



Acquisition d'un objet d'étude

Ljudmila PETKOVIC

Introduction aux humanités numériques (L1HN001)
Mineure « Humanités numériques », licence Lettres
Paris, le 28 septembre 2023, année 2023-2024

Les projets en humanités numériques

Quatre principales étapes :

1

2

3

4

Les projets en humanités numériques

Quatre principales étapes :

1 Acquisition d'un objet d'étude

2

3

4

Les projets en humanités numériques

Quatre principales étapes :

- 1 Acquisition d'un objet d'étude
- 2 Traitement d'un objet d'étude
- 3
- 4

Les projets en humanités numériques

Quatre principales étapes :

- 1 Acquisition d'un objet d'étude
- 2 Traitement d'un objet d'étude
- 3 Exploitation d'un objet d'étude
- 4

Les projets en humanités numériques

Quatre principales étapes :

- 1 Acquisition d'un objet d'étude
- 2 Traitement d'un objet d'étude
- 3 Exploitation d'un objet d'étude
- 4 Publication des résultats

Les projets en humanités numériques

Quatre principales étapes :

- 1 Acquisition d'un objet d'étude
- 2 Traitement d'un objet d'étude
- 3 Exploitation d'un objet d'étude
- 4 Publication des résultats

⚠ NB:

Les frontières entre ces étapes sont très poreuses et l'ordre des opérations ne respecte pas toujours cette progression logique.

Les projets en humanités numériques

Quatre principales étapes :

- 1 Acquisition d'un objet d'étude
- 2 Traitement d'un objet d'étude
- 3 Exploitation d'un objet d'étude
- 4 Publication des résultats

⚠ **NB:**

Les frontières entre ces étapes sont très poreuses et l'ordre des opérations ne respecte pas toujours cette progression logique.

Acquisition d'un objet d'étude

Objet d'étude

Source de travail des humanistes

Texte

Types

- littérature, articles scientifiques, documents techniques, documents publicitaires ou politiques, notes de cours, lettres, mails, SMS, blogs, etc.

Formes

- manuscrite, imprimée, numérique mode image, numérique mode texte

Son

Types

- discours, dialogues et conversations, musique, bruitages divers, etc.

Formes

- enregistrements, transcriptions du texte

Image

Types

- fixe (photo, dessin, esquisse), ou animée (film, captation vidéo, dessin animé, etc.)

Formes

- analogique, numérique

Artefact ou œuvre d'art

Types

- tableau, sculpture, bas-relief, *ready-made*, installation, poterie, bâtiment, ustensiles divers, etc.

Formes

- objets, images (→ forme analogique ou numérique), mentions ou descriptions (→ texte : manuscrit, imprimé, numérique)

Données

Types

- quantitatives (fermées)
 - tout ce qui peut être compté ou mesuré
 - résultats exprimés en chiffres (statistiques, données numériques)
- qualitatives (ouvertes)
 - descriptives, se référant aux phénomènes observables, mais pas mesurables (p. ex. couleurs, émotions)
- mixtes (qualitatives + quantitatives)
- recueillies : enquête de terrain, entretiens, extractions de bases de données
- reconstituées (recherches d'archives, études géomorphologiques, interprétation de photographies aériennes)

Formes : voir formes du texte, du son, de l'image

Archives

Archives

1. Collection de documents anciens, classés à des fins historiques

Archives

1. Collection de documents anciens, classés à des fins historiques
2. Lieu où les archives sont conservées

Archives

1. Collection de documents anciens, classés à des fins historiques
2. Lieu où les archives sont conservées

Corpus

- collection numérique de textes, fragments de texte et/ou transcriptions
- écrit ou parlé
- sélection : la *meilleure* représentation possible d'une langue, d'un dialecte ou d'un type de texte particulier
- rend la collection dans son ensemble une source fiable pour la recherche linguistique

Enjeu principal de l'acquisition des données

Format natif → format numériquement et pleinement exploitable

- *machine readable*, lisible / analysable par la machine, texte dynamique

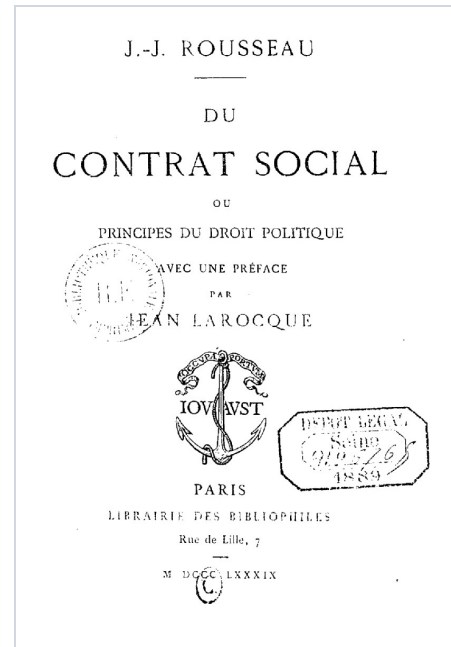
Numérisé

Numérisé

L'archive **numérisée**, à l'inverse de l'archive numérique, n'est pas nativement électronique. Il s'agit du résultat de la reproduction d'une archive initialement physique au moyen d'un outil numérique, tel qu'un scanner ou un appareil photo.

Numérisé

L'archive **numérisée**, à l'inverse de l'archive numérique, n'est pas nativement électronique. Il s'agit du résultat de la reproduction d'une archive initialement physique au moyen d'un outil numérique, tel qu'un scanner ou un appareil photo.



Numérique

[...] tout type de document produit aujourd'hui sous une forme électronique et dématérialisée. [...] toute archive « nativement » (c'est-à-dire dès sa création) **numérique**. Cette archive est faite à la fois de données (le contenu du document à proprement parler) et de métadonnées (les informations sur le document, telles que sa date de création, son auteur, etc.).

Numérique

[...] tout type de document produit aujourd'hui sous une forme électronique et dématérialisée. [...] toute archive « nativement » (c'est-à-dire dès sa création) **numérique**. Cette archive est faite à la fois de données (le contenu du document à proprement parler) et de métadonnées (les informations sur le document, telles que sa date de création, son auteur, etc.).

- contenu généré par l'utilisateur (angl. *user-generated content*, *UGC*)
 - issue des réseaux de communication, c'est-à-dire Internet, la messagerie électronique et les réseaux sociaux

OCR

OCR

Reconnaissance optique de caractères (angl. *optical character recognition*)

OCR

Reconnaissance optique de caractères (angl. *optical character recognition*)

Processus qui consiste à convertir un ensemble de signes graphiques, le plus souvent alphanumériques (mais aussi les ponctuations, espacements...), encodés sous la forme d'une image, en mode texte. L'OCR désigne à la fois un processus (d'OCR) et un logiciel (d'OCR).

(Camps & Perreaux, 2021)

OCR

Reconnaissance optique de caractères (angl. *optical character recognition*)

Processus qui consiste à convertir un ensemble de signes graphiques, le plus souvent alphanumériques (mais aussi les ponctuations, espacements...), encodés sous la forme d'une image, en mode texte. L'OCR désigne à la fois un processus (d'OCR) et un logiciel (d'OCR).

(Camps & Perreaux, 2021)

HTR

OCR

Reconnaissance optique de caractères (angl. *optical character recognition*)

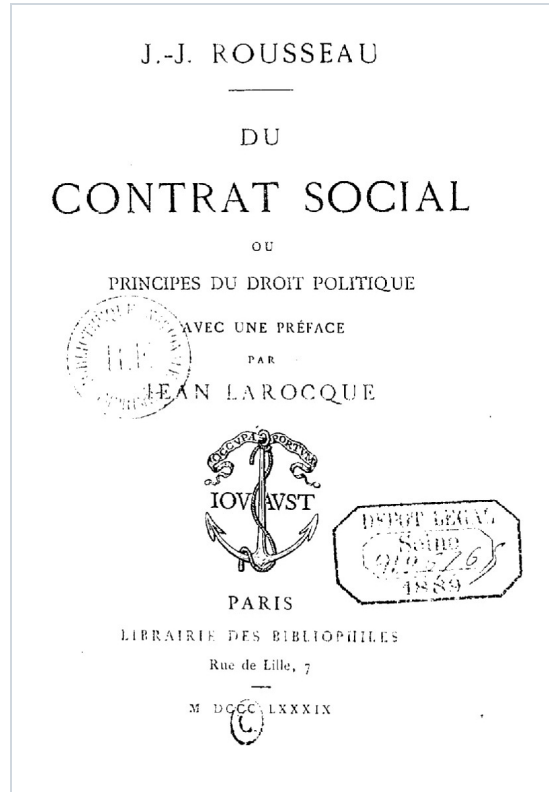
Processus qui consiste à convertir un ensemble de signes graphiques, le plus souvent alphanumériques (mais aussi les ponctuations, espacements...), encodés sous la forme d'une image, en mode texte. L'OCR désigne à la fois un processus (d'OCR) et un logiciel (d'OCR).

(Camps & Perreaux, 2021)

HTR

Reconnaissance de l'écriture manuscrite (angl. *handwritten text recognition*)

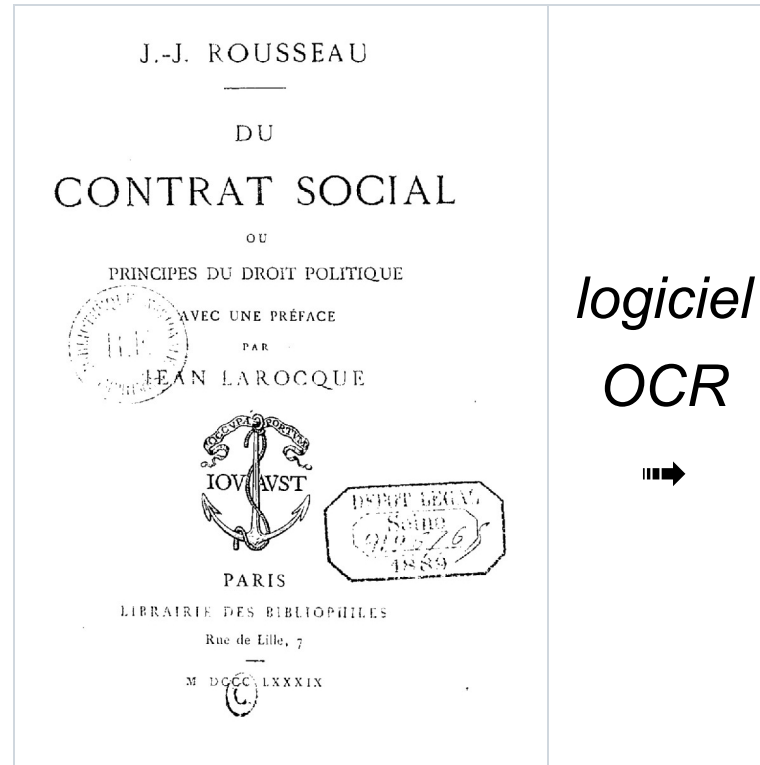
Du numérisé vers le numérique — exemple d'une sortie d'OCR



PDF, JPG, PNG, TIFF...

Source : Gallica (Bibliothèque nationale de France) <https://gallica.bnf.fr/ark:/12148/bpt6k5564953b>

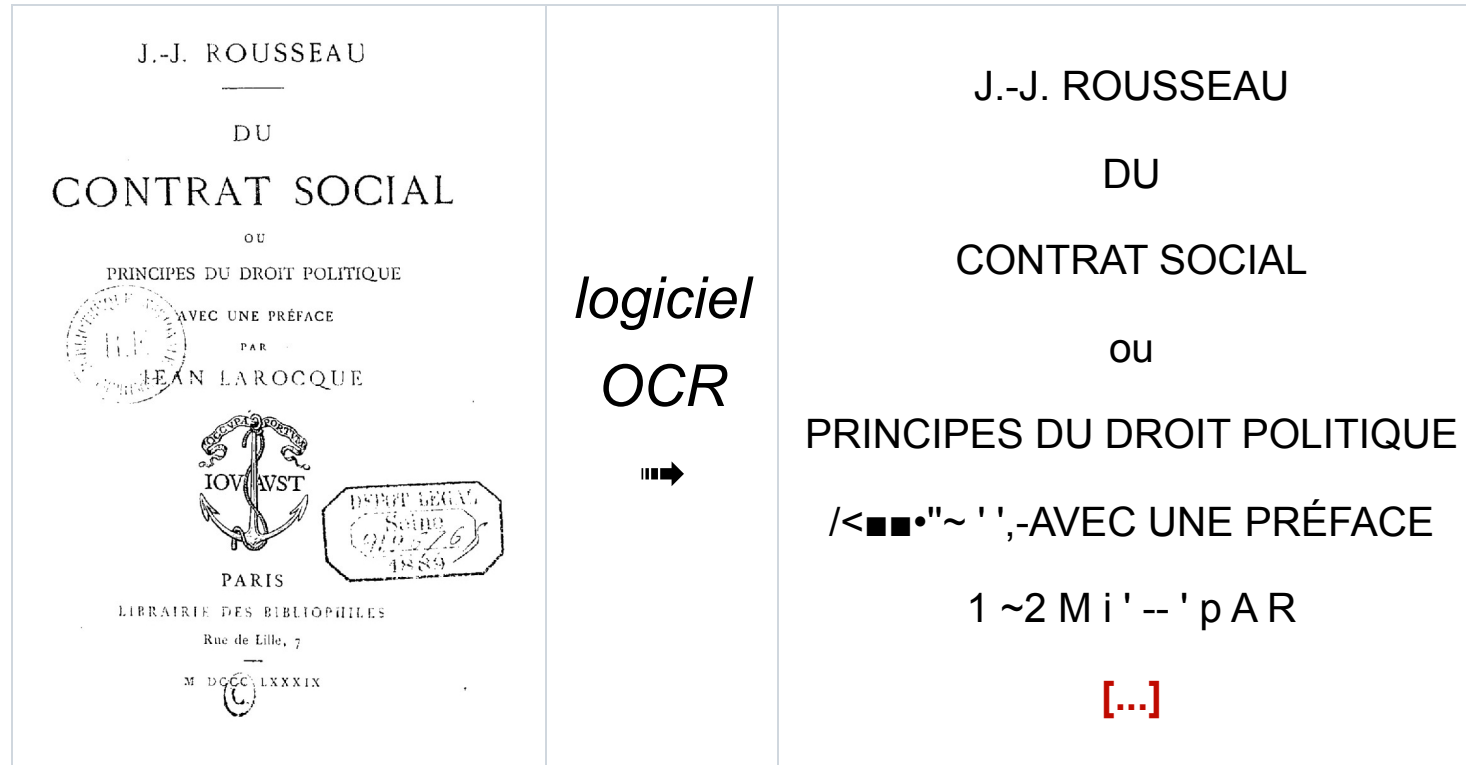
Du numérisé vers le numérique — exemple d'une sortie d'OCR



PDF, JPG, PNG, TIFF...

Source : Gallica (Bibliothèque nationale de France) <https://gallica.bnf.fr/ark:/12148/bpt6k5564953b>

Du numérisé vers le numérique — exemple d'une sortie d'OCR



PDF, JPG, PNG, TIFF...

Transcription automatique (.txt)

Source : Gallica (Bibliothèque nationale de France) <https://gallica.bnf.fr/ark:/12148/bpt6k5564953b>

Wikisource

- bibliothèque de 367 836 textes gratuits, librement disponibles et téléchargeables
- <https://fr.wikisource.org/wiki/Wikisource:Accueil>
 - exemple : J.-J. Rousseau, *Du contrat social* (1762)

The screenshot shows the Wikisource interface for the text 'Du contrat social/Édition 1762' by Jean-Jacques Rousseau. The page layout includes a left sidebar with navigation links, a top header with user status and search, and a main content area with the title and a table of contents.

Left Sidebar:

- Wikisource la bibliothèque libre
- Accueil
- Index des auteurs
- Portails thématiques
- Aide au lecteur
- Contacteur Wikisource
- Texte au hasard
- Auteur au hasard
- Contribuer
- Scriptorium
- Forum des nouveaux
- Aide
- Communauté
- Livre au hasard
- Modifications récentes
- Faire un don
- Imprimer / exporter
- Version imprimable
- Télécharger en EPUB
- Télécharger en MOBI
- Télécharger en PDF

Top Header:


- Non connecté(e) | Discussion | Contributions | Créer un compte | Se connecter
- Texte | Source | Discussion | Orthographe originale ▼
- Lire | Modifier | Voir l'historique
- Rechercher sur Wikisource

Main Content:

- Du contrat social/Édition 1762** (with a star icon and a 'Télécharger' button)
- < Du contrat social
- Pour les autres éditions de ce texte, voir *Du Contrat social*.
- Jean-Jacques Rousseau**
- Du contrat social**
- Marc Michel Rey, 1762.
- Texte sur une seule page
- TABLE**
- DES LIVRES**
- ET DES**
- CHAPITRES.**

Pandore

<https://pandore-toolbox.isir.upmc.fr/>

 Pandore toolbox


PandorePandoreOutilsProjetDocumentationFR
EN


Pandore : une boîte à outil pour les humanités numériques


Projet


Pandore offre un ensemble de modules permettant d'effectuer automatiquement les tâches les plus courantes liées au traitement de corpus pour la recherche en humanités numériques. Des chaînes de traitement permettant d'automatiser un ensemble de tâches sont également proposées.


Voir la démo


**OCR/HTR**
Conversion d'images en texte


**Conversion de formats**
Formatage XML-TEI, conversion de divers formats de fichiers.

**Fouille et annotation de texte**
Reconnaissance d'entités nommées, étiquetage morphosyntaxique, analyse de sentiments

**Visualisation**
Tanagra (entités de lieux sur carte)
Minerva (réseaux de cooccurrences)
Ariane (polarités textuelles)

**Collecte de corpus**
Scraping personnalisé des corpus
Wikisource

**Correction textuelle**
Correction d'erreurs et normalisation pour corpus à la graphie non standard

**Chaines de traitement**
Traitement automatique depuis l'OCR jusqu'à la reconnaissance d'entités et leur visualisation.

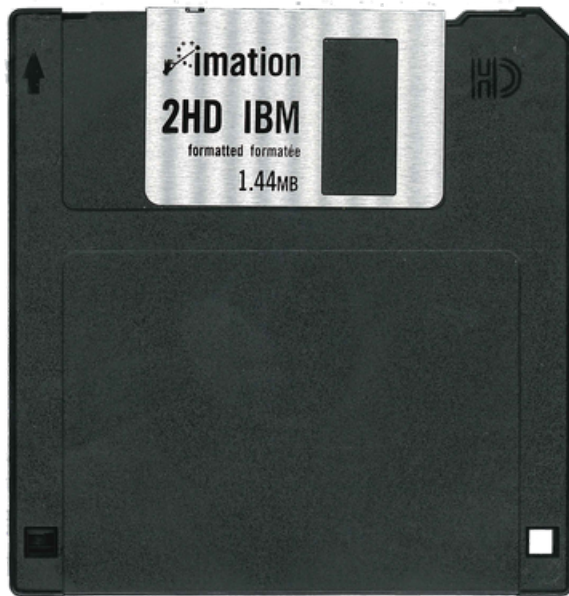
Gallica (BnF)

- bibliothèque numérique de la Bibliothèque nationale de France et de ses partenaires
- en ligne depuis 1997, elle s'enrichit chaque semaine de milliers de nouveautés et offre aujourd'hui accès à plusieurs millions de documents

<https://gallica.bnf.fr>

Obsolescence d'un support de stockage de données

- tout ce qui est numérique ne peut pas constituer un objet d'étude
- des opérations supplémentaires (remédiation, récupération, transformation) doivent être entreprises



Références

- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*].
- **Barbier, J., & Mandret-Degeilh, A.** (2018). 8. Les archives numériques et numérisées. Dans : , J. Barbier & A. Mandret-Degeilh (Dir), *Le travail sur archives: Guide pratique* (pp. 195-222). Paris: Armand Colin. <https://www.cairn.info/le-travail-sur-archives--9782200621056-page-195.htm>
- **Camps, J.-B., & Perreaux, N.** (2021). Reconnaissance optique des caractères et des écritures manuscrites · Projet E-NDP [*diapositives*] https://outils.lamop.fr/lamop/mp3/E-Ndp/JBC-NP_e-NDP_OCR-et-HTR.pdf