OPEN DATA

GFT

Esteban Chiner

Curso 2020/2021 - Edición 2

Fecha 24/10/2020

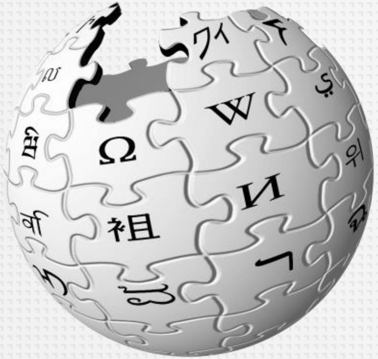
Agenda

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

Government

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

Pure “Open Data”.

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

Statistics

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

Provides economic or statistical information.

Examples: INE, World Bank Open Data…

Public Datasets

Usually used for testing purposes or in Artificial Intelligence training.

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

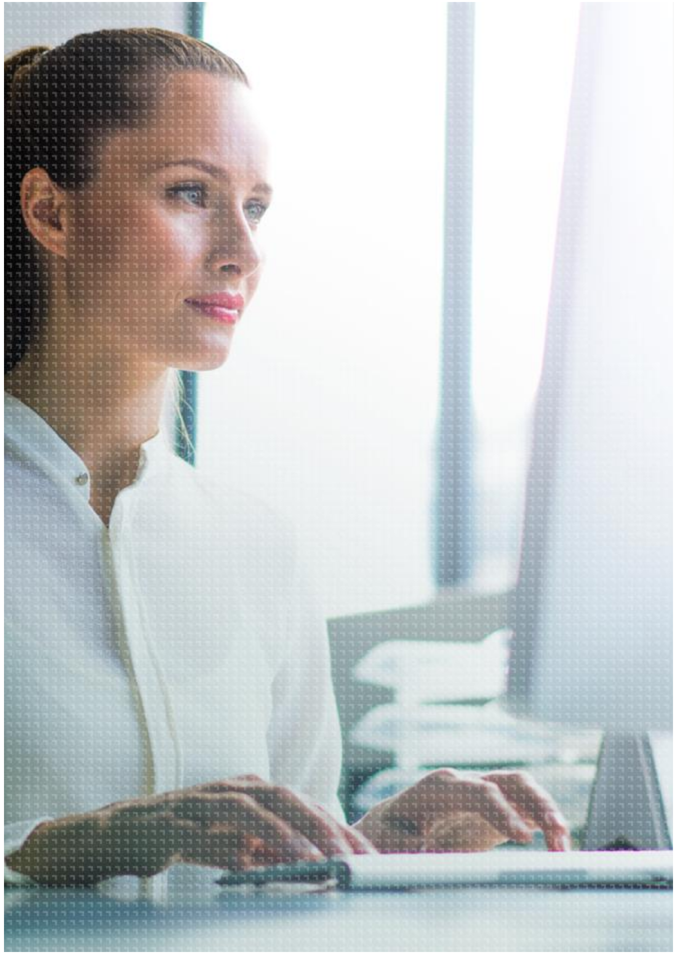
Examples: Google Dataset Search

Commercial

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), Financial Data (IEX)



What other examples do you know?

- Government
- Statistics
- Public datasets
- Commercial

Technical Background

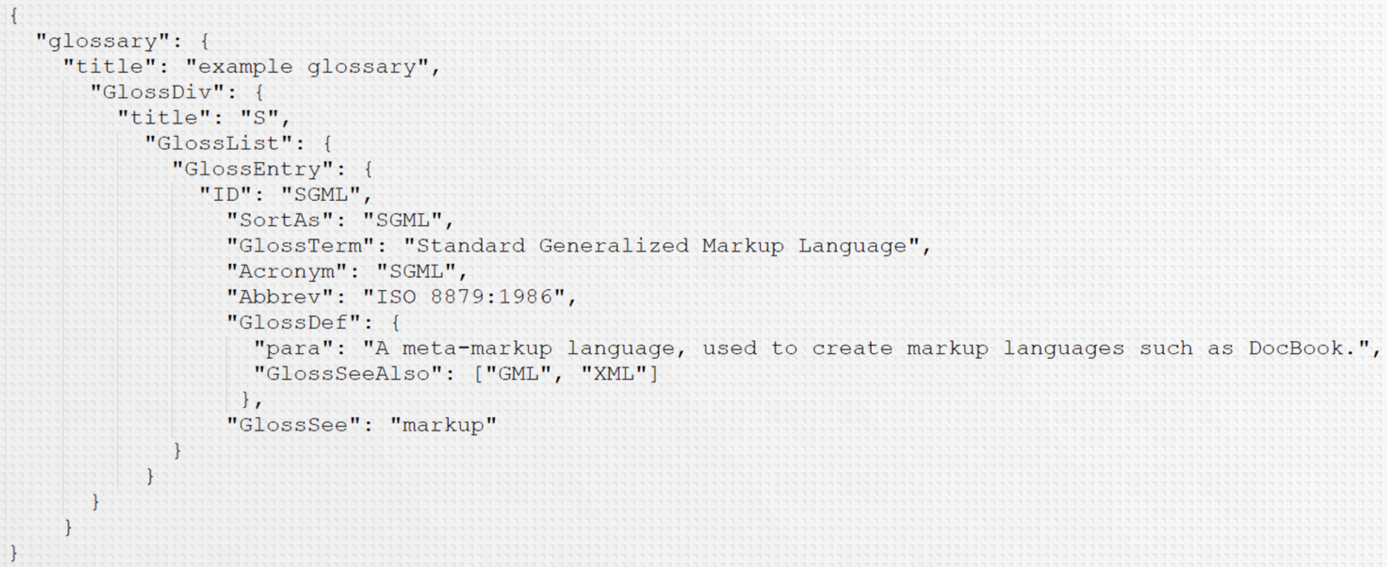
Data formats (I)

CSV (Comma Separated Values)

```
"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
```

Data formats (II)

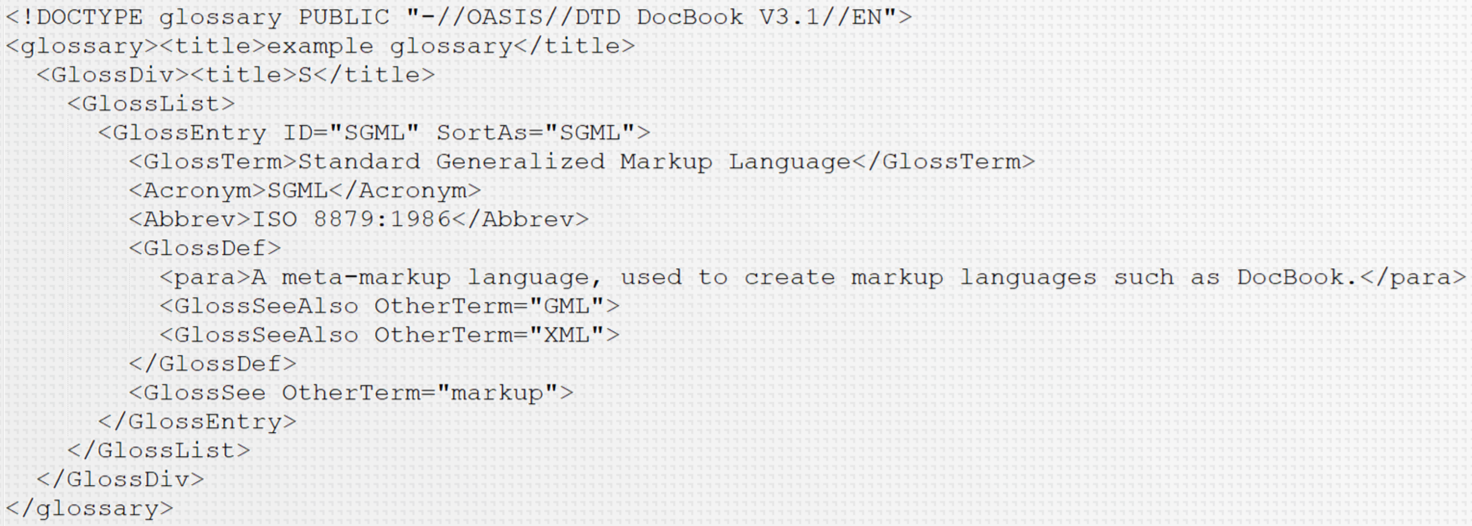
JSON (JavaScript Object Notation)



```
{
  "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"
            }
          }
      }
    }
}
```

Data formats (III)

XML (Extensible Markup Language)

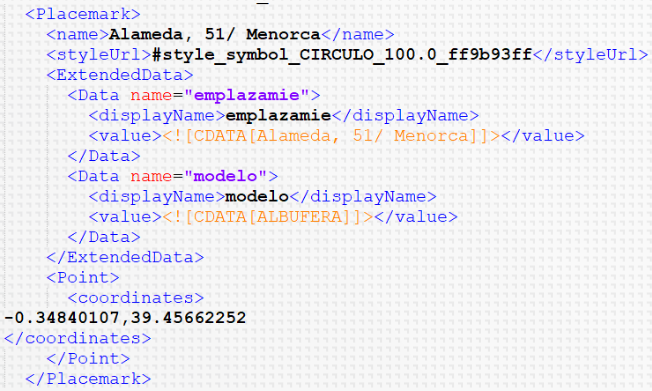


```
<!DOCTYPE glossary PUBLIC "-//OASIS//DTD DocBook V3.1//EN">
<glossary><title>example glossary</title>
  <GlossDiv><title>S</title>
    <GlossList>
      <GlossEntry ID="SGML" SortAs="SGML">
        <GlossTerm>Standard Generalized Markup Language</GlossTerm>
        <Acronym>SGML</Acronym>
        <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>
```

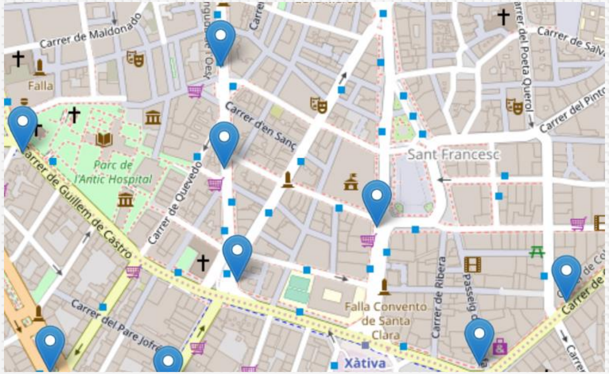
Data formats (III)

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



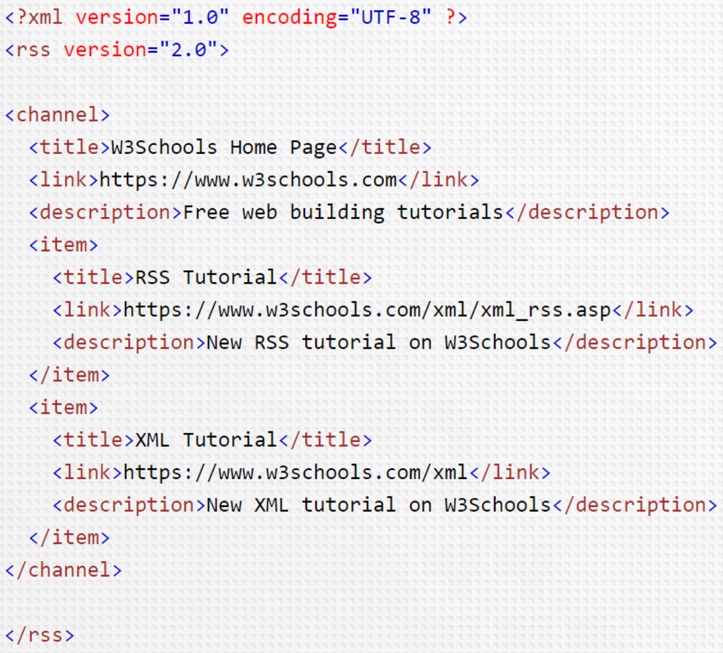
```
  <Placemark>
    <name>Alameda, 51/ Menorca</name>
    <styleUrl>#style_symbol_CIRCULO_100.0_ff9b93ff</styleUrl>
    <ExtendedData>
      <Data name="emplazamie">
        <displayName>emplazamie</displayName>
        <value><![CDATA[Alameda, 51/ Menorca]]></value>
      </Data>
      <Data name="modelo">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

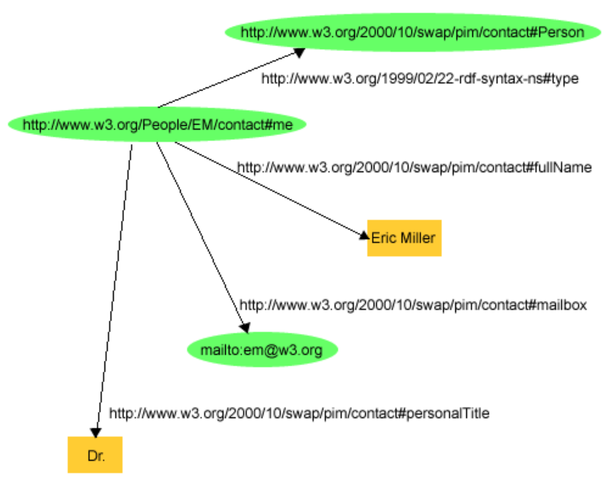


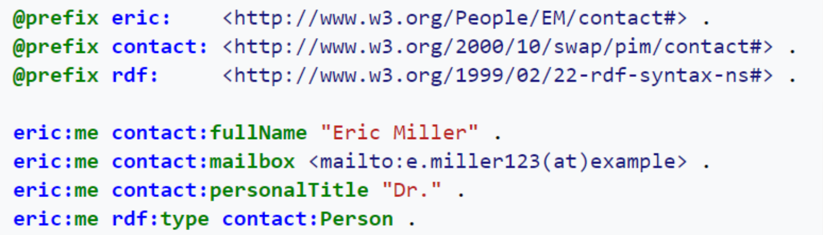
```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

<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…)


http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.



```
@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

A

DIRECT DOWNLOAD

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

B

REAL-TIME

Data is provided in real-time as an unbounded dataset (stream of data). It can be done via REST endpoint or messaging.

C

REST API

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).

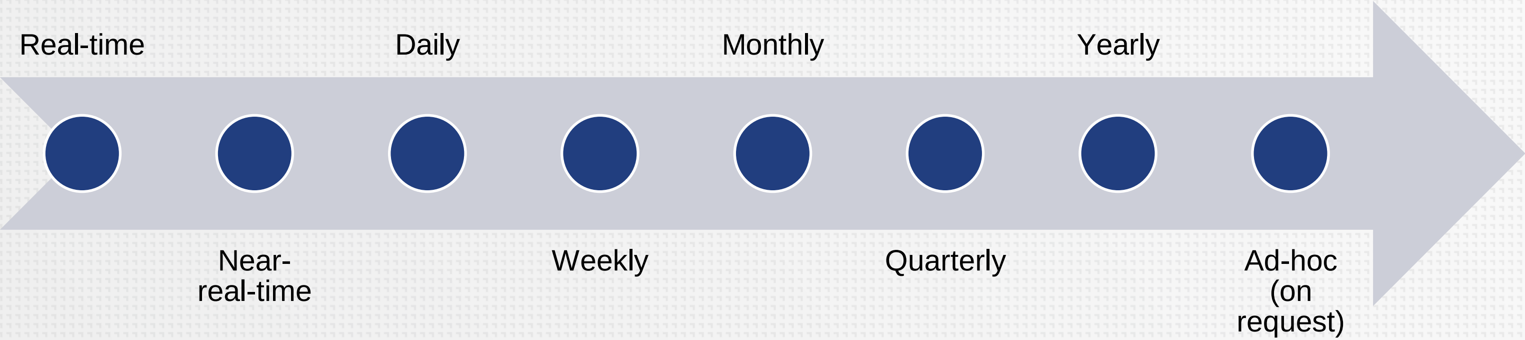
D

FEEDS

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

Frequency

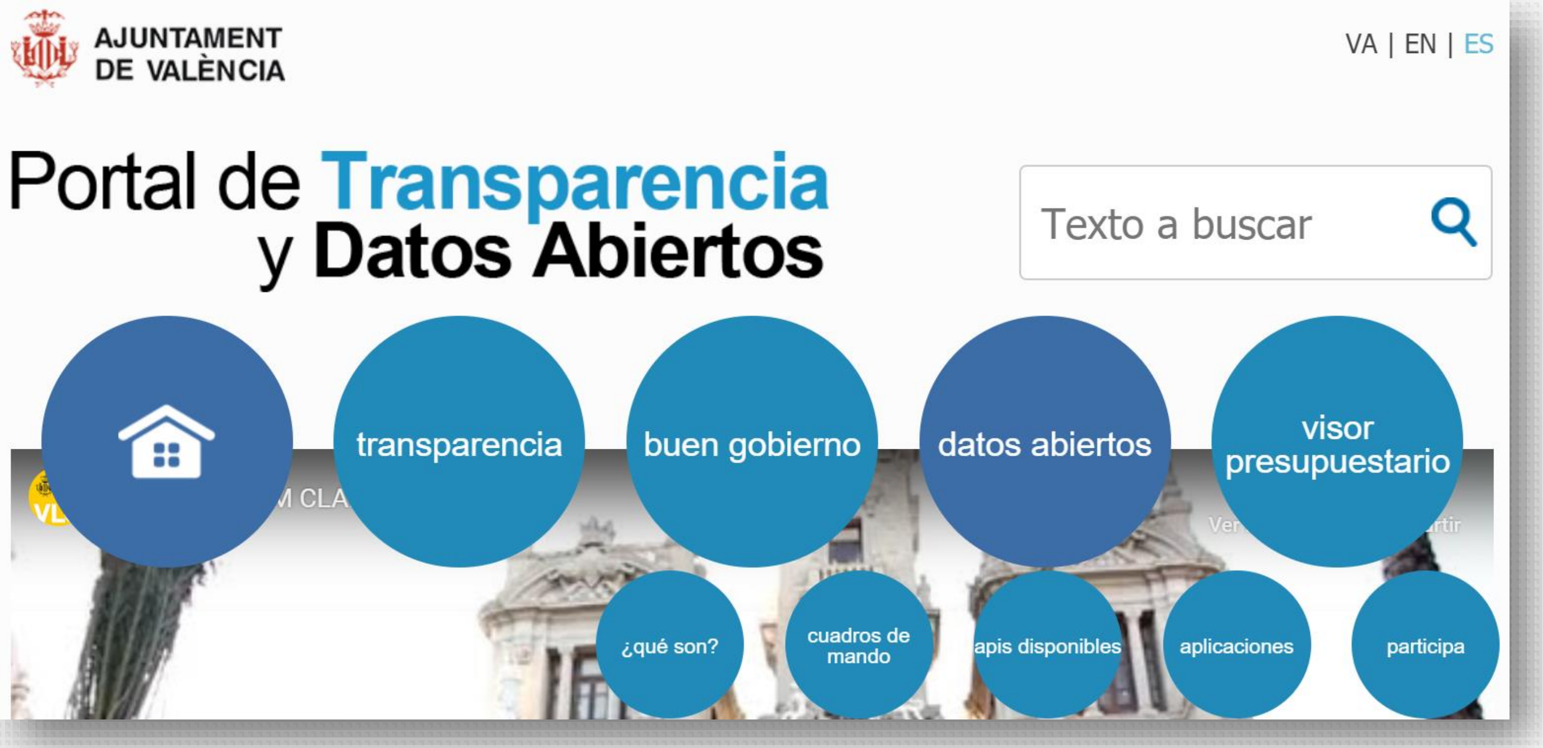
- Depending on the **nature of the data** and the **frequency it is updated**, it can be delivered in many different time spans.



Data Sources

Government

Government – Ajuntament de València



AJUNTAMENT
DE VALÈNCIA

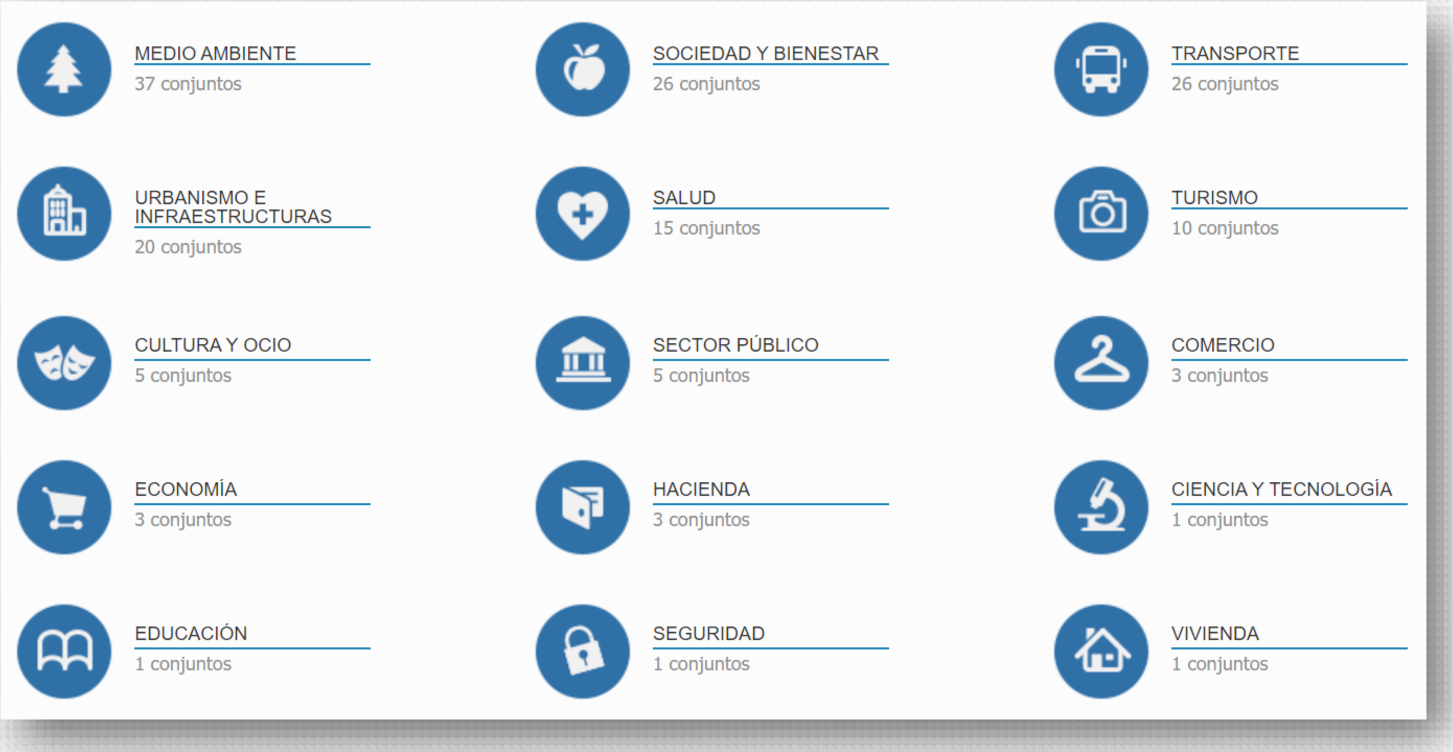
VA | EN | ES

Portal de Transparencia
y Datos Abiertos

Texto a buscar

http://gobiernoabierto.valencia.es/es/

Government – Ajuntament de València



MEDIO AMBIENTE
37 conjuntos

SOCIEDAD Y BIENESTAR
26 conjuntos

TRANSPORTE
26 conjuntos

URBANISMO E INFRAESTRUCTURAS
20 conjuntos

SALUD
15 conjuntos

TURISMO
10 conjuntos

CULTURA Y OCIO
5 conjuntos

SECTOR PÚBLICO
5 conjuntos

COMERCIO
3 conjuntos

ECONOMÍA
3 conjuntos

HACIENDA
3 conjuntos

CIENCIA Y TECNOLOGÍA
1 conjuntos

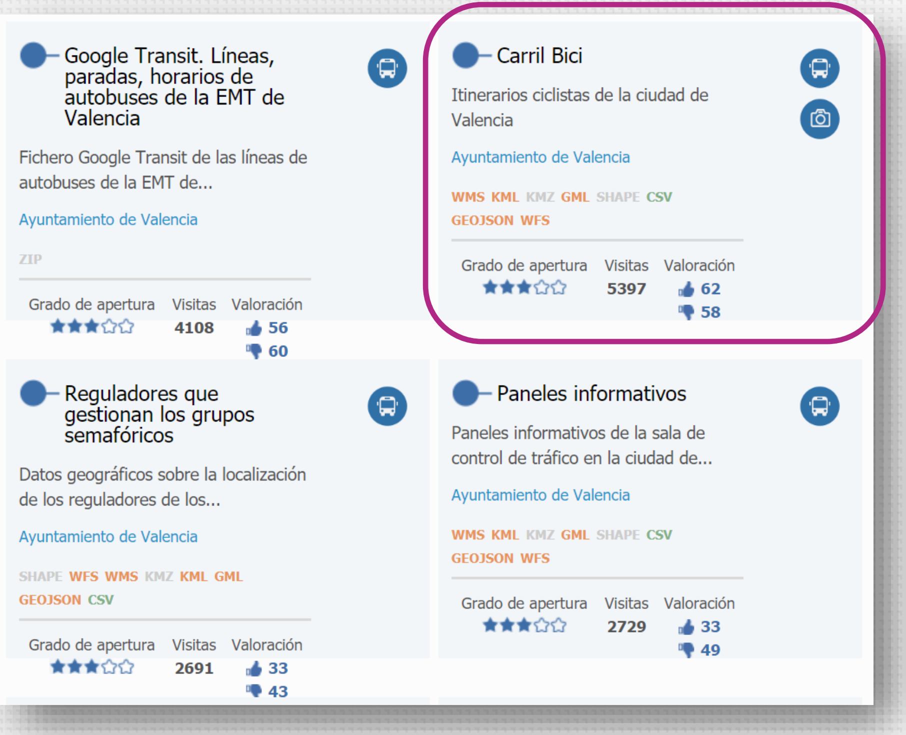
EDUCACIÓN
1 conjuntos

SEGURIDAD
1 conjuntos

VIVIENDA
1 conjuntos

http://gobiernoabierto.valencia.es/es/

Government – Ajuntament de València



Google Transit. Líneas, paradas, horarios de autobuses de la EMT de Valencia

Fichero Google Transit de las líneas de autobuses de la EMT de...

Ayuntamiento de Valencia

ZIP

Grado de apertura	Visitas	Valoración
★★★☆☆	4108	👍 56 👎 60

Carril Bici

Itinerarios ciclistas de la ciudad de Valencia

Ayuntamiento de Valencia

WMS KML KMZ GML SHAPE CSV GEOJSON WFS

Grado de apertura	Visitas	Valoración
★★★☆☆	5397	👍 62 👎 58

Reguladores que gestionan los grupos semafóricos

Datos geográficos sobre la localización de los reguladores de los...

Ayuntamiento de Valencia

SHAPE WFS WMS KMZ KML GML GEOJSON CSV

Grado de apertura	Visitas	Valoración
★★★☆☆	2691	👍 33 👎 43

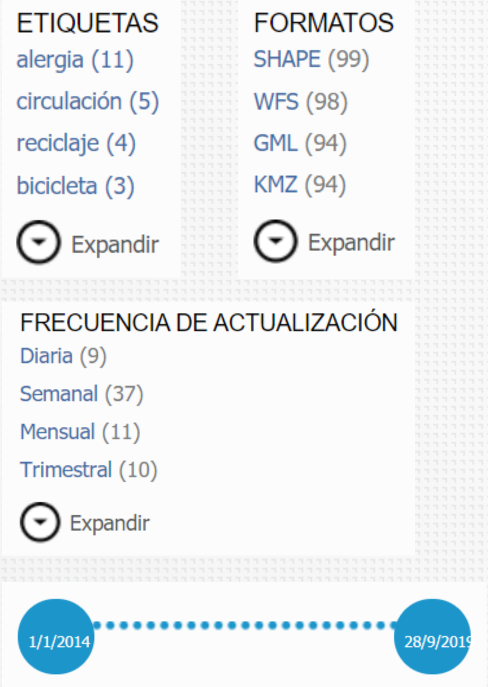
Paneles informativos

Paneles informativos de la sala de control de tráfico en la ciudad de...

Ayuntamiento de Valencia

WMS KML KMZ GML SHAPE CSV GEOJSON WFS

Grado de apertura	Visitas	Valoración
★★★☆☆	2729	👍 33 👎 49



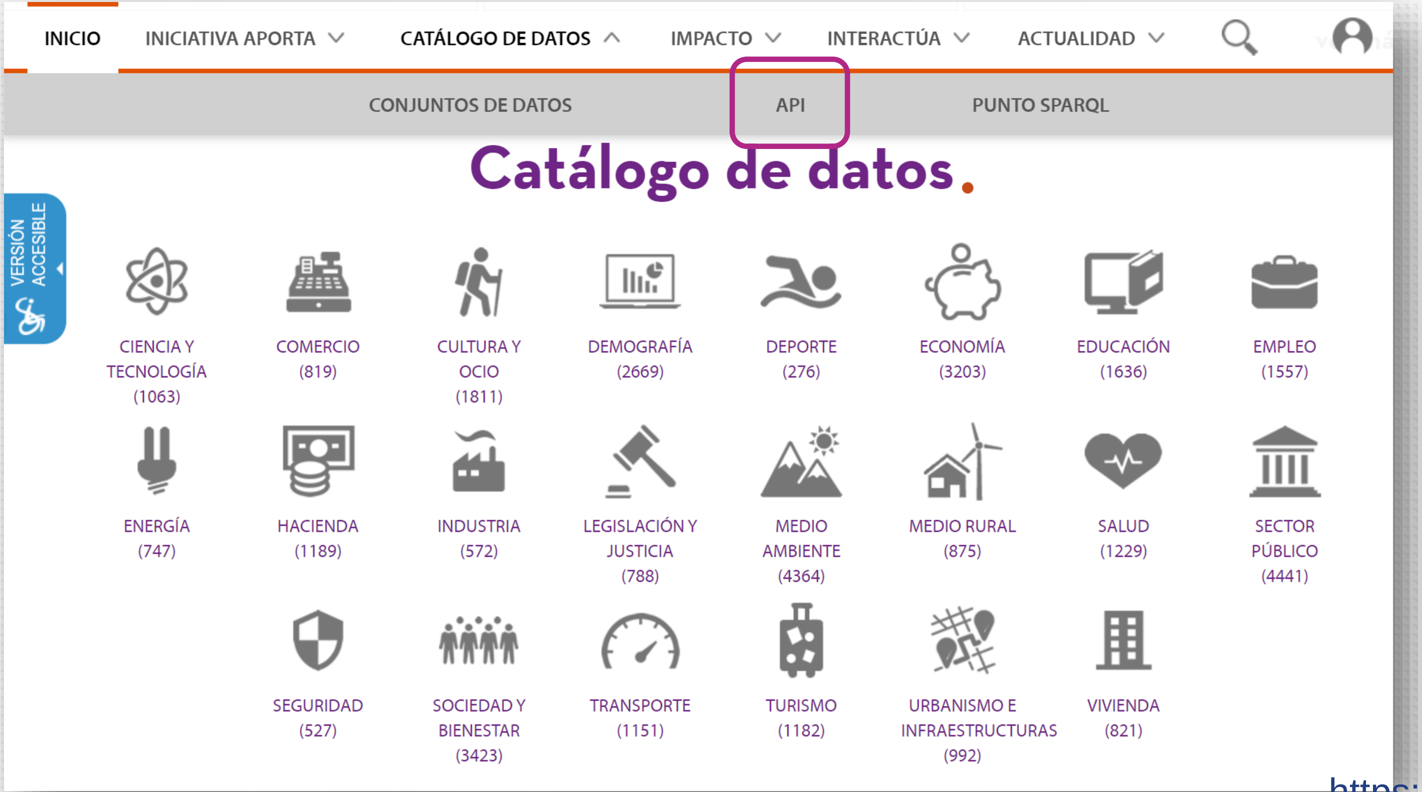
ETIQUETAS
- alergia (11)
- circulación (5)
- reciclaje (4)
- bicicleta (3)
- Expandir

FORMATOS
- SHAPE (99)
- WFS (98)
- GML (94)
- KMZ (94)
- Expandir

FRECUENCIA DE ACTUALIZACIÓN
- Diaria (9)
- Semanal (37)
- Mensual (11)
- Trimestral (10)
- Expandir

http://gobiernoabierto.valencia.es/es/

Government – Spanish Government

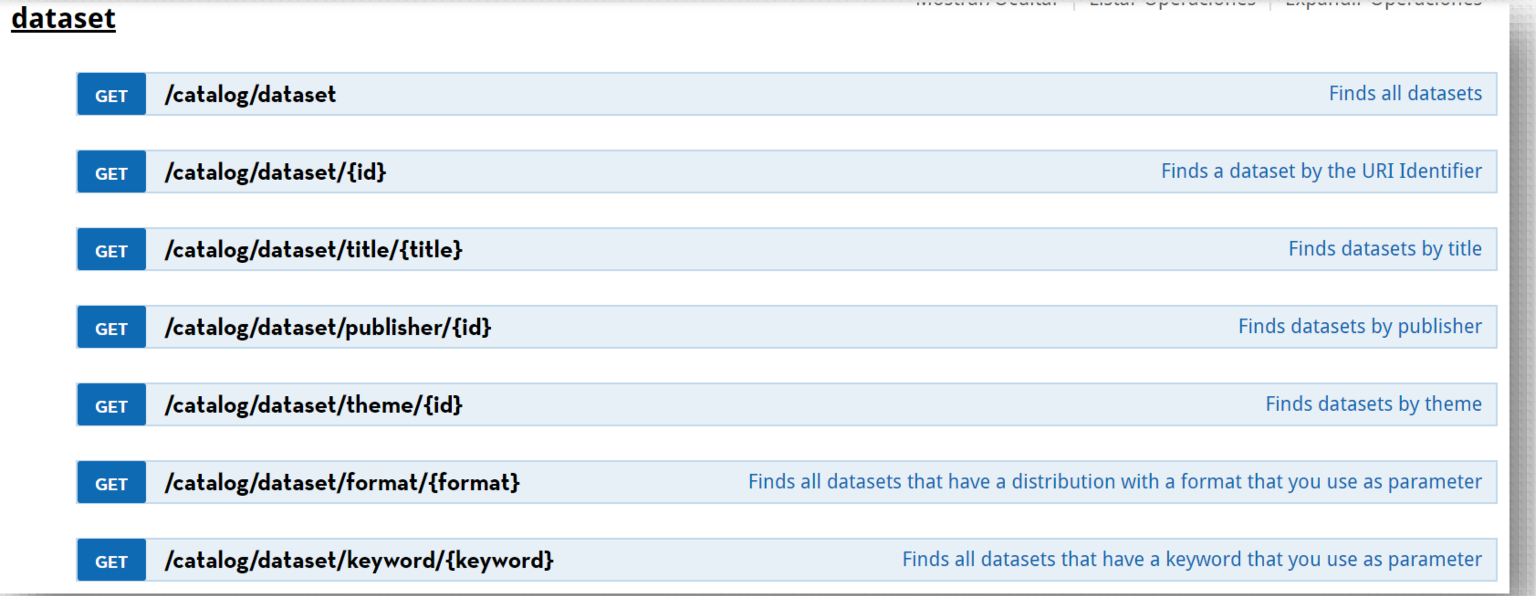


Catálogo de datos.

CIENCIA Y TECNOLOGÍA (1063)
COMERCIO (819)
CULTURA Y OCIO (1811)
DEMOGRAFÍA (2669)
DEPORTE (276)
ECONOMÍA (3203)
EDUCACIÓN (1636)
EMPLEO (1557)
ENERGÍA (747)
HACIENDA (1189)
INDUSTRIA (572)
LEGISLACIÓN Y JUSTICIA (788)
MEDIO AMBIENTE (4364)
MEDIO RURAL (875)
SALUD (1229)
SECTOR PÚBLICO (4441)
SEGURIDAD (527)
SOCIEDAD Y BIENESTAR (3423)
TRANSPORTE (1151)
TURISMO (1182)
URBANISMO E INFRAESTRUCTURAS (992)
VIVIENDA (821)

https://datos.gob.es

Government – Spanish Government



Mostrar/Ocultar | Listar Operaciones | Expandir Operaciones

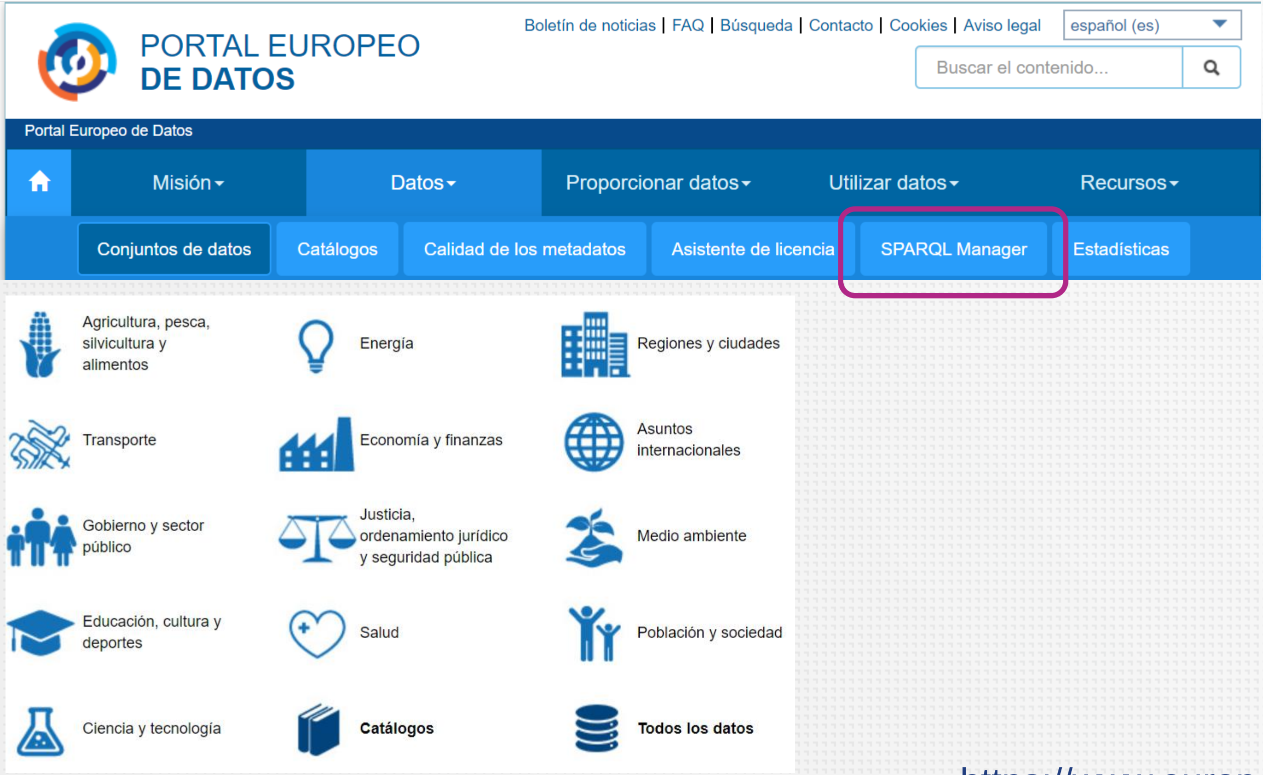
dataset

GET	/catalog/dataset	Finds all datasets
GET	/catalog/dataset/{id}	Finds a dataset by the URI Identifier
GET	/catalog/dataset/title/{title}	Finds datasets by title
GET	/catalog/dataset/publisher/{id}	Finds datasets by publisher
GET	/catalog/dataset/theme/{id}	Finds datasets by theme
GET	/catalog/dataset/format/{format}	Finds all datasets that have a distribution with a format that you use as parameter
GET	/catalog/dataset/keyword/{keyword}	Finds all datasets that have a keyword that you use as parameter

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

https://datos.gob.es

Government – European Union



Boletín de noticias | FAQ | Búsqueda | Contacto | Cookies | Aviso legal

español (es)

Buscar el contenido...

PORTAL EUROPEO
DE DATOS

Portal Europeo de Datos

Misión ▾ | Datos ▾ | Proporcionar datos ▾ | Utilizar datos ▾ | Recursos ▾

Conjuntos de datos | Catálogos | Calidad de los metadatos | Asistente de licencia | SPARQL Manager | Estadísticas

Agricultura, pesca, silvicultura y alimentos

Energía

Regiones y ciudades

Transporte

Economía y finanzas

Asuntos internacionales

Gobierno y sector público

Justicia, ordenamiento jurídico y seguridad pública

Medio ambiente

Educación, cultura y deportes

Salud

Población y sociedad

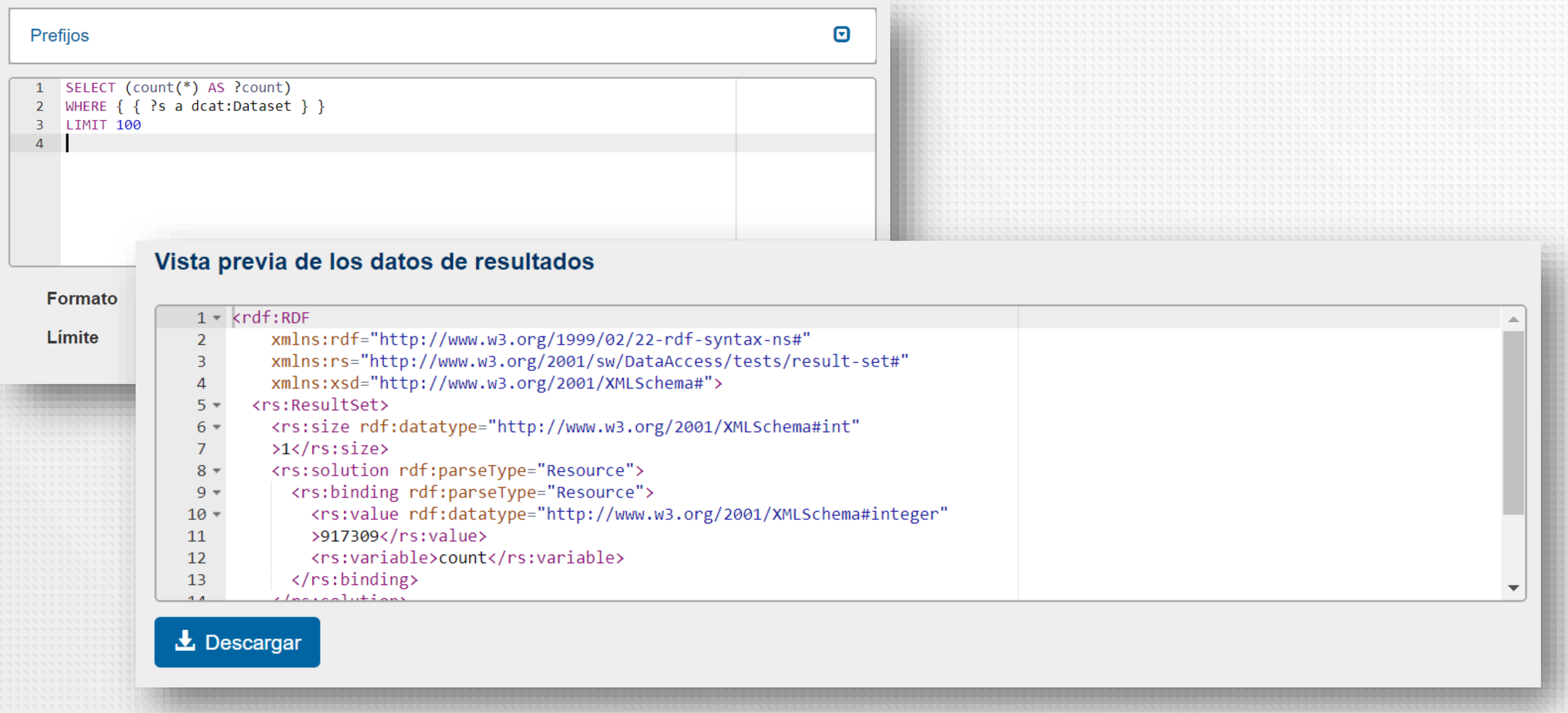
Ciencia y tecnología

Catálogos

Todos los datos

https://www.europeandataportal.eu/es/

Government – European Union

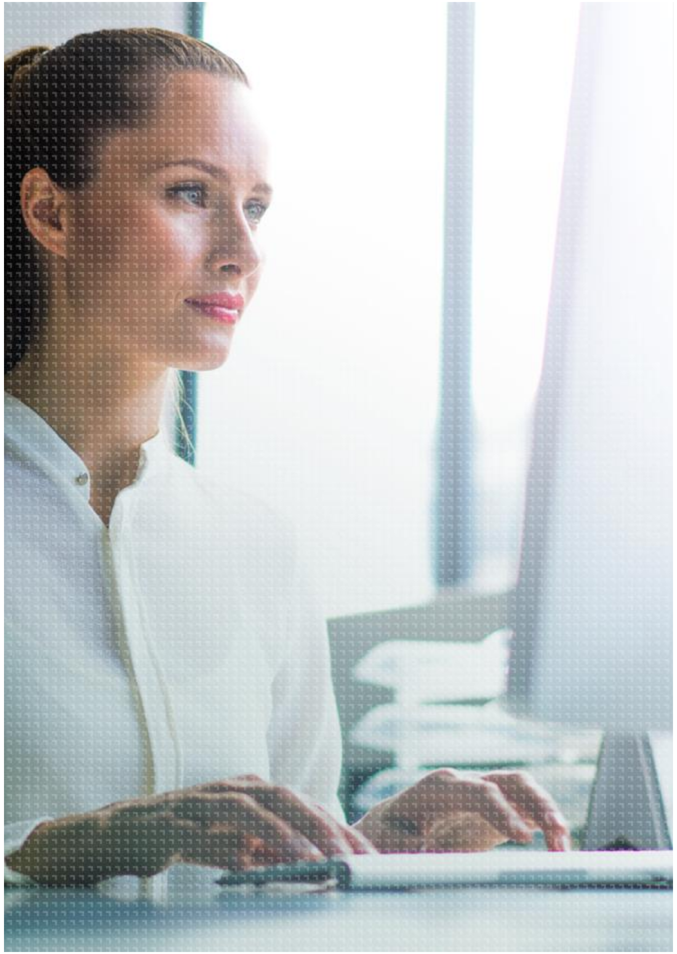


```
SELECT (count(*) AS ?count)
WHERE { { ?s a dcat:Dataset } }
LIMIT 100
```

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>
    </rs:solution>
```

https://www.europeandataportal.eu/es/



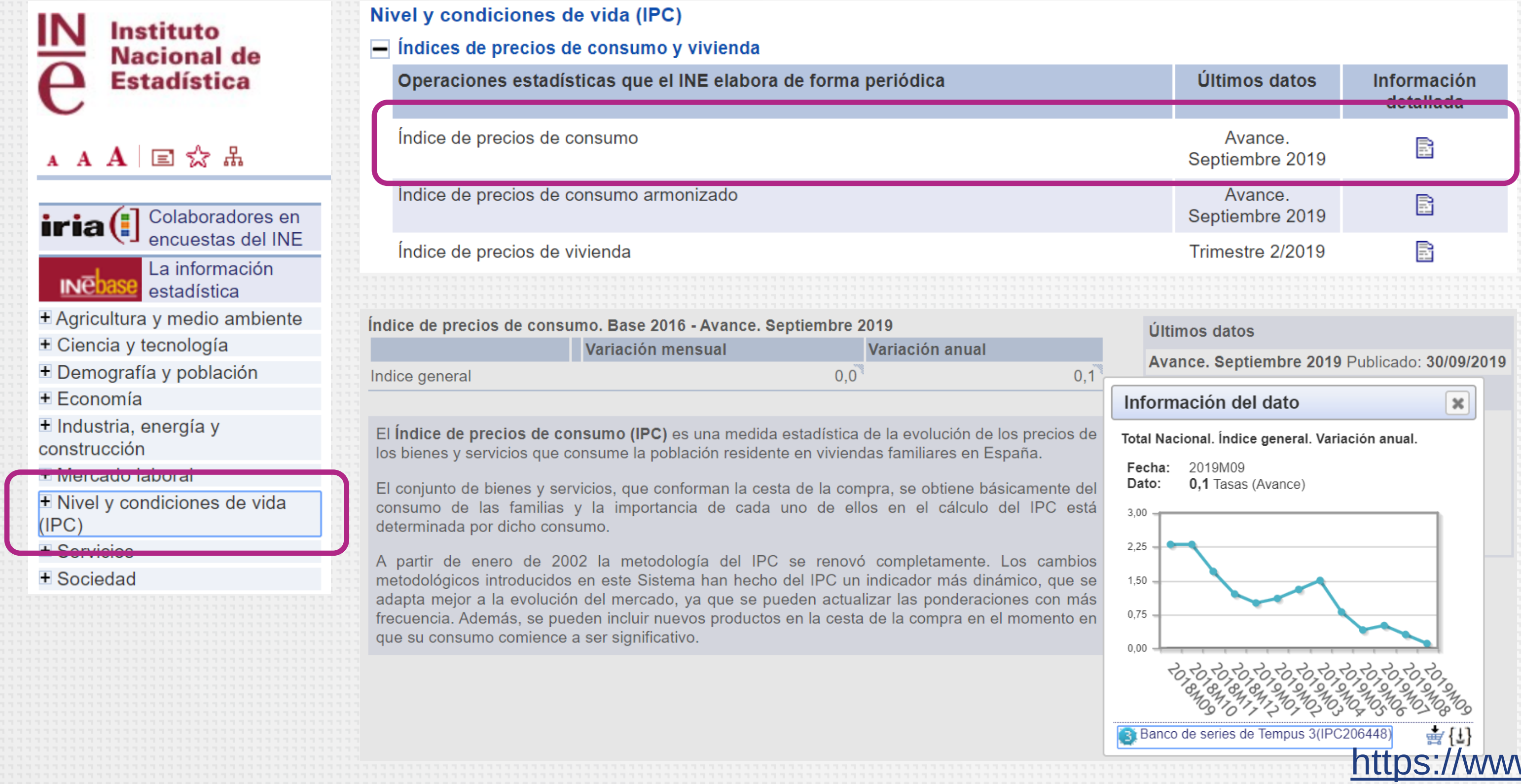
Analysing a dataset

- 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)



Instituto Nacional de Estadística

Colaboradores en encuestas del INE

La información estadística

- + Agricultura y medio ambiente
- + Ciencia y tecnología
- + Demografía y población
- + Economía
- + Industria, energía y construcción
- + Mercado laboral
- + Nivel y condiciones de vida (IPC)
- + Servicios
- + Sociedad

Nivel y condiciones de vida (IPC)

Índices de precios de consumo y vivienda

Operaciones estadísticas que el INE elabora de forma periódica	Últimos datos	Información detallada
Índice de precios de consumo	Avance. Septiembre 2019	🗎
Índice de precios de consumo armonizado	Avance. Septiembre 2019	🗎
Índice de precios de vivienda	Trimestre 2/2019	🗎

Índice de precios de consumo. Base 2016 - Avance. Septiembre 2019

	Variación mensual	Variación anual
Indice general	0,0	0,1

El **Índice de precios de consumo (IPC)** es una medida estadística de la evolución de los precios de los bienes y servicios que consume la población residente en viviendas familiares en España.

El conjunto de bienes y servicios, que conforman la cesta de la compra, se obtiene básicamente del consumo de las familias y la importancia de cada uno de ellos en el cálculo del IPC está determinada por dicho consumo.

A partir de enero de 2002 la metodología del IPC se renovó completamente. Los cambios metodológicos introducidos en este Sistema han hecho del IPC un indicador más dinámico, que se adapta mejor a la evolución del mercado, ya que se pueden actualizar las ponderaciones con más frecuencia. Además, se pueden incluir nuevos productos en la cesta de la compra en el momento en que su consumo comience a ser significativo.

Últimos datos

Avance. Septiembre 2019 Publicado: **30/09/2019**

Información del dato

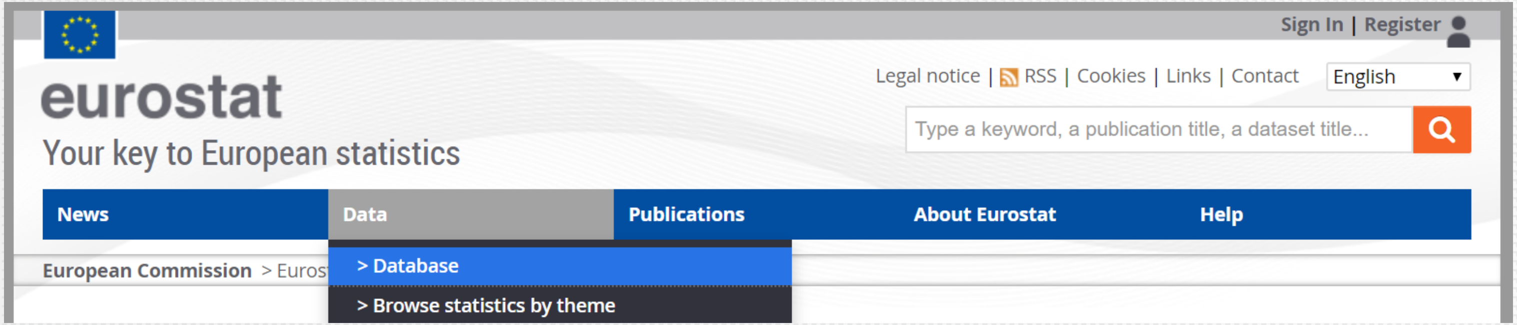
Total Nacional. Índice general. Variación anual.

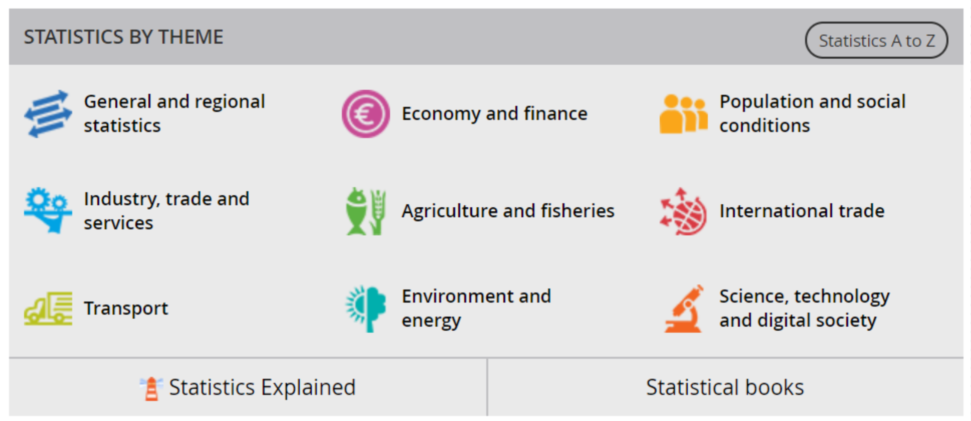
Fecha: 2019M09
Dato: **0,1** Tasas (Avance)

Banco de series de Tempus 3(IPC206448)

https://www.ine.es/

Statistics – Eurostat


Sign In | Register
eurostat
Your key to European statistics
Legal notice | RSS | Cookies | Links | Contact
English
Type a keyword, a publication title, a dataset title...
News
Data
Publications
About Eurostat
Help
> Database
> Browse statistics by theme
European Commission > Euros


STATISTICS BY THEME
Statistics A to Z
General and regional statistics
Economy and finance
Population and social conditions
Industry, trade and services
Agriculture and fisheries
International trade
Transport
Environment and energy
Science, technology and digital society
Statistics Explained
Statistical books

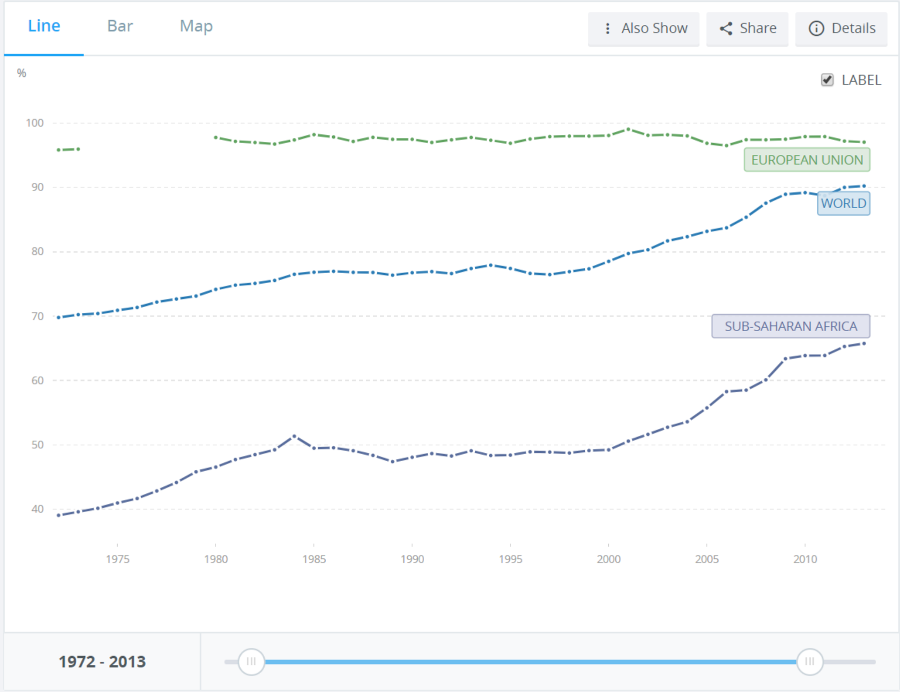
https://ec.europa.eu/eurostat

Statistics – World Bank

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

AREAS:

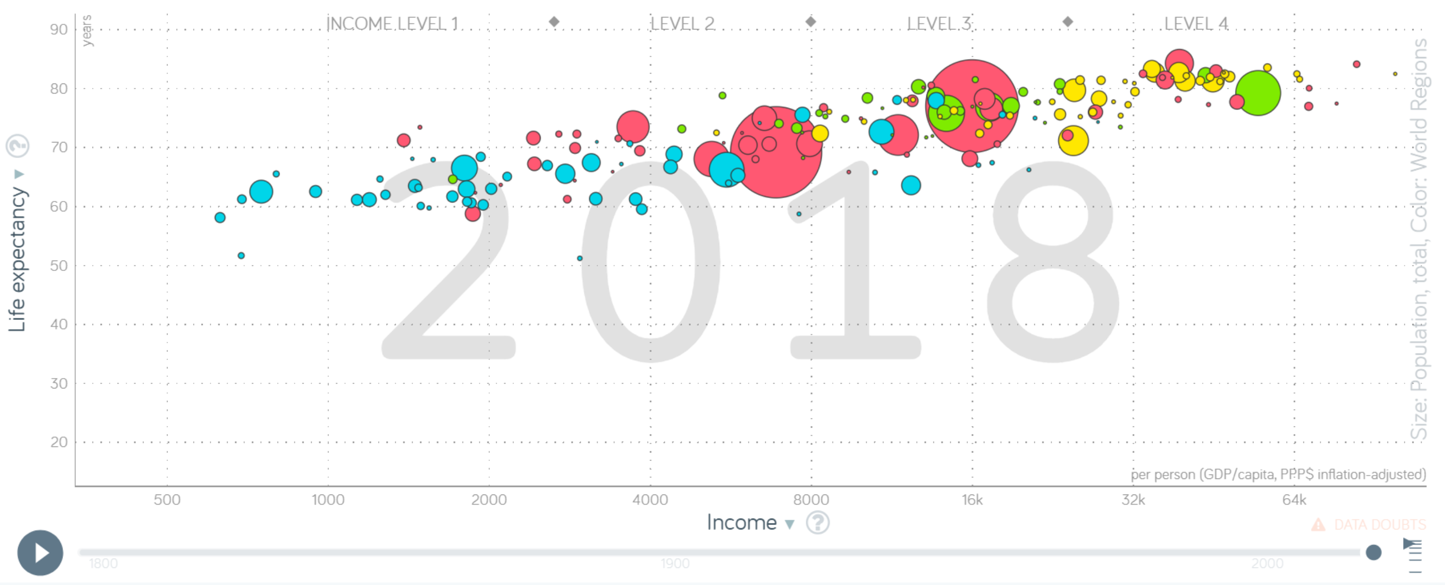
- Agriculture & Rural Development
- Aid Effectiveness
- Climate Change
- Economy & Growth
- Education
- Energy & Mining
- Environment
- External Debt
- Financial Sector
- Gender
- Health
- Infrastructure
- Poverty
- Private Sector
- Public Sector
- Science & Technology
- Social Development
- Social Protection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



https://data.worldbank.org/

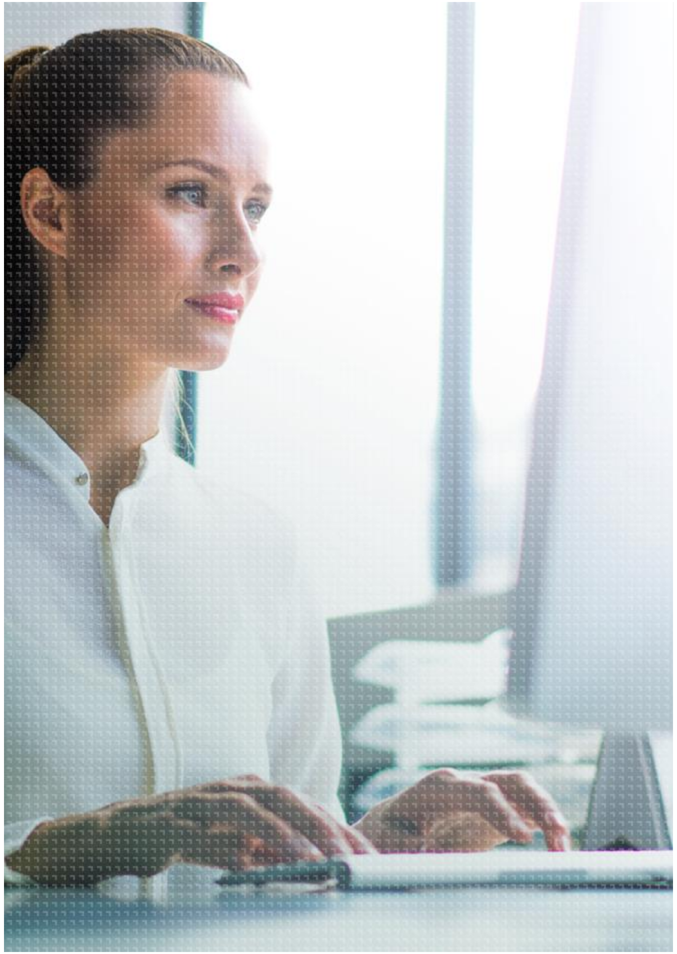
Statistics – Gapminder

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



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

https://www.gapminder.org/tools



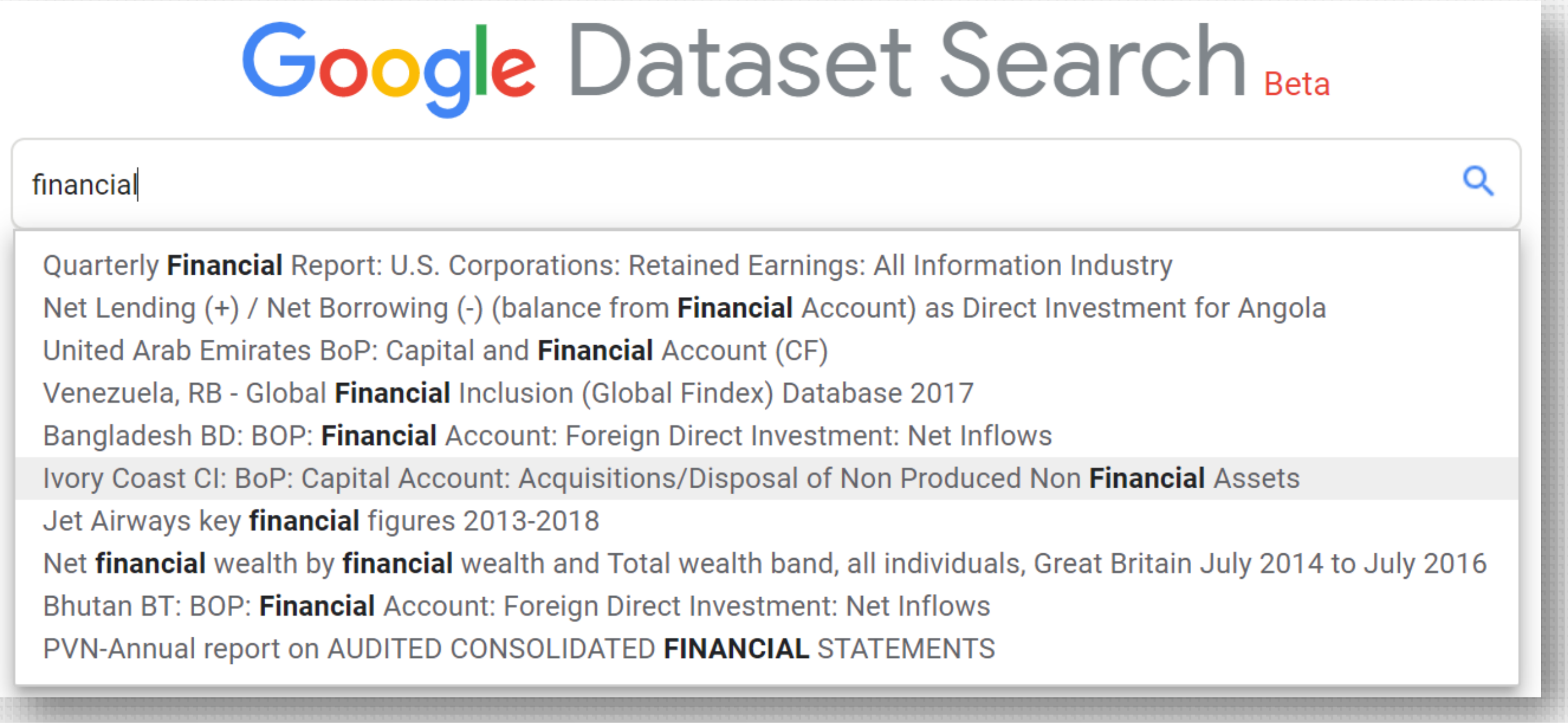
Socio-economic analysis

- 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.
- In **Gapminder**:
  - Child Mortality vs Income → What is the evolution over time
- (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



financial

Quarterly **Financial** Report: U.S. Corporations: Retained Earnings: All Information Industry

Net Lending (+) / Net Borrowing (-) (balance from **Financial** Account) as Direct Investment for Angola

United Arab Emirates BoP: Capital and **Financial** Account (CF)

Venezuela, RB - Global **Financial** Inclusion (Global Findex) Database 2017

Bangladesh BD: BOP: **Financial** Account: Foreign Direct Investment: Net Inflows

Ivory Coast CI: BoP: Capital Account: Acquisitions/Disposal of Non Produced Non **Financial** Assets

Jet Airways key **financial** figures 2013-2018

Net **financial** wealth by **financial** wealth and Total wealth band, all individuals, Great Britain July 2014 to July 2016

Bhutan BT: BOP: **Financial** Account: Foreign Direct Investment: Net Inflows

PVN-Annual report on AUDITED CONSOLIDATED **FINANCIAL** STATEMENTS

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

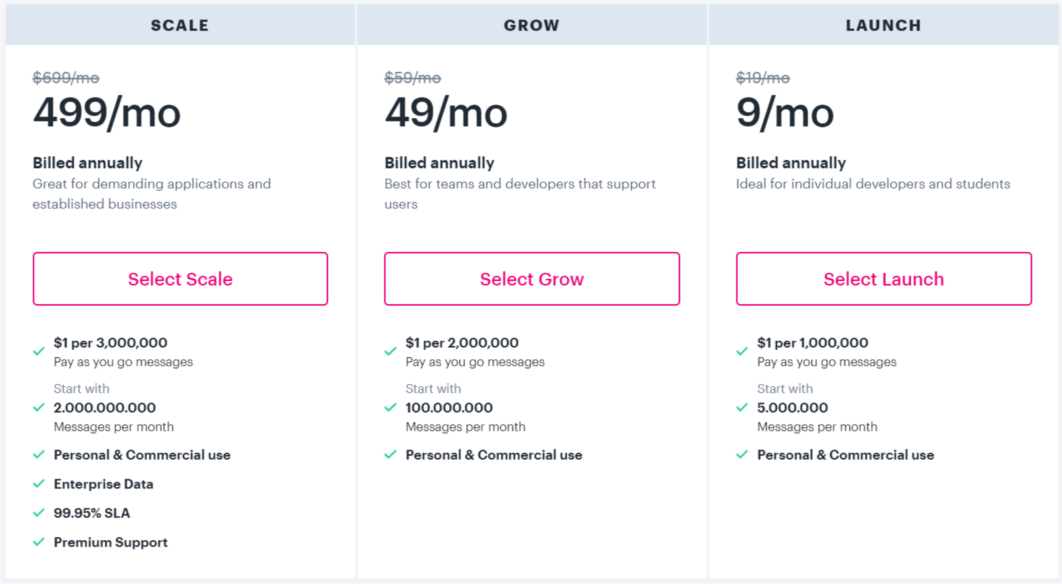
Examples: GCP Big Query datasets

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 Great for demanding applications and established businesses	**Billed annually** Best for teams and developers that support users	**Billed annually** 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:

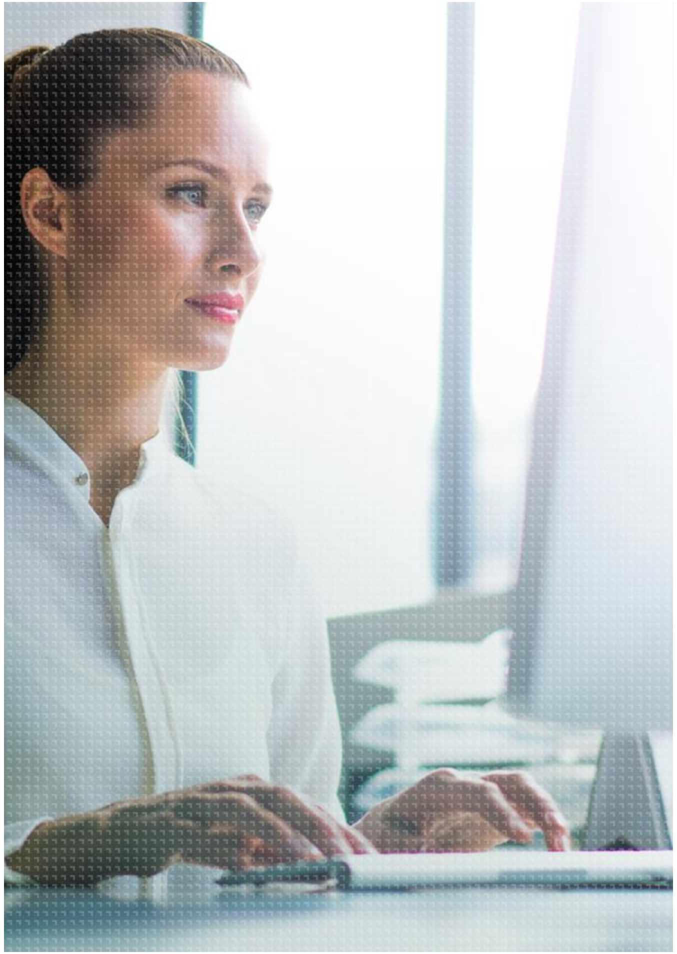
Account & Users:

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

Tweets:

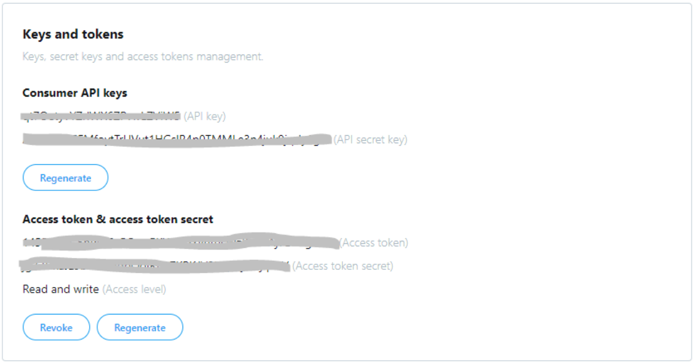
- Get user timeline
- Post tweets
- **Real-time tweets (sample)**
- Search

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:



Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

[redacted] (API key)

[redacted] (API secret key)

Access token & access token secret

[redacted] (Access token)

[redacted] (Access token secret)

Read and write (Access level)

Shaping the future of digital business

Esteban Chiner Sanz

Senior Architect

GFT – Data Practice

esteban.chiner@gft.com