une décision en fonction des performances mesurées.



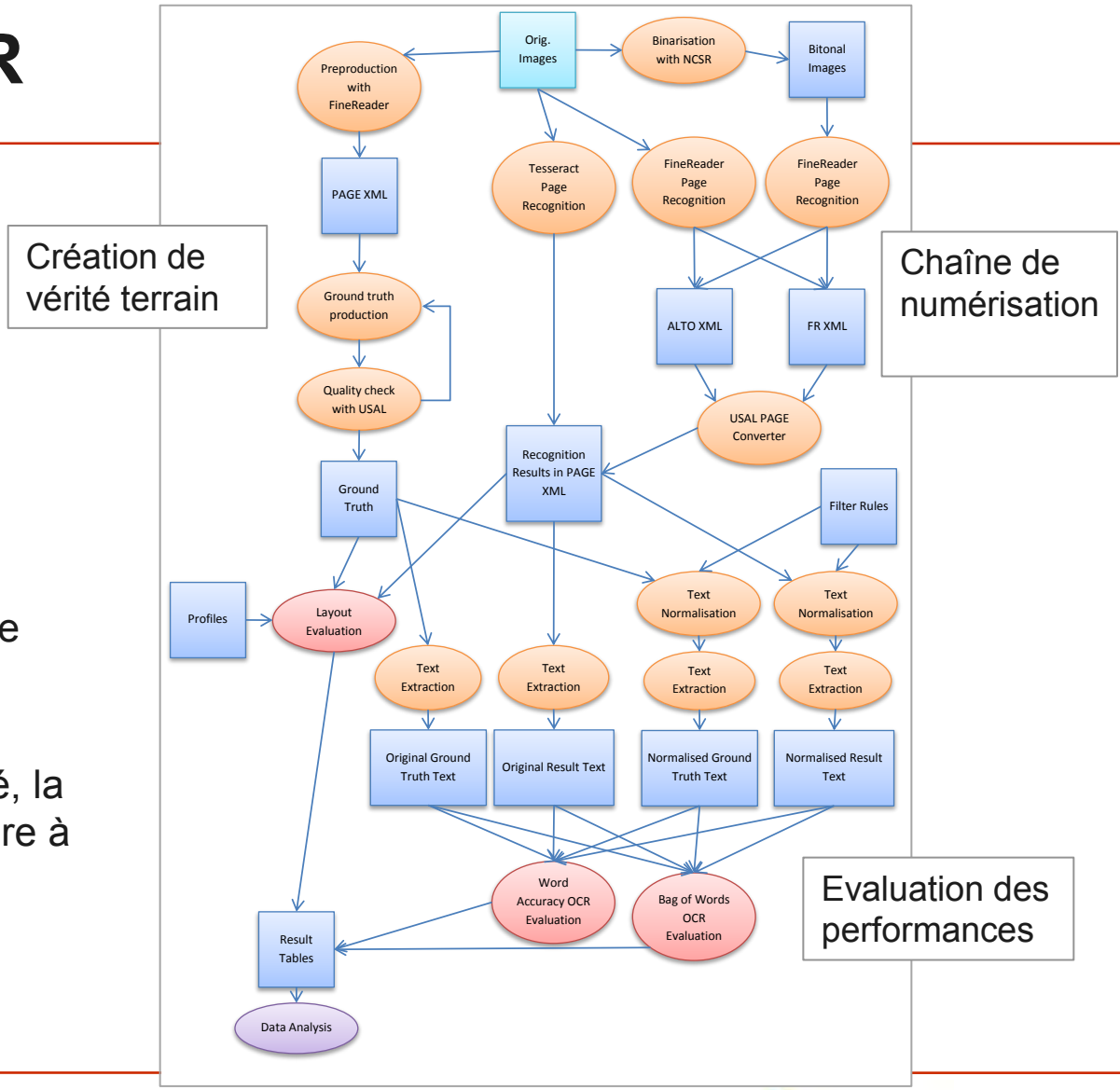
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contrôle de l'OCR

- Workflow

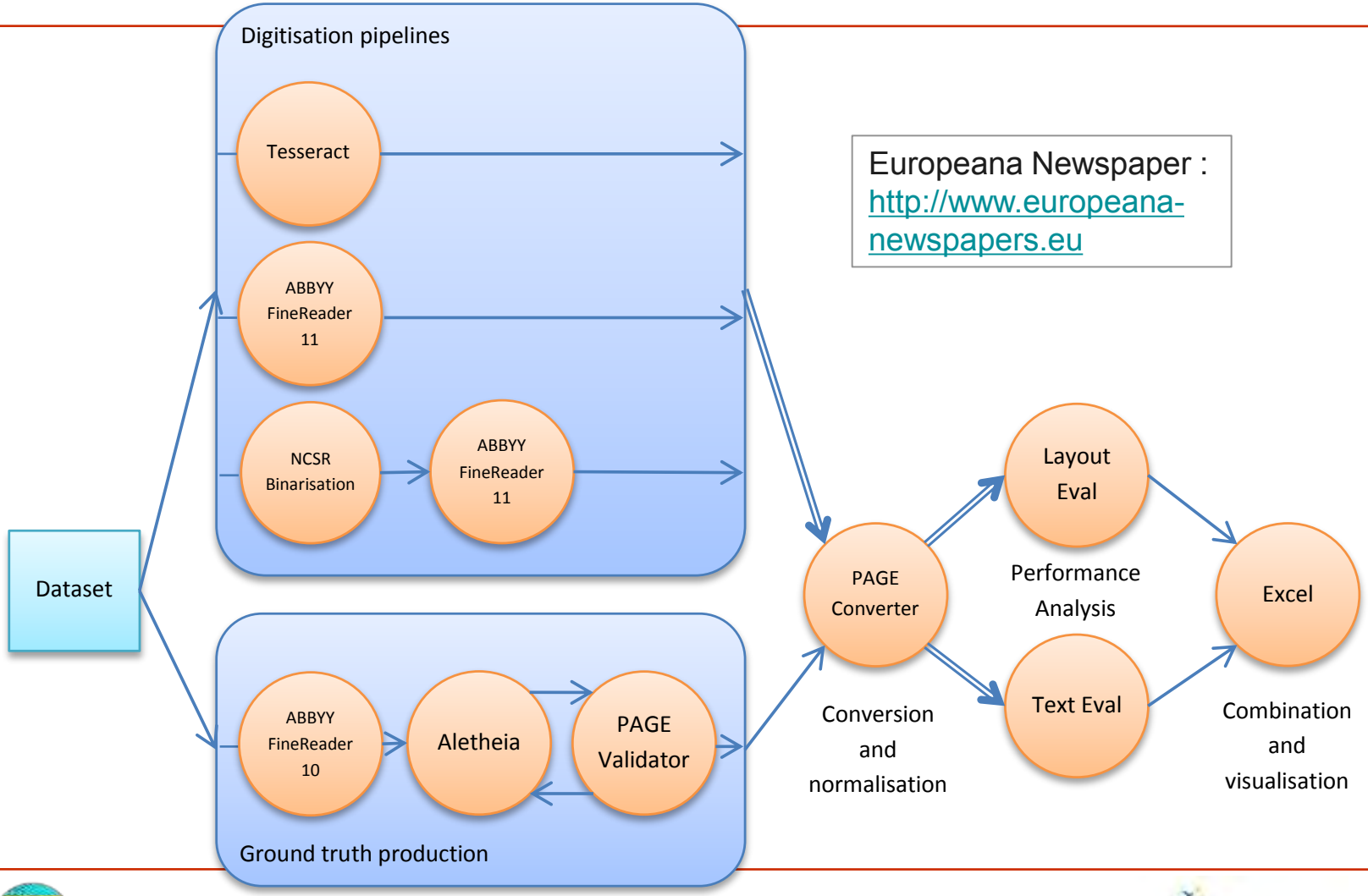
NB :

- Dans le cas d'une étude avant-projet, une chaîne de numérisation devra être mise en place (OCR *open source* ou autre).
- Pour un projet ou un marché, la chaîne est celle du prestataire à évaluer.



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

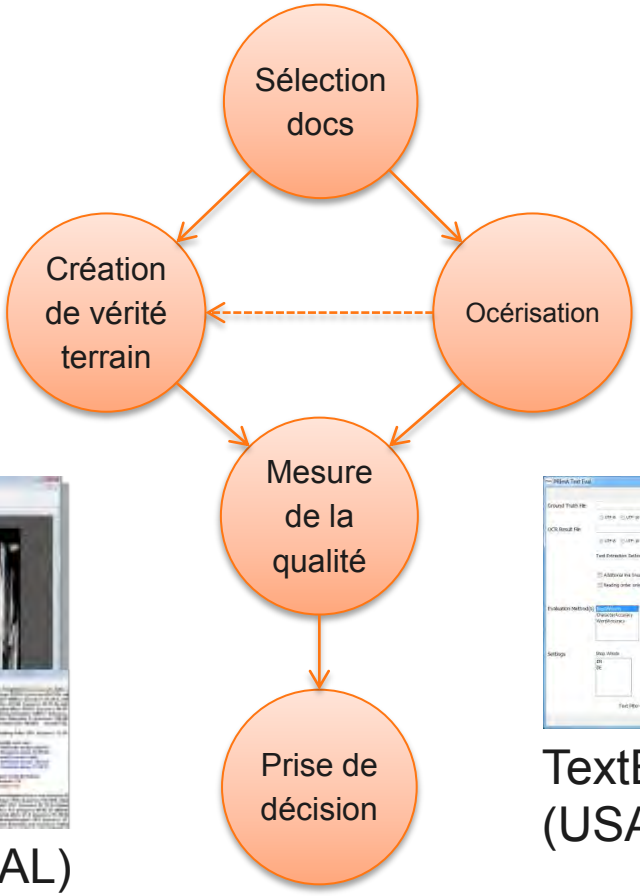
Exemple de workflow (Europeana Newspapers)



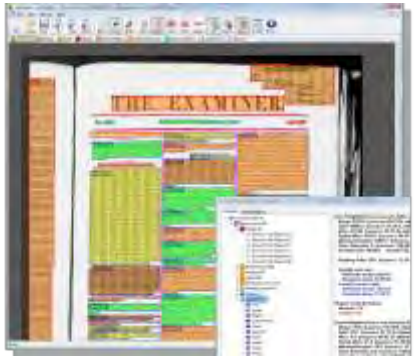
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contrôle de la qualité de l'OCR – **Outils**

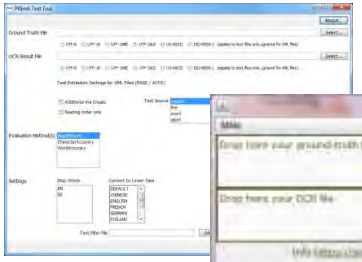
Aletheia



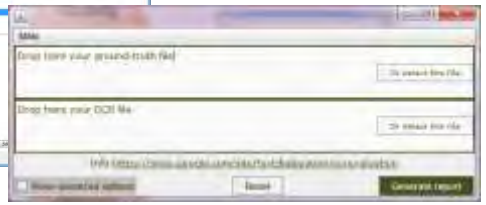
ABBYY FineReader
Tesseract
…



LayoutEval (USAL)



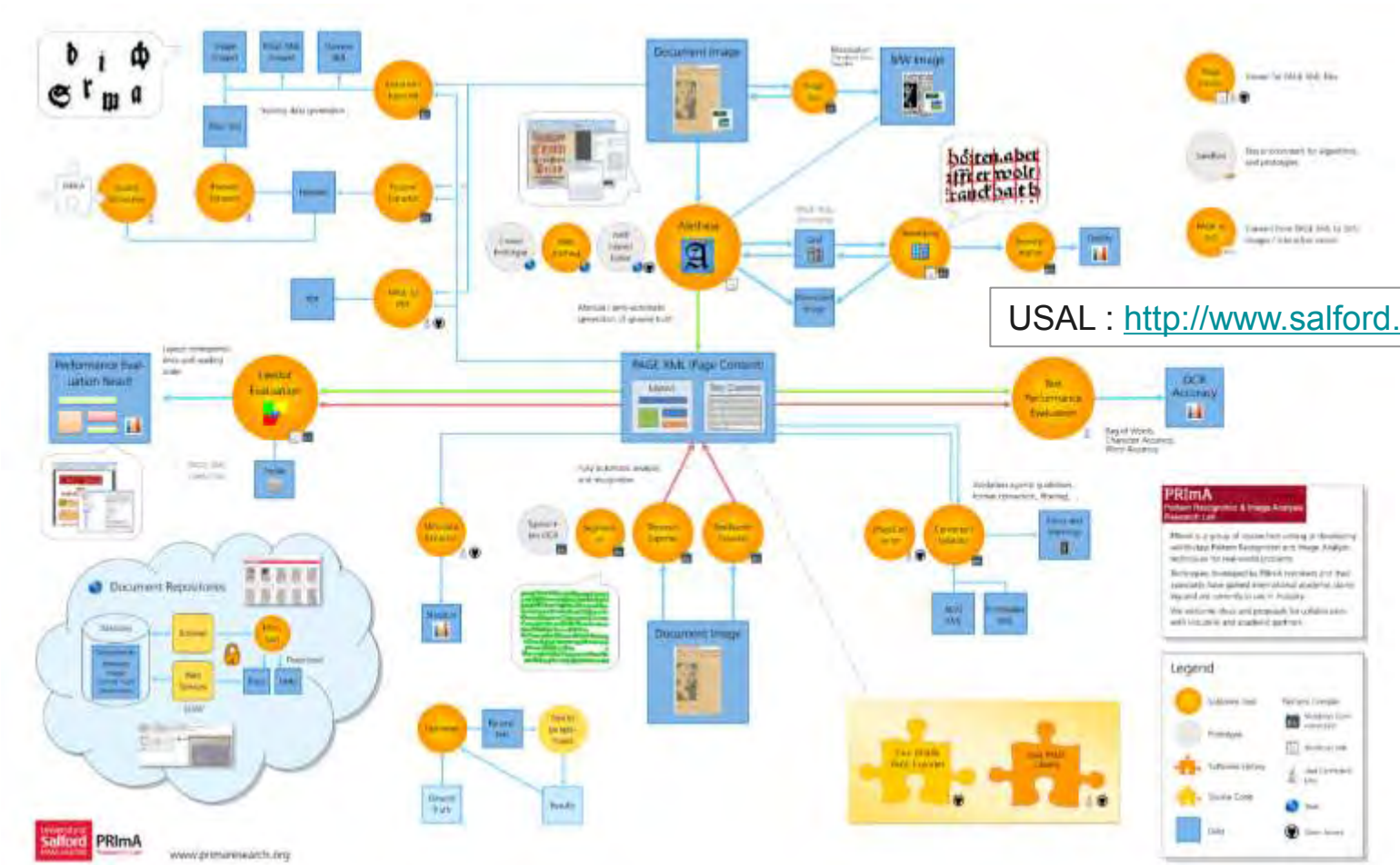
TextEval (USAL)



OCREvaluation (Univ. Alicante)

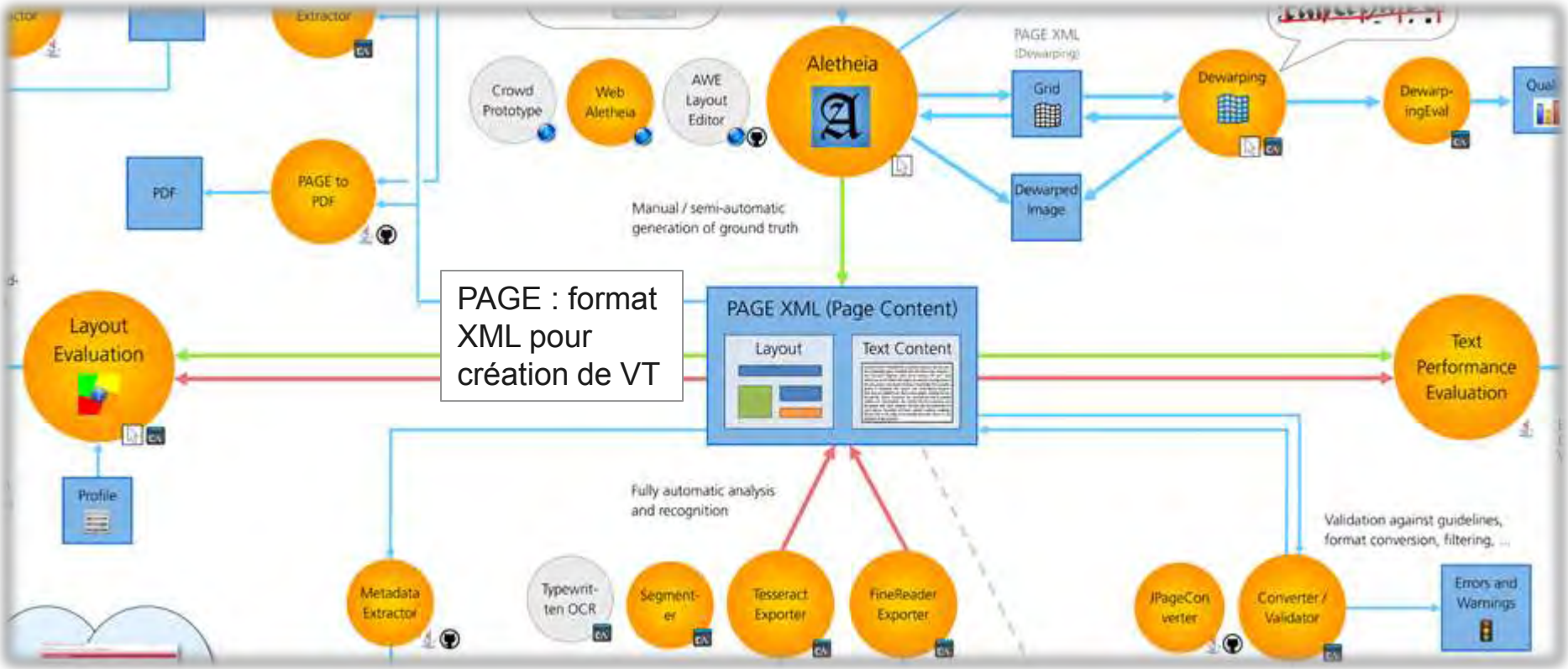
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contrôle de la qualité de l'OCR – **Outils et formats (USAL)**



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contrôle de la qualité de l'OCR – **Format PAGE**



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

europeana
newspapers

Partie 2 : Océrisation

Techniques et outils

OCR – **Analyse de page et reconnaissance**

Image

Numérisation

- Segmentation
- Classification des natures de contenu
- OCR
- …

OCR

ABSENT YET PRESENT. by Gilberta M. F Lyon, 3 vols., London; Digby, Long, and Co. We were able to speak in very high praise of Miss Lyon's former story, "For Good or Evil." This, however, is far better and must be classed amongst the best books of the season. The writer has conspicuous talents; in her de- scription of the scenery whither she guides us in the portraiture of the characters with which she peoples her story, in the varied pleasantries and troublous periods which go to make up the delightful romance which she unfolds in her pages, we are enchained by the most sympathetic ties, until the fortunes of hero and heroine are spun out to their

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICTPSP

OCR – **Alternatives**

- ABBYY FineReader Engine (FRE)
- Tesseract (*open source*)
- Aletheia (avec Tesseract intégré)

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

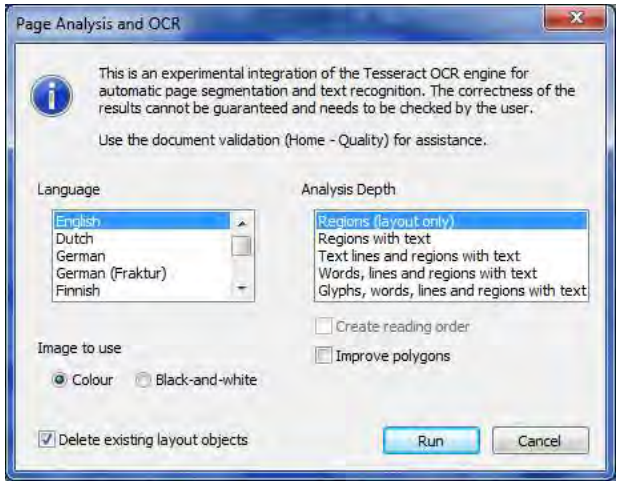
OCR – **En pratique avec Aletheia**

1. Créer un nouveau document.
2. Sélectionner IMG.tif
3. Confirmer "… without B/W image".
4. Zoom "Full page".
5. Pour lancer l'OCR : "Analyse Page".

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

OCR – **En pratique avec Aletheia**

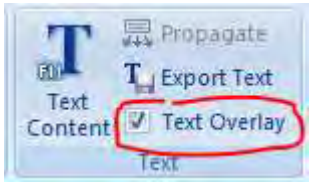
1. Paramétrer.
2. Cliquer sur “Run”.
3. Attendre…
4. Visualiser.

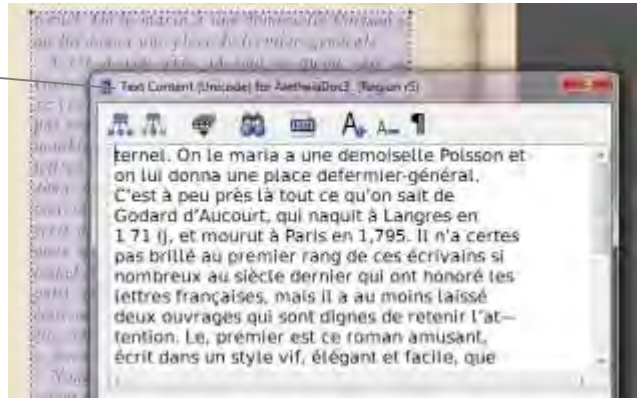


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

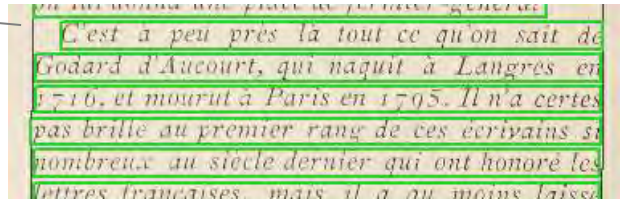
OCR – **En pratique avec Aletheia**

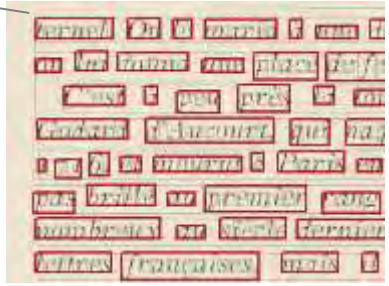
1. Consulter le texte océrisé : F11 ou activer “Texte overlay”





2. Visualiser ligne de texte (F7), mots (F8) et glyphes (F9)

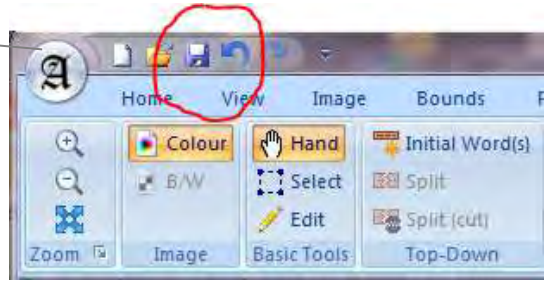




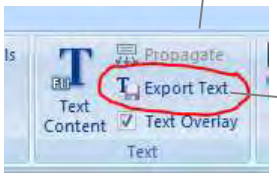
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

OCR – **En pratique avec Aletheia**

1. Enregistrer le document au format PAGE.



2. Si besoin, exporter le texte.



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

europeana
newspapers

Partie 3 : Création de vérité terrain

Cas d'usage, méthodes, outils

L’OCR est-il bon ?

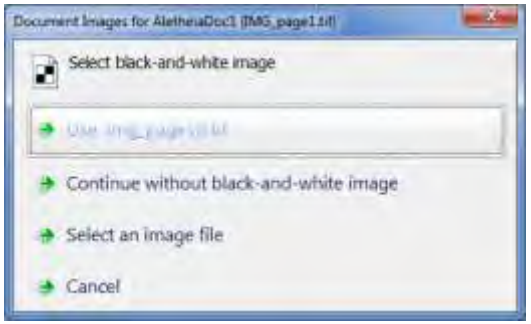
- Un document a été océrisé.
- Mais est-il de bonne qualité ?
- Un contrôle visuel est possible mais subjectif…
- Solution : comparer avec une référence

➔ Vérité terrain

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Création de vérité terrain – **En pratique avec Aletheia**

- Aletheia permet de créer un OCR de référence à partir d'une image couleur
- et éventuellement d'une image déjà binarisée :

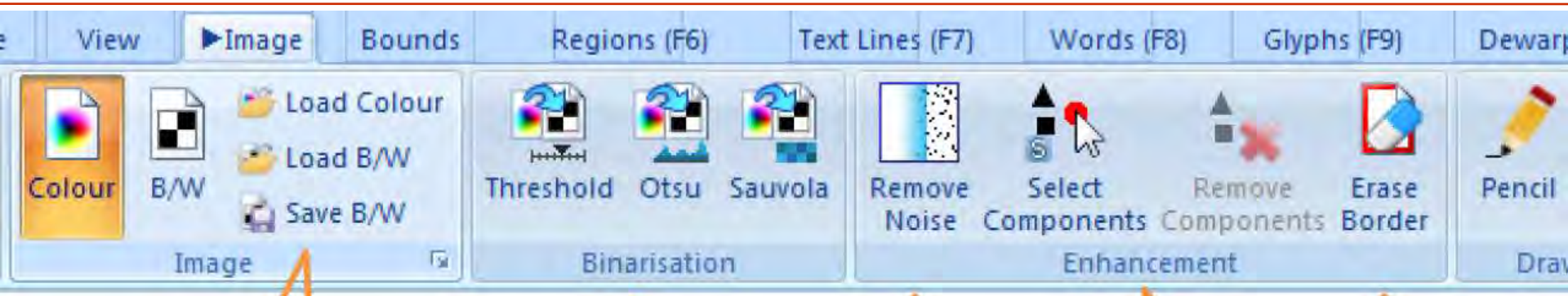


- NB : il est aussi possible de créer l'image binarisée dans Aletheia.

Aletheia :
http://www.primaresearch.org/tools

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Outils niveau image**

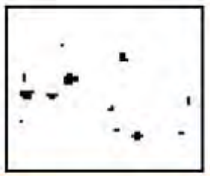


Chargement d'images (TIFF, PNG, JPEG)

Binarisation d'image :

- seuillage
- Otsu (automatique)
- Sauvola (automatique, adaptative)

Suppression du bruit



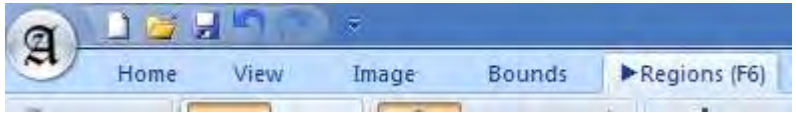
Suppression manuelle du bruit

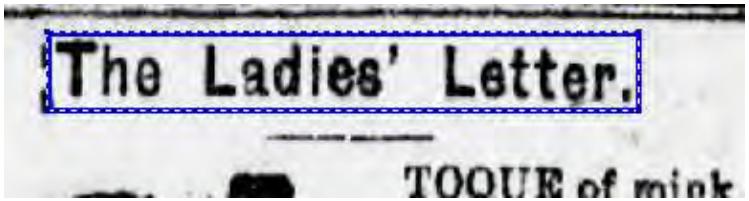
Suppression des bordures (sur image N&B)

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création manuelle de régions**

1. Onglet “Regions” (F6)
2. Zoomer sur une zone.
3. Outil “Rectangle”.
4. Tracer une boîte (ou un polygone) autour d’un bloc de texte.





This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

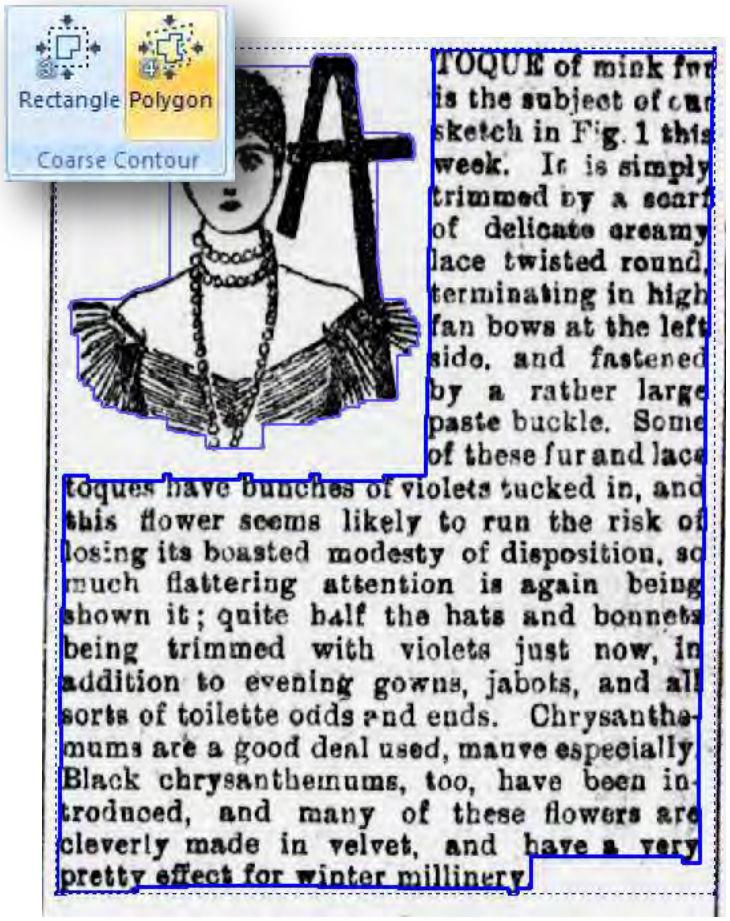
Aletheia – **Création semi-automatique de régions**

1. Outil “Fine Contour” - “Rectangle” (*contour fin*)
2. Tracer une boîte autour de l’image (tout inclure, mais pas au-delà)
3. Annuler et essayer de nouveau si nécessaire.

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création semi-automatique de régions**

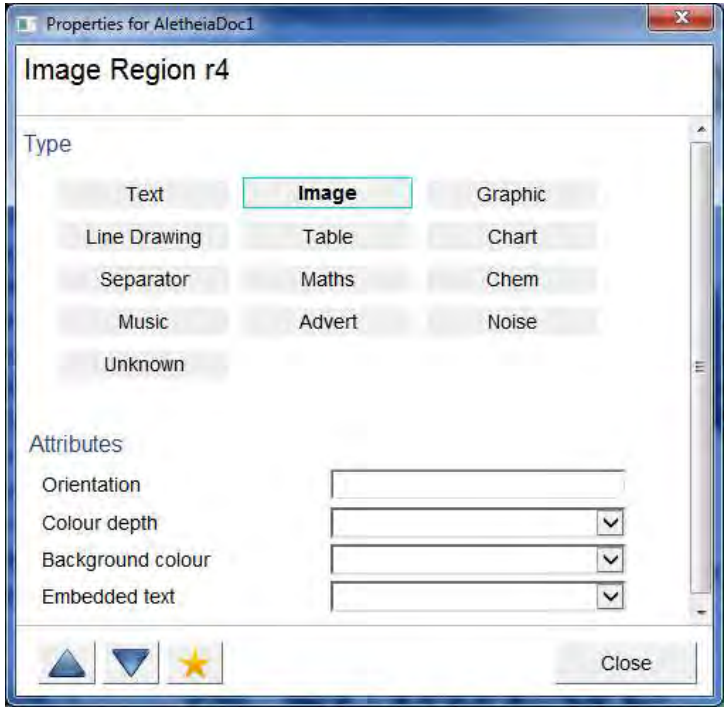
1. Outil “Coarse Contour” - “Polygon” (*contour grossier*)
2. Tracer une boîte autour d’un paragraphe de texte (un clic ajoute un point ; clic droit pour finir)



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Description des régions / illustration**

1. Double-clic sur une région illustration (ou F10).
2. Choisir “Image” dans la boîte de dialogue pour changer le type de région.
3. Renseigner les attributs en fonction du type de région.



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

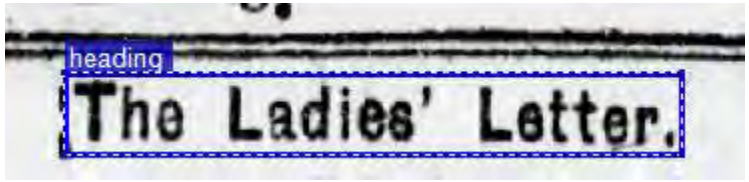
Aletheia – **Description des régions / texte**

1. Cliquer sur un titre. F10.
2. Changer “Text Type” de “Paragraph” en “Heading”
3. Les icônes Haut et Bas permettent de naviguer de bloc en bloc.
4. Fermer le dialogue.

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Saisie du texte**

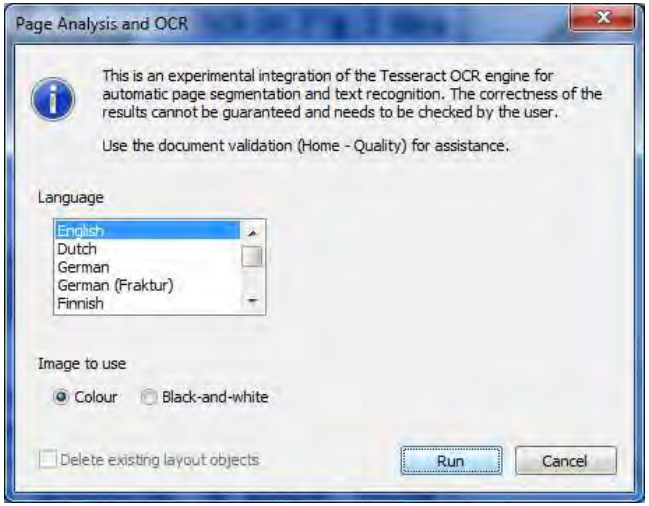
1. Cliquer sur un titre.
2. Ouvrir la boîte de dialogue “Text” (F11)
3. Saisir le texte du bloc.
4. Des caractères spéciaux sont disponibles via le clavier virtuel.
5. Cliquer sur “Next” pour passer à la saisie du bloc suivant.

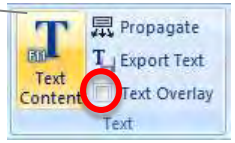


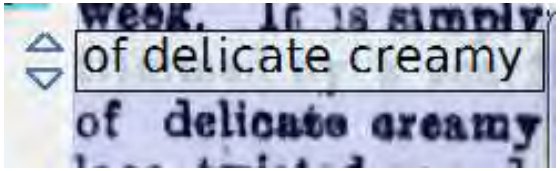
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création de régions – OCR**

1. Cliquer sur “OCR Region”.
2. Choisir l’image couleur ou N&B.
3. En fonction de la qualité, il est plus rapide de corriger ou de ressaisir…
4. L’option “Text overlay” permet de visualiser le texte au survol de souris.
5. Les touches Haut/Bas permettent un contrôle ligne à ligne.





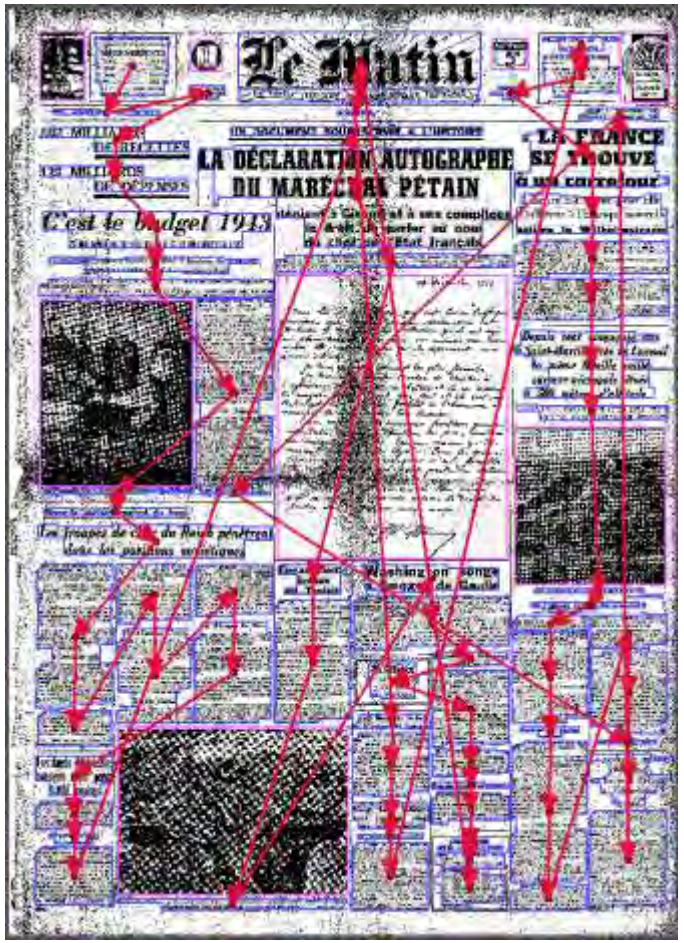


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Ordre de lecture**

L'ordre de lecture logique doit être défini dans la VT. Il est important pour certains usages du document numérique : recherche de phrases complète, lecture écran ou en synthèse sonore, conversion en livre numérique, etc.

Aletheia permet de créer l'ordre de lecture manuellement ou par analyse de la page.

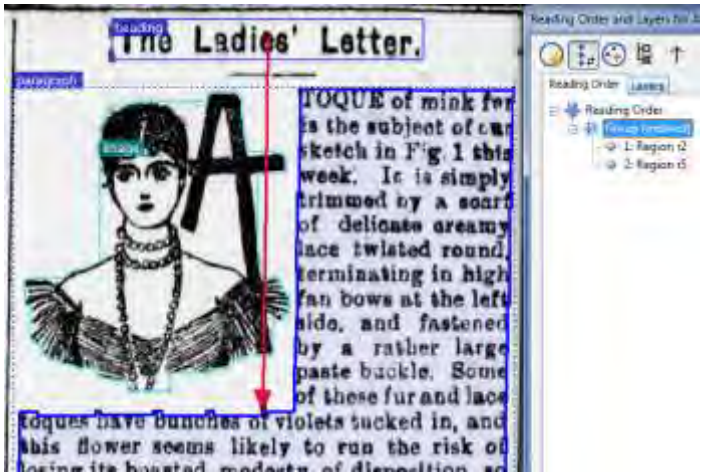


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création manuelle de l'ordre de lecture**

1. Ouvrir la fenêtre "Reading order and Layers" (F12)
2. Sélectionner "Group (ordered)" – C'est la racine du premier groupe de contenu du document.
3. Double-clic sur le premier bloc de contenu.
4. Double-clic sur le bloc suivant pour créer l'ordre de lecture. Etc.

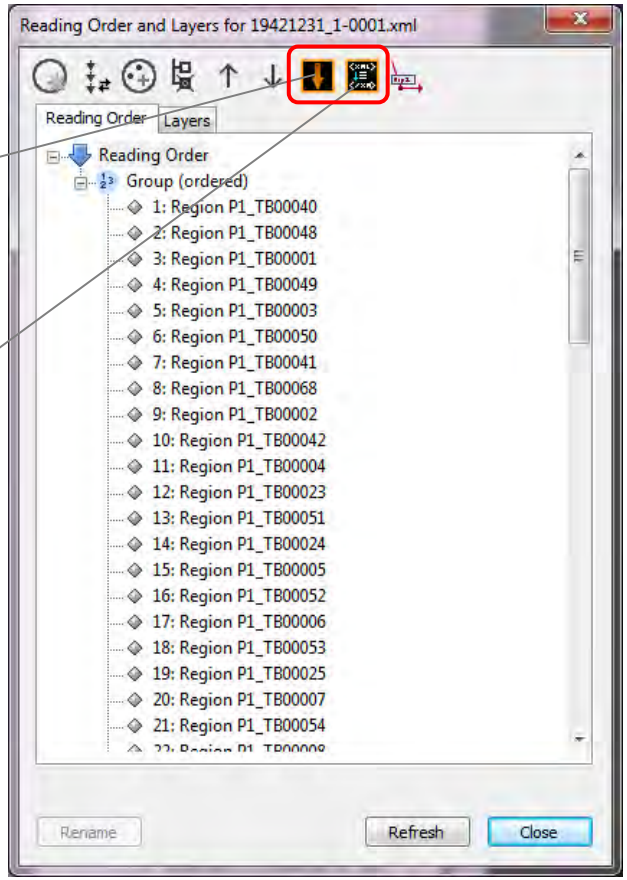
Alternative : placer tous les contenus dans un groupe ou créer des groupes et de définir leur ordre de lecture.



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création automatique de l'ordre de lecture**

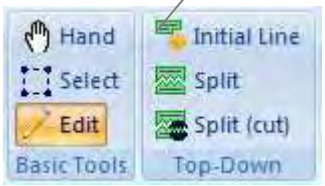
1. Dans la fenêtre "Reading order and Layers" (F12), deux options :
   - "Create top-to-bottom reading order" – Analyse descendante et gauche-droite des blocs de la page. NB : Ne convient pas pour le multicolonnage.
   - " Create reading order from internal region list (as in XML)" – Si l'ordre de lecture est déjà présent dans le fichier (cas de la création de VT à partir d'un OCR, voir page 34).
2. L'ordre peut être corrigé (flèches ↑↓)

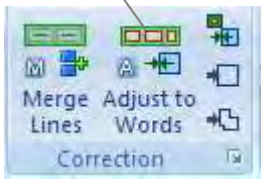


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Création des lignes de texte**

1. Sélectionner l'onglet "Text Lines" (F7).
2. Clic sur "Detect Lines".
3. Clic sur un paragraphe de texte. Les lignes sont détectées et créées.
4. Pour retoucher les lignes : outils manuels ou automatiques

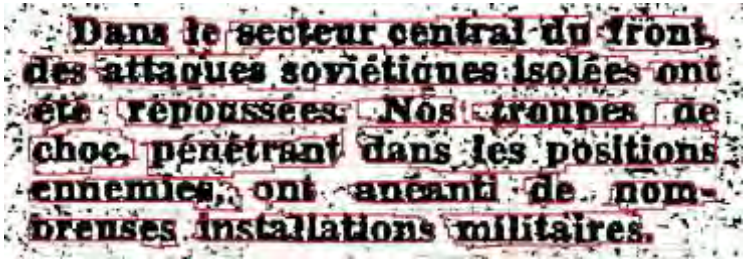


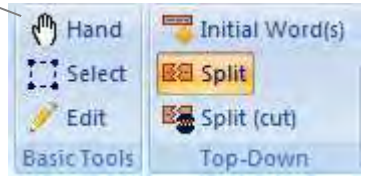


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Mot, caractère**

1. Sélectionner l'onglet "Words" (F8).
2. Clic sur "Detect Words".
3. Clic sur un paragraphe de texte. Les mots sont détectés et créés. La segmentation peut alors être corrigée.
4. Même principe avec les caractères (onglet "Glyphs", F9).

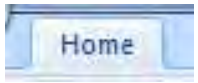

Dans le secteur central du front, des attaques soviétiques isolées ont été repoussées. Nos troupes de choc, pénétrant dans les positions ennemies, ont anéanti de nombreuses installations militaires.



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Validation**

1. Sélectionner l'onglet "Home".
2. Clic sur "Validation".
3. Clic sur "Validate". Il est possible de cibler les contrôles (bloc, ligne, mot, caractère)
4. Les résultats de validation sont affichés :
   - Information
   - Avertissement



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Utiliser un OCR**

La création ex nihilo d'une vérité terrain étant laborieuse, il est possible d'utiliser un document OCR comme base. Aletheia lit les formats Abbyy FineReader XML et ALTO XML.

1. Pour ouvrir un fichier OCR : "Open document"
2. Sélectionner l'image associée.
3. Enregistrer le document au format PAGE.

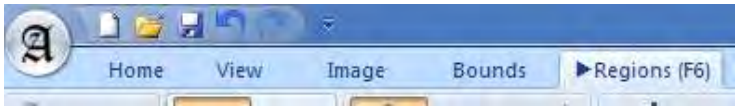
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

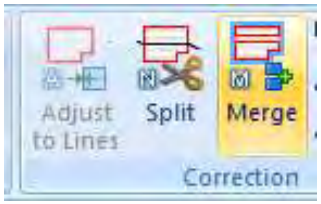
Aletheia – **Utiliser un OCR**

La segmentation des blocs peut ensuite être corrigée, ainsi que le texte.

- Edition des contours de bloc, découpage et fusion de blocs.
- Correction du texte.
- Correction de l'ordre de lecture.

NB : selon l'usage attendu pour la vérité terrain, la granularité de la segmentation peut varier. Par exemple pour le contrôle de la transcription du texte, le niveau bloc suffit.





This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Aletheia – **Conversion de formats : OCR vers PAGE**

ABBYY FineReader Engine (FRE)

- FR XML ▸ PageConverter (USAL)
- ALTO ▸ PageConverter (USAL)
- FineReader Integration

Tesseract (*open source*)

- TesseractToPage (USAL)

PAGE XML

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

europeana
newspapers

Partie 4 : Évaluation de la reconnaissance du texte

Cas d'usage, méthodes, outils

Evaluation des performances

Vérité terrain

OCR à évaluer

Mesure de la qualité

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Evaluation de la reconnaissance du texte – **Objectif**

- Comparer le texte de la page avec le texte de référence.
- Différentes mesures classiques :
  - Précision niveau “caractère” (nombre des caractères erronés)
  - Précision niveau “mot” (nombre des mots erronés)
  - “Sac de mots” : l’ordre de lecture ne compte pas

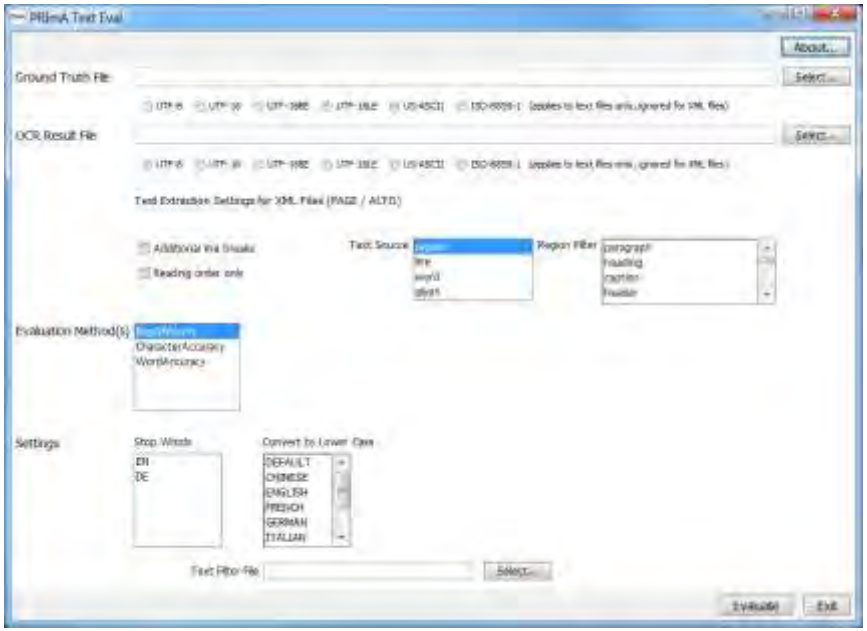
Note : normalisation du texte

- Pour une évaluation moins stricte, remplacer certains caractères par leur version “normalisée”, par ex. :
  - Accent é ➔ e
  - Ligature æ ➔ ae
  - Tiret long – ➔ -
  - S long (ſ) ➔ s

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Reconnaissance du texte – **En pratique/Text Eval**

- Outil **Text Eval** (USAL)
- Fichiers (texte ou XML) :
  - vérité terrain (“Ground truth”)
  - OCR à évaluer (“OCR result”)
- Choix de la méthode d’évaluation
- Options :
  - filtre d’éléments XML
  - liste de mots vides
  - ignorer la casse des caractères

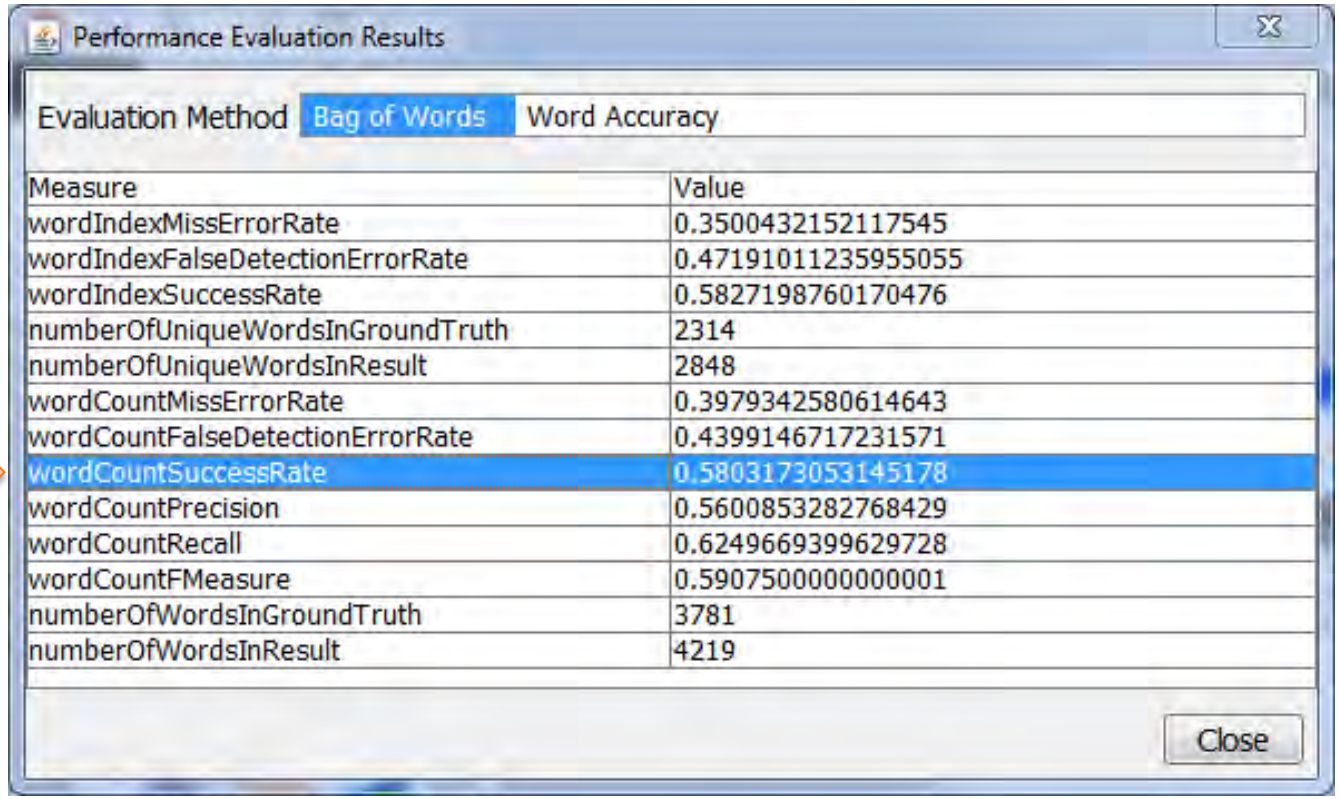


Text Eval : http://www.primaresearch.org/tools/PerformanceEvaluation

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Reconnaissance du texte – **En pratique/Text Eval**

Evaluation en mode « Sac de mots »

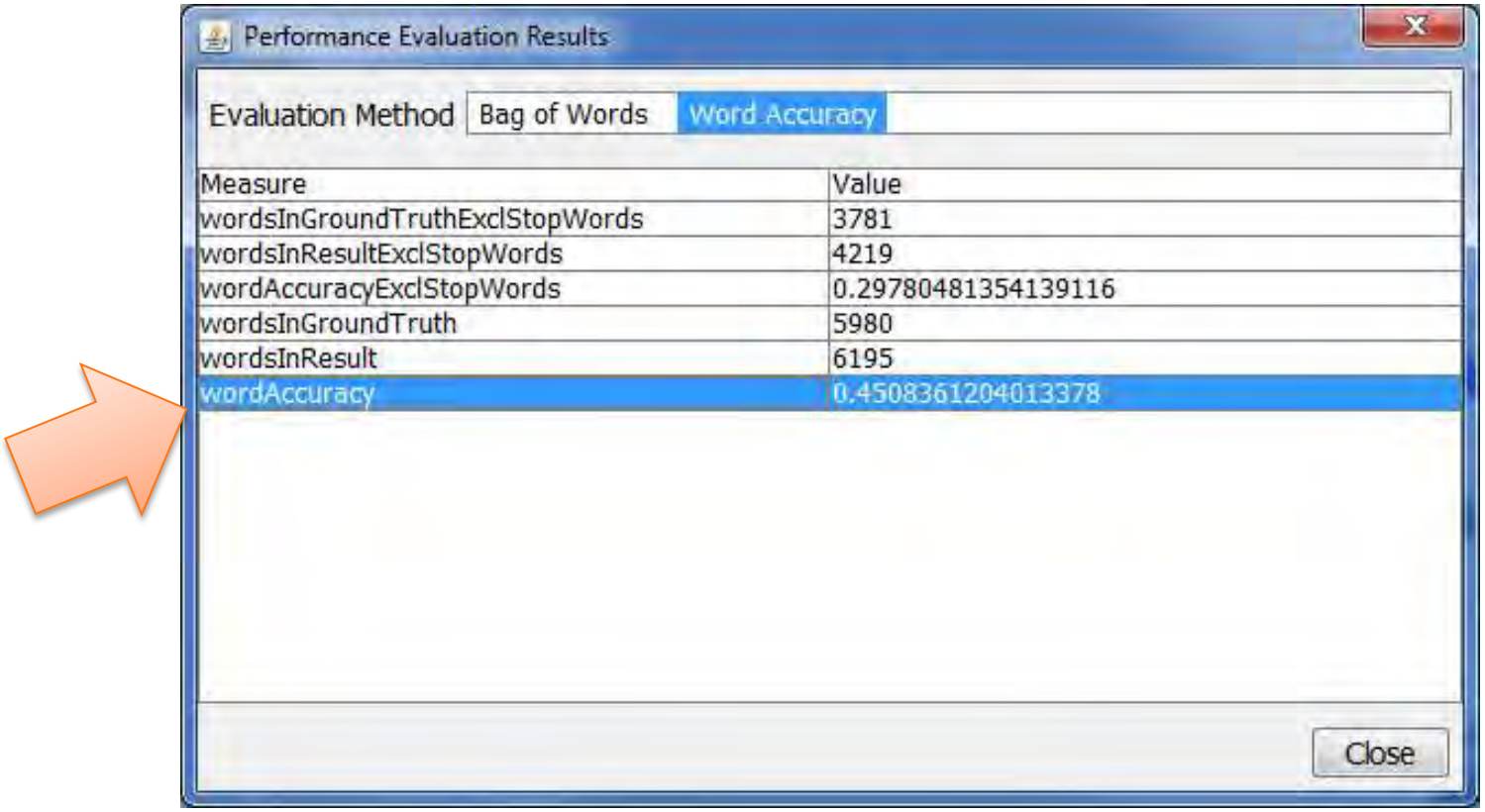


taux qualité
= 58 %

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Reconnaissance du texte – **En pratique/Text Eval**

Evaluation « mots »

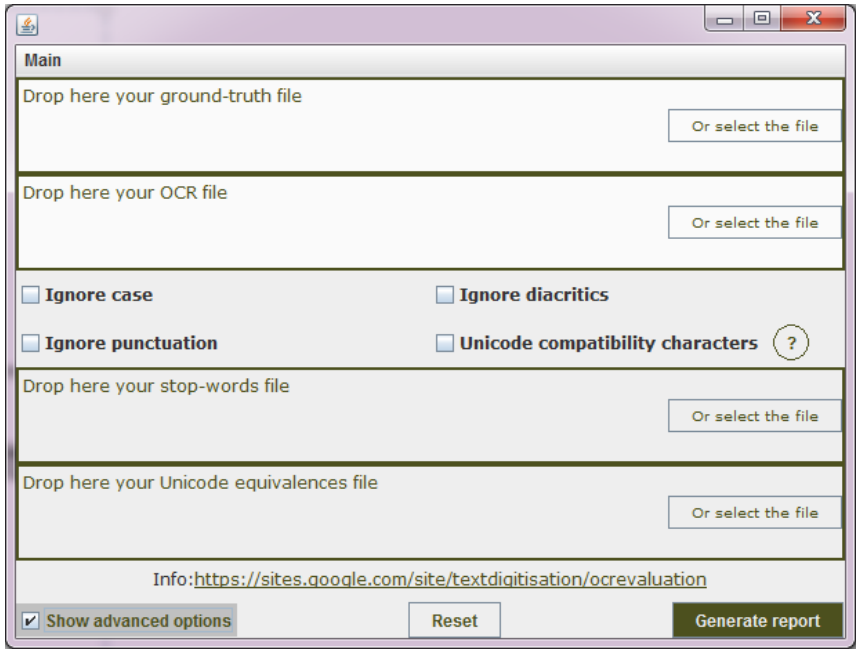


taux qualité
= 45,1 %

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Reconnaissance du texte – **En pratique/ocrevalUAtion**

- Outil **ocrevalUAtion** (université d'Alicante)
- Fichiers (texte ou XML)
- Options :
  - filtre d'éléments XML
  - liste de mots vides
  - équivalence de caractères
  - ignorer la casse et l'accentuation des caractères, les caractères de ponctuation

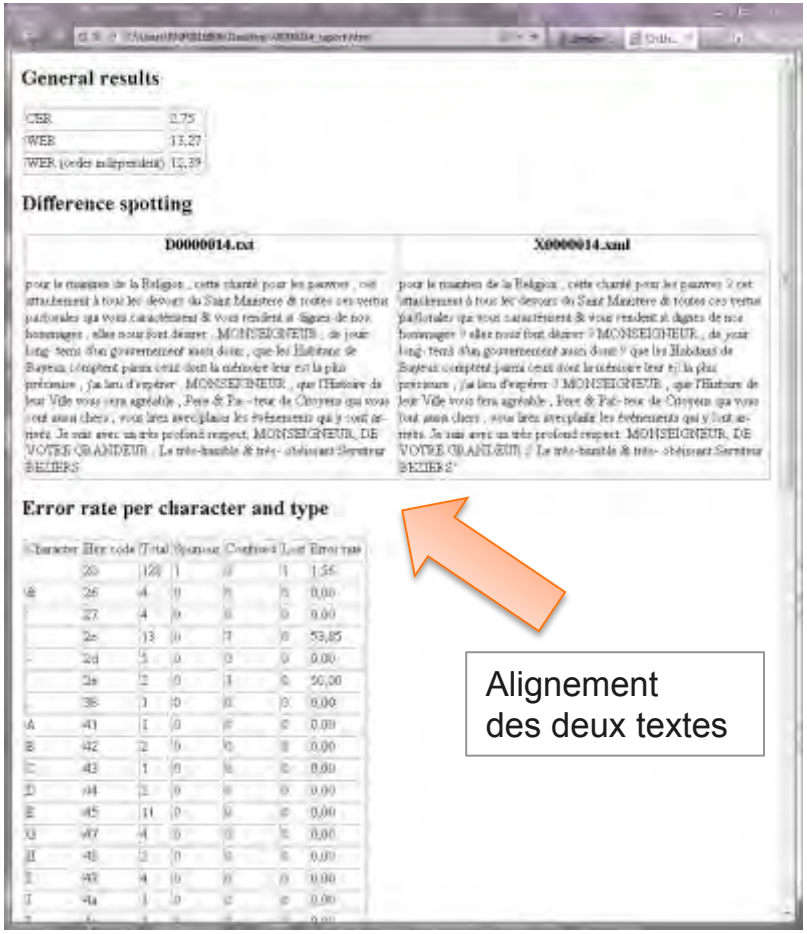


OCREvaluation:
https://github.com/impactcentre/ocrevalUAtion

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Transcription du texte – **En pratique/ocrevalUAtion**

CER : taux erreur caractères (2,75 %)
WER : taux erreur mots (13,27 %)
WER : taux erreur « sac de mots » (12,39 %)



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

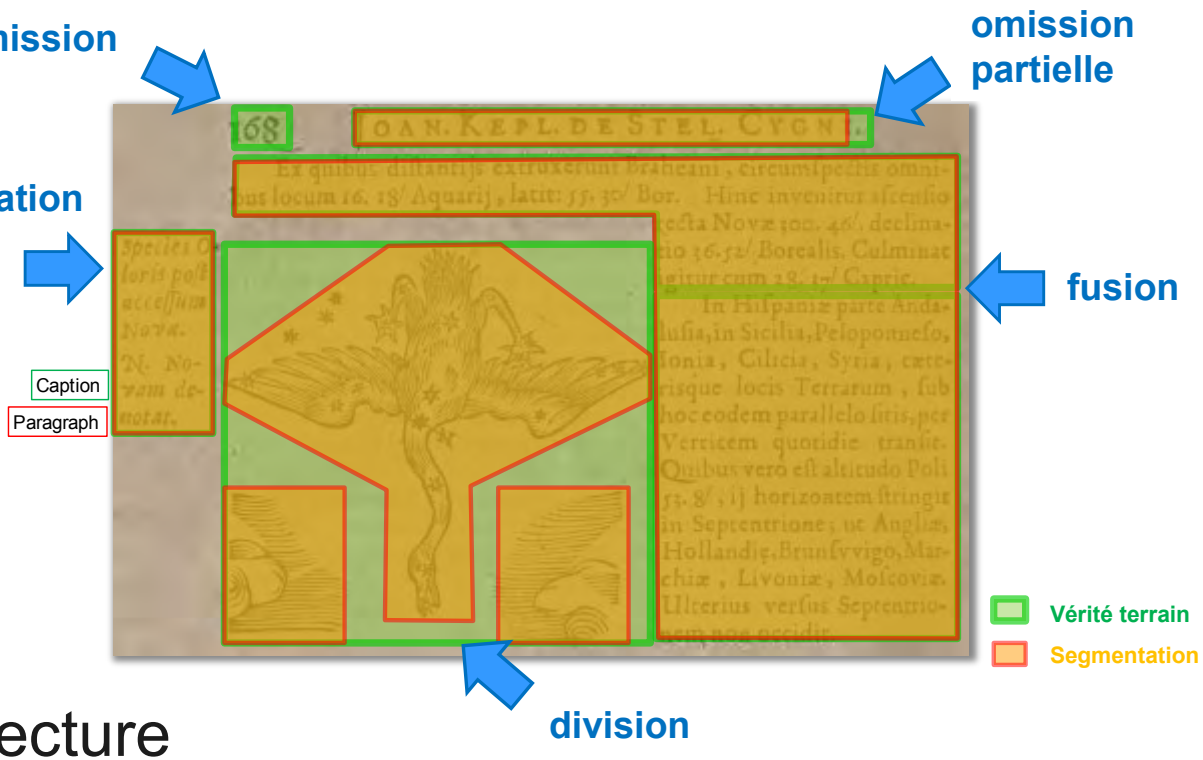
europeana
newspapers

Partie 5 : Évaluation de la segmentation

Cas d'usage, méthodes, outils

Évaluation de la segmentation

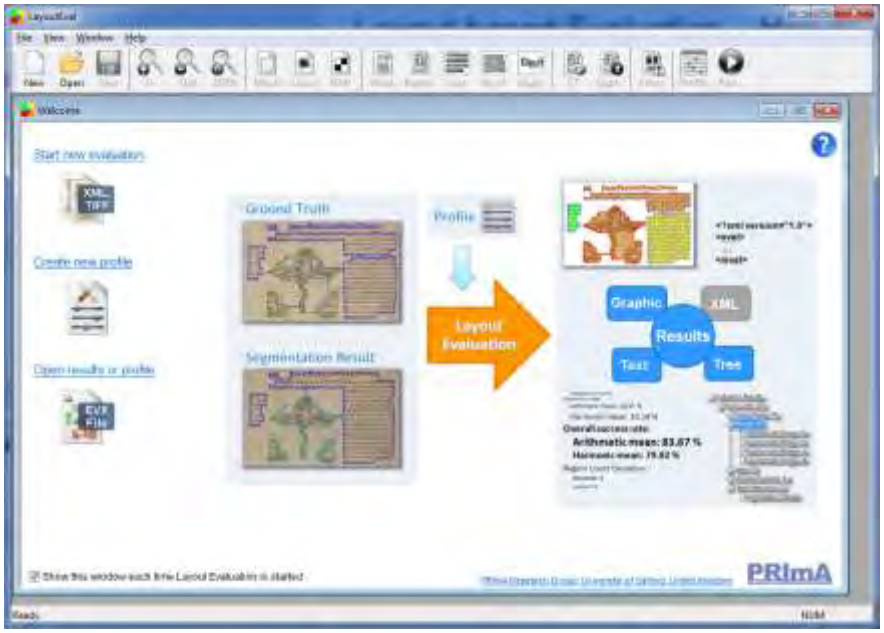
- Six types d'erreur de segmentation :
  - omission
  - omission partielle
  - division
  - fusion
  - défaut de classification
  - fausse détection
- Défaut d'ordre de lecture



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

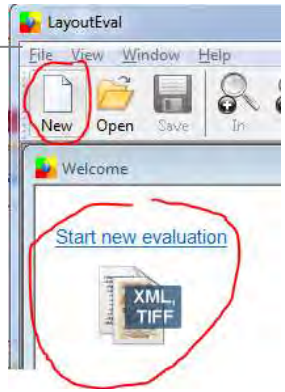
- Outil : LayoutEval (USAL)
- Fichiers :
  - Vérité terrain (format PAGE)
  - Résultat de l'OCR (ALTO, Fine Reader, Tesseract)
  - Image binarisée
  - Image couleur (optionnelle)

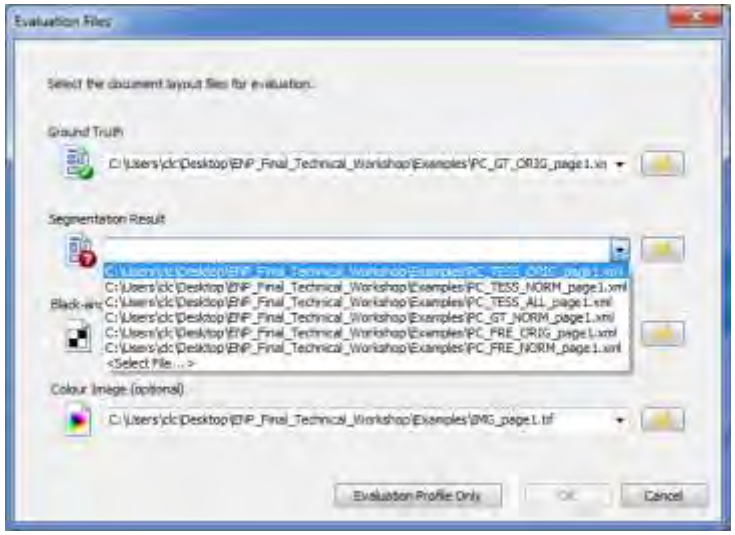


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Démarrer une nouvelle évaluation.
2. Sélectionner le fichier de vérité terrain (format PAGE) et les images.
3. Sélectionner le fichier OCR.
4. Clic sur “OK”

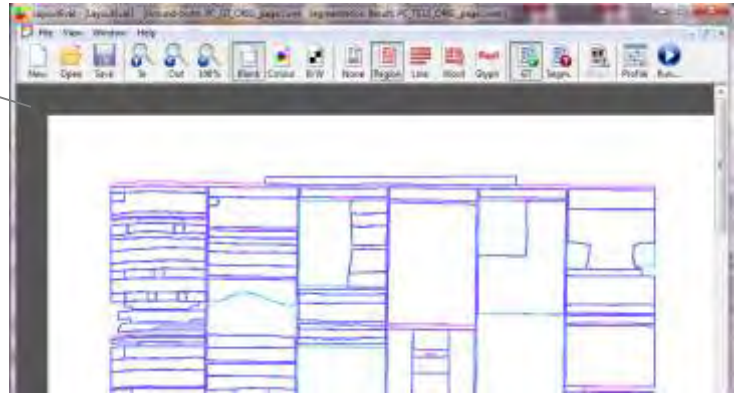




This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Activer la vue “Region”
2. Se déplacer dans la page
3. Clic sur les boutons “GT” (vérité terrain) et Segm. pour visualiser vérité terrain, segmentation, ou les deux
4. Blank, Color et B/W : changer le mode de visualisation (Blank = segmentation seule)

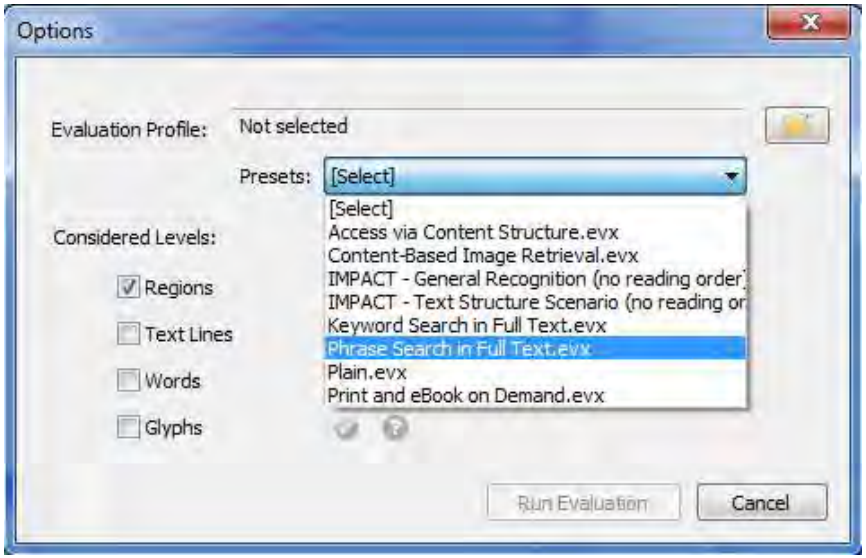


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Clic sur “Run…” 

2. Sélectionner le profil d’évaluation prédéfini “Phrase search in full text” et le niveau “Regions”
3. Clic sur “Run Evaluation”
4. Attendre…

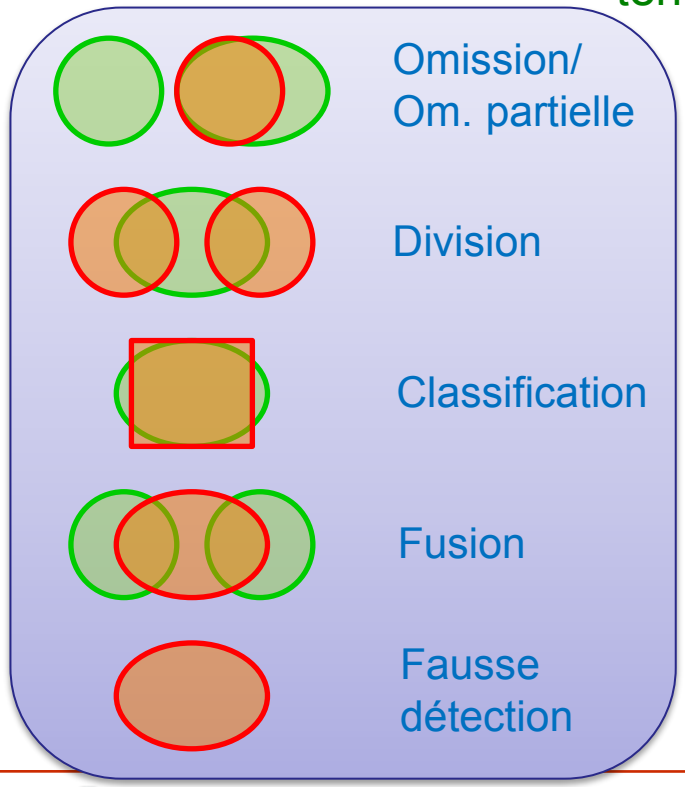


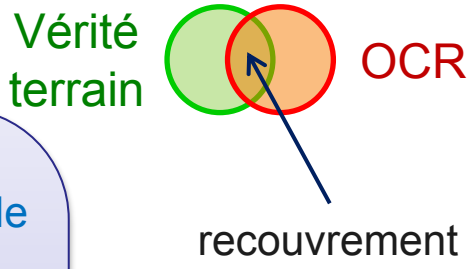
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

While you are waiting…

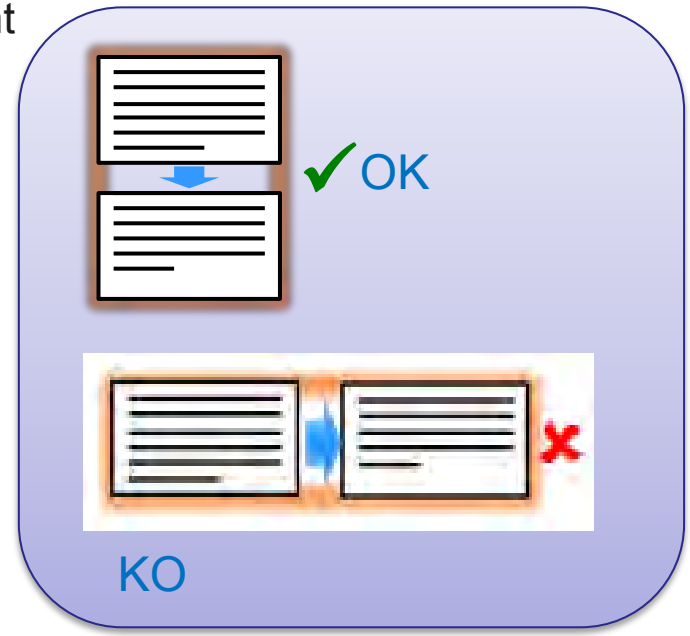
Les métriques de l'évaluation sont basées sur l'analyse du recouvrement des régions

Types d'erreurs





La différentiation des erreurs s'appuie sur l'ordre de lecture

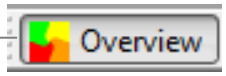


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

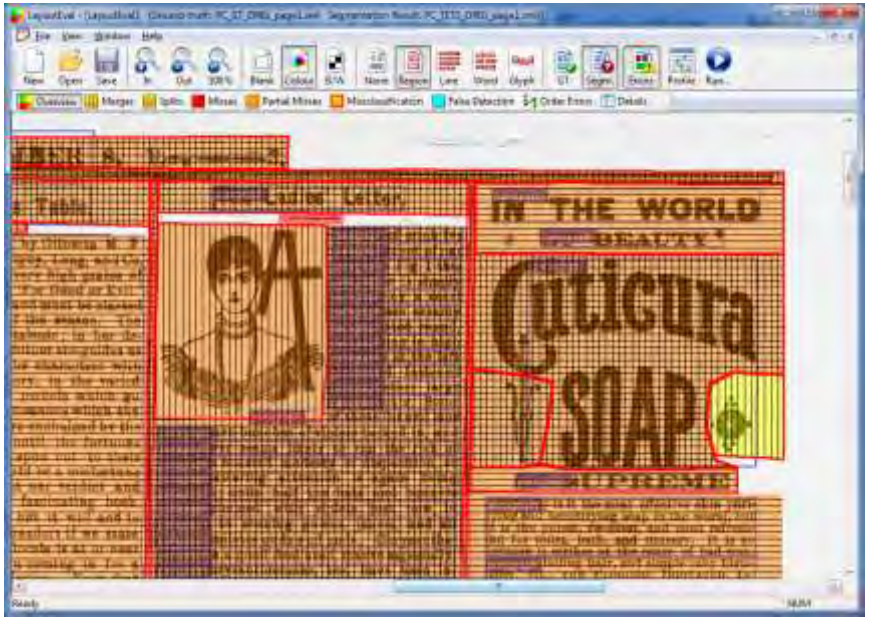
Évaluation de la segmentation – **En pratique**

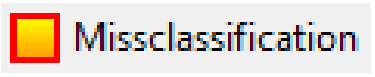
1. Activer l'affichage de la segmentation OCR

2. Chercher les erreurs (par code couleur dans le mode "Overview")



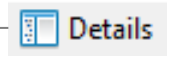
3. Visualiser les erreurs par catégorie (en surbrillance orange)



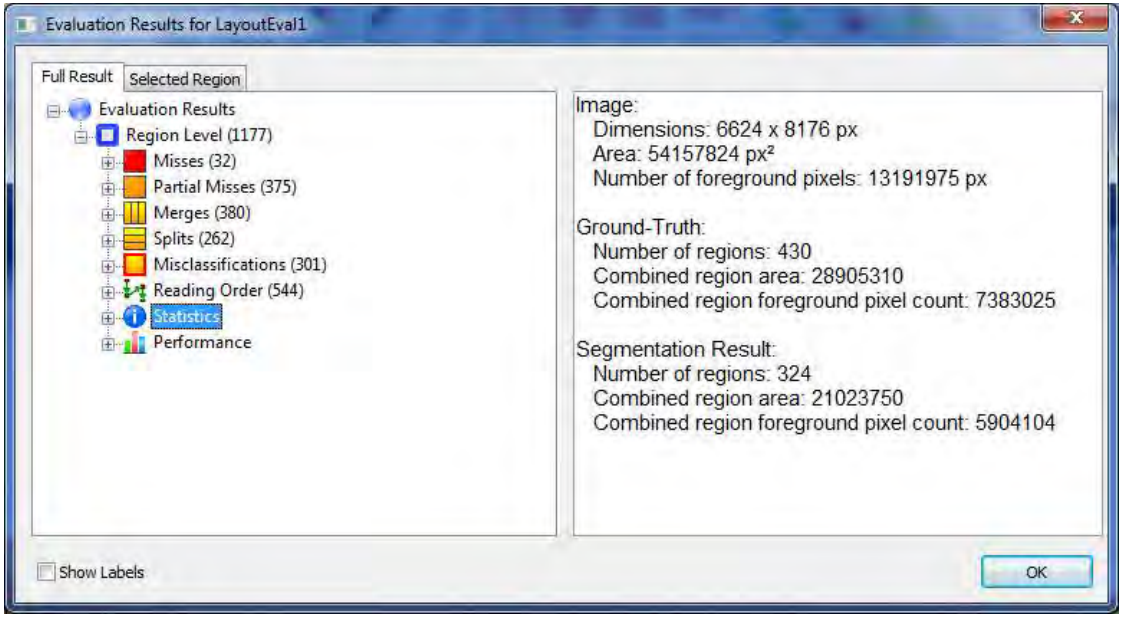


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Clic sur le bouton “Details” 

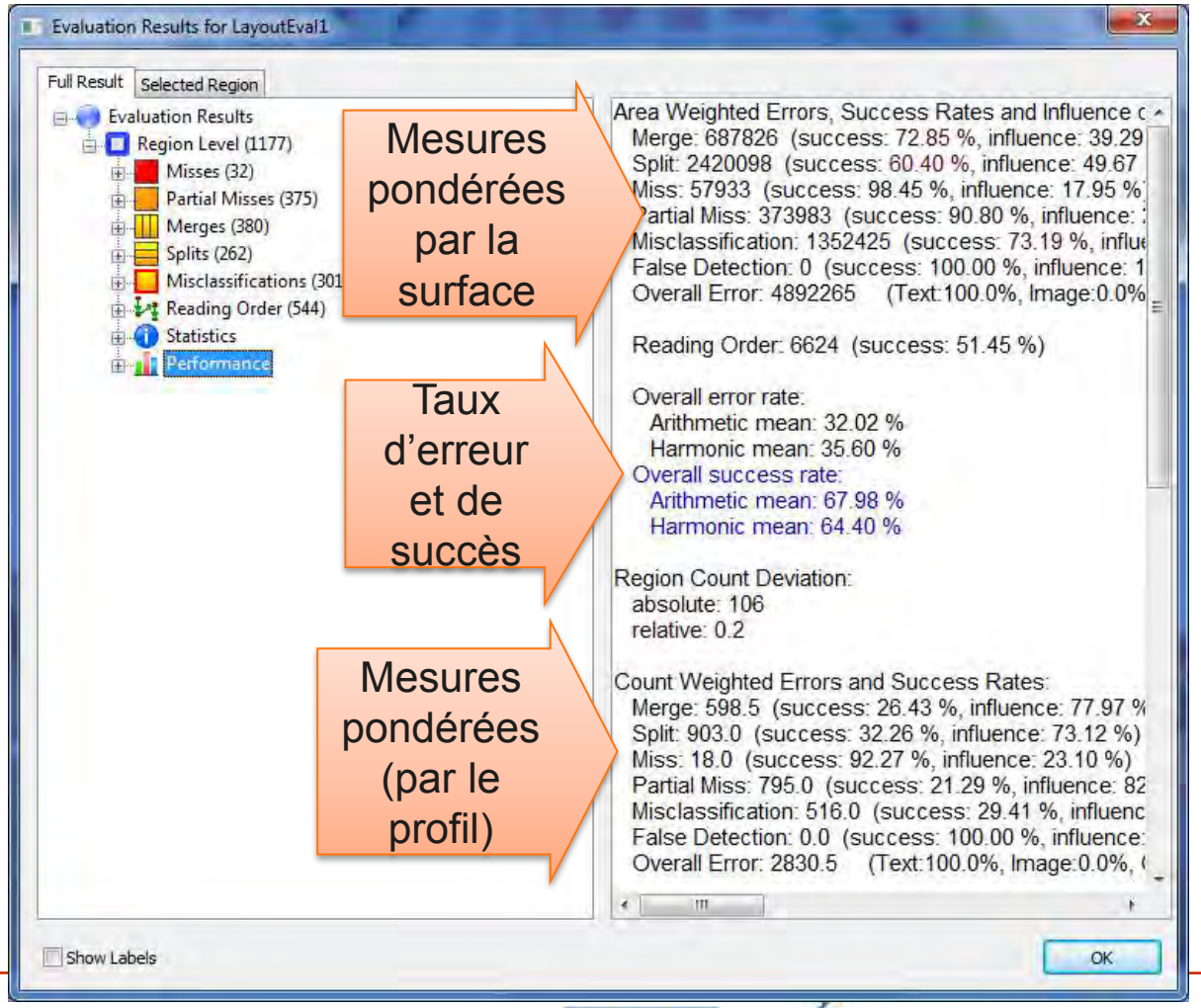
2. Clic sur “Statistics”



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Clic sur “Performance”
2. Consulter les taux d’erreur et de succès (en bleu les plus importants)

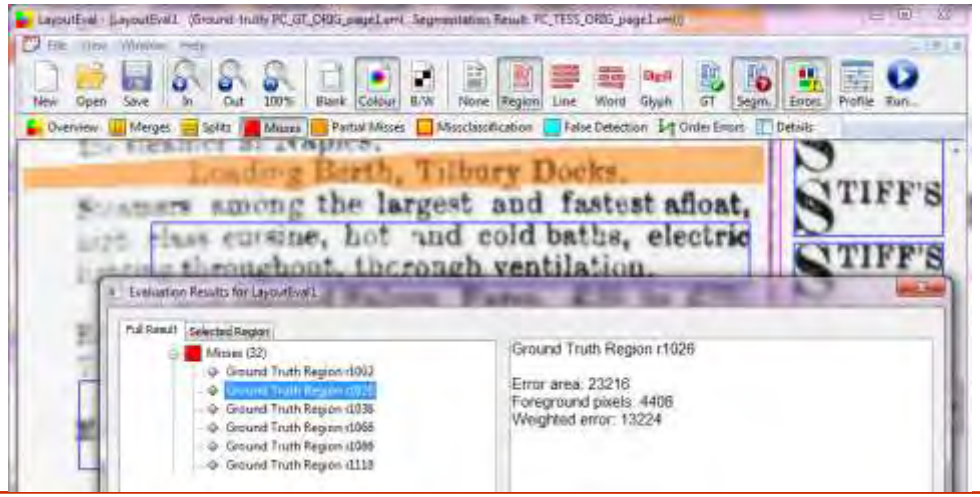


This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Activer la vue Vérité terrain (GT)

2. Analyser les résultats en détail, par ex. les omissions :
   1. Clic sur “Misses”
   2. Clic sur un élément de résultat.



Quantification

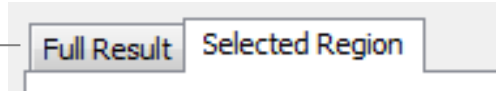
Ground Truth Region r1026

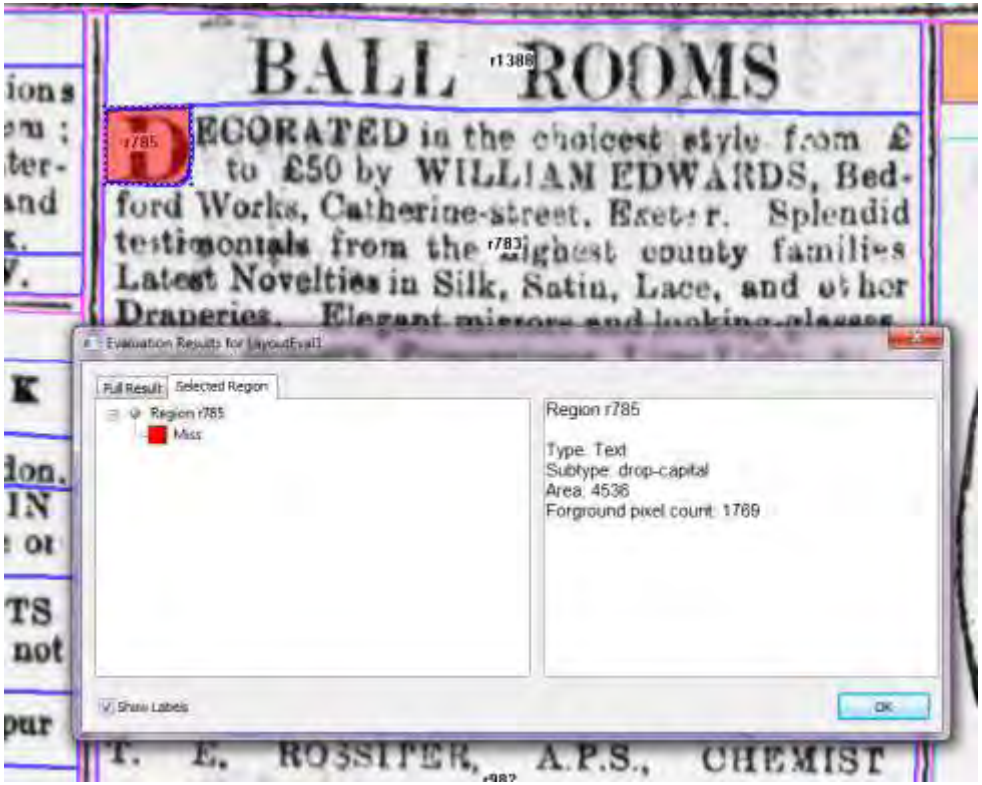
Error area: 23216
Foreground pixels: 4408
Weighted error: 13224

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

1. Sélectionner l'onglet "Selected region"
2. Sélectionner différentes regions : les types d'erreurs sont affichés

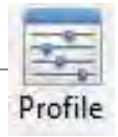




This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Évaluation de la segmentation – **En pratique**

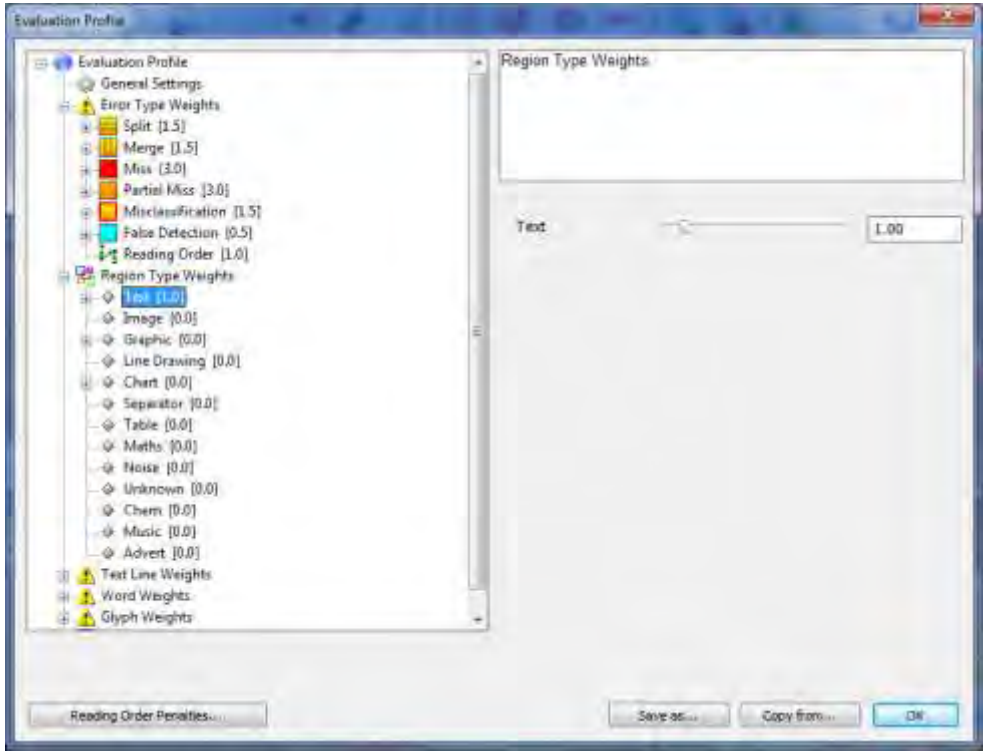
1. Clic sur “Profile”
2. “Region Type Weights” donne les poids associés aux natures de contenus



- Les poids des zones non textuels sont à 0
  ➔ les erreurs sur ces régions seront ignorées
- Même principe pour les catégories d’erreur (“Error Type Weights”)

Profils d’évaluation

- Paramétrer une évaluation pour un cas d’usage donné



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

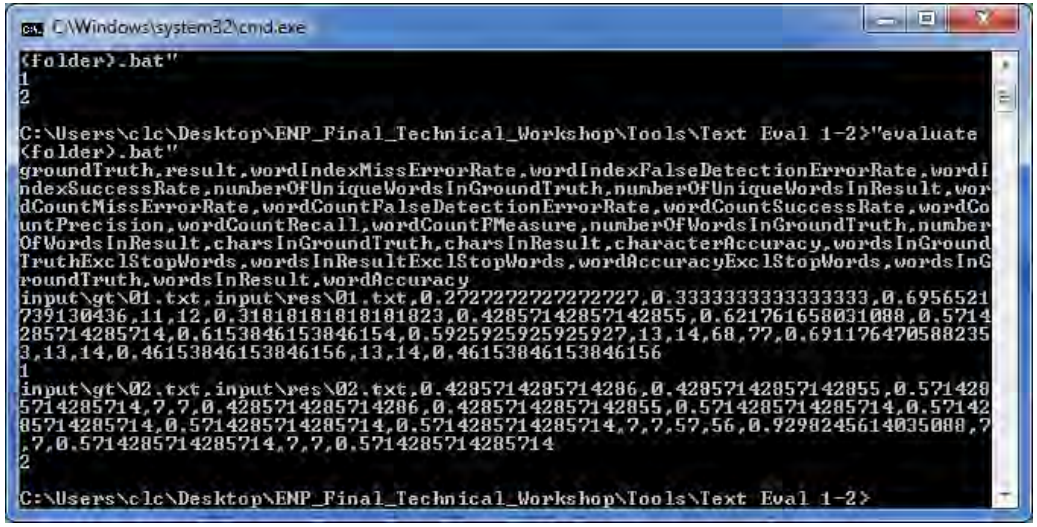
europeana
newspapers

Partie 6 : Analyse des performances

Méthodes, outils

Analyse des performances – **Automatisation**

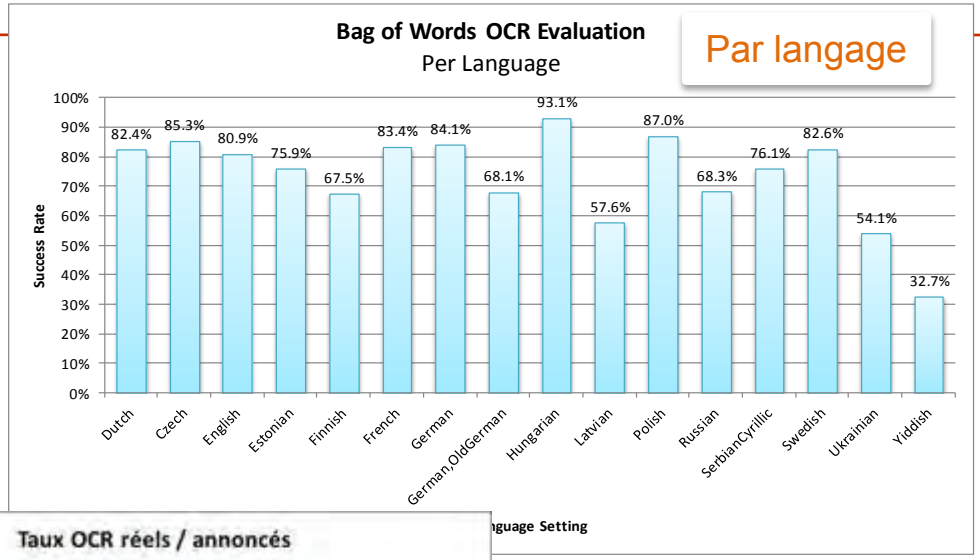
- Pour une campagne d'évaluation, nécessité de traiter de nombreux fichiers.
- Les outils sont disponibles en version ligne de commande.
- Traitement par boucle et sortie des résultats au format (CSV).



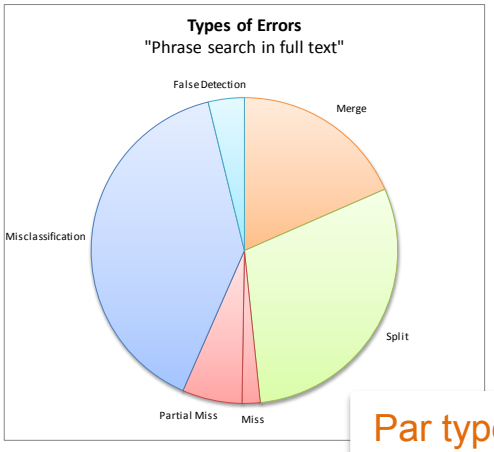
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Analyse des performances

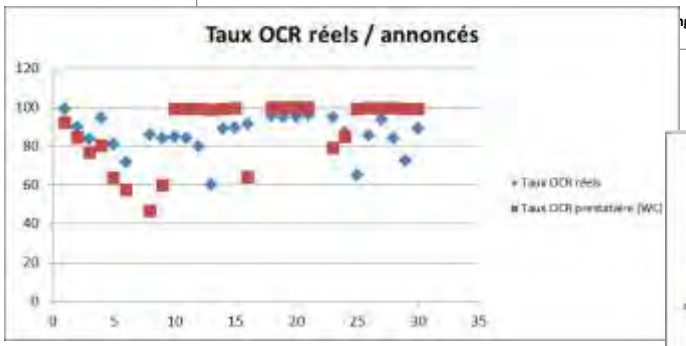
- Import dans un tableur des résultats CSV
- Création de moyennes, graphes, etc.



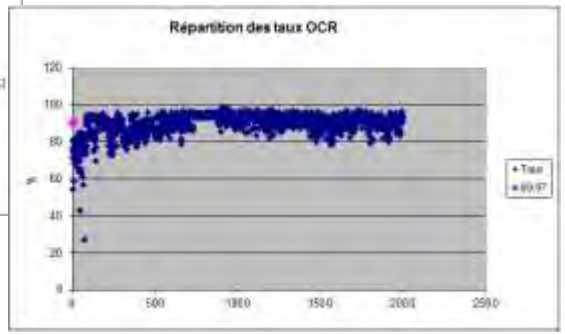
Par langage



Par types d'erreur



Par taux OCR



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Prise de décision – **Étude avant-projet**

Selon les résultats de l'analyse de performance, il peut être décidé :

- De lancer le projet de numérisation.
- De sélectionner un sous-ensemble du corpus pour favoriser la qualité OCR.
- D'améliorer le processus de numérisation.
- De choisir une combinaison des deux options précédentes.
- D'abandonner la numérisation du corpus.

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Prise de décision – **Contrôle qualité en production**

Selon les résultats de l'analyse de qualité, il peut être décidé :

- De demander une amélioration de qualité au prestataire ciblée sur les failles identifiées.
- D'auditer le processus de numérisation du prestataire afin de l'accompagner dans une démarche d'amélioration de la qualité.
- De rejeter les documents non conformes à la qualité attendue de la prestation.
- De dénoncer la prestation.

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Contact

Christian Clausner (USAL) / c.clausner@primaresearch.org

Jean-Philippe Moreux (BnF) / jean-philippe.moreux@bnf.fr