EDEM
Escuela de Empresarios

OPEN DATA

EDEM
Escuela de Empresarios

GFT

Esteban Chiner

Curso 2021/2022

Fecha 04/11/2021

Esteban Chiner

- **Computer Engineer** from the Universitat de València & **Master on Software Engineering** from Universitat Politècnica de València
- Working in GFT as a **Senior Architect** specialized in Big Data and Blockchain
- Part of **Offering Development** team, leading the **"Data"** and **"DLT & Crypto"** offerings
- **Lecturer**
  - Big Data & Analytics Master (**UPV**) – "Big Data Architectures in Financial Services"
  - Data Analytics for Enterprises (**EDEM**) + several seminars on Blockchain
  - Seminaris d'Empresa (**UPC**) – "Developing decentralized apps with Ethereum"
- **Hadoop Certified** in developer, administration and HBase
- Paper published: "**A Big Data Financial Information Management Architecture for Global Banking**"

@echiner

echiner

esteban.chiner@gft.com

Agenda

1. Introduction

- Introduction and types

2. Technical Background

- Data formats
- Delivery mechanisms
- Frequency

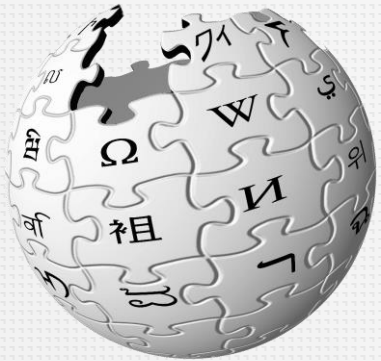
3. Datasources

- Government
- Statistics
- Public Datasets
- Commercial

Introduction

Introduction

“Open Data is the idea that some **data** should be **freely available to everyone** to use and republish as they wish, **without restrictions** from copyright, patents or other mechanisms of control.”



Types of Open Data

Provided by different governments levels (local, regional, national or supra national).

Pure “Open Data”.

Examples: Ajuntament de Valencia, European Data Portal, Data.gov…

Government

Can be considered a sub-type of government open data.

Provides economic or statistical information.

Examples: INE, World Bank Open Data…

Statistics

Usually used for testing purposes or in Artificial Intelligence training.

Cannot be really considered a data source, but a single dataset.

Examples: Google Dataset Search (Kaggle, DataCite…)

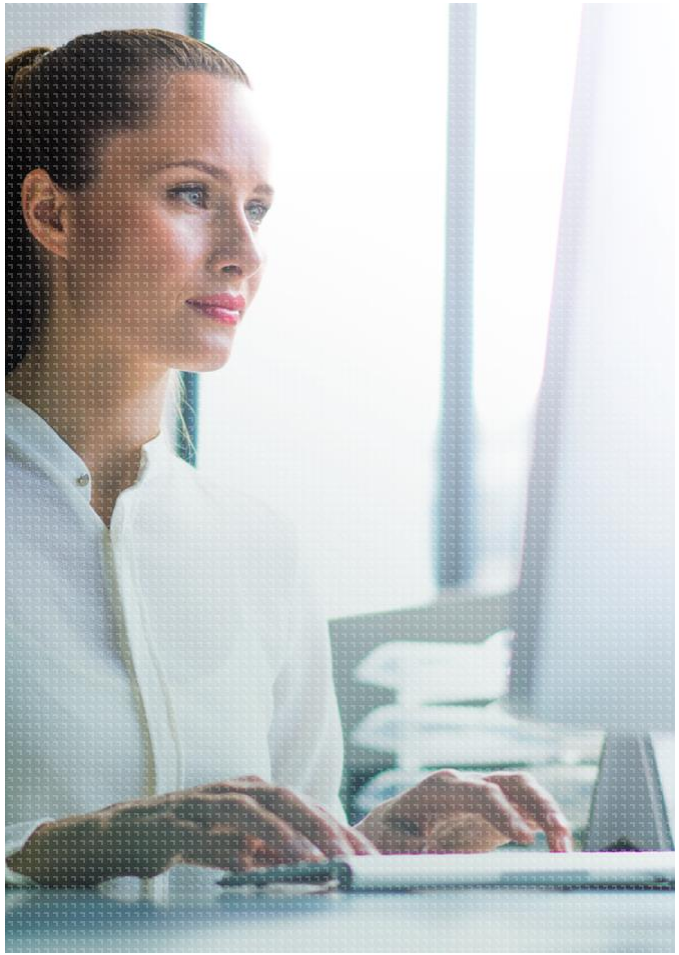
Public Datasets

Data is provided for commercial purposes.

Payment per use (calls, datasets, etc.), although in many cases they provide a free tier.

Examples: Social Networks (Twitter, Reddit…), Financial Data (IEX, Morning Star…)

Commercial



What other examples do you know?

Think about other examples of Open Data in the described areas:

- Government
- Statistics
- Public datasets
- Commercial

Technical Background

Data formats (I)

CSV (Comma Separated Values)

Header (optional)

```
"LatD", "LatM", "LatS", "NS", "LonD", "LonM", "LonS", "EW", "City", "State"
   41,     5,   59, "N",     80,    39,     0, "W", "Youngstown", OH
   42,    52,   48, "N",     97,    23,    23, "W", "Yankton", SD
   46,    35,   59, "N",    120,    30,    36, "W", "Yakima", WA
   42,    16,   12, "N",     71,    48,     0, "W", "Worcester", MA
   43,    37,   48, "N",     89,    46,    11, "W", "Wisconsin Dells", WI
```

Content

Data formats (II)

JSON (JavaScript Object Notation)

```
{
  "glossary": {
    "title": "example glossary",
      "GlossDiv": {
        "title": "S",
          "GlossList": {
            "GlossEntry": {
              "ID": "SGML",
                "SortAs": "SGML",
                "GlossTerm": "Standard Generalized Markup Language",
                "Acronym": "SGML",
                "Abbrev": "ISO 8879:1986",
                "GlossDef": {
                  "para": "A meta-markup language, used to create markup languages such as DocBook.",
                  "GlossSeeAlso": ["GML", "XML"]
                },
                "GlossSee": "markup"
            }
          }
      }
  }
}
```

Column Value (Array)

Column Value (String)

Column Name

Data formats (III)

XML (Extensible Markup Language)

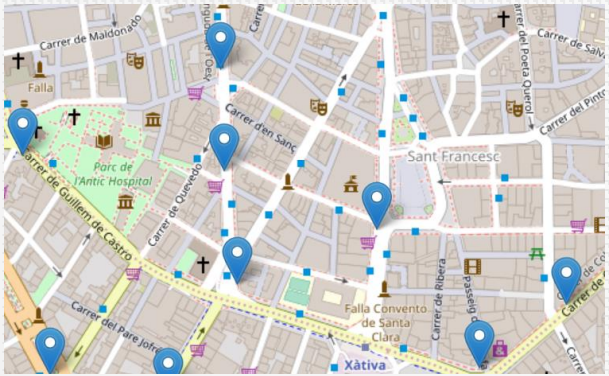
```
<!DOCTYPE glossary PUBLIC "-//OASIS//DTD DocBook V3.1//EN">
<glossary><title>example glossary</title>
  <GlossDiv><title>S</title>
    <GlossList>
      <GlossEntry ID="SGML" SortAs="SGML">
        <GlossTerm>Standard Generalized Markup Language</GlossTerm>
        <Acronym>SGML</Acronym>
        <Abbrev>ISO 8879:1986</Abbrev>
        <GlossDef>
          <para>A meta-markup language, used to create markup languages such as DocBook.</para>
          <GlossSeeAlso OtherTerm="GML">
          <GlossSeeAlso OtherTerm="XML">
        </GlossDef>
        <GlossSee OtherTerm="markup">
      </GlossEntry>
    </GlossList>
  </GlossDiv>
</glossary>
```

Data formats (III)

Geographical Information Formats

- **WMS** (Web Map Service)
  - International standard by OGC
  - Serves georeferenced map images
- **GML** (Geography Markup Language)
  - Developed by OpenGIS
  - Based on XML
- **KML** (Keyhole Markup Language)
  - Used for Google Earth
  - Similar to GML
  - Can be compressed as KMZ

```
<Placemark>
  <name>Alameda, 51/ Menorca</name>
  <styleUrl>#style_symbol_CIRCULO_100.0_ff9b93ff</styleUrl>
  <ExtendedData>
    <Data name="emplazamie">
      <displayName>emplazamie</displayName>
      <value><![CDATA[Alameda, 51/ Menorca]]></value>
    </Data>
    <Data name="modelo">
      <displayName>modelo</displayName>
      <value><![CDATA[ALBUFERA]]></value>
    </Data>
  </ExtendedData>
  <Point>
    <coordinates>
-0.34840107,39.45662252
</coordinates>
  </Point>
</Placemark>
```



Data formats (IV)

News feeds (Atom & RSS)

- Described in XML format
- Delivered via HTTP
- Many readers available (incl. Browser)
- Basically same features, although Atom more advanced

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

<channel>
  <title>W3Schools Home Page</title>
  <link>https://www.w3schools.com</link>
  <description>Free web building tutorials</description>
  <item>
    <title>RSS Tutorial</title>
    <link>https://www.w3schools.com/xml/xml_rss.asp</link>
    <description>New RSS tutorial on W3Schools</description>
  </item>
  <item>
    <title>XML Tutorial</title>
    <link>https://www.w3schools.com/xml</link>
    <description>New XML tutorial on W3Schools</description>
  </item>
</channel>

</rss>
```

Data formats (V)

RDF (Resource Description Framework)

- Defined by the W3C, used for modelling web resources
- Similar to Entity-Relationship or Class Diagrams
- Uses **SPARQL** as Query Language
- Can be serialized in different formats (Turtle, JSON, XML…)

```
@prefix eric:    <http://www.w3.org/People/EM/contact#> .
@prefix contact: <http://www.w3.org/2000/10/swap/pim/contact#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

eric:me contact:fullName "Eric Miller" .
eric:me contact:mailbox <mailto:e.miller123(at)example> .
eric:me contact:personalTitle "Dr." .
eric:me rdf:type contact:Person .
```

Delivery mechanisms

DIRECT DOWNLOAD

A

This is the most common option for Open Data and public datasets, since they just provide downloadable content.

REAL-TIME

B

Data is provided in real-time as an unbounded dataset (stream of data). It can be done via REST endpoint, messaging (e.g. broker, websocket…) or callback (e.g. webhook).

REST API

C

Representational State Transfer (REST) defines how web services should be defined. An URI (endpoint) represents a service and can come in any format (usually JSON).

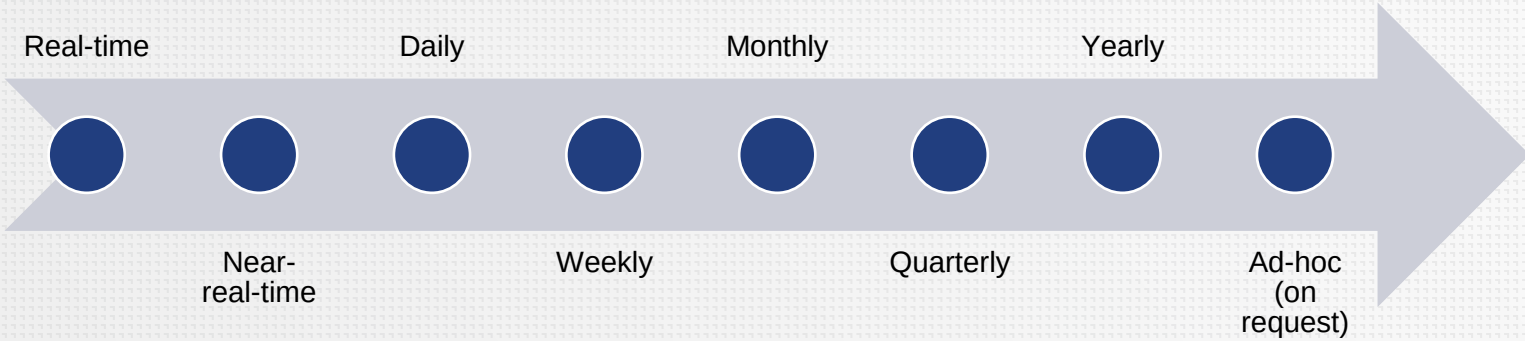
FEEDS

D

HTTP call which provide a result which gets updated more or less frequently. Examples being RSS and Atom.

Frequency

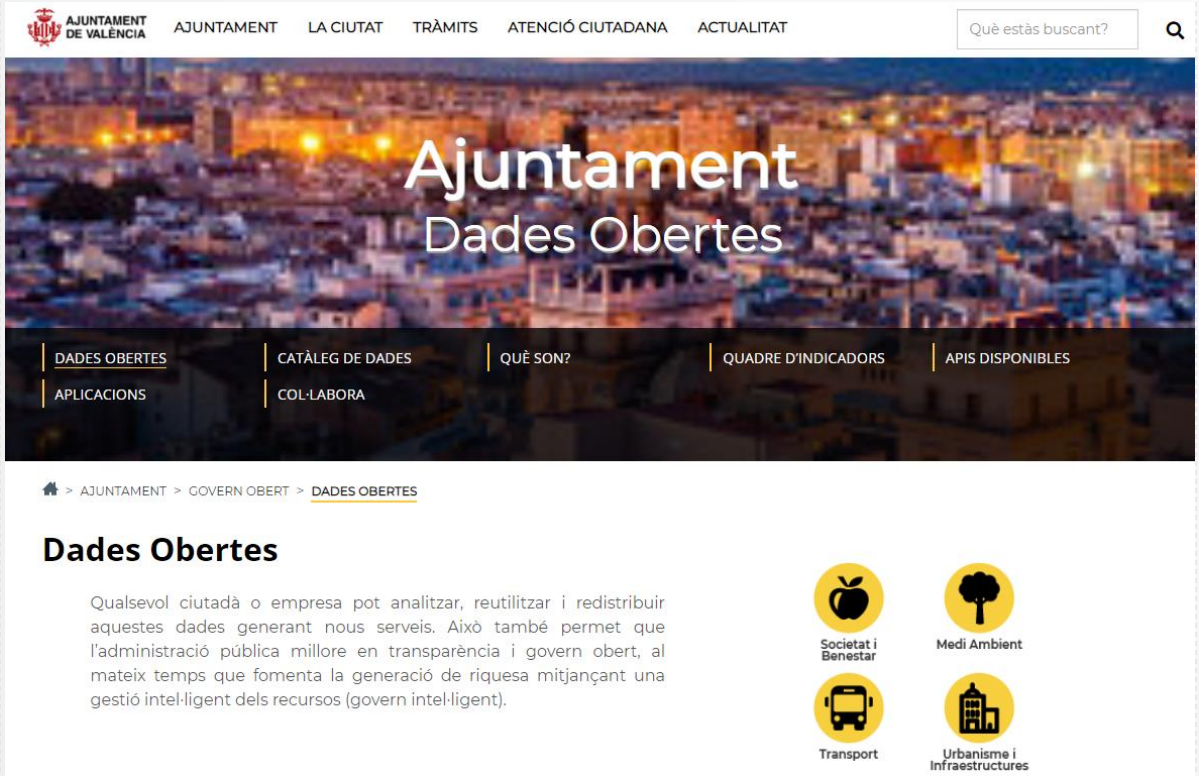
- Depending on the **nature of the data** and the **frequency it is updated**, it can be delivered in many different time spans.



Data Sources

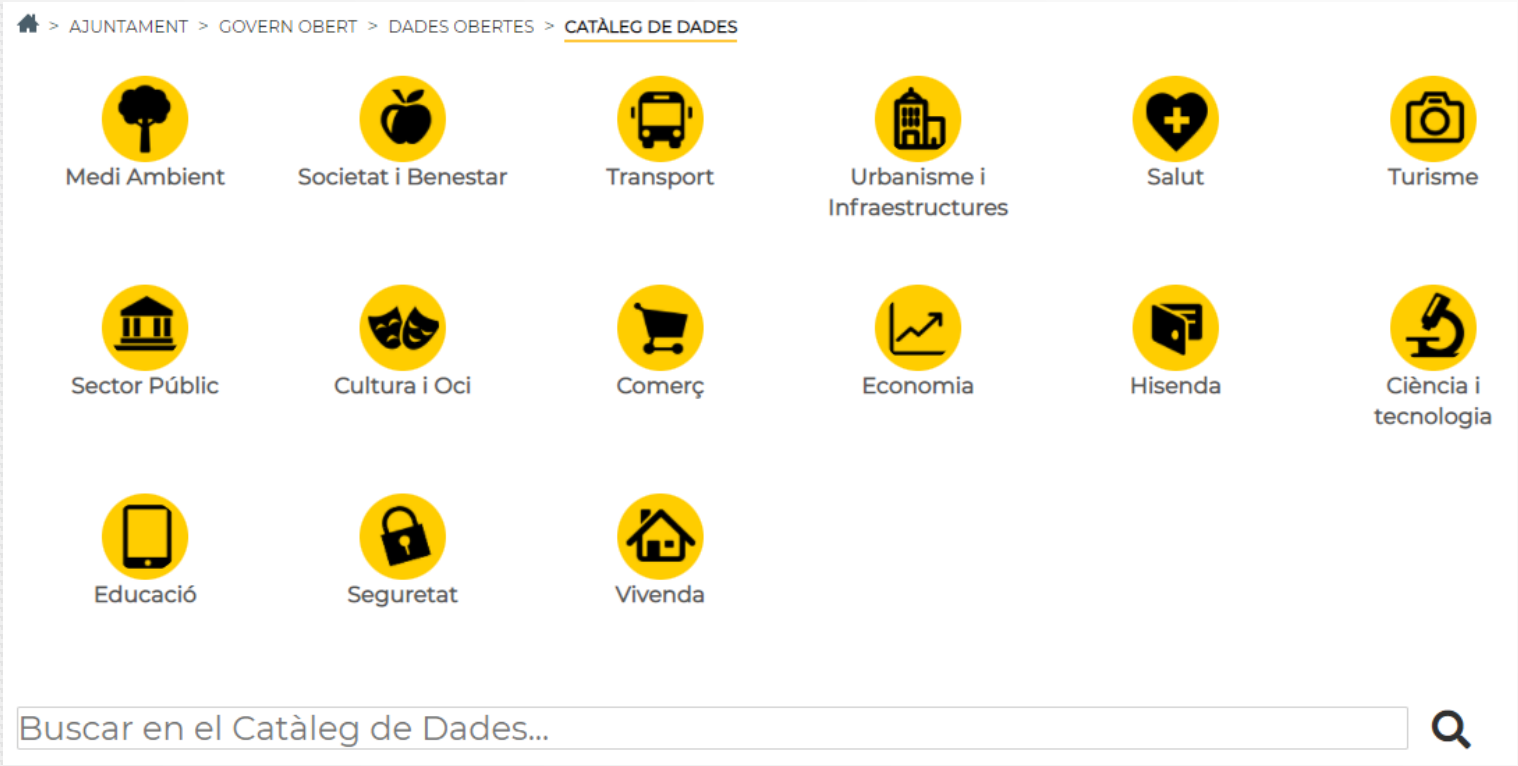
Government

Government – Ajuntament de València



https://www.valencia.es/dadesobertes/va/

Government – Ajuntament de València



https://www.valencia.es/dadesobertes/va/

Government – Ajuntament de València

Aparcament Mobilitat Reduïda

Aparcaments per a vehicles de persones amb mobilitat reduïda. Plazas: Nombre de places d'aparcament

Ajuntament de València

Visites **58**

Societat i Benestar Transport

Grau d'obertura ★★★☆☆

Formats de distribució: WFS, GEOJSON, SHP, GML, WMS, KML, KMZ, CSV

Estacions de Valenbisi

Localització de les estacions de valenbisi. Name: Nom de l'estació de valenbisi Number: Número de l'estació. Address: dirección. Open: Si està operativa o no. Available: Bicicletes...

Ajuntament de València

Visites **138**

Transport Turisme

Grau d'obertura ★★★☆☆

Formats de distribució: WFS, GEOJSON, SHP, GML, WMS, KML, KMZ, CSV

Carril Bici

Itineraris ciclistes de la ciutat de València

Ajuntament de València

Visites **160**

Transport Turisme

Grau d'obertura ★★★☆☆

Formats de distribució: WFS, GEOJSON, SHP, GML, WMS, KML, KMZ, CSV

Bol·lards extraïbles

Bol·lards extraïbles amb clau estàndard aprovada pel Servei de Bombers

Ajuntament de València

Visites **31**

Transport

Grau d'obertura ★★★☆☆

Formats de distribució: GML, GEOJSON, KML, KMZ, SHP, WFS, WMS, CSV

ETIQUETES

circulació (4)
parking (4)
aparcament (3)
vehicle (2)
vehícle (2)

+ Expandir etiquetes

FORMATS

CSV (120)
SHP (100)
WMS (100)
WFS (99)
GEOJSON (98)

+ Expandir formats

FREQÜÈNCIA D'ACTUALITZACIÓ

Diària (25)
Setmanal (36)
Mensual (15)
Trimestral (10)
Semestral (4)

+ Expandir

DARRERA ACTUALITZACIÓ

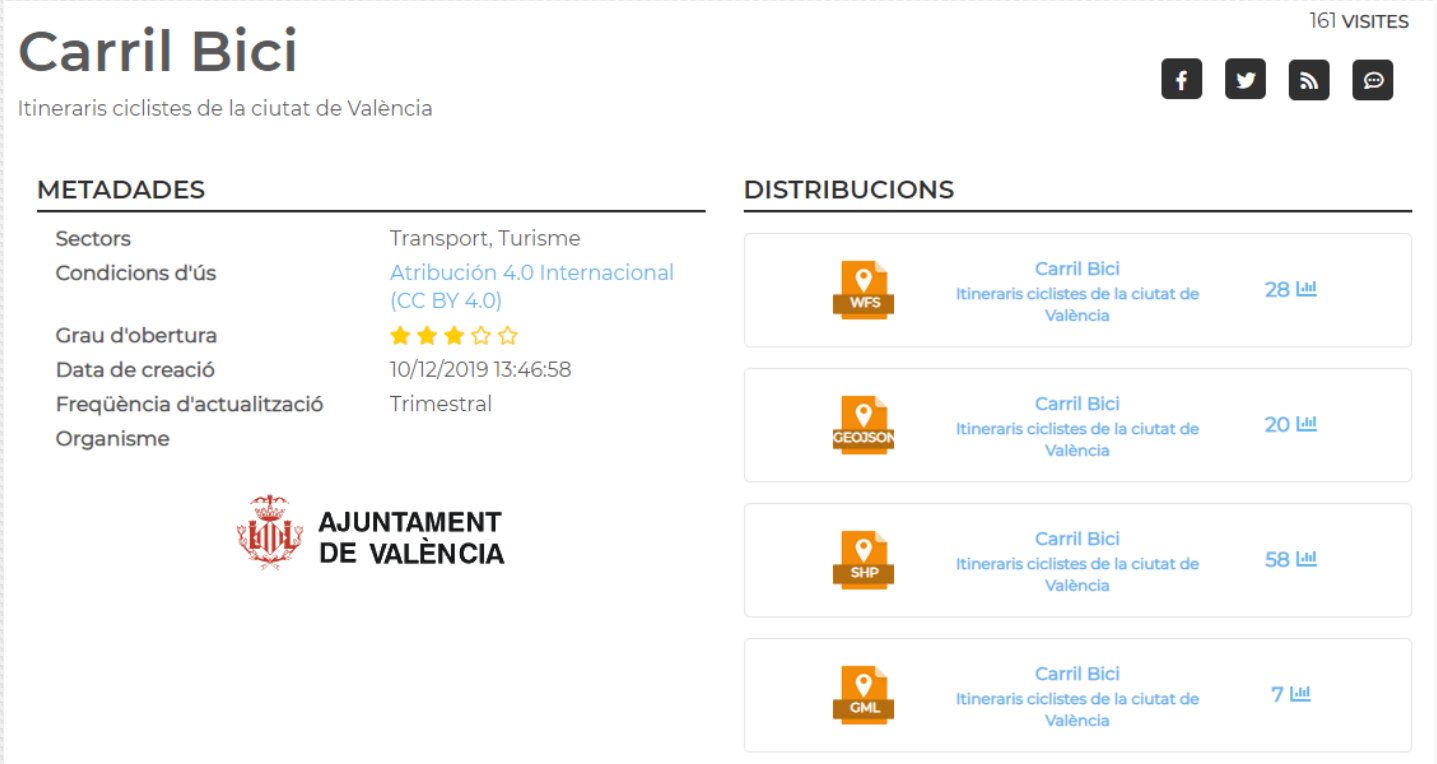
1/1/2013 31/10/2021

GRAU D'OBERTURA

★☆☆☆☆
★★☆☆☆
★★★☆☆
★★★★☆
★★★★★

https://www.valencia.es/dadesobertes/va/

Government – Ajuntament de València

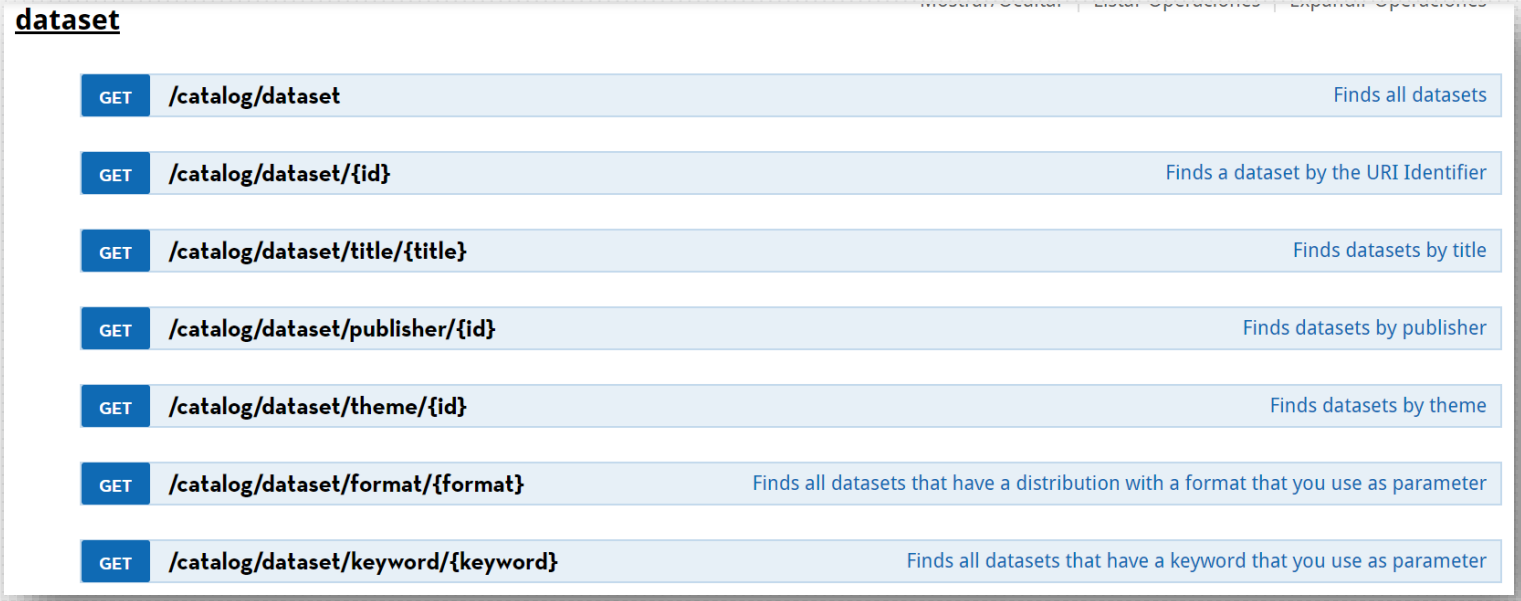


https://www.valencia.es/dadesobertes/va/

Government – Spanish Government

https://datos.gob.es_

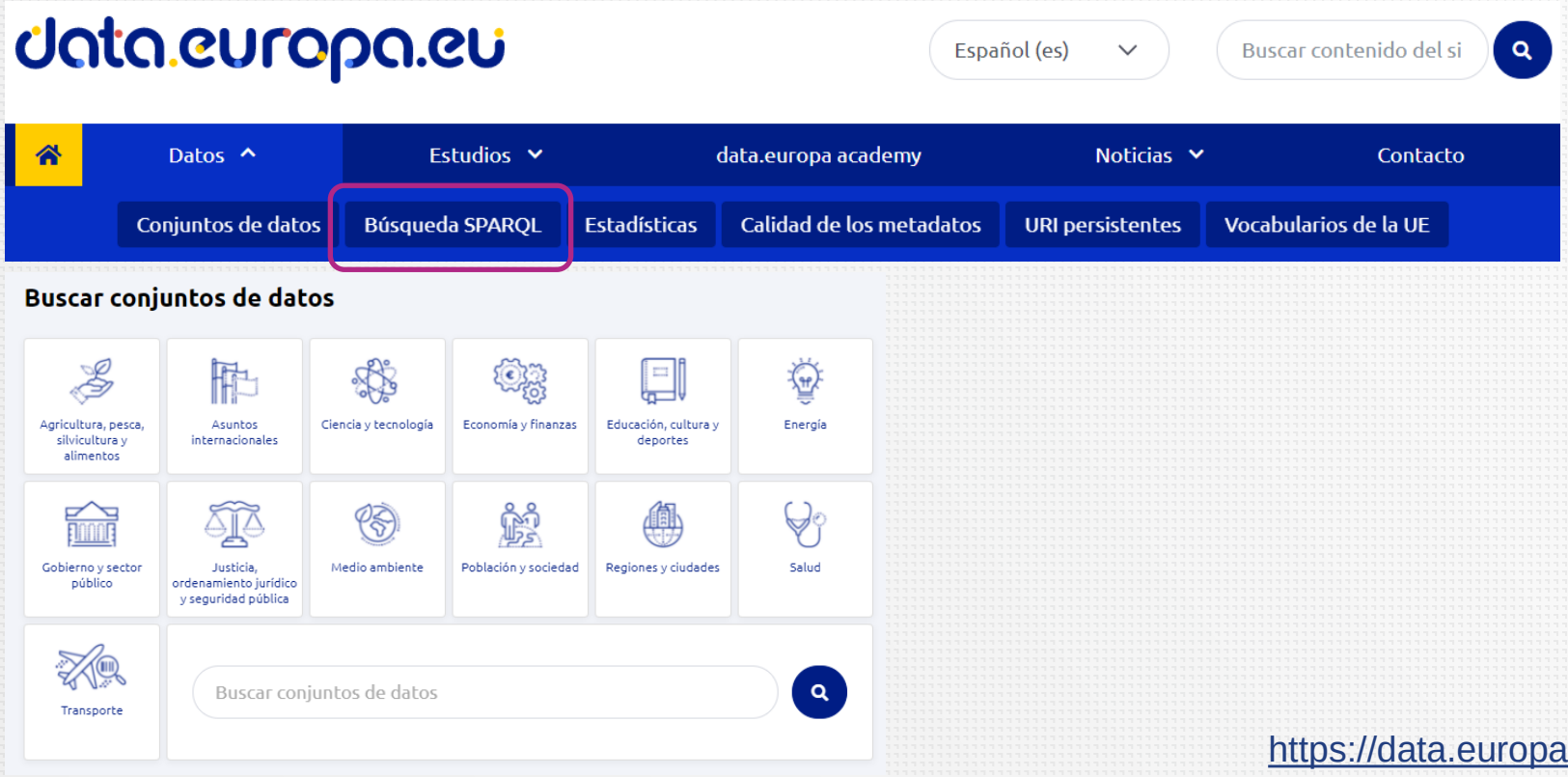
Government – Spanish Government



https://datos.gob.es/es/apidata#!/dataset/findDatasetById

https://datos.gob.es

Government – European Union



https://data.europa.eu/es

Government – European Union

Prefijos

```
SELECT (count(*) AS ?count)
WHERE { { ?s a dcat:Dataset } }
LIMIT 100
```

Formato

Límite

Vista previa de los datos de resultados

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rs="http://www.w3.org/2001/sw/DataAccess/tests/result-set#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <rs:ResultSet>
    <rs:size rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
    >1</rs:size>
    <rs:solution rdf:parseType="Resource">
      <rs:binding rdf:parseType="Resource">
        <rs:value rdf:datatype="http://www.w3.org/2001/XMLSchema#integer"
        >917309</rs:value>
        <rs:variable>count</rs:variable>
      </rs:binding>
```

Descargar

https://data.europa.eu/es

Government – European Union

Query +

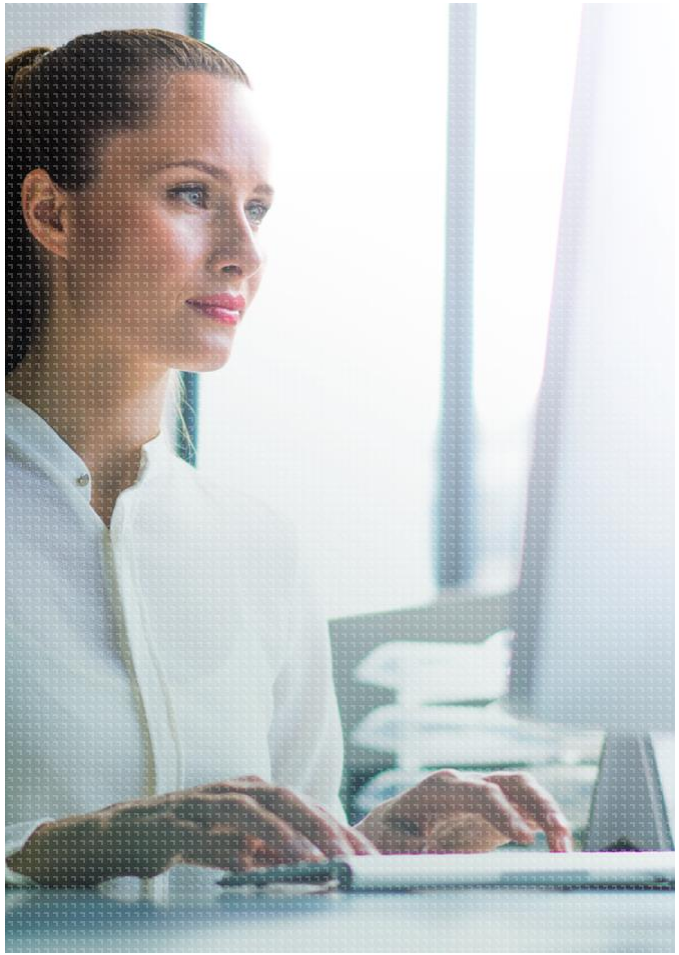
https://data.europa.eu/sparql

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX odp:  <http://data.europa.eu/euodp/ontologies/ec-odp#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT * WHERE { ?d a dcat:Dataset } LIMIT 10

```

Table Response 10 results in 0.176 seconds


```
{
  "head": {
    "link": [],
    "vars": [
      "d"
    ]
  },
  "results": {
    "distinct": false,
    "ordered": true,
    "bindings": [
      {
        "d": {
          "type": "uri",
          "value": "http://data.europa.eu/88u/dataset/5ig8huzrleoocpdue37xow"
        }
      },
      {
        "d": {
          "type": "uri",
          "value": "http://data.europa.eu/88u/dataset/honey-bee-seasonal-mortality-2012-2014-epilobee-analysis"
        }
      },
```


https://data.europa.eu/es



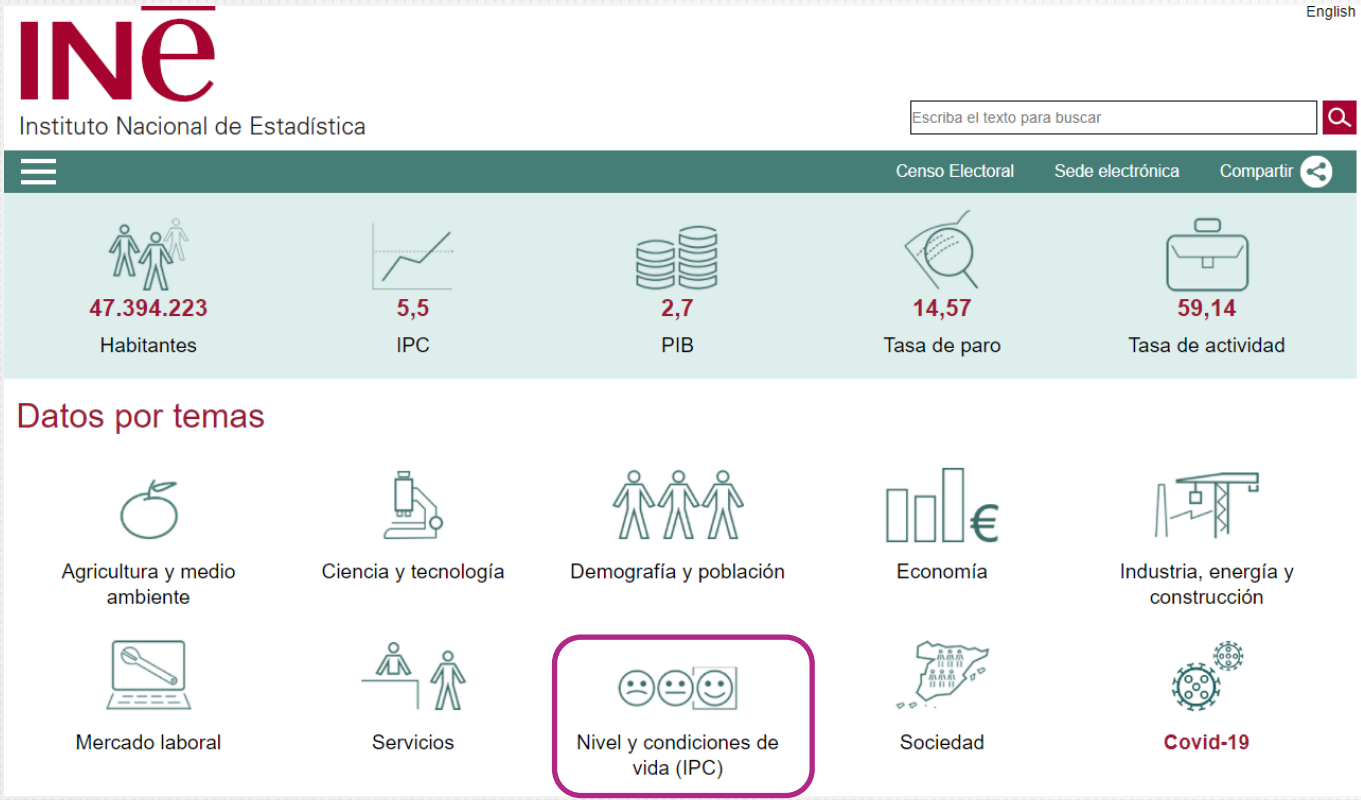
Analysing a dataset

- Search for **a couple of datasets** from two different sources of government Open Data
- Download in **CSV** format
- Load into **Excel** and “**analyse**”
- Share with the class

Data Sources

Statistics

Statistics – INE (Instituto Nacional de Estadística)



https://www.ine.es/

Statistics – INE (Instituto Nacional de Estadística)

INEbase / **Nivel y condiciones de vida (IPC)**

- Estadísticas A...Z
- Lista completa de operaciones
- Agricultura y medio ambiente
- Ciencia y tecnología
- Demografía y población
- Economía
- Industria, energía y construcción
- Mercado laboral
- Servicios
- **Nivel y condiciones de vida (IPC)**
- Sociedad
- Covid-19

Nivel y condiciones de vida (IPC)

Índices de precios de consumo y vivienda

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos
Índice de precios de consumo	Avance. Octubre 2021
Índice de precios de consumo armonizado	Avance. Octubre 2021
Índice de precios de vivienda	Trimestre 2/2021

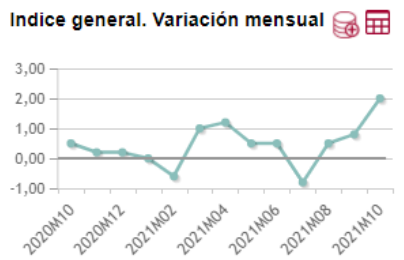
Operaciones elaboradas por otros organismos del sistema estadístico nacional

- Estadística de Valor Tasado de Vivienda
- Estadística de Precios del Suelo
- Estadística de Precios Públicos Universitarios

Condiciones de vida

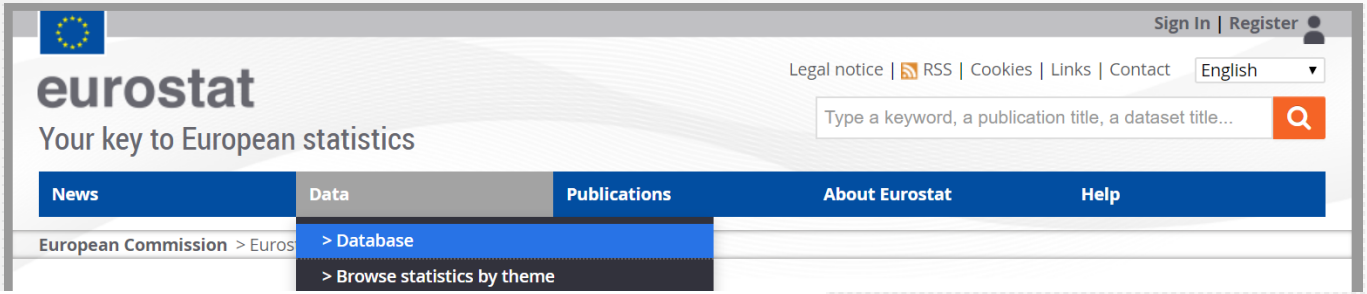
Índice de precios de consumo. Base 2016 - Avance. Octubre 2021

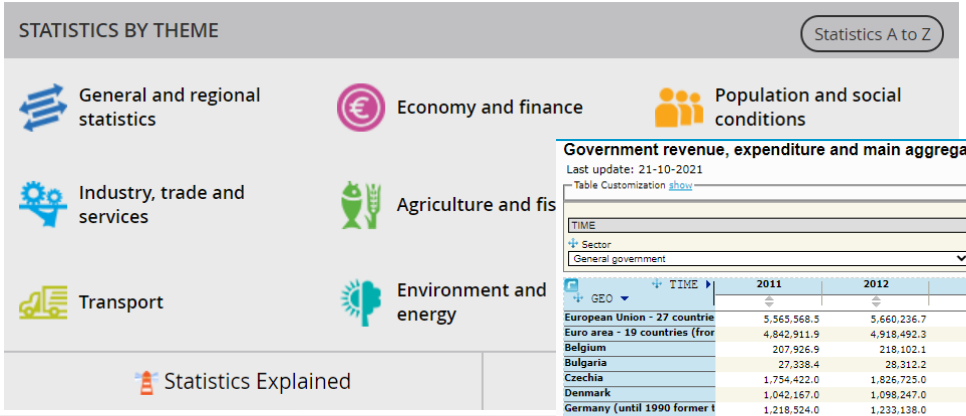
	Variación mensual	Variación anual
Indice general	2,0	5,5
Bienes COVID-19	1,1	1,3
Servicios COVID-19	3,6	16,2
Alimentos y bebidas no alcohólicas	1,3	1,6
Alimentos sin elaboración	2,5	0,8
Alimentos elaborados	0,5	1,8



https://www.ine.es/

Statistics – Eurostat





Government revenue, expenditure and main aggregates

Last update: 21-10-2021

Table Customization show

TIME | GEO

Sector: General government

National accounts indicator (ESA 2010): Total general government expenditure

GEO \ TIME	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
European Union - 27 countrie	5,565,568.5	5,660,236.7	5,708,987.2	5,769,961.1	5,874,538.1	5,942,485.0	6,105,908.5	6,295,747.9	6,520,492.5	7,117,763.7
Euro area - 19 countries (fror	4,842,911.9	4,918,492.3	4,960,345.4	5,012,825.5	5,095,520.6	5,163,760.7	5,289,225.4	5,436,975.4	5,622,014.7	6,129,127.1
Belgium	207,926.9	218,102.1	220,469.6	224,069.4	223,850.7	228,451.2	231,560.7	240,313.2	247,898.4	270,409.9 (p)
Bulgaria	27,338.4	28,312.2	31,057.0	36,329.2	36,172.9	33,155.3	35,724.8	40,622.8	42,737.3	50,139.0
Czechia	1,754,422.0	1,826,725.0	1,766,916.0	1,852,255.0	1,939,612.0	1,906,806.0	1,992,172.0	2,196,478.0	2,377,636.0	2,689,119.0
Denmark	1,042,167.0	1,098,247.0	1,077,153.0	1,093,954.0	1,110,377.0	1,106,149.0	1,108,531.0	1,139,013.0	1,147,822.0	1,244,789.0
Germany (until 1990 former t	1,218,524.0	1,233,138.0	1,263,544.0	1,296,940.0	1,335,789.0	1,390,374.0	1,443,273.0	1,492,813.0 (p)	1,562,698.0 (p)	1,712,131.0 (p)
Estonia	6,241.4	7,031.5	7,266.3	7,571.1	8,154.5	8,568.1	9,351.4	10,184.3	10,936.0	12,313.1
Ireland	81,157.6	74,615.5	72,809.7	73,241.6	76,355.6	75,791.2	77,802.2	82,586.6	86,285.6	102,032.8
Greece	112,064.0	106,844.0	112,926.0	89,913.0	95,336.0	87,154.0	85,871.0	87,137.0	87,758.0	98,871.0
Spain	490,976.0	501,688.0	467,649.0	465,650.0	472,962.0	472,733.0	478,804.0	501,774.0	524,037.0	588,279.0

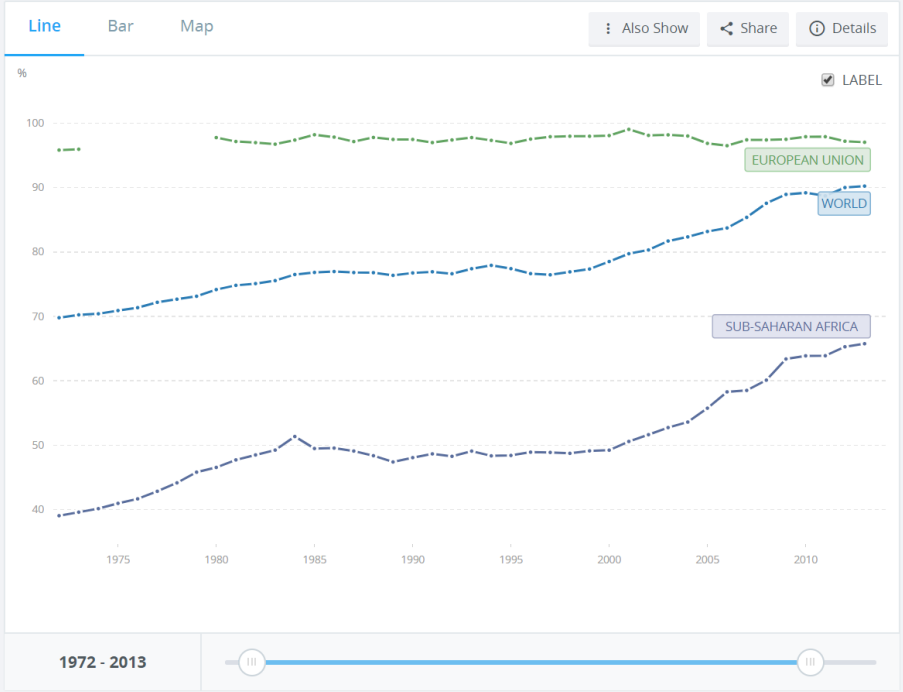
https://ec.europa.eu/eurostat

Statistics – World Bank

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

AREAS:

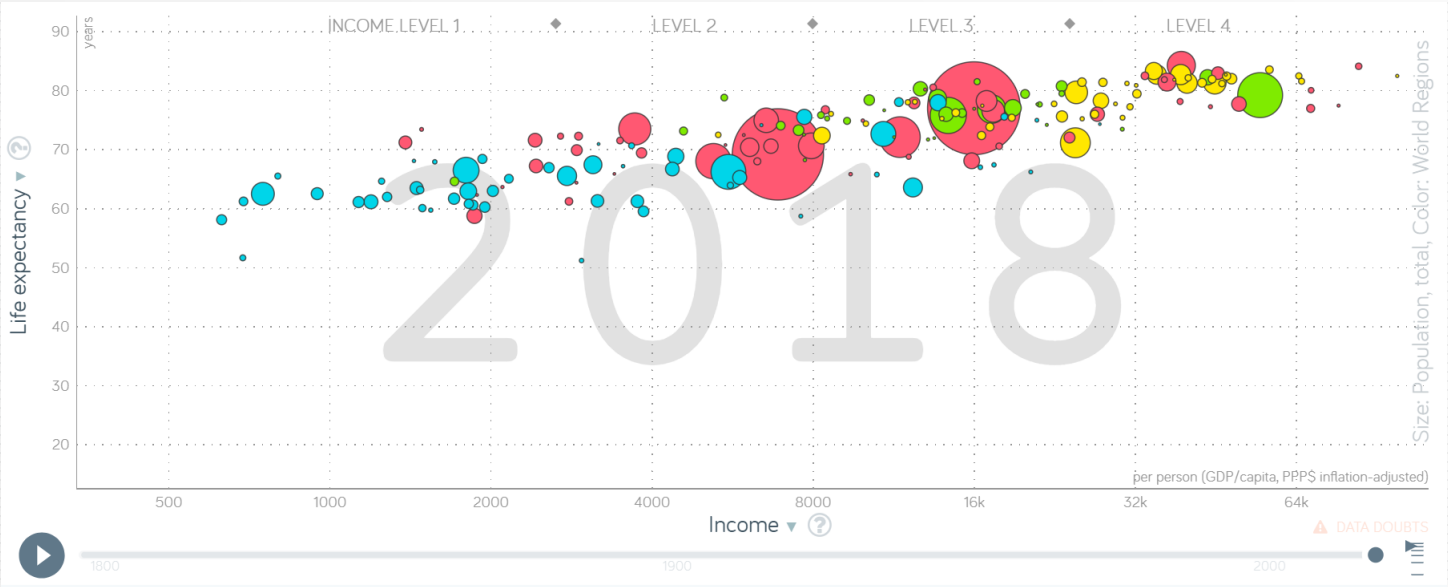
- Agriculture & Rural Development
- Aid Effectiveness
- Climate Change
- Economy & Growth
- Education
- Energy & Mining
- Environment
- External Debt
- Financial Sector
- Gender
- Health
- Infrastructure
- Poverty
- Private Sector
- Public Sector
- Science & Technology
- Social Development
- Social Protection & Labor
- Trade
- Urban Development



https://data.worldbank.org/

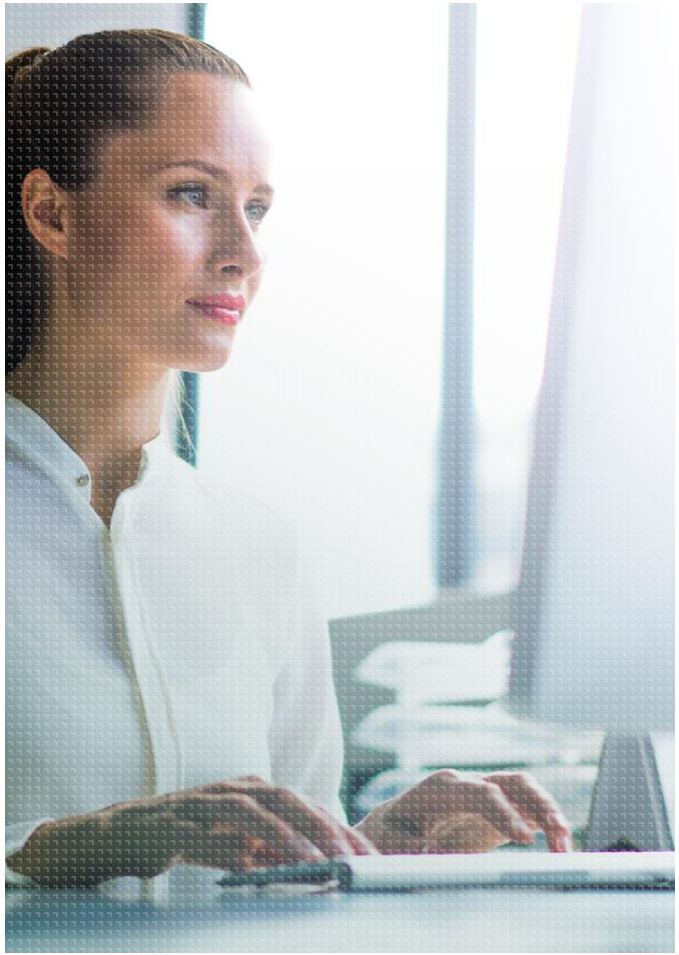
Statistics – Gapminder

- Great tool for **data crunching** and **visualization**



https://www.gapminder.org/tools/#$chart-type=bubbles

https://www.gapminder.org/tools



Socio-economic analysis

- In the **World Bank Data** site:
  - Compare the GDP Growth of “Low Income” vs “High Income” countries
  - Check the evolution of extreme poverty (1.90$/day) over time
  - What is the fertility rate in India? Compare to China, Spain and the rest of the world
  - CO2 emisions → compare China, USA and European Union
  - Now change to CO2 emisions per capita

- In **Gapminder**:
  - Child Mortality vs Income → What is the evolution over time
  - Sustainability Development Index vs Income
  - Human Development Index vs Income

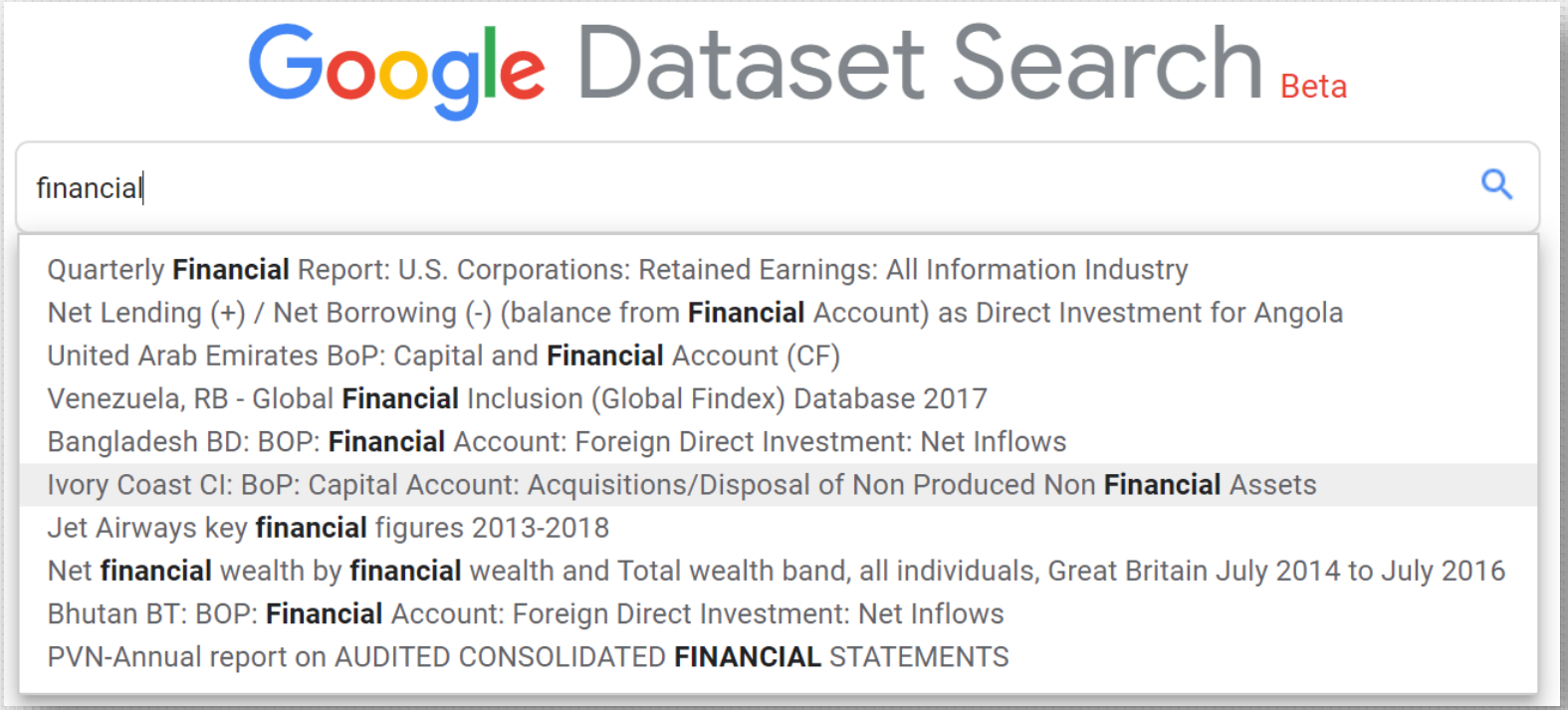
- (Optional) In **INE**, search for the evolution of the gap in GNP (PIB) between regions in Spain.

- Try to explain the results

Data Sources

Public Datasets

Public Datasources – Google Search



https://toolbox.google.com/datasetsearch

Public Datasources – Curated lists

70 free and amazing data sources for data visualization

- https://bigdata-madesimple.com/70-amazing-and-free-data-sources-for-data-visualization/

Big Data And AI: 30 Amazing (And Free) Public Data Sources For 2018

- https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/#1b8fed5f5f8a

70 Amazing Free Data Sources You Should Know

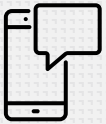
- https://www.kdnuggets.com/2017/12/big-data-free-sources.html

Data Sources

Commercial

Commercial – Introduction

- There are many companies which provide data in **different models** (DaaS, pay per use, downloadable content, indirect payment…) and they are provided by **data vendors**.



Social Networks

Examples: Twitter, Reddit or Meetup

Economy & Financial

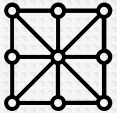
Examples: IEX Cloud or Quandl

Simulated Data

Examples: PubNub on sensor data

Cloud Datasets

Examples: GCP Big Query datasets



Many Others

Examples: Crypto exchanges, Google Analytics

Commercial – IEX Cloud

- **IEX Cloud** is a **source of curated financial data**, providing institutional grade data, including fundamentals, ownership, international equities, mutual funds, options, **real-time data**, and alternative data – all in one **API**.

SCALE	GROW	LAUNCH
~~$699/mo~~	~~$59/mo~~	~~$19/mo~~
499/mo	49/mo	9/mo
Billed annually	**Billed annually**	**Billed annually**
Great for demanding applications and established businesses	Best for teams and developers that support users	Ideal for individual developers and students
Select Scale	Select Grow	Select Launch
✓ **$1 per 3,000,000** Pay as you go messages	✓ **$1 per 2,000,000** Pay as you go messages	✓ **$1 per 1,000,000** Pay as you go messages
✓ Start with **2.000.000.000** Messages per month	✓ Start with **100.000.000** Messages per month	✓ Start with **5.000.000** Messages per month
✓ **Personal & Commercial use**	✓ **Personal & Commercial use**	✓ **Personal & Commercial use**
✓ **Enterprise Data**		
✓ **99.95% SLA**		
✓ **Premium Support**		

https://iexcloud.io

Commercial – Twitter

- **Twitter Developer API** allows you to **publish and analyze** Tweets, **optimize ads**, and create unique **customer experiences**.

- Provides the following **functionality** via API:

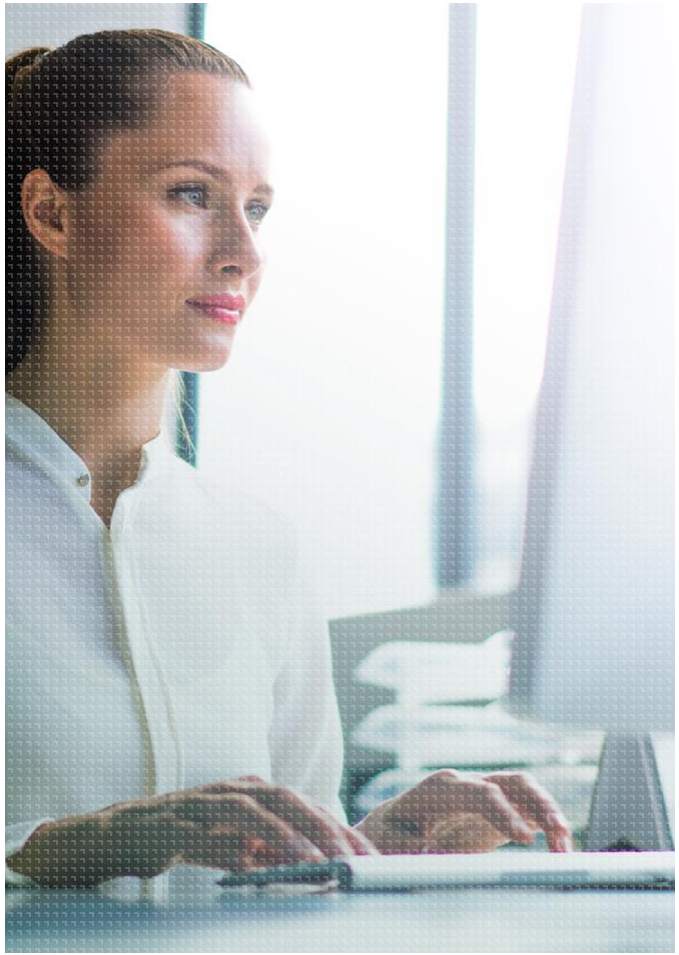
Account & Users:

- List followers and Friends
- Search users
- Manage account
- Mute, block and report

Tweets:

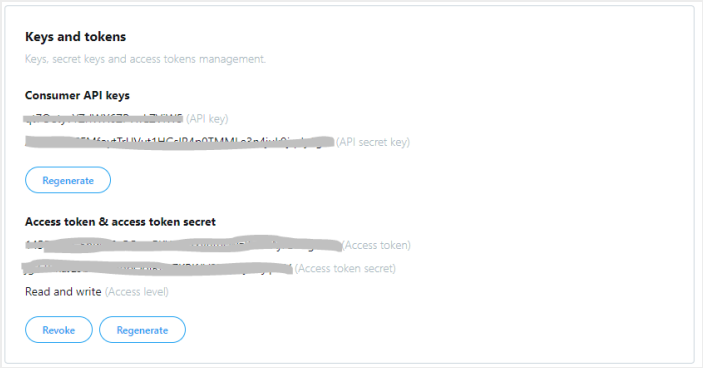
- Get user timeline
- Post tweets
- **Real-time tweets (sample)**
- Search

https://developer.twitter.com



Setting up Dev Twitter Account

- Sign up to Twitter Dev:
  - https://developer.twitter.com
- Go to “Apps”
- “Create an app” and once created, under “Keys and tokens” you will see the following:



GFT

Shaping the future of digital business

Esteban Chiner Sanz

Senior Architect

GFT – Data Practice

esteban.chiner@gft.com