Atelier
Outils de numérisation de documents: Transcription, OCR, HTR...

Motasem ALRAHABI, ObTIC - Scai, Sorbonne Université

28/10/2021

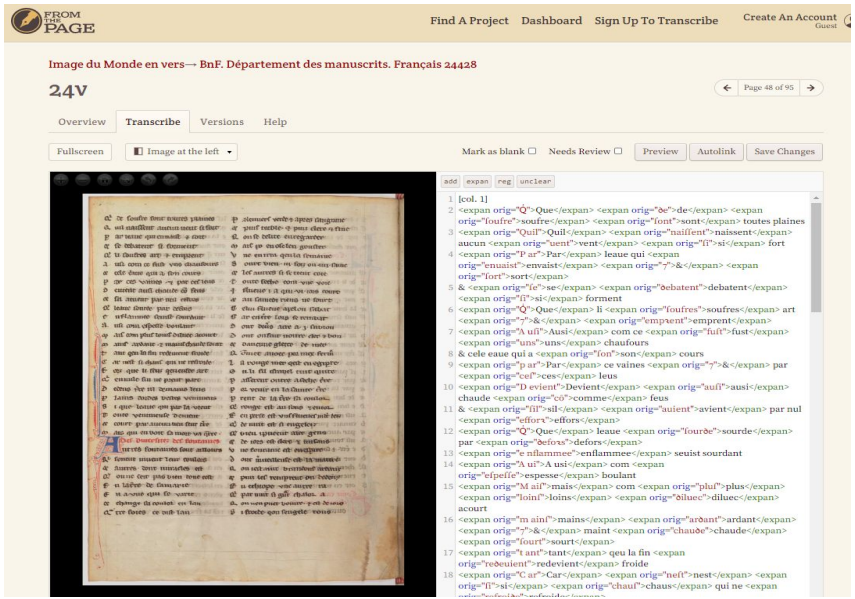
Plan

- Archives
- Transcription
- Numérisation de documents
- Prise en main de quelques outils: Abbyy, Kraken et Transkribus

Transcription

Transcription

- Il existe des outils pour aider à transcrire et annoter les images scannées.
- Ils proposant en vis-à-vis de l'image, un bloc texte généralement accompagné de fonctionnalités de mise en forme, de structuration et d'enrichissement.
- Exemples: From The Page
- Scribe, Transcrire, T-Pen…
  Plugins: e-man, Scripto
- Certains de ces outils permettent aussi le travail collaboratif: Recital, Transcrire, Tact, Transcribathon...



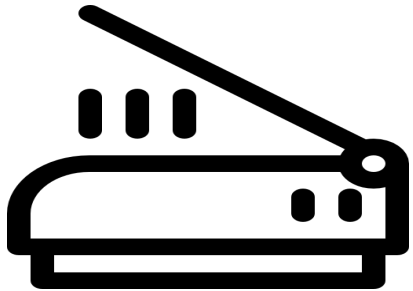
Numérisation

Numérisation et transformation digitale

- La numérisation de documents facilite l’analyse des contenus textuels ainsi qu’une meilleure qualité de lecture et de stockage.
  - Numérisation des fonds d'archives publiques, numérisation des fonds des bibliothèques, numérisation des documents d’entreprise (notes de frais, formulaires papier, factures, chèques, reçus…), reconnaissance automatique des plaques d’immatriculation, reconnaissance d’écriture manuscrite en temps réel (écran tactile...), dossiers hospitaliers, etc.

Numérisation

- La numérisation d’un texte s’effectue en trois phases distinctes :
  - Scanner les pages, reconnaître les chaînes de caractères et améliorer la qualité de la sortie.
- Photographie (scan) des pages:
  - La qualité de la photographie (jpeg, pdf…) a un impact sur l’étape suivante. Il est conseillé d’effectuer des images avec une résolution de 300 dpi (dots per inch).



Numérisation

- Océrisation: reconnaissance optique de la photographie de la page
  - Procédé informatique pour la transformation d'images de textes manuscrits (Handwritten Text Recognition), d'images imprimés ou dactylographiés (Optical Character Recognition) en fichiers de texte (texte brut, xml comme Alto...).
- Etapes:
  - Prétraitement : réalignement, corrections de contraste, binarisation (bicolore), détection de contours, suppression du bruit (déparasitage)…
  - Segmentation en lignes, en mots et en caractères, détection des blocs et des zones...
  - Reconnaissance proprement dite des caractères: comparaison avec une bibliothèque de formes connues et sélectionner les formes les plus proches, selon une distance ou une vraisemblance.

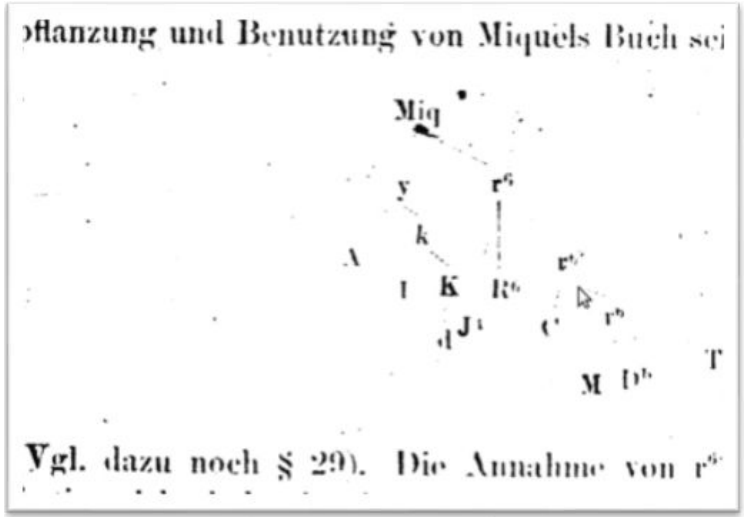
Numérisation

- Principe:
  - Comparer la forme à reconnaître avec une base de données de formes.
  - Le système choisit ensuite la forme la plus proche.
  - Techniques utilisées: apprentissage profond (réseaux de neurones), méthodes métriques (distance Levenshtein…), méthodes probabilistes (chaînes de Markov...)

Numérisation

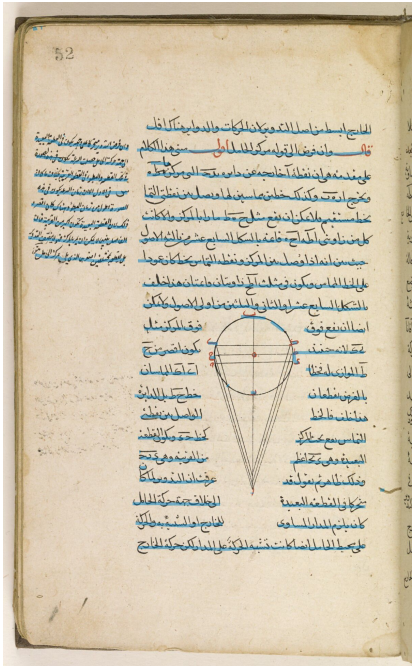
- Correction des erreurs: l'OCR ne permet pas toujours d'obtenir un résultat parfait.
- Utilisation de méthodes linguistiques et contextuelles pour réduire le nombre d'erreurs:
  - Systèmes à base de règles, méthodes statistiques basées sur des dictionnaires de mots, de syllabes, de N-grammes, éliminer des solutions incorrectes, etc.
- Exemples d'erreurs pour le français:

```
u / n
a / o
l / 1
! / 1
ponctuation
accentuation
...
```



Numérisation

- Les techniques d’OCR sont en progrès constant pour répondre à une demande très forte, mais la qualité de segmentation et de reconnaissance dépend de facteurs liés au document original et à sa numérisation.
- Difficultés pour l’OCR:
  - Différents styles typographiques (gras, italique, normal…) et de polices.
  - Documents historiques, anciennes graphies, noms obsolètes...
  - Diversité des langues, nombre variable du vocabulaire…
  - Défauts d’impression (caractères empâtés, bavures…)
  - Documents en colonnes ou illustrés, lecture non linéaire…
  - Polices (trop) petites ou grandes...
  - Résolution de l’image scannée, contraste et luminosité, qualité du support...
  - ...



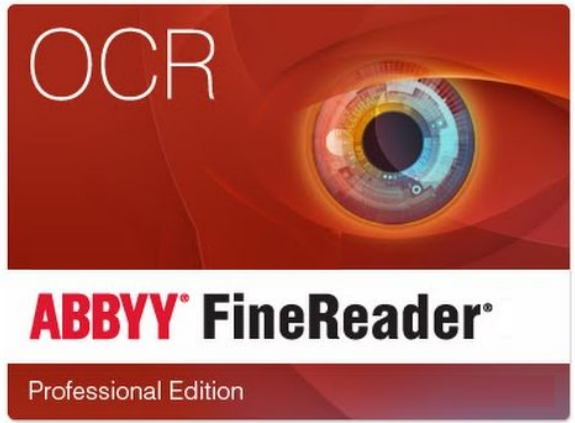
Numérisation

- Exemples d’outils:
  - Tesseract
  - Kraken
  - eScriptorium
  - Transkribus
  - Abbyy FineReader
  - Ocr4All
  - Google Cloud Platform : API Vision OCR
  - Microsoft Azure Cognitive Services : Vision par ordinateur OCR
  - Amazon Web Services : Amazon Textract
  - OCR Space
  - …

Abbyy FineReader

- Abbyy FineReader est un logiciel d’OCR propriétaire.
- Le processus se décompose en 3 étapes :
  - Ouvrir (numériser) le document, le reconnaître puis le sauvegarder dans un format courant (DOC, RTF, XLS, PDF, HTML, TXT, etc.) ou exporter les données directement vers une application de Microsoft Office telle que Microsoft Word, Excel ou Adobe Acrobat.

- Démonstration.



Kraken

- Développé en python (Kiessling, 2019), Kraken est un logiciel open source d'OCR et HTR qui fonctionne grâce à des réseaux de neurones récurrents.
- Kraken est en ligne de commande mais avec eScriptorium, il est en train de se doter d'une interface graphique.
- Il permet de binariser et de segmenter une page, de la transcrire (d'abord à la main pour créer les données d'entraînement, sinon automatiquement), d'effectuer l'entraînement d'un modèle d'OCR/HTR et la reconnaissance d'écriture.
- Export du résultat en texte brut ou au format XML ALTO.

Transkribus

- Transkribus est un logiciel avec interface graphique (GUI) qui permet de segmenter une page, de la transcrire (à la main ou automatiquement) et de l'annoter avant de l'exporter dans plusieurs formats possibles.
  - On doit d'abord transcrire manuellement une centaine de pages, afin de permettre d'entraîner un modèle de transcription. Le logiciel va alors reconnaître automatiquement le texte des pages suivantes et en proposer une transcription ligne à ligne.
- L'interface de Transkribus est open source, mais ce n'est pas le cas de son système d'entraînement pour la transcription automatique : les modèles de transcription ne sont donc accessible que par l'intermédiaire de Transkribus.