Project
Implementation of a (Big) Data Management Backbone

Use Case and P1 description

Big Data Management – FIB – UPC

Project's statement

- Descriptive and predictive analysis of data related to Barcelona's housing and the relationship with its economy
- Examples of descriptive analysis KPIs
  - Average number of new listings per day
  - Correlation of rent price and family income per neighborhood
- Examples of predictive analysis KPIs
  - Estimate the rental price for a new apartment
  - Evaluate the deviation of a predicted price with respect to the real average price in a neighborhood

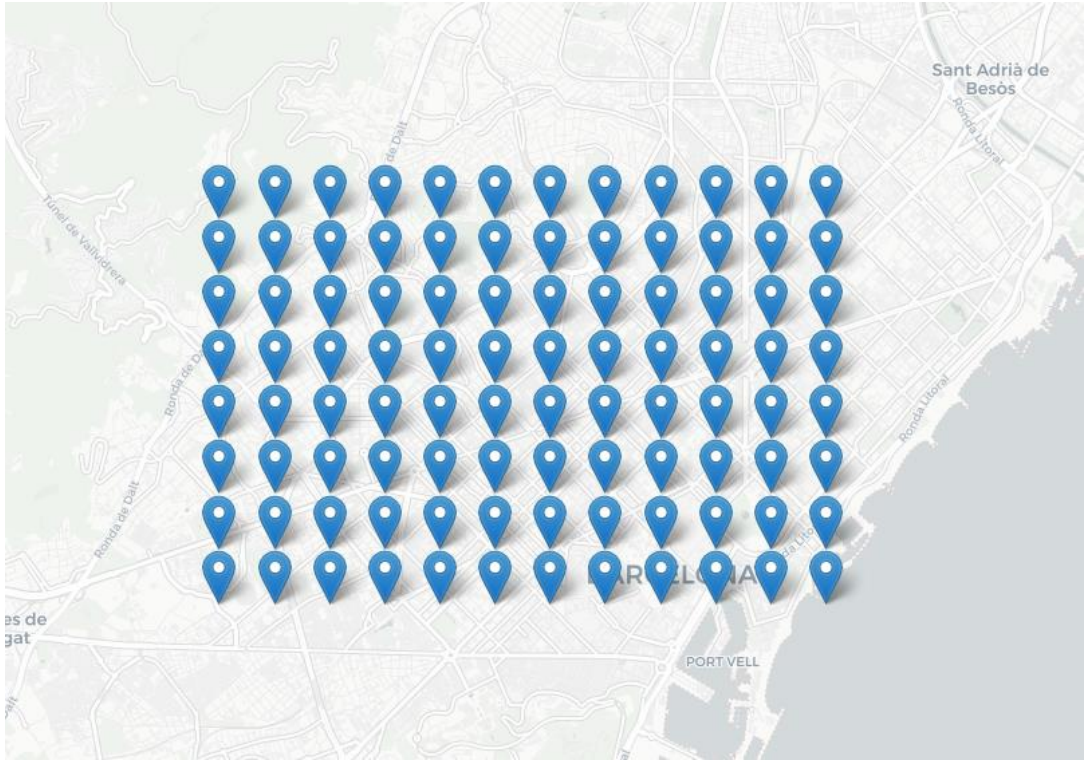
Mandatory datasets

- Barcelona rentals
  - idealista
- Territorial distribution of income
  - Open Data Barcelona
- Lookup tables

D1 - Barcelona rentals

idealista

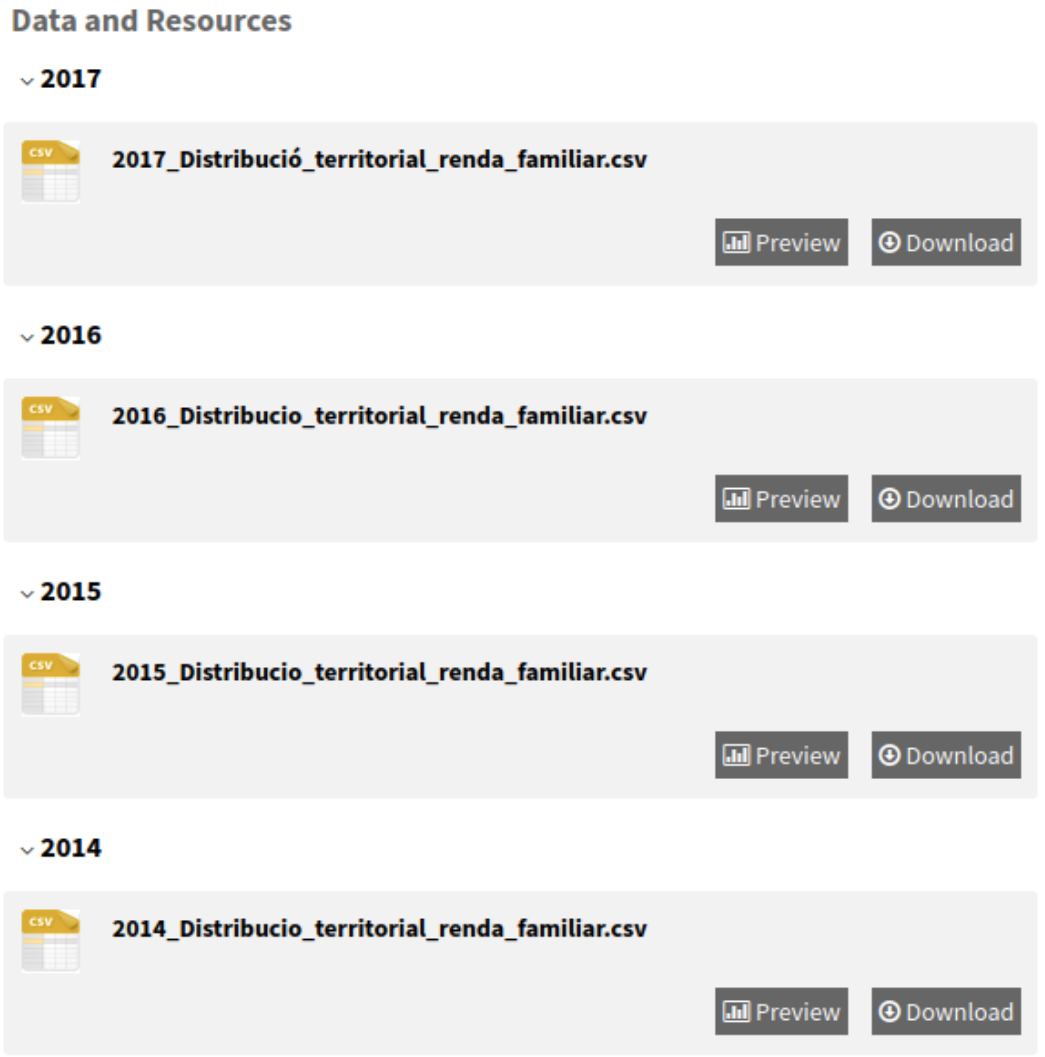
- JSON documents
- Listings for apartments downloaded once per day on a random point in Barcelona's grid (1.5 km radius)
- Ingestion date encoded in the filename



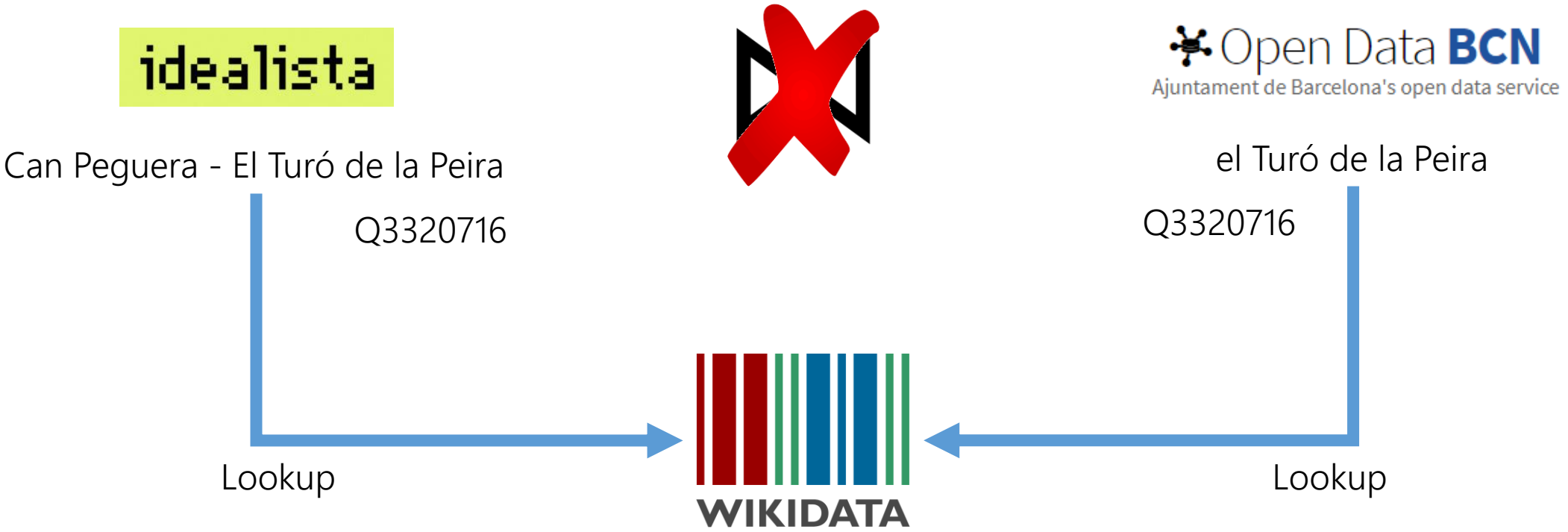
D2 - Territorial income distribution in the city of Barcelona

- CSV files
- Population and RFD (family income index
  - Per year (encoded in the filename)
  - Per neighborhood

Open Data **BCN**
Ajuntament de Barcelona's open data service



Data reconciliation



D3 - Lookup tables

- Two CSV files (D1 and D2)
- For each distinct district and neighborhood their Wikidata ID

Adding a third dataset

- You must propose the inclusion of a third dataset into the pipeline
  - It must require the implementation of a Data Collector (i.e., be external)
- You can check out OpenData BCN portal or other Open Data portals
- You might need to implement your own reconciliation process (in P2)

Technological choice

- Propose the right kind of storage, metadata and structure for each zone
- Some possibilities are:
  - A distributed file system with Big Data formats
    - Using some of the studied Big Data formats: SequenceFile, Avro, Parquet
  - A column-family key-value store
    - Apache HBase
  - A document store
    - MongoDB

- There is not a single correct solution
  - The most important part is how you **justify your choices**, and **discuss pros/cons**

Closing