

# Exploration des savoirs

## L'enquête, ses données

Émilien Schultz - Nicolas Benvegnu

# Objectifs

- ▶ Lien enquête/données
- ▶ Réfléchir aux données possibles
- ▶ Discuter bonnes pratiques

# La démarche d'enquête

1. Discussion 360°
2. Phase exploratoire / contexte de découverte
3. Consolidation de la question principale
4. Calendrier & objectifs & données pour l'argumentation
5. Démarche de collecte et de preuve
6. Analyse
7. Restitution

Parlons données

# Les données dans l'enquête

Connaître la réalité quel accès ?

- ▶ **Sources**
  - ▶ Résultats d'autres enquêtes
- ▶ Documenter directement les phénomènes
  - ▶ le langage des **données**

## Mise à l'épreuve des affirmations

The (real) scientific method.



# C'est quoi une donnée ?

Définition générique : *un ensemble d'informations issues par différentes transformations de traces laissées par un phénomène*<sup>1</sup>

- ▶ définition a-disciplinaire
- ▶ importance du lien avec un **phénomène**
  - ▶ *la donnée n'est pas le phénomène*
- ▶ des étapes intermédiaires
- ▶ nature de la "trace" recueillie : comment ? quoi ?

---

<sup>1</sup>Wikipédia : “Data, as a general concept, refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.”

Des données partout



Et sans fin

HEY, LOOK, WE HAVE A BUNCH  
OF DATA! I'M GONNA ANALYZE IT.

NO, YOU FOOL! THAT WILL  
ONLY CREATE MORE DATA!



# Problématique fil rouge

## **i** Cohabitation des mobilités en ville

Constat de transformations en cours et de conflits entre les différentes mobilités dans les zones urbaines.

- ▶ Quels sont les transformations en cours ?
- ▶ Quelle place pour les nouvelles mobilités ?
- ▶ Comment la cohabitation s'organise avec la croissance de l'usage du vélo ?

Une question interdisciplinaire et controversée

# Enquêter pour dépasser les stéréotypes



**Un petit rappel à tous les doués dans le domaine de la pédale**



**Les vaches y arrivent très bien, alors pourquoi pas les cyclistes !**

# Quelles sont les données possibles ?

Pour enquêter sur les cohabitations des mobilités autour de l'université et la place du vélo.

## Types de données

- ▶ Mon expérience de cycliste
- ▶ ...

# La place des données dans la démarche

- ▶ *Theory-driven*
  - ▶ Les données arrivent après la formulation d'une hypothèse
- ▶ *Data-driven*
  - ▶ Les réponses émergent des données (notamment, le big data)

En fait

- ▶ Aucune donnée ne peut répondre à toutes les questions
- ▶ Aucune donnée ne répond en soit à une question

L'enquête : un mouvement plus incrémental liant terrain & données

## Deux principaux usages

Les données interviennent à différents moments dans l'enquête

- ▶ Exploratoire/découverte : diversifier son rapport au monde
- ▶ Argumentation/justification : défendre un argument
- ▶ (Illustration)

## Mais des spécialités “régionales” de données

- ▶ Chaque domaine/secteur a ses données
  - ▶ Histoire/Physique/Biologie
  - ▶ Procès/Journalistes/...
- ▶ Pour l'enquête interdisciplinaire
  - ▶ Dépasser les habitudes
  - ▶ Relier des pratiques différentes
  - ▶ Se nourrir de la diversité

Importance d'un vocabulaire commun

# Le mythe du processus linéaire

DATA



SORTED



ARRANGED



PRESENTED  
VISUALLY



EXPLAINED  
WITH A STORY



## Un processus itératif

- ▶ Le lien entre question & données n'est pas immédiat
- ▶ Processus itératif :
  - ▶ Certaines données amènent à des questions
  - ▶ Certaines questions amènent à chercher des données
- ▶ Mais
  - ▶ Certaines données n'existent pas
- ▶ Accepter une certaine forme d'insatisfaction

**Un appariement pragmatique et toujours partiel ; passe souvent par expliciter le protocole**

## Est-ce que toutes ces données se valent ?

- ▶ Format : brut ou transformé (étapes intermédiaires)
- ▶ Nature différentes (valeur, texte, image, ...)
- ▶ Disponibilité (déjà collectées, à collecter)
- ▶ Taille
- ▶ Qualité

## Face à cette diversité

- ▶ Comment choisir les données les plus pertinentes ?
- ▶ Comment les manipuler et les stocker ?
- ▶ Comment passer des données à des résultats ?

Enjeu : prendre au sérieux les données pour des enquêtes plus riches et des affirmations plus solides.

# Corpus : construction & représentativité

Corpus = ensemble de données rassemblée suivant une logique

Pourquoi ?

- ▶ Décrire un ensemble d'éléments
- ▶ Situer des éléments par rapport à d'autres

## i Pratiques du vélo

Une donnée : un entretien avec un cycliste

Un corpus : des entretiens avec tous les cyclistes qui ont accepté de répondre à mes questions un samedi soir rue Rivoli

## Enjeu de la représentativité

A-t-on suffisamment de données ? Est-ce que l'on peut vraiment se dire que l'on peut caractériser l'entièreté du phénomène ?

- ▶ Dépend des domaines
- ▶ Définir l'entité (population, grandeur, etc.) à représenter et le passage du corpus à cette entité

Comment faire ?

- ▶ Entre méthodes & bricolage
- ▶ Important de justifier/mettre en discussion.

# Produire vs. réutiliser des données

- ▶ Données  **primaire**
  - ▶ construire ses données/corpus
  - ▶ un gradient allant de la production complète à la mise en forme
- ▶ Données  **secondaire**
  - ▶ données déjà constituées
  - ▶ permet d'explorer en amont de l'enquête
  - ▶ multiplication des données disponibles open data
- ▶ Entre les deux : **l'explosion des traces numériques**
  - ▶ accès à des bases de données/API

# Conséquences de la disponibilité

- ▶ Données primaires : possible de les adapter à sa question mais vulnérabilité du protocole
- ▶ Données secondaires : répondent souvent à une autre question que la votre

## Vélo & Justice

- ▶ Les données de comptage des vélos à Paris produits par la ville
- ▶ Compter moi-même les vélos dans une rue

## La forme compte : quels formats ?

- ▶ Données “désagrégées” permettant les recombinaisons:
  - ▶ structurées
    - ▶ tabulaires
  - ▶ peu structurée
    - ▶ un mélange de formats
  - ▶ pas structuré
    - ▶ un contenu dont il faut extraire l'information
- ▶ Données déjà agrégées
  - ▶ Tableaux déjà calculés
  - ▶ Graphiques

# Une pragmatique des données

## Des arbitrages

- ▶ Balance coût / avantage de l'accès à des données
- ▶ Adéquation avec les objets de la question
- ▶ Adéquation avec la nature de la question
- ▶ Facilité de retransformer les données
- ▶ Disponibilités des méthodes d'analyse pour obtenir un résultat

Où mettre son énergie ?

# Difficulté d'accès aux données

- ▶ Le coût des données :
  - ▶ ces données peuvent déjà exister
  - ▶ peuvent être collectées avec un effort raisonnable
  - ▶ sont couteuses à récolter

## **i** Des choix à faire

- ▶ Réaliser ma propre enquête par questionnaire, avec mes questions spécifiques
- ▶ Utiliser les données publiques disponibles à l'échelle des villes

## Volume des données

- ▶ Le volume des données n'est pas l'élément le plus important
- ▶ Adéquation avec la question

**Ne pas hésiter à restreindre la taille et privilégier l'adéquation**

# Qualité des données

Toutes les données ne se valent pas : une qualité multidimensionnelle.

- ▶ Tracabilité de la chaîne de production
- ▶ Fiabilités (producteurs douteux, ou au contraire légitime)
- ▶ Données manquantes (est-ce qu'il manque des éléments)
- ▶ Erreurs dans le corpus
- ▶ Ouverture versus fermeture de l'accès
- ▶ ...

## L'analyse dans tout ça ?

- ▶ La valeur des données vaut par la capacité d'en faire découler des résultats (actionner les données)
  - ▶ Interdépendance collecte/données/résultats (par ex. l'expérimentation)
- ▶ Certaines analyses sont complexes et posent la question de la faisabilité (ex. statistiques)
- ▶ Des outils spécifiques...

Par quelles données débuter

## Les données “accessibles”

- ▶ Toute enquête se pose la question:
  - ▶ De l'espace public
  - ▶ De la production des connaissances
- ▶ Avec quelques manières de les utiliser
  - ▶ Ce point sera traité dans une séance ultérieure plus en détail

**Il n'y a pas vraiment de limites, n'hésitez pas à être créatifs**

# Documenter l'espace public : la presse

L'article de presse comme contenu complexe :

- ▶ date (évolution temporelle)
- ▶ contenu textuel (sujets)
- ▶ cadre : journal, journaliste, ...

Permet une diversité d'approches :

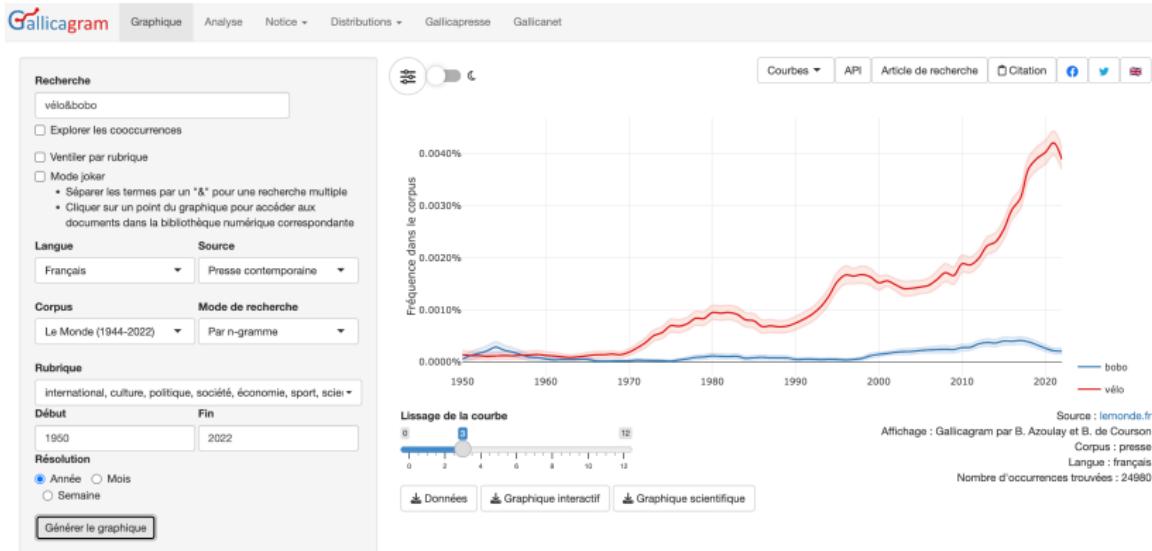
- ▶ l'article individuel comme source
- ▶ corpus d'articles sur un sujet/période
  - ▶ thèmes abordés
  - ▶ acteurs récurrents
  - ▶ évolutions temporelles

## Accès captifs

Une presse sous droit d'auteur.

- ▶ Par des outils dédiés comme Gallicagram
- ▶ Par l'interface de portails comme Europresse
- ▶ Certains médias sont libres

# Cadrage d'une tendance



# Construire et extraire un corpus d'Europresse

Accès Europresse à des données payantes

## Vélo & Justice

Comment aborde-t-on la question des inégalités quand on parle du vélo en région parisienne ?

Le contour du corpus :

- ▶ Définir la période
- ▶ Définir les mots-clés
- ▶ Définir les journaux concernés
  - ▶ Se concentrer sur un journal
  - ▶ Sur un bouquet?

## Construire des requêtes

- ▶ Chaque interface a sa grammaire
- ▶ Lire les résultats
- ▶ Ne pas hésiter à combiner les éléments
- ▶ Viser un nombre réaliste de résultats

## Consulter vs. conserver

Possibilité de consulter sur Europresse. Mais comment conserver ?

- ▶ Collecter à la main vs. télécharger
- ▶ Des transformations nécessaires

Puis aller vers l'analyse :

- ▶ Descriptive : compter
- ▶ Plus avancée
- ▶ Visualisations



### Statut des données

Ce sont des données sous droit d'auteur. Leur usage est très réglementé. Il y a une tolérance pour la recherche mais il est interdit des diffuser.

**EUROPRESSE** UNE SOLUTION DE CIBLAGE

RECHERCHER DOSSIERS PUBLICATIONS PDF

English ? Étudiant

Recherche simple | Recherche avancée | Recherche express | Recherche de biographies |

TEXT= vélo & (paris | parisien\*) Dans toutes les archives ▾ Monde, Le

Rechercher

Press Presse Télévision et radio Médias sociaux

50 sur 2 913 Pertinence

Le Monde L'empêcheur de pédaler en rond

2023-03-29 - 199 mots [PDF](#)

Pascale Krémer - Regardez ! » Bicyclette stoppée net, Stein van Oosteren observe ce qui n'attire le regard de personne. Les tours de roue hésitants d'un petit, entre mère anxieuse et chalands du centre ...

Aussi paru dans: ▾

Le Monde Vacances à vélo, la vérité à cru

2022-07-11 - 2071 mots [PDF](#)

Guillemette Faure - Et si on essayait le «slow tourisme»? Voilà trois jours que je suis partie en vacances à [vélo](#). Mon intention, avec ma fille et les amis qui nous rejoindront en route ...

Aussi paru dans: ▾

Le Monde A Paris, la fin des trottinettes de location

ÉTUDES ET RAPPORTS

Etudes et rapports 0% - 0 documents

Répertoires et références 0% - 0 documents

Répertoires et références

Positif 70% - 1337 documents

Neutre 1% - 14 documents

Négatif 30% - 566 documents

ÉVOLUTION

Pic médiatique : 11 documents le 27 juin 2016

DOCUMENTS

2002 2004 2006 2008 2010 2012 2014 2016 2018 2020 2022 2024

# Exemple d'analyse de la presse



ELSEVIER

Contents lists available at ScienceDirect

## Accident Analysis and Prevention

journal homepage: [www.elsevier.com/locate/aap](http://www.elsevier.com/locate/aap)



### Trends in local newspaper reporting of London cyclist fatalities 1992–2012: the role of the media in shaping the systems dynamics of cycling



CrossMark

Alex Macmillan<sup>a,\*</sup>, Alex Roberts<sup>b</sup>, James Woodcock<sup>c</sup>, Rachel Aldred<sup>d</sup>, Anna Goodman<sup>e</sup>

<sup>a</sup> Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand

<sup>b</sup> College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

<sup>c</sup> UK CRC Centre for Diet and Activity Research (CEDAR), MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, University of Cambridge, Cambridge, UK

<sup>d</sup> Department of Planning and Transport, Faculty of Architecture and the Built Environment, University of Westminster, London, UK

<sup>e</sup> Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

---

#### ARTICLE INFO

**Article history:**

Received 20 August 2014

Received in revised form 27 May 2015

Accepted 18 October 2015

Available online 10 November 2015

---

**Keywords:**

Cycling

Fatality

Injury

Media

Systems dynamics

Trends

---

#### ABSTRACT

**Background:** Successfully increasing cycling across a broad range of the population would confer important health benefits, but many potential cyclists are deterred by fears about traffic danger. Media coverage of road traffic crashes may reinforce this perception. As part of a wider effort to model the system dynamics of urban cycling, in this paper we examined how media coverage of cyclist fatalities in London changed across a period when the prevalence of cycling doubled. We compared this with changes in the coverage of motorcyclist fatalities as a control group.

**Methods:** Police records of traffic crashes (STATS19) were used to identify all cyclist and motorcyclist fatalities in London between 1992 and 2012. We searched electronic archives of London's largest local newspaper to identify relevant articles (January 1992–April 2014), and sought to identify which police-reported fatalities received any media coverage. We repeated this in three smaller English cities.

**Results:** Across the period when cycling trips doubled in London, the proportion of fatalities covered in the local media increased from 6% in 1992–1994 to 75% in 2010–2012. By contrast, the coverage of motorcyclist fatalities remained low (4% in 1992–1994 versus 5% in 2010–2012;  $p = 0.007$  for interaction between mode and time period). Comparisons with other English cities suggested that the changes observed in

## Mettre en forme les données

- ▶ Délimiter le corpus
- ▶ Téléchargement sous un format HTML
  - ▶ Gérer la limite
- ▶ Des fichiers HTML à dater + un document de métadonnées
- ▶ Transformer en un document tabulaire

## Étendre le domaine de la presse

Europresse n'est pas le seul portail sur lequel collecter des données.

Suivant la question :

- ▶ se concentrer sur les articles parus sur un site spécifique spécialisé
- ▶ récupérer les newsletters d'une organisation

### Statut des données

Récupérer les évènements de la FUB pour voir lesquels mentionnent la question de la justive :  
<https://www.fub.fr/fub/actualites>

## Parenthèse : la difficile question des médias sociaux

- ▶ Une source riche d'échanges
- ▶ Mais de plus en plus fermés (Twitter, Insta, etc.)
  - ▶ Difficulté de collecter massivement
  - ▶ Possibilités de constituer des petits corpus

Das, Subash, et al. "Extracting patterns from Twitter to promote biking." IATSS research 43.1 (2019): 51-59.

Collecter vos propres données

# Importance du protocole

- ▶ Identifier clairement le rôle des données dans votre question
  - ▶ Notamment les résultats que vous attendez
- ▶ Clarifier la population ciblée & les critères de représentativité
- ▶ Identifier les risques
  - ▶ Que se passe-t-il si vous n'arrivez pas à tout collecter ?

## i Vélo & Justice

Je veux m'intéresser aux images du vélo dans les films. Comment je construis la population de films (uniquement ceux qui le mentionne dans le résumé ? un échantillon aléatoire ?) ? Comment ensuite je code pour chaque film (je me limite au résumé de manière automatique ? Je les regarde ?)

## Importance du format de codage

- ▶ Quelle “trace” garder
  - ▶ Extensive : couteux
  - ▶ Condensé : perte d'information
- ▶ Construire des catégories en les testant
  - ▶ Ne jamais hésiter à faire des allers-retours

## Données “quantitatives” et “qualitatives”

- ▶ Division un peu artificielle
- ▶ Passages possibles du qualitatif au quantitatif
  - ▶ Codage des entretiens
  - ▶ Analyse du texte
- ▶ Importance de la structure des jeux de données / leur meta-information

## Exemple des entretiens

### Définir le dispositif de collecte

- ▶ Quel est le corpus (qui va être interrogé)
- ▶ Comment passer du discours à des données
  - ▶ Annotation : quoi ?
- ▶ Quels éléments doivent être présents dans l'entretien absolument ?
- ▶ Que se passe-t-il s'il manque un entretien ?

## Exemple des entretiens

Après : l'enregistrement/retranscription est une "trace" qu'il **faut mettre en forme**

- ▶ Mettre les entretiens dans un format unifié
  - ▶ .docx au mieux
- ▶ Nommer les fichiers de manière systématique
- ▶ Avoir un fichier avec les métadonnées (date, interlocuteur)
- ▶ Transformer les données :
  - ▶ Coder des éléments spécifiques des entretiens dans un fichier

## Gestion matérielle des données

## En pratique :

- ▶ Comment bien collecter et conserver les données ?
- ▶ Où les stocker ?
- ▶ Comment les décrire dans ses travaux ?
- ▶ Comment travailler collaborativement avec des données ?

## Bonnes pratiques numériques

- ▶ Collecter et conserver les données brutes
- ▶ Documenter le contenu dans des **métadonnées**
- ▶ Documenter chacune des transformations
- ▶ Consolider un corpus nettoyé avant l'analyse

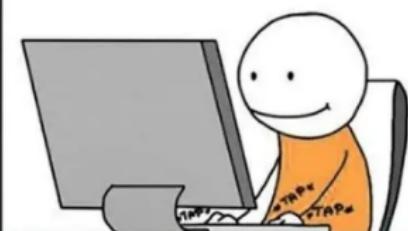
**Dans l'idéal, aller vers la reproductibilité**

# Documenter documenter documenter

Journal collectif des traitements

## UNFINISHED WORK

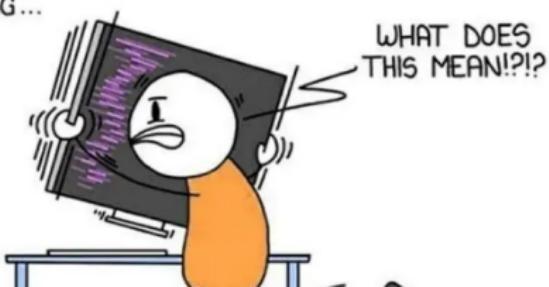
FRIDAY EVENING



PERFECT!  
I'LL FINISH  
THIS ON  
MONDAY



MONDAY MORNING...



## Format et stockage des données

- ▶ Simplifier le plus possible
  - ▶ .txt, .csv, ...
- ▶ Faire attention à la confidentialité
  - ▶ Anonymisation/pseudonymisation
- ▶ Noms clairs
- ▶ Dater les fichiers
- ▶ Système de dossier clair

# Ranger ses données

- ▶ Conserver les données brutes
- ▶ Mettre ensuite en forme les données pour faciliter les analyses

country	year	cases	population
Afghanistan	1990	145	1497071
Afghanistan	2000	666	20995360
Brazil	1999	3737	17206362
Brazil	2000	84488	17404898
China	1999	21258	127215272
China	2000	21066	128025583

variables

country	year	cases	population
Afghanistan	1990	145	1497071
Afghanistan	2000	666	20995360
Brazil	1999	3737	17206362
Brazil	2000	84488	17404898
China	1999	21258	127215272
China	2000	21066	128025583

observations

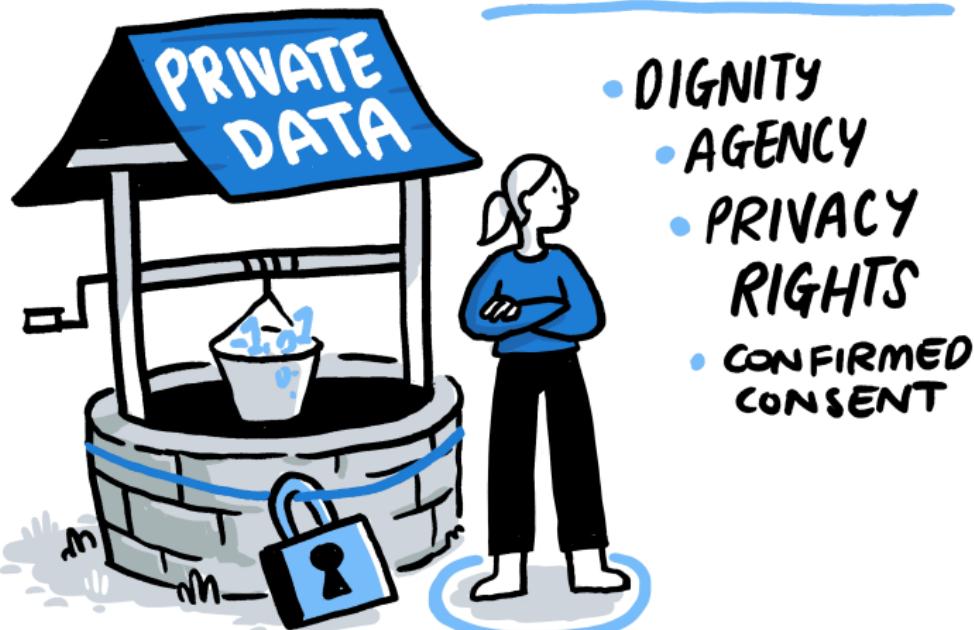
country	year	cases	population
Afghanistan	1990	145	1497071
Afghanistan	2000	666	20995360
Brazil	1999	3737	17206362
Brazil	2000	84488	17404898
China	1999	21258	127215272
China	2000	21066	128025583

values

## Sensibilité des données

- ▶ Limiter les données identifiantes
- ▶ Consentement et autorisation

PEOPLE DESERVE:



# Décrire les données dans ses travaux

Dans les travaux :

- ▶ Préciser la manière dont ont été constituées les données et.ou les sources
  - ▶ Pour chaque corpus
- ▶ Donner des informations descriptives sur le contenu
- ▶ Les transformations qui ont été faites
  - ▶ Ex. les filtrages
- ▶ Préciser les limitations s'il y en
  - ▶ Données manquantes
  - ▶ Artefacts/problèmes

## L'importance de décrire les données

In light of the significance of discursive construction in positioning environmental issues, I selected media presenting contrasting ideological views, since they shape public views towards key issues in different ways. *Le Figaro* and *Le Monde*, with their respective rightist and leftist readerships, were selected for their broad spectrum of views (see Cohen 2012) and high circulation rates.

Articles were identified from the newspapers' online sites using the search terms *bicyclette* (bicycle) and *vélo* (bike) for the year 2017. This is indeed a pivotal year as it marks the end of Anne Hidalgo's 2015-2020 bicycle plan for Paris, because of the transition of the bike-sharing system from JCDecaux to new operator Smovengo, the emergence of electric bikes on the streets of French cities, and the boom in sales of electric bikes (and the subsequent subsidies (and their early withdrawal in September 2017). Only articles related to cycling in the context of the study were selected. This means that articles dealing with sports events, cycling in the countryside, cycling in the Paris region, cycling in the rest of France or bicycle touring were discarded, even though they may contribute to the general view of cycling. The final corpus contains 68 articles from *Le Figaro* and 32 from *Le Monde*, totalling 100 articles and 56,868 words, as shown in Table 1. Although the two datasets contain many articles in *Le Monde* as there are in *Le Figaro*, *Le Monde* articles are longer than those in *Le Figaro*: 811 words in average vs. 474 words in *Le Figaro*, thus totalling 24,633 words for *Le Monde* and 32,235 words for *Le Figaro* dataset. The articles from the 2017 *Le Figaro* dataset were collected between January 1st and December 31st, 2017.

## Aller vers l'analyse

- ▶ La valeur des données dépend des possibilités d'analyse
- ▶ Les analyses dépendent du type de données
- ▶ Trois approches :
  - ▶ “à la main” : manipulation (quasi) directe des données (Excel, codage manuel, ...)
  - ▶ “pipeline automatisé” : usage de logiciels spécialisés (Iramuteq, VosViewer, Jamovi, ...)
  - ▶ “programmation” : scripts de traitement de données (Python, R)

# Les données : un ingrédient (central)

