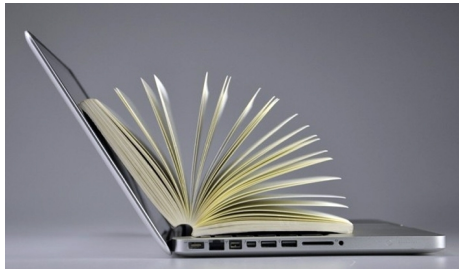
Sorbonne Nouvelle
université des cultures



Méthodes de numérisation
Formats

Ljudmila PETKOVIC

Introduction aux humanités numériques (L1HN001)
Mineure « Humanités numériques », licence Lettres
Paris, le 5 octobre 2023, année 2023-2024

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.

Méthodes de numérisation

Enjeu principal de l'acquisition des données

Format natif → format numériquement et pleinement exploitable

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

Numérisation

> Processus qui consiste à convertir des informations d'un support (texte, image, audio, vidéo) ou d'un signal électrique en données numériques.

Glossaire humanités numériques :

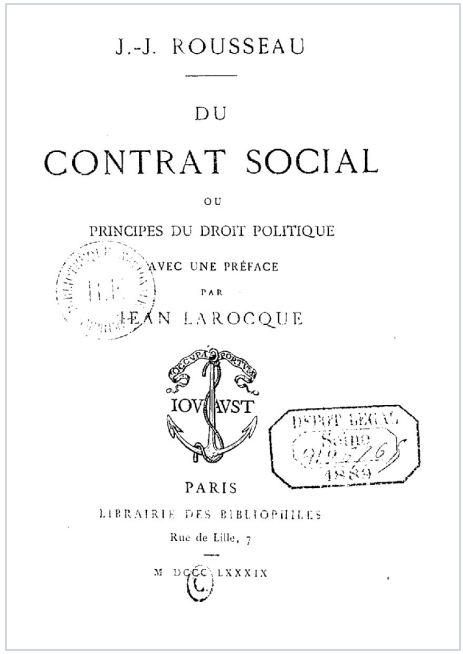
- PDF : http://docplayer.fr/199284324-Glossaire-introduction-aux-humanites-numeriques-version-du-31-01-2020.html
- graphe : https://www.arthurperret.fr/digithum-glossaire-hn.html

Rétro-numérisation : la numérisation d'une œuvre conçue et publiée à une époque antérieure

- p. ex. SDS online https://www.ssrq-sds-fds.ch/fr/projets/retronumerisation/

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.


J.-J. ROUSSEAU

DU

CONTRAT SOCIAL

OU

PRINCIPES DU DROIT POLITIQUE

AVEC UNE PRÉFACE

PAR

JEAN LAROCQUE

IOVI AVST

PARIS

LIBRAIRIE DES BIBLIOPHILES

Rue de Lille, 7

M DCCC LXXXIX

Du texte sur support matériel au texte numérique

Supports matériels du texte : papier, papyrus, pierre, bois, tablettes d'argile, etc.

Acquisition du document numérique

Mode image

- modes d'acquisition : photographie, scannage, radiographie, IRM
- formats : jpg, png, tiff, iiif, pdf, etc.
- cas de figure : campagnes de rétro-numérisation en bibliothèque et centres de documentation ; numérisations « maison »
- objectif : mise en œuvre, publication et partage rapide de ressources
- inconvénients : format peu exploitable (lecture)
    - p. ex. : Fonds Charcot https://patrimoine.sorbonne-universite.fr/fonds/item/3597-manuscrit-n-22-lecons

Du texte sur support matériel au texte numérique

Mode texte · saisie manuelle

- formats : docx, txt, odt, xml
- types de documents : manuscrits, imprimés anciens
- cas de figure :
  - chercheurs individuels + quantité limité de texte
  - *crowdsourcing* : production participative de grands volumes de données
- avantages : très grande fiabilité du texte saisi, analyse concomitante de sa structuration, éventuellement annotation
- inconvénients : grande consommation de ressources, chronophage
  - p. ex. : projet « Bentham papers » https://blogs.ucl.ac.uk/transcribe-bentham/

Du texte sur support matériel au texte numérique

Mode texte · OCR

- formats : txt, xml, pdf multicouche
- types de documents : grands volumes, manuscrits ou imprimés ; projets d'envergure
- avantages : rapidité du traitement
- inconvénients : erreurs typographiques de reconnaissance des caractères, caractères manquants, interversion de l'ordre de lecture, etc.
    - p. ex. : Destouches *Le Philosophe marié* https://gallica.bnf.fr/ark:/12148/bpt6k5654522c/f2.item.texteImage

Du texte nativement numérique au corpus

Textes nativement numériques

- fichiers texte produits avec différents matériels et logiciels ⇨ préservation et récupération de données « anciennes »
- Rétro-informatique (*retrocomputing*, analyse forensique)
  - utiliser du matériel et des logiciels informatiques obsolètes
  - remédiation : problème de l'accès aux données (conservation sur des supports inexistants, périphériques disparus, mises à jour)
  - histoire des techniques, p. ex.
    - Pathfinders : http://dtc-wsuv.org/wp/pathfinders/description/
    - AGRIPPA : http://agrippa.english.ucsb.edu/category/the-book-subcategories/the-poem-running-in-emulation

Du texte nativement numérique au corpus

- production écrite dans la sphère numérique : blogs, pages internet, tweets, commentaires, etc. ⇨ collecte sur le web

Méthodes de recueil :

1. **transcription manuelle** (ou « copier-coller »)
    - chronophage, biais de sélection, problèmes d'encodage de caractères
2. ***(web)scraping*** (moissonnage, *harvesting*, utilisation d'un *spider / bot*…)
    - rapide et puissant, mais complexe (programmation)
    - obstacles : code HTML de la page envisagée mal formaté, page peu structurée, barrières techniques (systèmes d'authentification, *cookies*, blocage d'accès en masse, codes CAPTCHA, *paywalls*, etc.), problèmes juridiques
    - librairies Python : Scrapy, Beautiful Soup

Formats

Qu’est-ce qu’un format numérique pleinement exploitable ?

- Format informatique = convention pour représenter une donnée sous forme numérique
- Un format peut être :
  - spécifié
  - ouvert
  - normalisé
  - standardisé
  - propriétaire

CINES

Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles

Spécifié

- format suffisamment décrit pour en développer une implémentation complète
- sous la forme d'un fichier au format pdf, text ou xml, en une ou plusieurs langues

Non spécifié

- format que l’on peut déduire de la forme produite par un logiciel mais dont la description n’est pas explicitement donnée et est souvent intimement mêlée au code du logiciel
- quand aucune forme ne peut être appliquée, on parle généralement de fichier binaire car c’est la seule connaissance que l’on a du fichier
    - tous les fichiers informatiques sont binaires, car c’est le mode d’écriture de l’information sur tous les supports numériques jusqu’à présent

Ouvert

> On entend par standard ouvert tout protocole de communication, d'interconnexion ou d'échange et tout format de données **interopérable** et dont les spécifications techniques sont **publiques** et **sans restriction** d'accès ni de mise en œuvre.

- légalement exempté de droits d'utilisation
- **compréhensible** : sa description ou spécification est publique, tout le monde peut prendre connaissance de la manière dont les informations sont organisées au niveau de ce format
- **intéropérable** : il est alors possible de créer une variété non limitée de programmes qui l'exploitent

Normalisé

- un format est normalisé quand la description du format est adoptée par un organisme de normalisation, comme p. ex. :
  - ISO : *Organisation Internationale de normalisation* http://www.iso.org/
  - IEEE : *Institute of Electrical and Electronics Engineers* http://www.ieee.org/
  - W3C : *World Wide Web Consortium* http://www.w3.org/

Standardisé

- format conforme à un standard

  ***XML* (angl. eXtensible Markup Language)**

  - format standard pour structurer les données sous forme de fichiers en entrée et de sortie pour les définitions stockées
  - langage de balisage généraliste recommandé par le consortium W3C
  - sous-ensemble du langage *SGML* (angl. *Standard Generalised Mark-up Language*)
  - XML n'est pas prédéfini, vous devez définir vos propres balises
  - objectif : partage de données entre différents systèmes, tel qu'Internet

Propriétaire

- son cadre d'utilisation est contrôlable par une personne ou une entité juridique
  - p. ex. via le droit d’auteur, le brevet ou le *copyright*
  - format dont l’utilisation est fortement restreinte par les droits que possède son propriétaire, et si la spécification n’est même pas consultable
  - ⚠️ format PDF est ouvert (ses spécifications sont libres d’accès et les programmes tiers peuvent réutiliser son format), même s'il est propriétaire (Adobe Systems)

Qu’est-ce qu’un format numérique pleinement exploitable ?

- format qui permet le partage et l’interopérabilité, garantit la propriété des données
- un format pérenne
  - qui permet de conserver et d’accéder au document sur le long terme
  - quatre types de risque:
    - obsolescence matérielle
    - obsolescence logicielle
    - obsolescence du format de fichier
    - la perte de signification du contenu

Qu’est-ce qu’un format numérique pleinement exploitable ?

- Un format qui permet d’envisager de multiples opérations :
  - lecture : lecture traditionnelle (cursive), lecture simultanée (multiples supports – multiples utilisateurs), lecture adaptée (handicaps)
  - édition : modifier le texte en vue d’en proposer une nouvelle version (enrichie, augmentée, abrégée, fragmentée, intercalée, interpolée, etc.)
  - comptage : unités prédéterminées vs. non-prédéterminées (Michael Charles)
  - exploration : mots dans leur contexte, récurrences, régularités, motifs, etc.

Formats ouverts

- privilégier les formats ouverts, standardisés ou normalisés, quel que soit le type de source à traiter (texte, image, son, données)
- en pratique : focus sur les formats de gestion du texte
- trois familles de logiciels :
  - éditeurs de texte
  - logiciels de traitements de texte
  - éditeurs xml

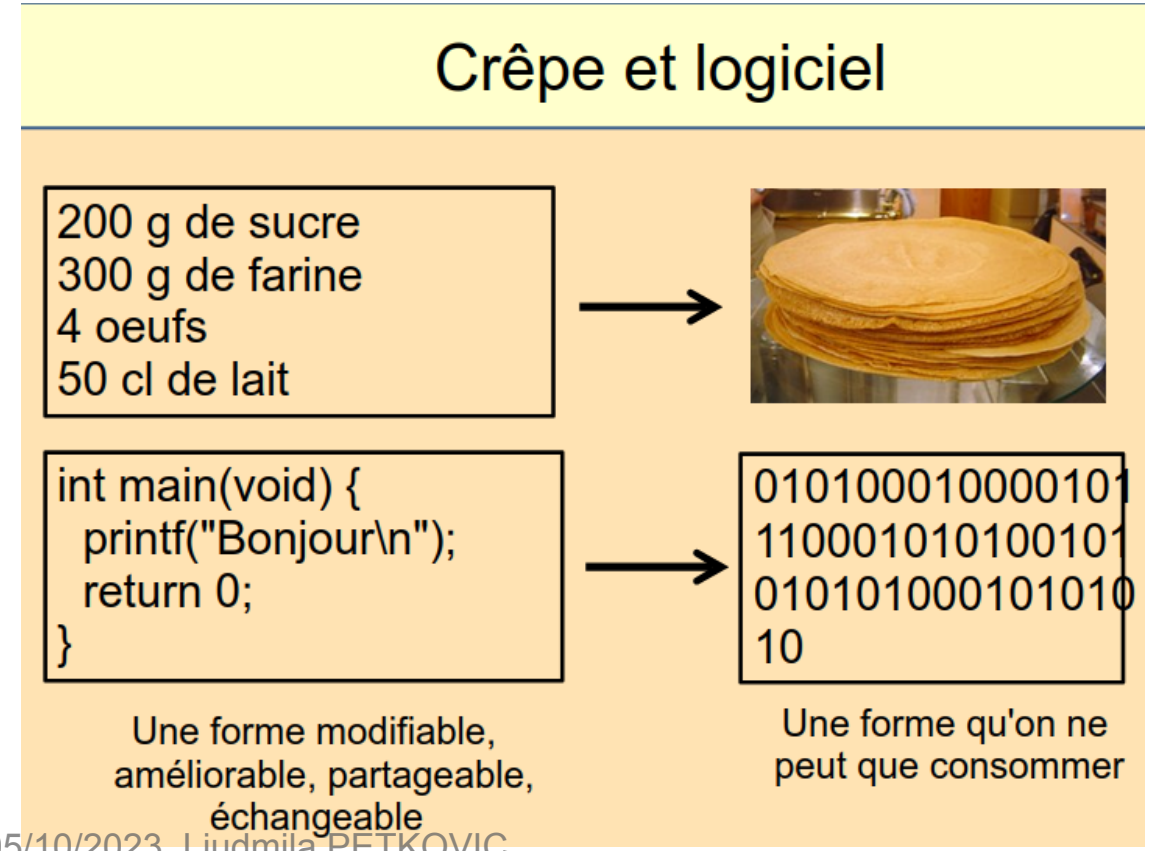
	Traitement de texte	Éditeur de texte	Éditeur XML
Logiciels	Word, Libre Office, Open Office...	Notepad, BlocNotes, TextEdit, SublimeText...	XMLMind, oXygen
Formats	doc, docx, odf (odt, ods, odp…)	txt	xml
Orientation	Présentation, impression	Écriture de contenu non mis en forme	Gestion de contenu
Avantages	Simple d'utilisation, nombreuses fonctionnalités de mise en forme, mise en pages, ajout d'images...	Pas de balises cachées	Larges possibilités d'annotation et de structuration
Inconvénients	Balises cachées	Aspect austère, peu de fonctionnalités de mise en pages...	Contraintes syntactiques fortes, complexité

Une analogie de Thomas Petazzoni (crêpes)

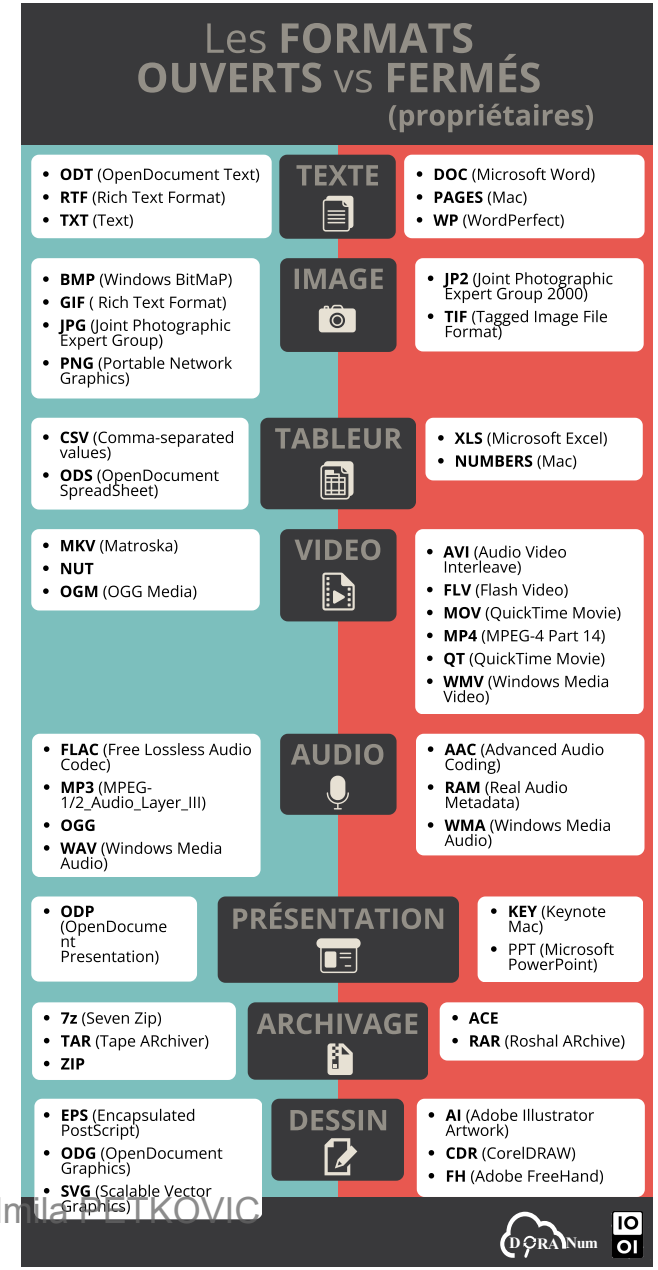
- Imaginez un monde
  - où les crêpes ne sont disponibles que toutes prêtes
  - où la recette des crêpes n'est pas disponible
  - où il ne viendrait à personne l'idée d'avoir la recette
- Ce monde existe : la face « visible » du monde du logiciel depuis les années '80

Crêpe et logiciel

- Deux formes pour un logiciel
  - code source : équivalent de la recette dans notre histoire de crêpe
  - exécutable : équivalent du plat cuisiné



Formats ouverts vs. fermés



Retour sur Pandore

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

Exercice :

- OCRiser *Du contrat social* avec Tesseract (*cf.* iCampus)
- ouvrir le fichier OCRisé dans un éditeur de texte et essayer de faire une première recherche d'un mot souhaité en tapant sur Ctrl + F (ou aller dans l'onglet Rechercher)

Références

- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*].
- **Perret, A.** (2020). Glossaire. Introduction aux humanités numériques. Version du 31/01/2020 http://docplayer.fr/199284324-Glossaire-introduction-aux-humanites-numeriques-version-du-31-01-2020.html.
- **Projet TGE-Adonis** (2010). Guide méthodologique pour le choix de formats numériques pérennes dans
  un contexte de données orales et visuelles
  https://francearchives.gouv.fr/file/d5e3b871af9d7b8b5000f8ffb1260646bcc7675b/static_3774.pdf.
- **Petazzoni, T.** (*s.d.*). « Les Logiciels Libres Principes et enjeux » [*diapositives*] https://slideplayer.fr/slide/11135877/.