Quelques perspectives actuelles en humanités numériques

Reconnaissance de l’écriture manuscrite, transcription collaborative, annotation de documents patrimoniaux avec IIIF

Pierre Willaime

Archives Henri-Poincaré (Nancy)

2022-09-28

Les humanités numériques ?

http://whatisdigitalhumanities.com/

> an opportunity to ask new questions, try new methods, engage in new conversations (ssenier)

Une définition (parmi d'autres)

> « Le développement des humanités numériques comme application du développement de l'informatique dans un champ spécifique permet à la fois d'en éclairer les ressorts profonds et d'interroger la place que peuvent occuper les humanités dans des sociétés sous l'influence des technologies numériques. »

Pierre Mounier, *Les humanités numériques : Une histoire critique*, 2018

Une (des) histoire(s)

Quelques jalons :

- Robert Busa (1948) et l'index thomisticus
- Linguistique computationnelle, *Humanities computing*
- Text Encoding Initiative (1987)
- A Companion to Digital Humanities (2004)
- THATCamp Paris (2010)

En histoire :

- Article d'Adeline Daumard et François Furet *Méthodes de l'Histoire sociale : les Archives notariales et la Mécanographie* (1959)
- Le Médiéviste et l'ordinateur (1979)

Des communautés, des manifestes

- Stéphane Pouylau, "Les digital humanities ont-elles existé ? (2014)
- Éric Guichard, "Les humanités numériques n’existent pas" (2019)

Lecture distante / Lecture proche (Franco Moretti, 2000)

- Une histoire globale, "longue durée",
- permise par des moyens informatiques,
- qui ne remplace pas mais complète l'analyse "proche", manuelle

Variété des usages, des métiers, évolutions

Rôle et responsabilités de l'éditeur numérique

D'une autorité à un facilitateur

> But the social edition is something different. The idea here is to offer the text to the community not only for contributions such as annotation, comments, and translations but also for the editing of existing texts or the addition of new texts. The meaning of ‘social’ is then twofold: social in the way a text is edited (social editing) or in the way it is released to the public (‘social edition’).

Pierazzo, Elena. “Digital Scholarly Editing: Theories, Models and Methods,” 2014. https://hal.univ-grenoble-alpes.fr/hal-01182162, p.24.

Outils et technologies

- Annoter (le texte, les images).
- Transcrire,
- manuellement et à plusieurs,
- collégialement, participativement.

International Image Interoperability Framework (IIIF)

- Constat que la diffusion des images sur le web était « trop lente, trop coûteuse, trop disjointe et trop complexe ».
- lancée en 2015 par la British Library, la Bibliothèque Nationale de France, Die Bayerische Staatsbibliothek, Nasjonalbiblioteket, Artstor, Wellcome Trust et les universités Cornell, Oxford, Princeton, Stanford et Yale
- Un consortium.
- Un cadre technique commun et standardisé,
- pour la diffusion mais aussi la manipulation des images.

Les images deviennent véritablement accessibles, consultables, comparables, manipulables, citables, annotables et mixables par n’importe quelle application compatible capable de se « brancher » sur les entrepôts des uns et des autres

Concrètement

- Deux APIs (Image, Presentation) = spécifications

https://gallica.bnf.fr/iiif/ark:/12148/btv1b6001280q/manifest.json

```
{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}{.format}
```

```
https://gallica.bnf.fr/iiif/ark:/12148/btv1b10224708f/f1/full/full/0/native.jpg
```

https://iiif.io/api/image/2.1/#identifier

https://demos.biblissima.fr/chateauroux/osd-demo/

Web Annotation Data Model

Transcrire à la main mais à plusieurs

Pourquoi transcrire collectivement ?

Des raisons "mécaniques"

- corpus trop gros
- difficulté de lecture et d'interprétation
- nécessité d'une expertise

ou d'autres "choisies"

- éviter les biais
- rendre explicite tous les choix de transcription (même ceux qui semblent anodins)
- l'occasion pour une équipe de s'approprier un corpus

Défis d'une transcription collective

1. Pouvoir échanger autour de la transcription
    - "guidelines"
    - "la sagesse des foules" (Surowiecki)
2. Pouvoir agir sur le même document (quasiment) en même temps
3. Pouvoir voir qui a fait quoi
4. Pouvoir revenir à une version antérieure de la transcription
5. Élaborer une processus de relecture et de validation des transcriptions.

Omeka (Classic ou S)

Un garant d'interopérabilité.

- Système de gestion de contenus conçu pour la publication de documents d’archives
- Omeka S : version qui s’inscrit dans l’esprit du Web sémantique, avec l’utilisation d’ontologies et la création de liens entre des ressources
- Logiciel libre, modulable, développé aux USA par Digital Scholar
- Mutisites (extension du projet au-delà de la correspondance)
- Interropérable (protocoles (API REST, OAI-PMH, IIIF), schémas de métadonnées)
- une communauté active (https://omeka.fr), journées Omeka 2016 (Paris), 2019 (Poitiers), 2020 (Nancy), 2021 (Paris), 2022 (Grenoble)

Scripto

- une installation Omeka (Classic ou S)
- une installation mediawiki *sur le même serveur*
- un module/plugin à installer côté Omeka
- qui permet l'interconnexion via l'api de mediawiki

Démonstration

WarDepartmentPaper

https://transcrire.huma-num.fr/

https://testaments-de-poilus.huma-num.fr/

http://test.ahp-numerique.fr/omeka-s/admin/scripto/1/item

- Tact

https://tact.demarre-shs.fr/

1. Sous quelles conditions transcrire peut être une activité collective ?
2. Le doit-elle seulement ?

(comparaison avec les instruments de recherche en archivistique)

Reconnaissance automatique de l'écriture

OCR, HTR, ICR, quésako ?

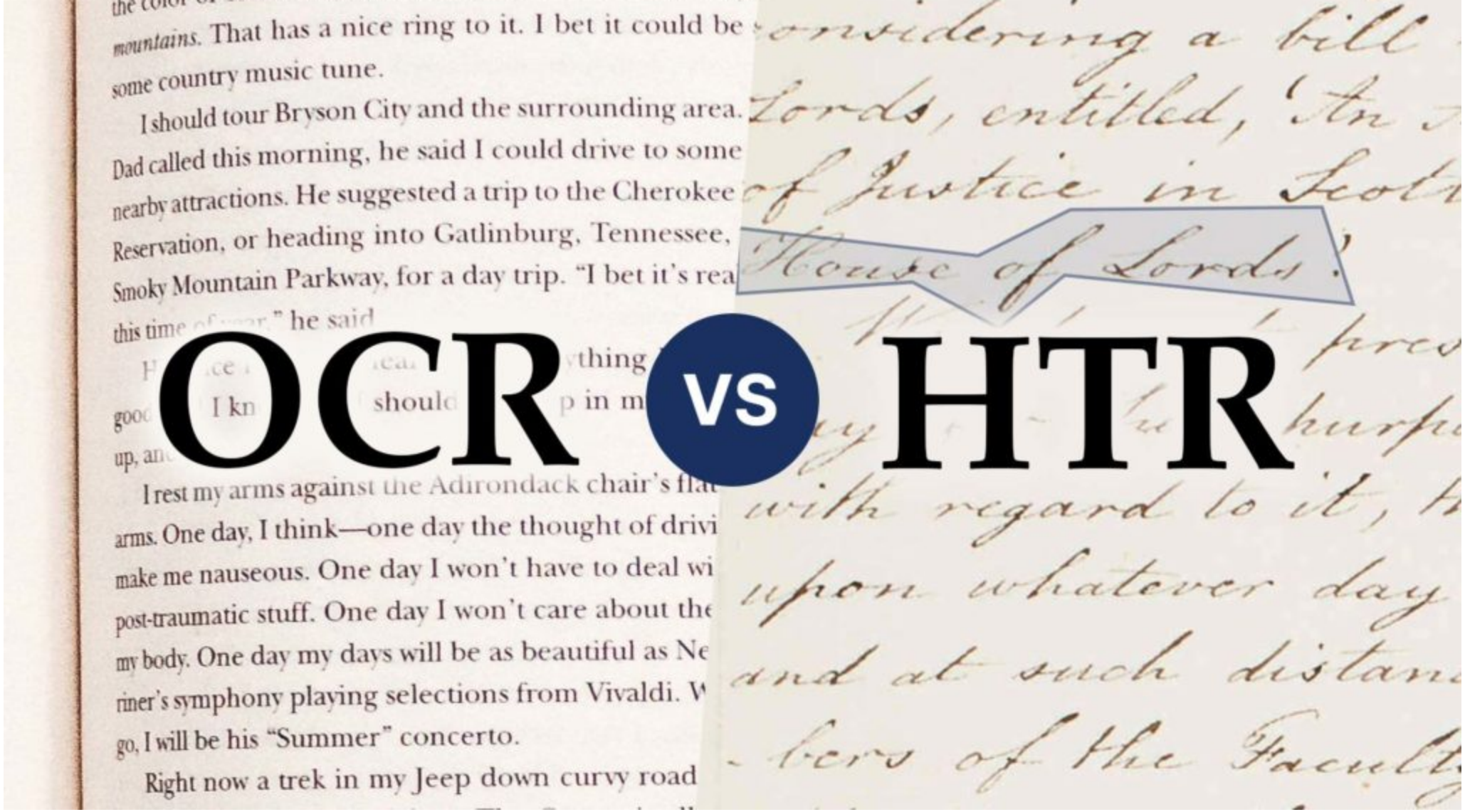
> Optical Character Recognition
> ~~Intelligent Character Recognition~~ (daté)
> Handwritten Text Recognition

L'OCR travaille au niveau des glyphes, l'HTR a un niveau plus global (souvent la phrase).

Étapes :

1. Analyse de l'image et pré-traitement
2. Segmentation
3. Reconnaissance
4. Post-traitement

HTR : basée sur l'apprentissage et la construction de modèles (pour la segmentation *et* pour la reconnaissance).



Opposition trompeuse car il s'agit plutôt de deux technologies (une vieillissante et une en construction).

On peut faire de l'"HTR" sur des documents dactylographiés...

Logiciels

Ancienne génération

- ABBYY (propriétaire)
- Tesseract < 4

Nouvelle génération (réseaux de neurones)

- Tesseract >= 4 (Architecture 'Long short-term memory' (LSTM))
- Kraken
- Transkribus (modèle économique basé sur des services payants)
- Calamari (Kraken)
- eScriptorium (kraken)

Liste plus complète

eScriptorium

http://localhost:8080/project/bbk/documents/

> HN : Usage réflexif du numérique pour tenter de répondre aux problématiques de recherche en Humanités.

- Pas simplement une mise à disposition d'outils !
- Un renouvellement (et questionnement) des pratiques de recherche en Humanités.
- Une communauté de pratiques
- Des dynamiques, des interactions entre disciplines, corps de métiers, ...
- Cenhtor https://cenhtor-msh-lorraine.cnrs.fr/
- OLIO https://olio.hypotheses.org/