

Atelier 1 – Reconnaissance automatique de texte (ATR)

Atelier “Créer une édition scientifique numérique”

Floriane Chiffolleau
Ingénieure en humanités numériques à l'ObTIC
DataLab - BNF (Salle 70)
17 janvier 2025



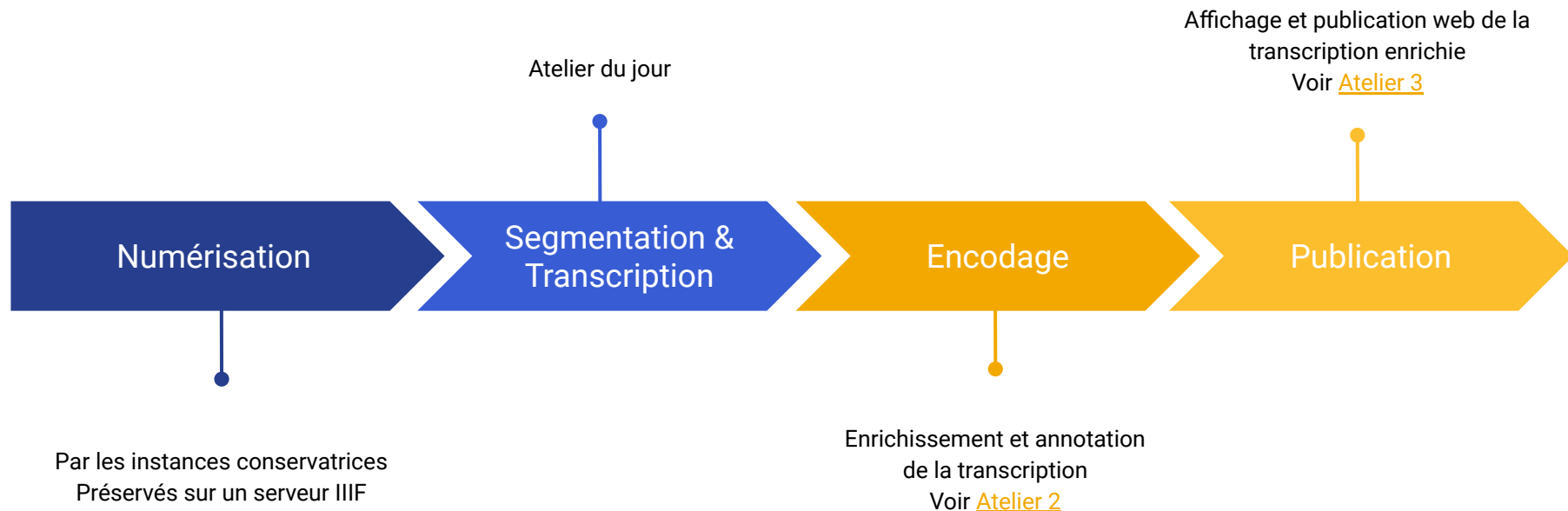
Sommaire

- ❑ L'ATR, une étape de la création d'éditions numériques
- ❑ Qu'est-ce qu'est l'ATR ?
- ❑ Quelques termes clés
- ❑ Kraken/eScriptorium
- ❑ Exercices pratiques
- ❑ Ressources

L'ATR, une étape de la création d'éditions numériques



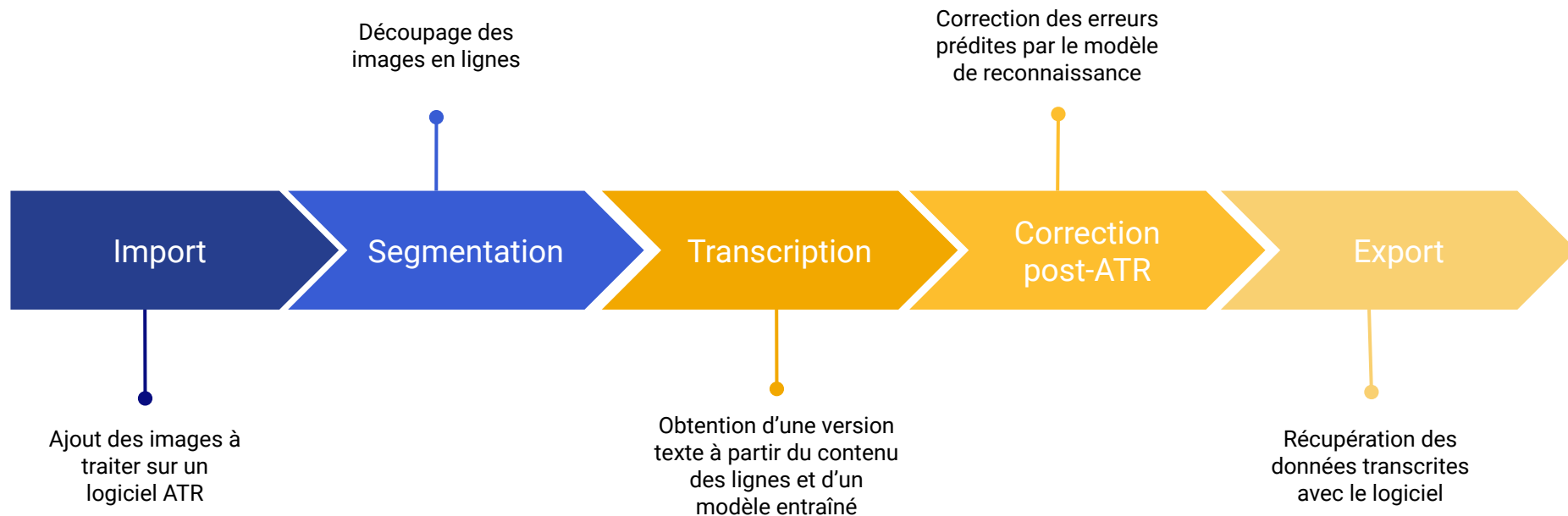
La création d'éditions numériques



Qu'est-ce qu'est l'ATR ?



Qu'est-ce que l'ATR ?



Quelques termes clés



Généralités

- ❑ **Automatic Text Recognition (ATR)** : Processus d'acquisition automatique, généralement avec des technologies d'apprentissage machine, de données textuelles numériques depuis un document analogue numérisé
- ❑ **Handwritten Text Recognition (HTR)** : Processus de reconnaissance et d'extraction de texte manuscrit depuis des images scannées d'écriture, en utilisant un système informatique
- ❑ **Optical Character Recognition (OCR)** : Conversion d'images d'un texte imprimé en un texte lisible par la machine. De nos jours, la plupart des systèmes s'appuient sur des réseaux de neurones, similairement aux techniques de l'HTR

L'entraînement de modèles

- ❑ **Entraînement** : Processus d'apprentissage d'un outil ATR pour produire un modèle
- ❑ **Fine-tuner** : Technique permettant de spécialiser un modèle d'apprentissage machine pré-entraîné sur une tâche spécifique
- ❑ **Modèle** : Fichier créé à la suite d'un processus d'entraînement, qui contient les paramètres utilisés pour produire une transcription depuis une image
- ❑ **Vérités de terrain** : Information que l'on sait vraie, qui, dans le contexte de l'ATR, fait référence à la transcription manuelle ou vérifiée d'un texte, et qui sert de donnée d'entraînement

La segmentation

- ❑ **Baseline/Topline** : Ligne virtuelle, passant par au moins deux points, sur laquelle du texte est écrit, qui sert de base à la reconnaissance de texte
- ❑ **Masque** : Polygone défini par au moins trois points de coordonnées, qui délimite la zone de pixels contenant le texte, pour une *baseline* ou une *topline* donnée.
- ❑ **Ordre de lecture** : Décompte, de haut en bas, des lignes créées par la segmentation
- ❑ **Segmentation** : Processus qui consiste à diviser une image en des régions ou des segments distincts, utilisé pour faciliter l'analyse, et identifier et tracer des objets ou des zones d'intérêt dans une image. Dans le cas de l'ATR, la segmentation peut être appliquée aux zones ou aux lignes d'un document textuel.

La transcription

- ❑ **Correction post-ATR** : Processus de correction automatique ou manuelle de la prédiction produite par l'ATR.
- ❑ **Gold corpus** : Données créées et vérifiées exclusivement par des humains, pour obtenir une transcription parfaite
- ❑ **Prédiction** : Acte d'utilisation d'un modèle pour générer la reconnaissance de la structure d'une image ou de son texte, en utilisant une image et un modèle de segmentation ou de transcription.
- ❑ **Silver corpus** : Données acquises par la prédiction du modèle réalisé à partir du *gold corpus*
- ❑ **Transcription** : Processus d'acquisition de données textuelles, lisibles par une machine, depuis un document analogue numérisé

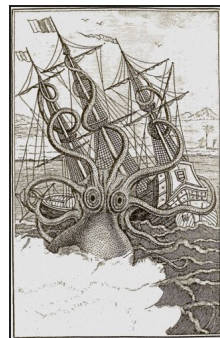
L'export de la transcription

- ❑ **Fichier de sortie** : Type de fichiers finaux utilisés pour présenter et partager la donnée après l'avoir traitée en suivant une méthode telle que l'ATR par exemple
- ❑ **Format de fichier** : Standardisation de la structure de l'information encodée et conservée dans les fichiers informatiques, qui permet l'organisation de la donnée, ainsi que la compatibilité, l'accessibilité et l'interprétation par des logiciels spécifiques
- ❑ **PAGE XML** : Standard XML qui permet d'encoder des documents numérisés et qui peut être utilisé pour afficher la structure d'une page et son contenu
- ❑ **XML ALTO** : Standard XML qui permet de rendre compte de la structure physique et logique d'un texte transcrit par ATR et qui conserve toutes les coordonnées géographiques du contenu (texte, illustrations, graphiques) dans l'image et permet à l'image et au texte d'être superposés

Kraken/eScriptorium

Un logiciel de reconnaissance et son interface

- ❑ Utilisation de **Kraken** et de son interface web **eScriptorium**
 - ❑ **Kraken** → **logiciel de reconnaissance de texte** qui permet également l'entraînement de modèles spécifiques pour des textes imprimés, tapuscrits et manuscrits
 - ❑ **eScriptorium** → **interface web** pour des projets collaboratifs de transcription automatique



Un logiciel de reconnaissance et son interface

- ❑ Plusieurs **types de transcriptions** possibles
 - ❑ **Automatique** → un modèle de transcription adapté au corpus existe déjà
 - ❑ **Semi-automatique** → création de *vérités de terrain* avec la transcription manuelle de votre corpus et ensuite, entraînement d'un modèle, soit *from scratch* (avec seulement les données créées), soit *finetune* (en utilisant ses vérités de terrain avec un modèle déjà existant)
 - ❑ **Manuelle** (non conseillé)

Importation des images

Train 

 Binarize
  Segment
  Transcribe
  Align

Statut de la binarisation, segmentation et transcription

16

eScriptorium Home Contact

All Search in new me

My Projects My Models Hello Floriane

Description Ontology Images Edit Models Reports new me

Element 1 - bpt6k1281160s_0011.jpg - (2202x3787) - 628.93 KB

manual

Différentes transcriptions disponibles

Texte segmenté

Transcription

LA FOIRE
S. GERMAIN.
COMEDIE.
ACTE PREMIER.

LA FOIRE
S. GERMAIN,
COMEDIE.
ACTE PREMIER.


Exercices pratiques

Import des images

- ❑ Sur le GitHub, allez au document intitulé [ark_gallica.txt](#)
- ❑ Copiez le contenu du document
- ❑ Allez sur Pandore Toolbox → Collecte de corpus → Gallica
- ❑ Dans la section approprié, coller les identifiants ARK et “Envoyer”
- ❑ Dézipper le fichier téléchargé

- ❑ Allez sur eScriptorium
- ❑ Créez un nouveau document
- ❑ Dans “Description”, mettez un titre et choisissez “*Latin*” comme script
- ❑ Une fois dans Documents, cliquez sur “*Click to upload*” et sélectionnez les images acquises depuis Gallica

Segmentation des images

- ❑ Sélectionnez tous les documents en cliquant sur “Select all”
- ❑ Cliquez sur “Segment”
 - ❑ Choisissez le modèle ???
 - ❑ Vérifiez que les options sont bien “Lines and regions” et “Horizontal l2r”
- ❑ Une fois fini, allez dans le premier document du dossier pour vérifier la segmentation
 - ❑ Rajoutez des lignes si certains éléments nécessaires n'ont pas été segmentés
 - ❑ Supprimez les lignes qui n'ont pas lieu d'être (par ex: décoration)
 - ❑ Vérifier le bon ordre des lignes en cliquant sur 

Transcription des images

- ❑ Allez sur le dépôt “[HTR/OCR Models](#)” de Zenodo
- ❑ Récupérez les modèles listés dans le GitHub dans le document [modeles.md](#)
- ❑ Sur eScriptorium, allez dans “My Models”
- ❑ Ajoutez les modèles sur l’instance

- ❑ Retournez dans votre dossier
- ❑ Sélectionnez les documents selon le modèle qui sera utilisé (Astuce : sélectionnez le premier document que vous voulez, appuyer sur Maj et sélectionnez le dernier)
- ❑ Cliquez sur “Transcribe”
 - ❑ Choisissez le modèle correspondant à votre sélection
 - ❑ Vérifiez que le choix sélectionné pour “Select a transcription” est bien “-- New --”

Correction post-ATR des transcriptions

- ❑ Cliquez sur le premier document du dossier
- ❑ Choisissez le modèle qui correspond aux documents sur lequel vous travaillez
- ❑ Cliquez sur la première ligne du document
- ❑ Comparez l'image avec le texte et corrigez selon le besoin
- ❑ Réitérez avec les autres types de documents

Export des transcriptions

- ❑ Cliquez sur “Select all” ou sélectionnez manuellement les transcriptions à exporter
- ❑ Cliquez sur “Export”
- ❑ Sélectionnez la prédiction que vous souhaitez exporter
- ❑ Deux versions :
 - ❑ Choisissez “Text” et appuyez sur “Export” → Un nouvel onglet va s’ouvrir avec votre export en version texte
 - ❑ Choisissez “ALTO” ou “PAGE” et appuyez sur “Export” → Un fichier zip se télécharge. Il contient les versions XML des transcriptions avec les coordonnées de segmentation conservés, ainsi que des métadonnées dans un fichier METS

Ressources

Publications

- ❑ Chagué, Alix, Floriane Chiffolleau, and Hugo Scheithauer (Aug. 2024), “*Collaboration and Transparency: A User-Generated Documentation for eScriptorium*”, in: DH2024 Reinvention & Responsibility, Alliance of Digital Humanities Organizations, Washington D. C., United States, url: <https://hal.science/hal-04594142>
- ❑ Chagué, Alix and Thibault Clérice (July 2023), “*“I’m here to fight for ground truth”: HTR-United, a solution towards a common for HTR training data*”, in: Digital Humanities 2023: Collaboration as Opportunity, Alliance of Digital Humanities Organizations and University of Graz, Graz, Austria, url: <https://inria.hal.science/hal-04094233>
- ❑ Gabay, Simon and Ariane Pinche (2021), “*SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)*”, in: url: <https://hal.science/hal-03336528/>
- ❑ Kiessling, Benjamin (Dec. 2019), *Kraken - a Universal Text Recognizer for the Humanities*, fr, doi: 10.34894/Z9G2EX, url: <https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX>
- ❑ Kiessling, Benjamin et al. (Sept. 2019), “*eScriptorium: An Open Source Platform for Historical Document Analysis*”, in: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 2, pp. 19–19, doi: [10.1109/ICDARW.2019.10032](https://doi.org/10.1109/ICDARW.2019.10032)

Ressources web

- ❑ Kraken : <https://kraken.re/main/index.html>
- ❑ eScriptorium : LIEN
- ❑ Documentation eScriptorium : <https://escriptorium.readthedocs.io>
- ❑ XML ALTO : <https://www.loc.gov/standards/alto/>
- ❑ PAGE XML : <https://github.com/PRIImA-Research-Lab/PAGE-XML>
- ❑ HTR-United : <https://htr-united.github.io/index.html>
- ❑ Modèles HTR/OCR : https://zenodo.org/communities/ocr_models/

Merci de votre attention

Des questions ?

[https://github.com/FloChiff/
AtelierObTIC-creer-une-edition-scientifique-numerique](https://github.com/FloChiff/AtelierObTIC-creer-une-edition-scientifique-numerique)
[chiffolleau.floriane\[at\]gmail.com](mailto:chiffolleau.floriane[at]gmail.com)