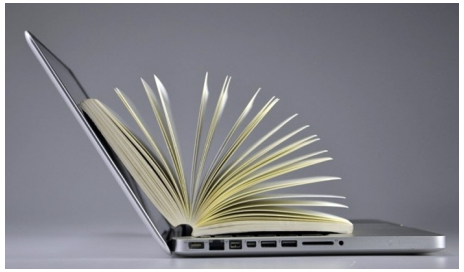
Sorbonne Nouvelle
université des cultures



Introduction aux humanités numériques

Ljudmila PETKOVIC

Introduction aux humanités numériques (L1HN001)
Mineure « Humanités numériques », licence Lettres
Paris, le 21 septembre 2023, année 2023-2024

Informations pratiques

Formation	mineure Humanités numériques, licence Lettres (L1)
Enseignante	Ljudmila PETKOVIC
Semestre	automne
Salle	C202
Horaire	jeudi 08h-10h
Matériel	https://icampus.univ-paris3.fr/course/view.php?id=41883

Organisation du cours

Cours

- séances 1 à 6 (21 septembre – 26 octobre)
- congés / ateliers : 2 novembre
- séances 7 à 12 (9 novembre – 14 décembre)

Évaluation : contrôle continu

```
1. écrit (50%) : 26 octobre
2. évaluation de pratique numérique (50%) : 14 décembre
```

Plan du cours

- Que sont les humanités numériques ? Définitions et exemples
- Évolution du domaine, projets de recherche
- Acquisition des objets d'étude : (rétro-)numérisation, collecte sur le web
- Traitement des objets d'études : nettoyage, métadonnées, annotation
- Exploitation des objets d'étude : éditions, extraction d'annotation, fouille de texte (fouille de données), alignement, traduction automatique, visualisations
- Publication des objets d'études : site web, application, plateforme, archives

Que sont les humanités numériques ?

Définition des humanités numériques (HN)

- domaine de recherche au croisement de l'informatique, des arts, lettres, sciences humaines et sociales (SHS)
- communauté de pratiques autour d'approches informatisées (ensembles de pratiques utilisant le numérique) pour l'analyse des données dans ces domaines
- science ou non ? (tous les secteurs des lettres et SHS potentiellement concernés)
- les HN s'intéressent au « numérique » de deux manières différentes :
  - comme objet d'étude
  - comme outils et méthodes (pour étudier le numérique ou d'autres objets)

Côté humanités (approche I)

- mise à disposition (de préférence grâce à des éditions de référence) des *corpus* fondamentaux pour la culture d'aujourd'hui
  - données attestées et méthodiquement assemblées sur support informatique
  - textes littéraires, manuscrits historiques, collections d'images, partitions de musique...

L'utilisation du numérique a une part secondaire lors du choix des documents, de leur édition et de leur consultation.

→ mise au point d'éditions numériques avancées pour l'enseignement ou la recherche

Centrée sur l'édition et la mise à disposition des données

Côté numérique (approche II)

- utiliser les données massives aujourd'hui disponibles dans les domaines cités → faire émerger des tendances et des faits nouveaux, qu'il serait quasiment impossible de découvrir sans ordinateur
- recours à des algorithmes et des techniques d'analyse inédites pour :
  - étudier de façon originale les données
  - vérifier des hypothèses, difficilement validées par une analyse humaine
- *humanités computationnelles* : analyses fines nécessitant, autant que possible, des corpus soigneusement encodés et annotés
- les deux approches sont complémentaires

Centrée sur l'exploitation (fouille) des données

Évolution du domaine

Projets de recherche

Avènement des HN

- informatique : outil puissant pour mener à bien des tâches fastidieuses
- chercheurs travaillant sur des textes ont ainsi souvent besoin de *concordances*
  - relevé exhaustif de toutes les occurrences d'un mot ou d'une expression donnée, pertinente pour une question de recherche précise, dans un corpus de référence
  - tâche fastidieuse si elle est menée à la main
- XX siècle : « machines mécaniques »
  - indexer les textes, retrouver les différentes occurrences d'un mot ou produire automatiquement des concordances
- essor de l'informatique après la Seconde Guerre mondiale

Tradition de l'édition électronique des textes

- précurseur des HN : jésuite italien Roberto Busa (1913-2011)
  - **1949** : projet de création d'index autour de l'œuvre de saint Thomas d'Aquin (*Index thomisticus*) en partenariat avec la société IBM
  - projet bénéficiant d'une très importante (mais ignorée) main-d'œuvre féminine
- utilisation du numérique pour la mise au point des éditions de référence d'œuvres majeures, des index, des concordanciers ainsi que les outils nécessaires pour produire des analyses statistiques à partir de textes
  - **1972** : projet *Thesaurus Linguae graecae* (l'université de Californie à Irvine)
  - mise à disposition d'un ensemble très important de textes grecs sur support informatique (l'équivalent a été réalisé indépendamment pour le latin)
  - les textes n'étaient pas numérisés mais saisis à la main

Tradition de la linguistique de corpus

1959 : Randolph Quirk, projet *Survey of English Usage*

- collection d'enregistrements et de transcription de différentes variétés d'anglais → *British National Corpus*, corpus de référence pour l'analyse de la langue anglaise
- mise au point de corpus en propre pour conduire des études linguistiques
- examiner le sens des mots à partir d'exemples réels (en lexicologie ou pour la mise au point de dictionnaires)
- mettre en lumière les différences linguistiques entre groupes d'individus (sociolinguistique)

Traitement automatique des langues (TAL)

- angl. *natural language processing*
- champ disciplinaire qui se développe en parallèle mais de manière largement indépendante
- fournit les outils d'analyse permettant d'interroger les corpus de manière plus précise (à partir des formes de surface, des lemmes ou des catégories morphosyntaxiques)
  - retrouver toutes les occurrences d'un mot dans un texte à partir de sa forme canonique, quelle que soit la forme employée dans le texte (*fait*, *faisons*, *fera*... → *faire*)
  - tâche non triviale qui continue de susciter de multiples problèmes

Text Encoding Initiative (TEI)

Traitements informatisés des données textuelles (tradition de linguistique de corpus)

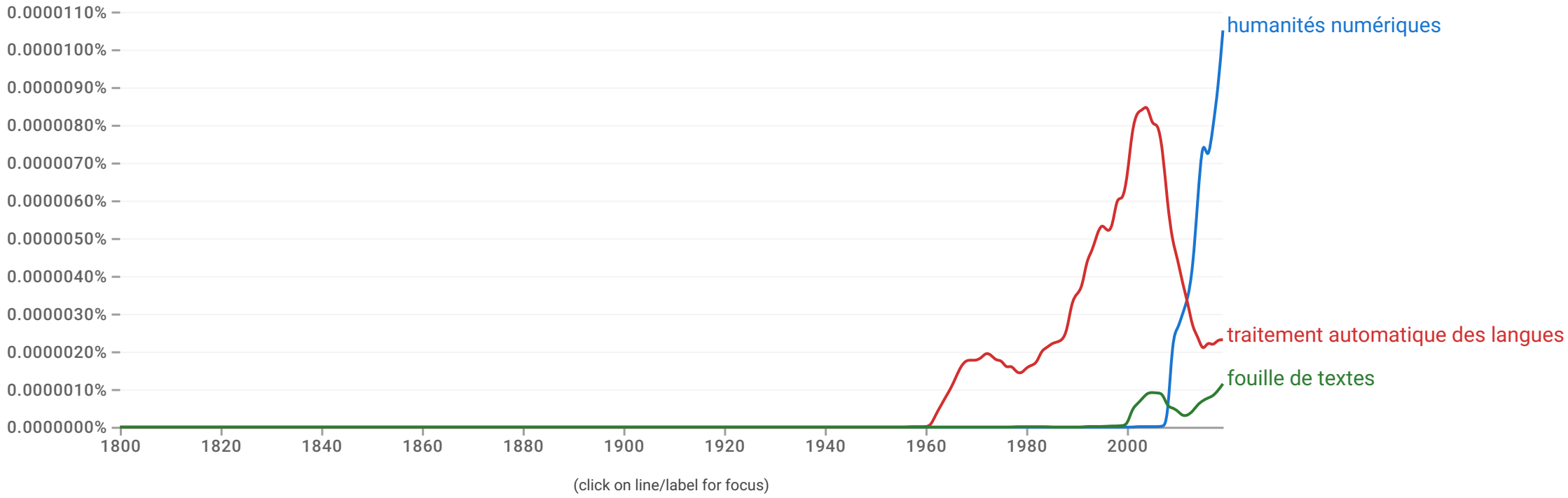
- requiert des formats adaptés, des standards d'encodage avancés permettant de tenir compte du type et de la nature des textes étudiés
- **1987** : *Text Encoding Initiative (TEI)*
  - norme d'encodage des textes permettant de rendre compte de manière très fine de nombreux types de documents : poésie, théâtre, documents historiques, etc.
  - encodage suivant le langage de description documentaire *SGML* (*Standard Generalized Markup Language*, désormais obsolète)
  - langage *XML* (*eXtensible Markup Language*) : standard pour la structuration de documents
    - manipulable avec des outils de transformation de format

Données massives (angl. *big data*)

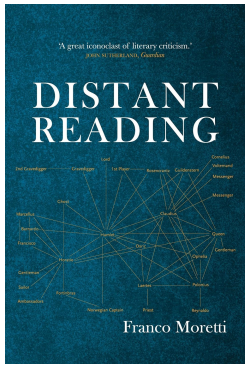
- **1990** : développement du Web et des ordinateurs personnels, augmentation exponentielle de la puissance de calcul
  - accès à des données massives, avec un ordre de magnitude jamais imaginé jusque-là
  - nouveaux modèles informatiques plus performants
- **2004** : initiative Google Books
  - vise à numériser tous les livres disponibles au monde, grâce à des accords avec les éditeurs et avec les grandes bibliothèques nationales ou régionales
  - collection unique, par son contenu et par sa taille

Google Books N-gram Viewer

- Fréquence relative d'apparition des expressions dans un corpus (démo)



Franco Moretti



2005 : introduit le concept de la « lecture distante » (*distant reading*)

- processus de « compréhension de la littérature non pas en étudiant des textes particuliers, mais en regroupant et en analysant des quantités massives de données »
- concept assez controversé dans le domaine des humanités numériques → changement de paradigme
- ≠ lecture proche (angl. *close reading*) : attention est portée sur l'œuvre qui est lue et analysée en détails
    - incapable de saisir toute l'envergure de la littérature

Voyant Tools

Interface graphique de Voyant Tools (Rockwell & Sinclair, 2016).

Remarques conclusives

Points forts des HN

- domaine de recherche transdisciplinaire, à la croisée de plusieurs champs académiques
- domaine qui se revendique en faveur de la « diffusion, du partage et de la valorisation du savoir »
- au-delà de l'objet ou des méthodes particulières du champ, c'est peut-être toute la recherche, son fonctionnement et sa visée qui sont questionnés par les HN
- à partir de simples manipulations opérées sur des données massives, il devient possible d'observer des faits signifiants sur le plan de l'histoire et de la culture
- il faut savoir poser les bonnes questions et interpréter les données en corpus, tout en conservant un point de vue d'ensemble pour éviter les recherches trop parcellaires

Références

- **Doduik, N.** (2017). « Les humanités numériques, une révolution ? », *Hypothèses*. https://doctlames.hypotheses.org/77
- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*].
- **Poibeau, T.** (*s.d.*). « HUMANITÉS NUMÉRIQUES », Encyclopædia Universalis. https://www.universalis.fr/encyclopedie/humanites-numeriques/
- **Rockwell, G. & Sinclair, S.** (2016.) *Hermeneutica. Computer-Assisted Interpretation in the Humanities*, Cambridge, Massachusetts, MIT Press, 2016