

Méthodes Quantitatives

Séance 7 - Construction de bases de données à partir de documents non structurés

Sciences Po

2024-2025

Plan de la séance

- 1 Introduction : Des documents aux données
- 2 Principes de construction d'une base de données
- 3 Exemple détaillé : Le registre de la Commune
- 4 Méthodes de collecte

Introduction : Des documents aux données

Le défi de la quantification

- **Problématique centrale :**

- Comment transformer des informations qualitatives en données quantitatives ?
- Comment standardiser des formats très différents ?
- Comment assurer la comparabilité ?

- **Enjeux :**

- Scientifique : produire des connaissances fiables
- Méthodologique : assurer la rigueur du processus
- Pratique : gérer de grands volumes d'information

Qu'est-ce qu'une catégorie ?

- **Définition** : Ensemble d'éléments regroupés selon certains critères
- **Propriétés nécessaires** :
 - Exhaustivité (tout élément doit appartenir à une catégorie)
 - Exclusivité (un élément ne peut appartenir qu'à une seule catégorie)
 - Pertinence (les catégories doivent avoir du sens)
- **Enjeux** :
 - Théoriques : quelle vision du monde ?
 - Méthodologiques : comment opérationnaliser ?
 - Pratiques : comment coder efficacement ?

Exemple introductif : Les pays dans le monde

- **Différents décomptes :**

- ONU : 193 États membres
- FIFA : 211 associations membres
- CIO : 206 comités nationaux
- ISO : 249 codes de pays

- **Cas ambigus :**

- Taiwan
- Kosovo
- Palestine
- Territoires d'outre-mer

Implications

La définition d'un "pays" dépend du :

- Contexte institutionnel
- Objectif de l'étude
- Période historique

Diversité des sources

Sources historiques

- Archives manuscrites
- Registres administratifs
- Journaux d'époque
- Correspondances
- Documents notariés

Sources contemporaines

- Réseaux sociaux
- Sites web
- Bases de données en ligne
- Documents numérisés
- Traces numériques

Point commun

Nécessité de transformer des informations non structurées en données analysables

Exemple 1 : Les annonces matrimoniales

Hier (1921-1939)

"Jeune homme, 28 ans, situation stable, recherche jeune fille sérieuse de bonne famille, en vue mariage."

Aujourd'hui (Meetic)

"H28, cadre sup, sportif, aime voyages et cinéma, recherche F25-30 pour relation sérieuse."

• Variables potentielles :

- Genre de l'annonceur
- Âge
- Profession/situation
- Critères recherchés
- Type de relation souhaité

Exemple 2 : La communication politique

Journal Officiel (1881)

- Format texte dense
- Structure formelle
- Langage soutenu
- Contexte institutionnel

TikTok politique (2024)

- Format vidéo court
- Style informel
- Multimédia
- Interaction directe

Variables comparables

- Thèmes abordés
- Registre de langage
- Public visé
- Impact/réception

Principes de construction d'une base de données

Les "10 commandements" Lemercier-Zalc

① Conserver les formulations d'origine

- Permet de revenir aux données brutes
- Autorise différentes interprétations
- Facilite les vérifications

② Segmenter l'information au maximum

- Une information = une variable
- Éviter les variables composites
- Faciliter les recodages ultérieurs

Les "10 commandements" (suite)

③ Standardiser la saisie

- Définir des règles claires
- Les appliquer systématiquement
- Les documenter

④ Éviter la pré-catégorisation

- Garder les nuances
- Permettre différentes analyses
- Reporter les choix théoriques

Les "10 commandements" (suite)

5 Documenter ses choix

- Justifier les décisions
- Décrire les procédures
- Faciliter la réplication

6 Vérifier régulièrement

- Contrôles de cohérence
- Détection des erreurs
- Corrections systématiques

Les "10 commandements" (fin)

- 7 **Prévoir les usages futurs**
- 8 **Faciliter les modifications**
- 9 **Assurer la traçabilité**
- 10 **Penser à la réutilisation**

Application pratique

Ces principes s'appliquent à tout type de source :

- Archives historiques
- Données web
- Contenus multimédia

Exemple détaillé : Le registre de la Commune

Le registre des enfants de la Commune de Paris

Contexte historique :

- Document de 1871
- Registre administratif
- Source semi-structurée

Informations disponibles :

- État civil des enfants
- Situation familiale
- Circonstances du placement
- Suivi administratif

Structure du document original

Exemple d'entrée

"Martin (Jules), né le 15 mars 1865 à Paris, fils de Pierre (décédé) et de Marie née Durand, placé le 3 mai 1871 suite aux événements, domicilié rue Saint-Antoine"

● Informations à structurer :

- Nom et prénom
- Date et lieu de naissance
- Parents (nom, statut)
- Date de placement
- Motif
- Adresse

Transformation en base de données

nom	prenom	date_naiss	lieu_naiss	pere_nom	pere_statut	mere_nom
Martin	Jules	15/03/1865	Paris	Pierre	décédé	Durand

mere_prenom	date_placement	motif	adresse
Marie	03/05/1871	événements	rue Saint-Antoine

• Principes appliqués :

- Segmentation maximale
- Conservation des formulations
- Standardisation des dates
- Variables distinctes

Méthodes de collecte

Saisie manuelle vs. automatisée

Saisie manuelle

- Pour petits volumes
- Données complexes
- Sources historiques
- Nécessite vérification

Web scraping

- Pour grands volumes
- Données standardisées
- Sources numériques
- Automatisation possible

Focus sur le web scraping

• Définition

- Extraction automatique de données web
- Transformation en format structuré
- Stockage systématique

• Outils disponibles

- APIs (quand disponibles)
- Bibliothèques Python
- Services spécialisés

• Contraintes

- Légales (conditions d'utilisation)
- Techniques (anti-scraping)
- Éthiques (vie privée)

Exemple de processus de scraping

1 Identification des données

- Structure du site
- URLs des pages à scraper
- Localisation des données dans le HTML
- Format des données souhaitées

2 Extraction des données

- ...

3 Traitement et stockage

- Nettoyage des données
- Structuration
- Export (CSV, JSON, Base de données...)