Universidad de Lleida; Massive Data Processing

OPEN DATA

February 19, 2023

bluetab

an IBM Company

OPEN DATA

Definition

- We refer to a philosophy and practice that proposes that data should be freely accessible to the public, able to be used without the restrictions imposed by copyright, patents, etc.
- This movement is aligned with other 'open' movements
  - Free Open Source Software
  - Open Hardware
  - Open Science/ Open Access
  - Open Content / Creative Commons
- We will consider data to be open when anyone is free to use them, reuse them, redistribute them, subject only, at most, to the requirement of attributing and/or sharing alike

OPEN DATA

Motivation

- In the age of instant, ubiquitous, universal, and nearly free access to all knowledge/content produced by humanity, **what do you think about laws that focus on preventing the dissemination of this type of knowledge (data)?**

OPEN DATA

Motivation

- In the age of instant, ubiquitous, universal, and nearly free access to all knowledge/content produced by humanity, **what do you think about laws that focus on preventing the dissemination of this type of knowledge (data)?**
- To counteract this movement towards compartmentalization, privatization, and data lockdown that is being driven by the current legal situation and future trends, a set of 'open' initiatives has emerged.
- Some data are generated in processes funded with public money, which is why many believe they should be accessible to the public:
  - Public research data
  - Data generated/collected by the government (at all levels).
- Advocates of open data argue that providing public access to these data is, in fact, giving owners access to data they already have.

OPEN DATA

Motivation

- Regarding data funded with public funds, there are arguments against publication:
  - Charging for access to the data will help recover costs.
  - Public funding should not challenge private activities.
  - Non-profit organizations can obtain funding by charging for data access.
  - Privacy issues.
- Some argue that charging for access to data will be a more efficient use of taxpayer funds.

OPEN DATA

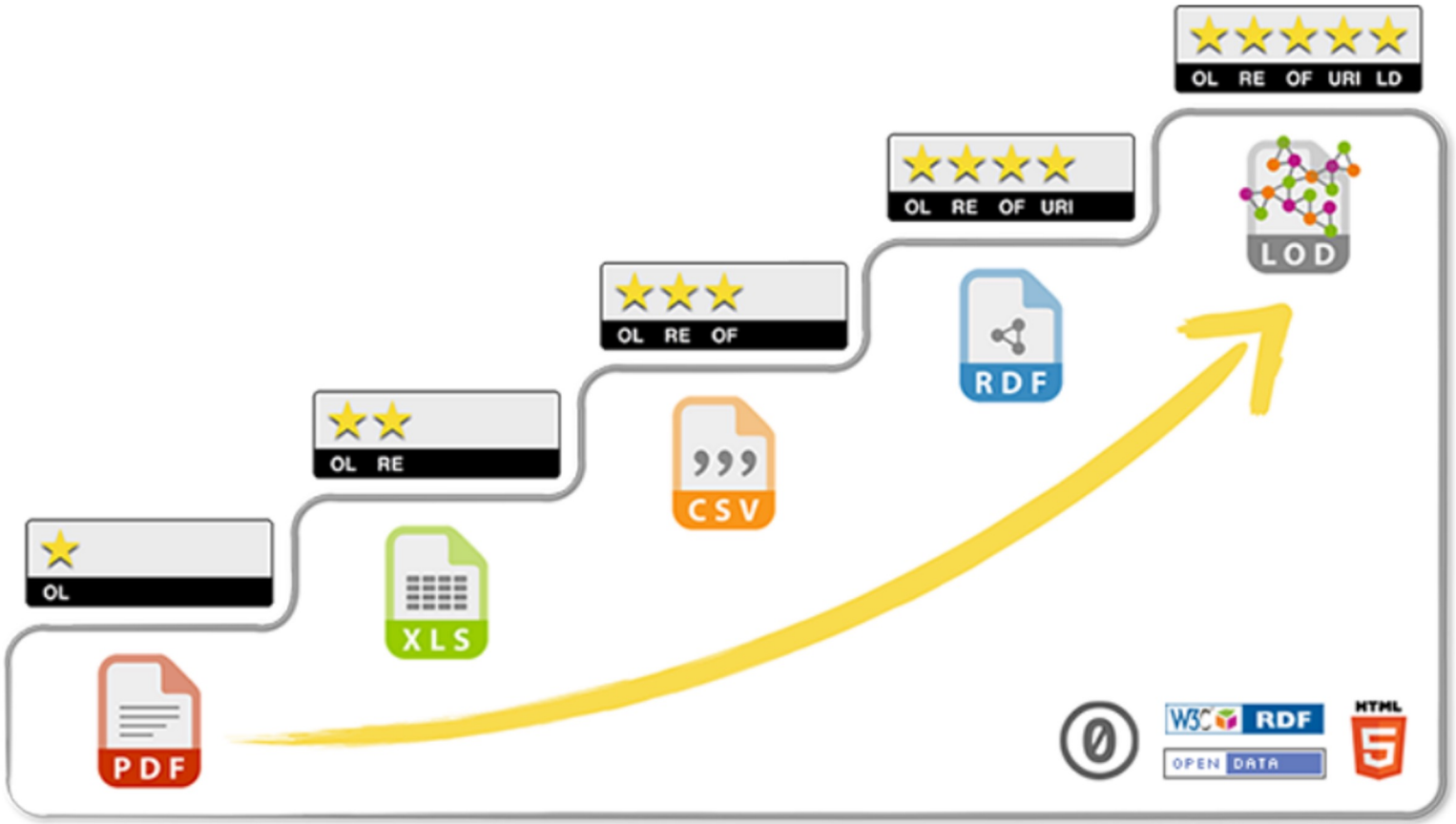
Legality

- In the United States, they follow common law practices.
  - Everything paid for with public funds should default to the public domain.
- The European Union adopted Directive 2003/98/EC on the reuse of public sector information (PSI Directive) to encourage EU member states to make public information available for reuse by the general public.
- In 2008, it became a mandatory implementation, and in 2013, it was amended to align with the principles of OpenData/OpenGovernment.

OPEN DATA

5 star data

- **Tim Berners-Lee** proposed a classification 'by stars' according to how open the data and the linkability."



OPEN DATA

5 star data

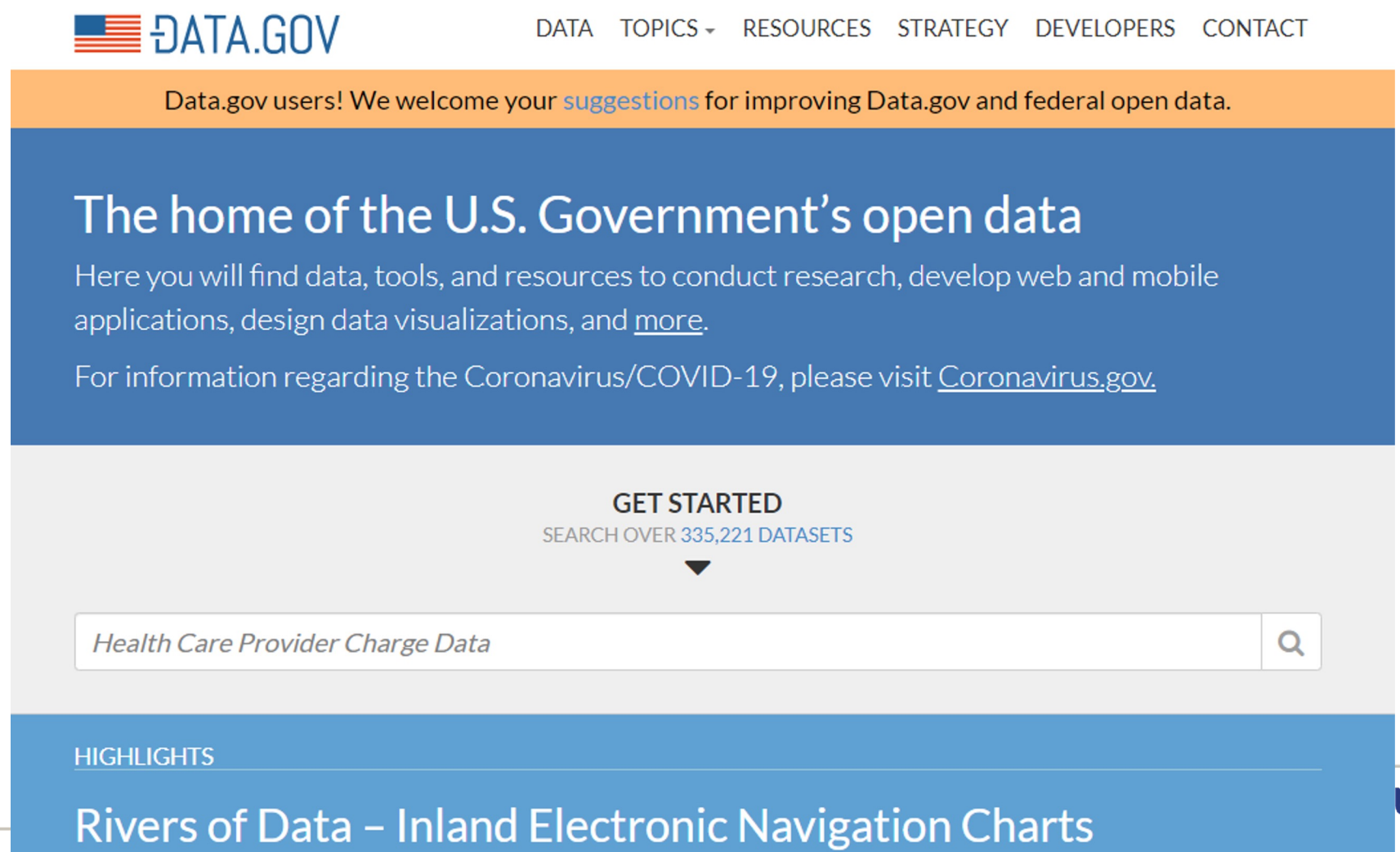
- **Tim Berners-Lee** proposed a classification 'by stars' according to how open the data and the linkability."

⭐ Data available on the web (in any format) under an open license.
⭐⭐ Structured data.
⭐⭐⭐ Data in a non-proprietary format.
⭐⭐⭐⭐ Use URIs to denote other things.
⭐⭐⭐⭐⭐ 'Link' our data to other data to provide context.

OPEN DATA -Examples

US Government Open data initiative:

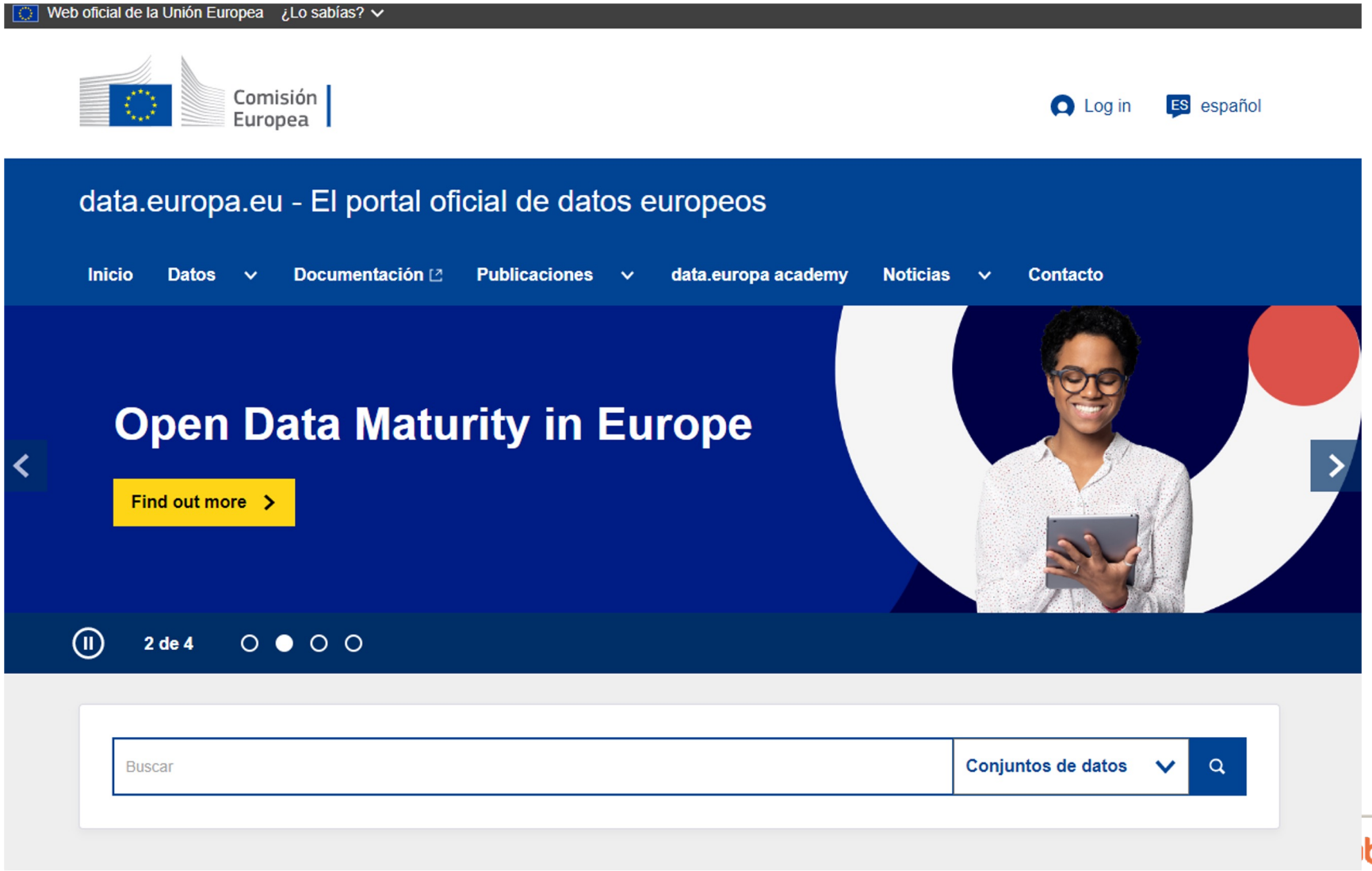
http://www.data.gov/



OPEN DATA -Examples

European Union Open Data Portal

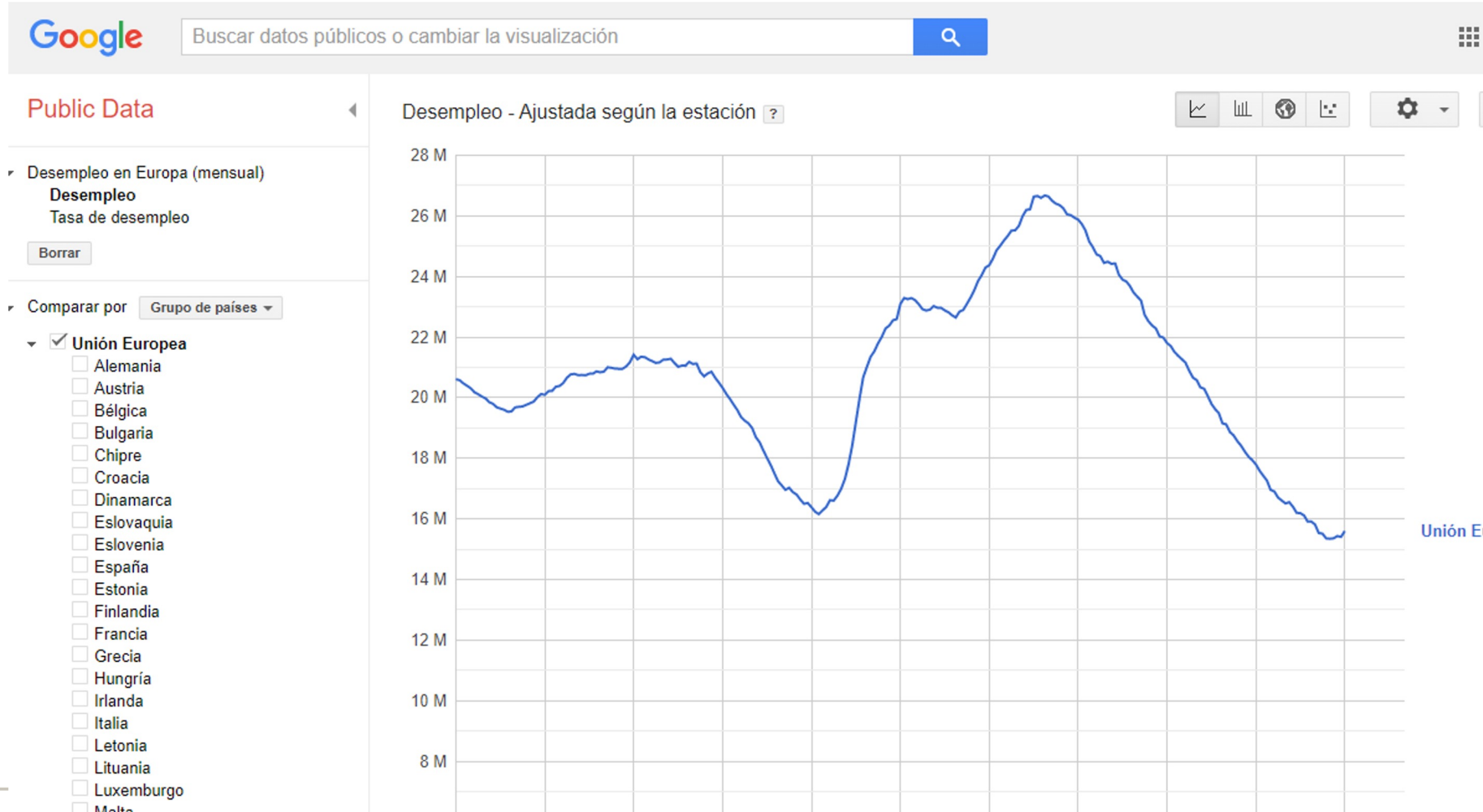
https://data.europa.eu/es



OPEN DATA -Examples

Google Public Data

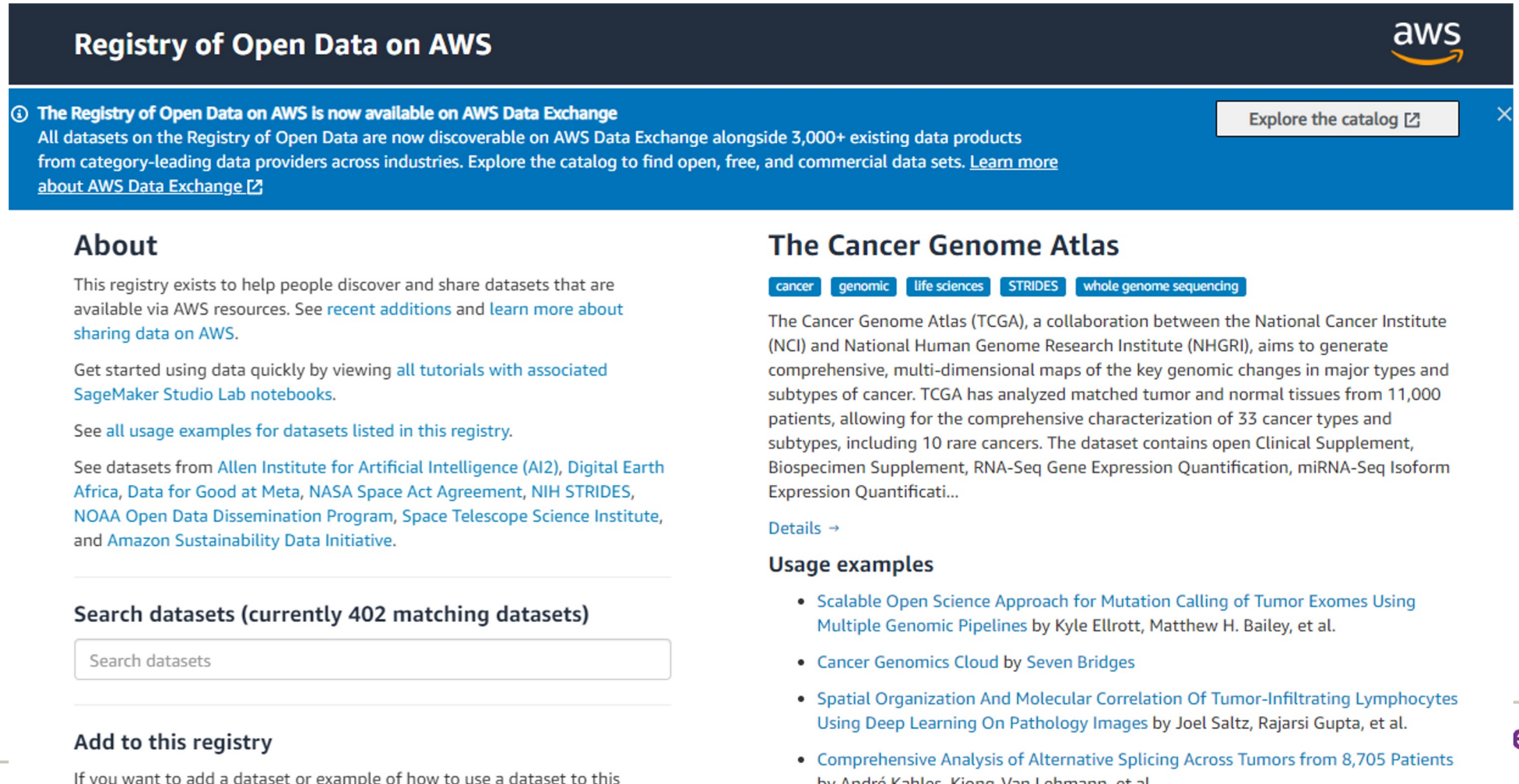
https://www.google.com/publicdata/directory



OPEN DATA -Examples

AWS Open Data

https://registry.opendata.aws/



Activity 1

Look for the following information

- Energy efficiency certificates for existing buildings in Lleida.
- List of accommodation facilities in the city of Barcelona.
- What is available openly at Lleida City Council (paeria.es) - three examples.
- Fuel consumption in Spain.
- License plate numbers in Spain.

Provide documentation with access links to where you found this data, assign how many stars according to TBL's classification, and justify.

Linked Data

Definition

- The linked data are those data published (on the web) in a structured format that
  - Facilitates linking between these data
  - Facilitates querying of these data through semantic queries.

Linked Data

Definition

- The linked data are those data published (on the web) in a structured format that
  - Facilitates linking between these data
  - Facilitates querying of these data through semantic queries.

- History of the web
  - Web 1.0: Documents from editors to people (unidirectional)
  - Web 2.0: Information from people to people
  - Web 3.0: Information for machines by machines for people

Linked Data

Definition

- The linked data are those data published (on the web) in a structured format that
  - Facilitates linking between these data
  - Facilitates querying of these data through semantic queries.

- History of the web
  - Web 1.0: Documents from editors to people (unidirectional)
  - Web 2.0: Information from people to people
  - Web 3.0: **Information for machines by machines for people**

Linked Data

Definition

- The linked data are those data published (on the web) in a structured format that
  - Facilitates linking between these data
  - Facilitates querying of these data through semantic queries.

- History of the web
  - Web 1.0: Documents from editors to people (unidirectional)
  - Web 2.0: Information from people to people
  - Web 3.0: **Information for machines by machines for people**

People is good at extracting information from web pages:

WE CAN READ!

Linked Data

Definition

- The linked data are those data published (on the web) in a structured format that
  - / Facilitates linking between these data
  - / Facilitates querying of these data through semantic queries.

- History of the web
  - / Web 1.0: Documents from
  - / Web 2.0: Inform
  - / We

We need machines to be able to read a page and extract information from it: CORRECTLY.

ting information from web pages:

WE CAN READ!

Linked Data

Ambiguity

- Search on Google:
  - MUSK
  - ATOM
  - RASPBERRY

Linked Data

Ambiguity - Where is the problem?

- In HTML (a markup language) we have two sets of information:

```
<H1>Elon Musk</H1>
<P>Elon Musk (June 28, 1971, - ), in Twitter:
<A HREF="https://twitter.com/elonmusk">@elonmusk</A>,
is a very successuful <I>entrepeneur</I> know for ....
```

We have the 'content': 'June 20, 1971[...]'

We have the 'markup' or 'tag': H1, A....

Linked Data

Ambiguity - What can we understand?

- From the content, we can 'extract' information: but in natural language (NLP is still a hot research topic, despite advances), so we have to deal with: ambiguity, errors, spelling mistakes, lack of context, etc.
- From the markup, we can also 'extract' information: importance (H1), relationships (paragraph format), other sources (A), **but we need more information**: context, etc.
- To solve the 'problem of lost information,' some proposals emerged, **and the semantic web** is one of them.
  - The semantic web proposes common data formats and protocols (especially **RDF**).
  - Definition according to W3C:
    - The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Linked Data

According to Tim Berners-Lee, it consists of:

- Using URIs to name (identify) things.
- Using HTTP URIs so that these things can be looked up (interpreted).
- Providing useful information about what a name identifies when looked up, using open standards like RDF, SPARQL, etc.
- Querying other things using their HTTP URI-based names when data are published on the web.

Linked Data

RDF

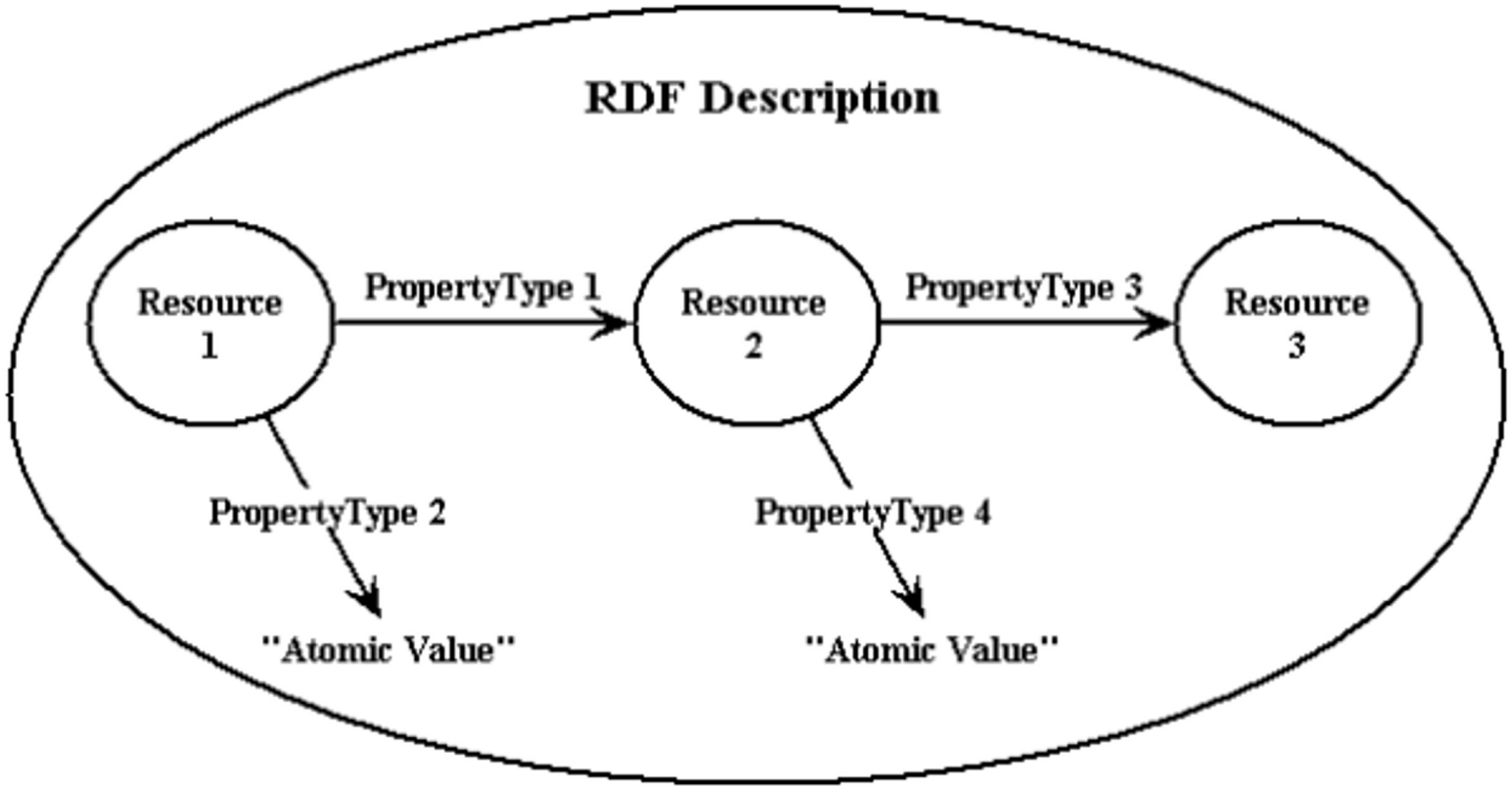
- You can use HTML, XML, etc., to publish data, but they require prior knowledge of the **metadata** of this data to be useful.
- **RDF**(Resource Description Framework) is a data model representation based on graphs that allows us to offer a standard representation of the data model adapted for the web (and Linked Data).
- RDF describes data **through a graph**, where nodes represent data (subject/object) and edges represent relationships.
- Thus, RDF represents data assertions as triplets, consisting of

```
subject - relationship (verb) - object
```

- On each of them is given as a URL/URI.

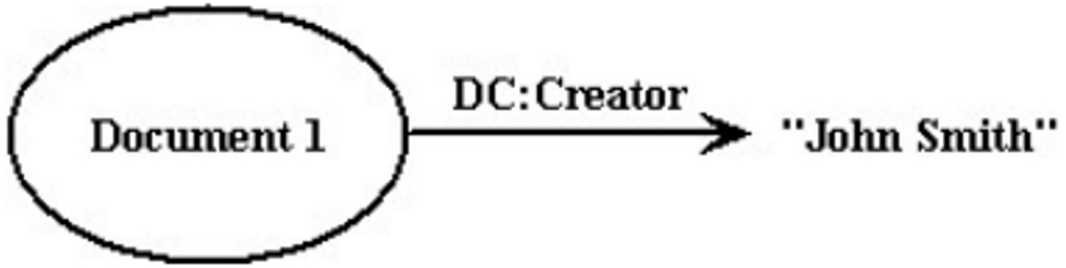
Linked Data

RDF - Model



Linked Data

RDF - Example




```
<?xml version="1.0"?>
    <rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.oclc.org/DC/">
        <rdf:Description RDF:HREF = "http://uri-of-Document-1">
        <dc:Creator>John Smith</dc:Creator>
        </rdf:Description>
    </rdf:RDF>
</xml>
```

Linked Data

According to Tim Berners-Lee, it consists of:

- Using URIs to name (identify) things.
- Using HTTP URIs so that these things can be looked up (interpreted).
- Providing useful information about what a name identifies when looked up, using open standards like RDF, SPARQL, etc.
- Querying other things using their HTTP URI-based names when data are published on the web.

Later he added a 5th component: **Open Content.**

And with that, we have **Linked Open Data**: data that are accessible, automatically processable, and open!

Bibliography

[1] Mateu, Carles. “Massive data processing”, Universitat de Lleida. Lleida. 2021.

QUESTIONS

/bluetab
an IBM Company

¡THANK YOU!

alba.lamas@bluetab.net

¡Síguenos!

https://bluetab.net/

https://www.linkedin.com/company/bluetab/