



Open Data and New Frontiers in Research Publication

Data papers, Code repositories, Open Access and Reproducible Research

Cristian Consonni

Eurecat - Eurecat Centre Tecnològic de Catalunya

Barcelona, 04/02/2020

0. Let's Know Each Other

Informal Poll (I)

- How many of you are:
 - **Scientists** (Professor, Post-docs, PhD students, ...)
 - **Academic Staff** (Administrators, Research supports, ...)
 - **Librarians** (Academic or not)
 - **Public servants** (Local government, Public institutions, ...)
 - **Software Developers**

- How many of you have heard about:
 - **Creative Commons licenses**
 - **GPL**
 - **ODbL**

Informal Poll (II)

- How many of you have hear about:
 - CSV
 - JSON
 - XML
 - RDF

- How many of you have hear about:
 - **Python, Jupyter**
 - **Version Control (Systems)**
 - **GitHub, GitLab**

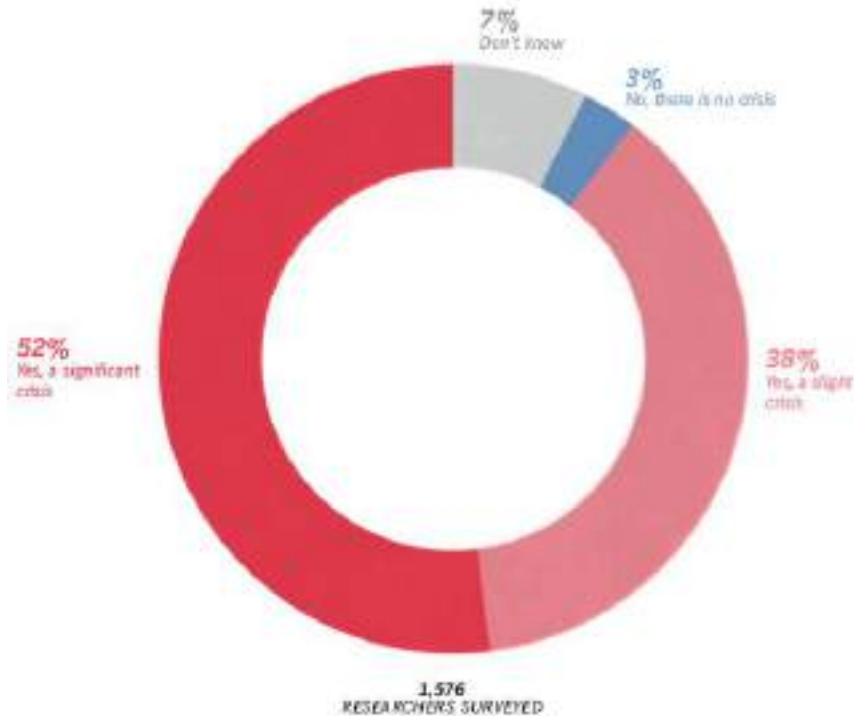
Informal Poll (III)

- How many of you have hear about **Open Access**?
- How many of you have hear about **the reproducibility crisis in science**?

1. Motivation

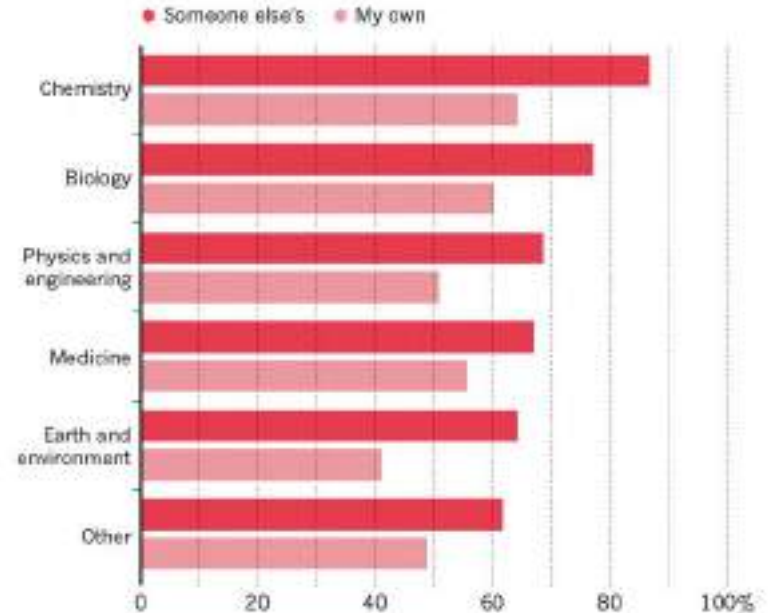
Is Science in crisis?

Reproducibility and replicability



HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

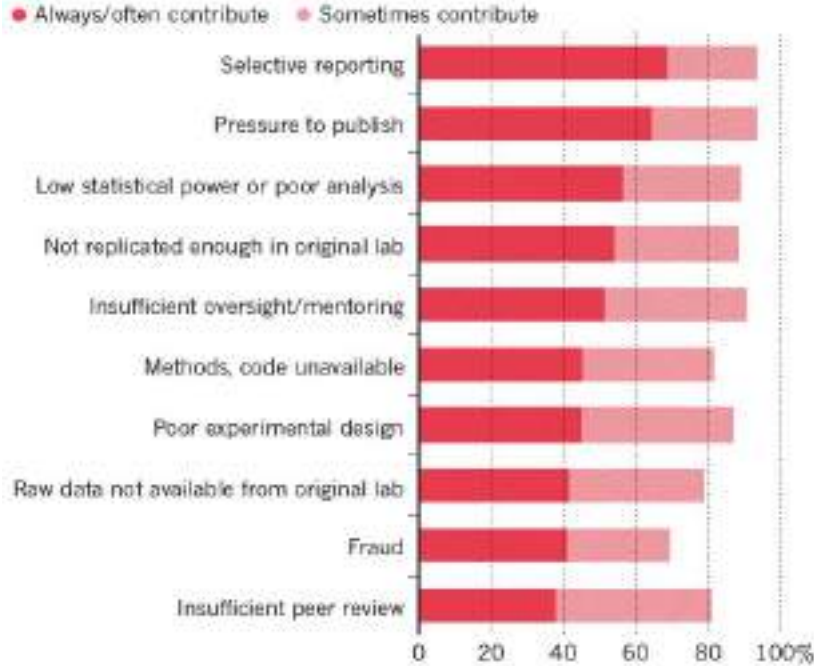
Most scientists have experienced failure to reproduce results.



Is Science in crisis? (II)

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



Additional material:

- Why most Published Research Findings Are False?
<https://journals.plos.org/plosmedicine/article%3Fid%3D10.1371/journal.pmed.0020124>
- The reproducibility crisis in science: A statistical counterattack
<https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2015.00827.x>

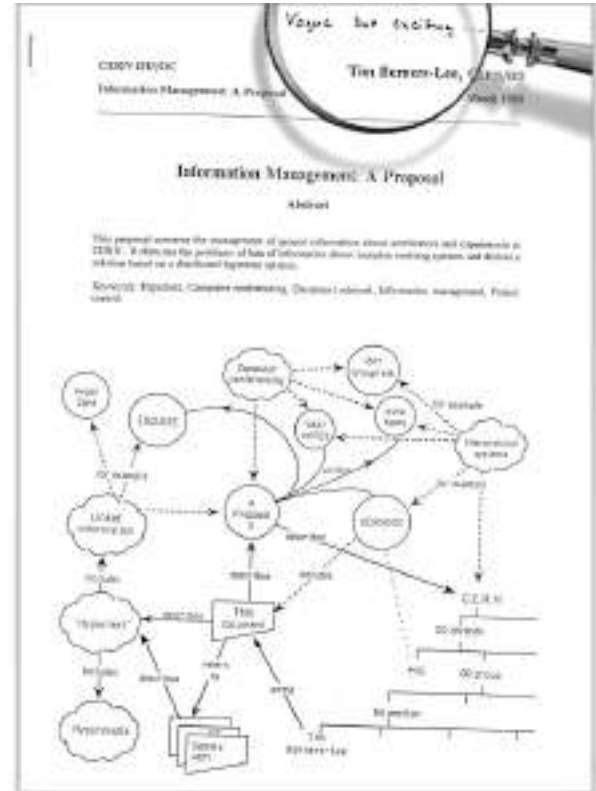
2. Introduction

A Universal Linked Information System

(“Weaving the Web” by T. Berners-Lee, 1999)

«We should work toward a universal linked information system [...] The aim [of which] would be to allow a place to be found for any information or reference which one felt was important, and a way of finding it afterwards.»

[“Information Management: A Proposal”](#), 1989



everything is data (or metadata)



"On the Internet, nobody knows you're a dog."

"The Big New Yorker Book of Dogs." (New York: The New Yorker Magazine, 2012), p. 135.

«Raw data now!»

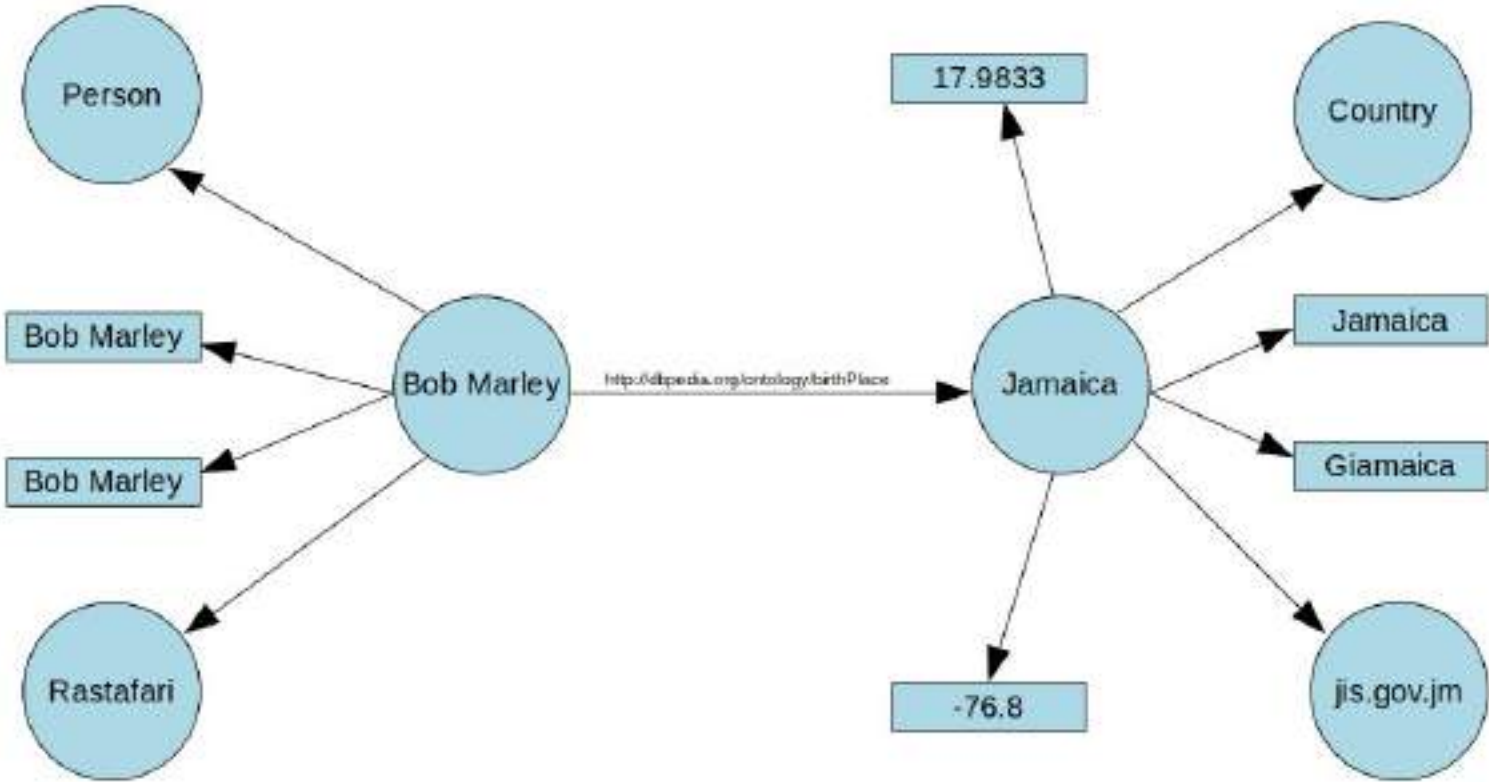
The Next Web - TED 2009



«[The next web](#)» by Tim Berners-Lee on [ted.com](#)

Sir Tim Berners-Lee by Paul Clarke via [Wikimedia Commons](#) [CC BY-SA 4.0]

Linked Data



3. What is Open Data?

«Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.»

Defining Open Data, Open Data Handbook

Open Data BCN

Servicio de datos abiertos del Ajuntament de Barcelona



SOBRE ESTE SITIO

CATÁLOGO DE DATASETS

ACTUALIDAD

VISUALIZACIONES Y APLICACIONES

ESTADÍSTICAS

DESARROLLADORES

RETOS BARCELONA DATOS ABIERTOS

WORLDWIDE CHALLENGE BARCELONA 2016

p.ej. medio ambiente

Hay un total de **460** datasets en el Catálogo de Open Data BCN



Territorio



Población



Ciudad y servicios



Administración



Economía y Empresa

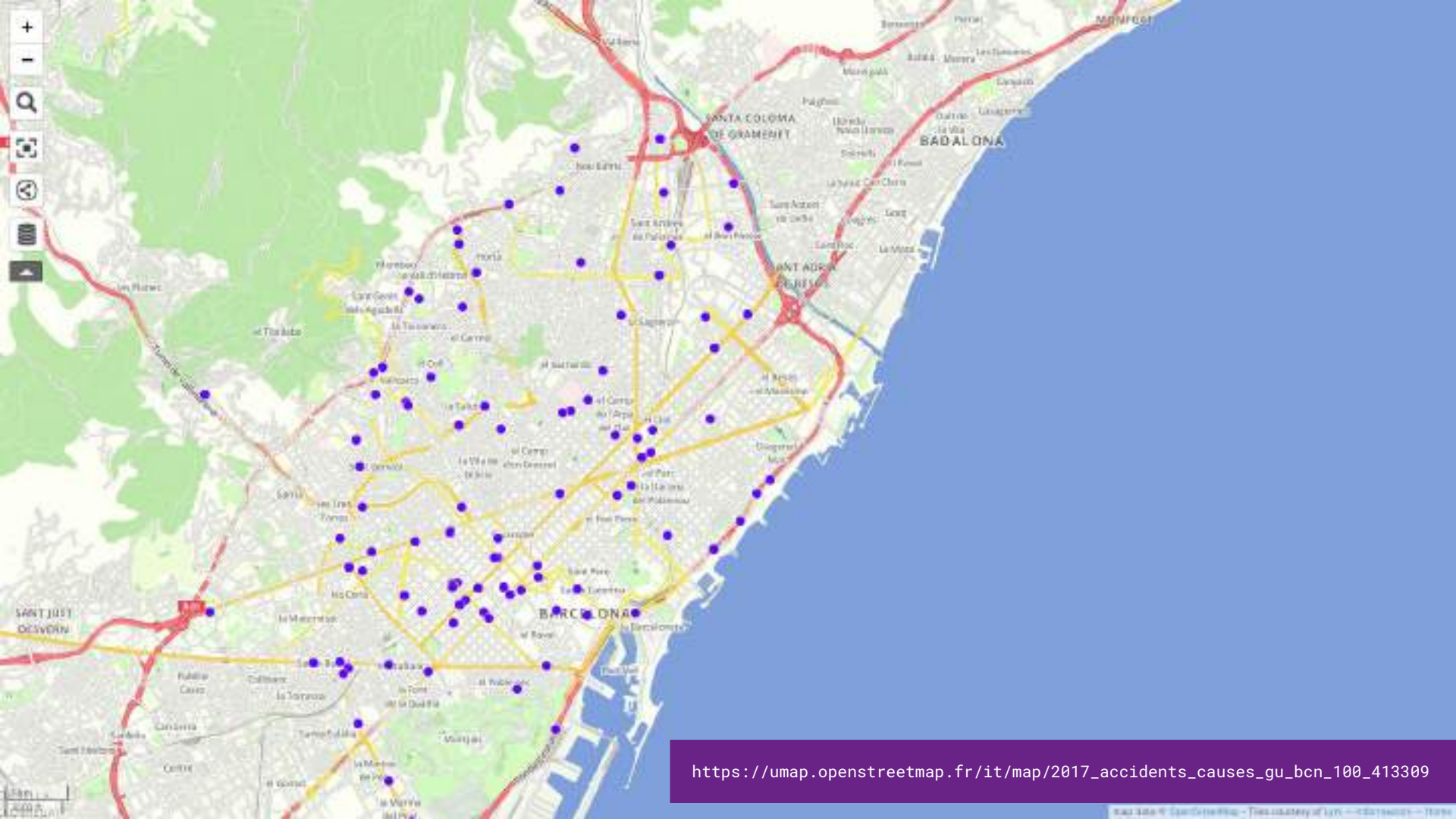
<https://opendata-ajuntament.barcelona.cat/>

Example

Listado de de la causalidad de los accidentes gestionados por la Guardia Urbana en la ciudad de Barcelona.

File: BCN/2017_accidents_causes_gu_bcn_.csv

| | Número d'expedient | Codi districte | Nom districte | Codi barri | Nom barri | Codi carrer | Nom carrer | Num postal | Descripció dia setmana | Dia setmana | Descripció tipus dia,NK | Any | Mes de any | Nom mes | Dia de mes | Mora de dia | Descripció torn | Descripció causa mediate | Coordenada UTM (X) | Coordenada UTM (Y) | Longitud | Let itud | | | | | | |
|----|--------------------|----------------|---------------|------------|----------------------------|-------------|-----------------------------|------------|------------------------|-------------|-------------------------|------|------------|---------|------------|-------------|------------------------|--------------------------|----------------------|--------------------|----------|----------|------|---------|-----------|---------|---------|------|
| 1 | 20175000198 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 346482 | Trinxant / Ruiz de Padrón | | | | | 2017 | Novembre | 7 | 14 | Tarda | No hi ha causa mediate | 431940.75,4585401.48 | 2.184483,41.415436 | | | 8090 | 0890 | Dimarts | Dm | Laboral | 2017 | |
| 2 | 20175008278 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 05598 | Conca / Industria | | | | | 2017 | Octubre | 7 | 10 | | | No hi ha causa mediate | 431576.37,4585215.2 | 2.180149,41.41373 | | | 0828 | 0828 | Dissabte | Ds | Laboral | 2017 |
| 3 | 20175008462 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 00592 | Còrsega | | | | | 2017 | Octubre | 7 | 30 | | | No hi ha causa mediate | 431384.99,4584870.55 | 2.177891,41.410600 | | | 0655 | 0655 | Dissabte | Ds | Laboral | 2017 |
| 4 | 20175001826 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 268083 | RO55ELLO / Dos de Maig | | | | | 2017 | Març | 6 | 4 | Nit | Alcoholèmia | 431703.96,4584700.47 | 2.181716,41.409913 | | | 0573 | 0573 | Dilluns | DI | Laboral | 2017 | |
| 5 | 20175009733 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 18505 | Aragó | | | | | 2017 | Novembre | 11 | 23 | | | No hi ha causa mediate | 431877.83,4584413.1 | 2.183841,41.406532 | | | 0537 | 0537 | Dijous | Dj | Laboral | 2017 |
| 6 | 20175001624 | 18 | Sant Martí | 64 | el Camp de l'Arpa del Clot | 109101 | ENAMORATS | | | | | 2017 | Febrer | 7 | 2 | | | No hi ha causa mediate | 431890.19,4584524.82 | 2.183984,41.407533 | | | 0121 | 0123 | Dilluns | DI | Laboral | 2017 |
| 7 | 20175005133 | 18 | Sant Martí | 72 | Sant Martí de Provençals | 60007 | Cantàbria / Guipúscoa | | | | | 2017 | Juny | 12 | 10 | | | No hi ha causa mediate | 433330.95,4585862.62 | 2.201063,41.419785 | | | 0829 | 0829 | Dilluns | DI | Laboral | 2017 |
| 8 | 20175001560 | 18 | Sant Martí | 65 | el Clot | 161161 | INDEPENDÈNCIA / Arago | | | | | 2017 | Febrer | 25 | 15 | | | No hi ha causa mediate | 431950.88,4584412.8 | 2.184834,41.40653 | | | 0261 | 0261 | Dissabte | Ds | Laboral | 2017 |
| 9 | 20175007531 | 18 | Sant Martí | 65 | el Clot | 169489 | Glòries Catalanes / Badajoz | | | | | 2017 | Setembre | 11 | 17 | | | No hi ha causa mediate | 432229.7,4584197.87 | 2.188876,41.404617 | | | 8823 | 8823 | Dilluns | DI | Laboral | 2017 |
| 10 | 20175010616 | 18 | Sant Martí | 65 | el Clot | 18565 | Aragó | | | | | 2017 | Desembre | 23 | 11 | | | No hi ha causa mediate | 431964.93,4584467.51 | 2.184875,41.407026 | | | 0560 | 8560 | Dissabte | Ds | Laboral | 2017 |
| 11 | 20175008974 | 18 | Sant Martí | 72 | Sant Martí de Provençals | 3486 | Huelva / Agricultura | | | | | 2017 | Octubre | 30 | 14 | | | No hi ha causa mediate | 433169.99,4585881.12 | 2.199134,41.419063 | | | 0264 | 0268 | Dilluns | DI | Laboral | 2017 |
| 12 | 20175002287 | 18 | Sant Martí | 65 | el Clot | 113506 | PL GLÒRIES CATALAN | | | | | 2017 | Març | 22 | 9 | | | No hi ha causa mediate | 432296.82,4584356.47 | 2.188850,41.406854 | | | 0044 | 0044 | Dimecres | Dc | Laboral | 2017 |
| 13 | 20175003886 | 18 | Sant Martí | 65 | el Clot | 161101 | Independència | | | | | 2017 | Maig | 2 | 15 | | | No hi ha causa mediate | 432296.82,4584356.47 | 2.188850,41.406854 | | | 0239 | 0239 | Dimarts | Dm | Laboral | 2017 |
| 14 | 20175002040 | 18 | Sant Martí | 65 | el Clot | 161101 | INDEPENDÈNCIA | | | | | 2017 | Març | 14 | 10 | | | No hi ha causa mediate | 432296.82,4584356.47 | 2.188850,41.406854 | | | 0243 | 0243 | Dimarts | Dm | Laboral | 2017 |
| 15 | 20175010357 | 18 | Sant Martí | 65 | el Clot | 222408 | Democràcia | | | | | 2017 | Desembre | 15 | | | | No hi ha causa mediate | 432002.91,4584366.13 | 2.185341,41.406119 | | | 0001 | 0003 | Divendres | Dv | Laboral | 2017 |



https://umap.openstreetmap.fr/it/map/2017_accidents_causes_gu_bcn_100_413309

Making Public Transit Fairer to Women Demands Way More Data

Most transit systems aren't designed for women, who tend to run errands and care for children. But cities can't fix a problem they don't understand.



«Big data is, ultimately, political. It's about asking the right questions and also acting on the answers.»

An artist wheeled 99 smartphones around in a wagon to create fake traffic jams on Google Maps

BI

aholmes@businessinsider.com (Aaron Holmes), Business Insider • February 3, 2020



Simon Weckerl

«There is no such thing as neutral data. Data is always collected for a specific purpose, by a combination of people, technology, money, commerce, and government»

Note: some doubts were raised about the performance, i.e. if the phones were real, if the effect was really obtained, etc. We want to highlight the reflexion that the performance sparked, not the performance in itself.

Benefits of Open Data

«The benefits of Open Data are diverse and range from **improved efficiency** of public administrations, **economic growth** in the private sector to wider **social welfare**»



Open Data Barometer

Home Report Country Sheets Compare Previous Editions Contact About



GROUP

Choose

COUNTRY

Type a country

SEARCH

The Open Data Barometer

A global measure of how governments are publishing and using open data for accountability, innovation and social impact.

The Leaders Edition looks at the 30 governments that have adopted the Open Data Charter and those that, as OS20 members, have committed to OS20-Amb-Compliant Open Data Principles.

See the updated methodology for more. Open Data Barometer - 4th Edition is the latest full edition.



- REPORT
- LEADERS EDITION
- INFO
- LEADS



Leaders

May 6th 2017 edition >

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Data is **NOT** the new oil

- Where the metaphor works:
 - Data extractivism
 - Exploitation
 - Data as an asset
- Where the metaphor doesn't work:
 - Surveillance capitalism
 - Value of data versus value of information

DIKW pyramid



«[DIKW Pyramids and Car Crashes](#)» by Shôn Ellerton, October 16, 2019 on Medium

DIKW pyramid



«[DIKW Pyramids and Car Crashes](#)» by Shôn Ellerton, October 16, 2019 on Medium



DIKW pyramid



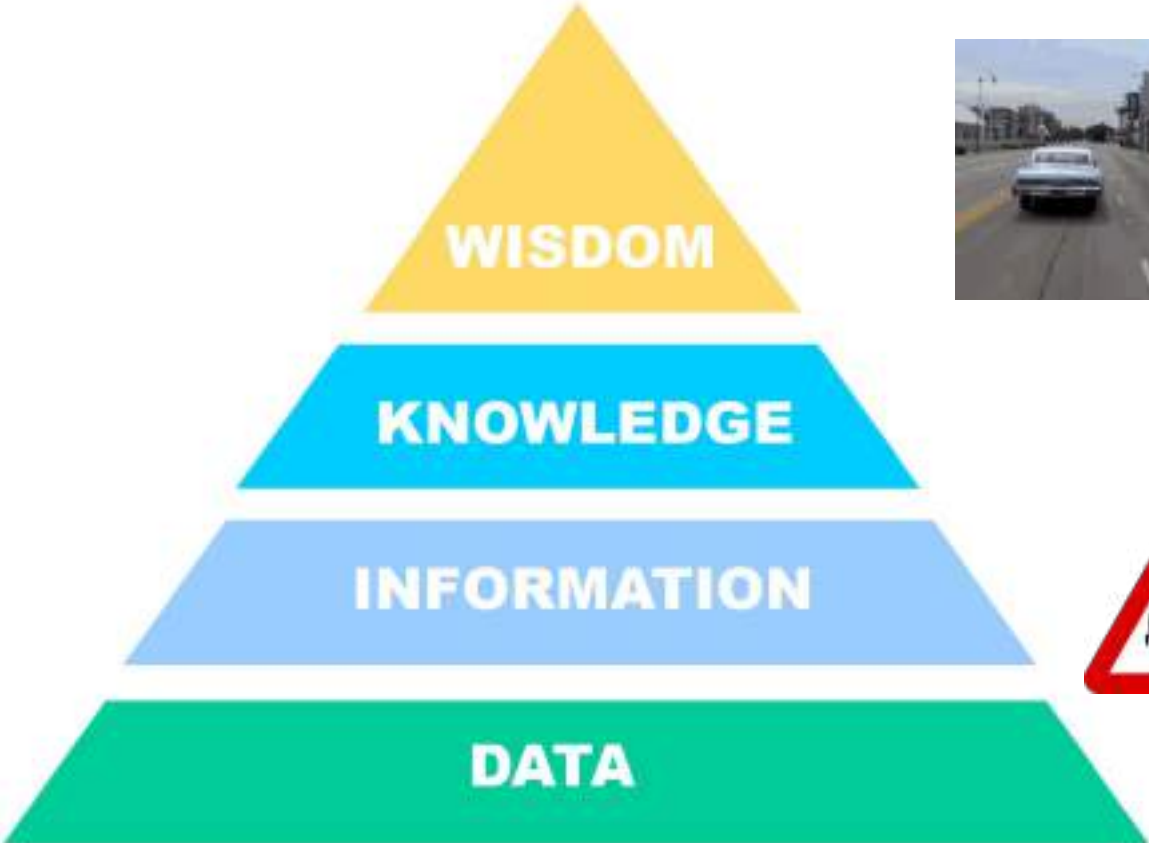
«[DIKW Pyramids and Car Crashes](#)» by Shôn Ellerton, October 16, 2019 on Medium

DIKW pyramid



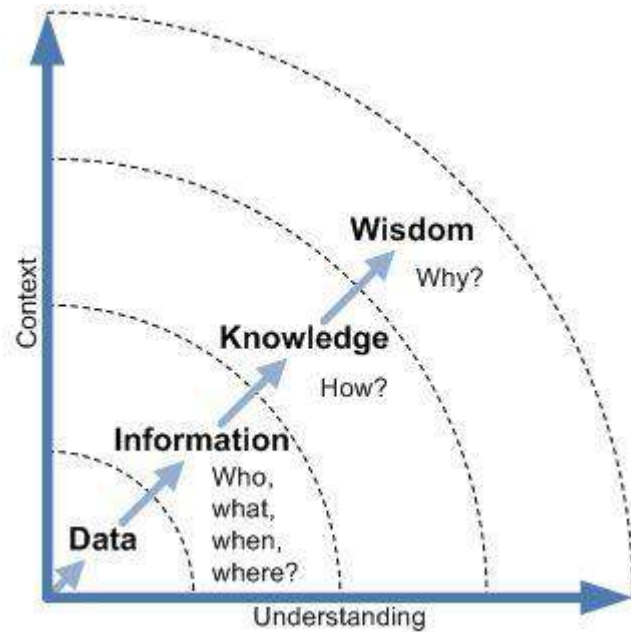
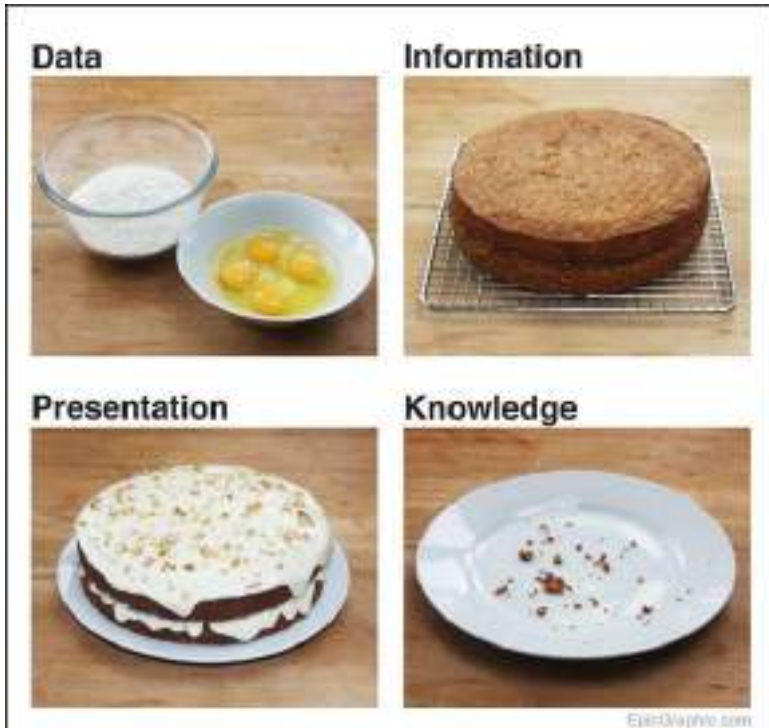
«[DIKW Pyramids and Car Crashes](#)» by Shôn Ellerton, October 16, 2019 on Medium

DIKW pyramid



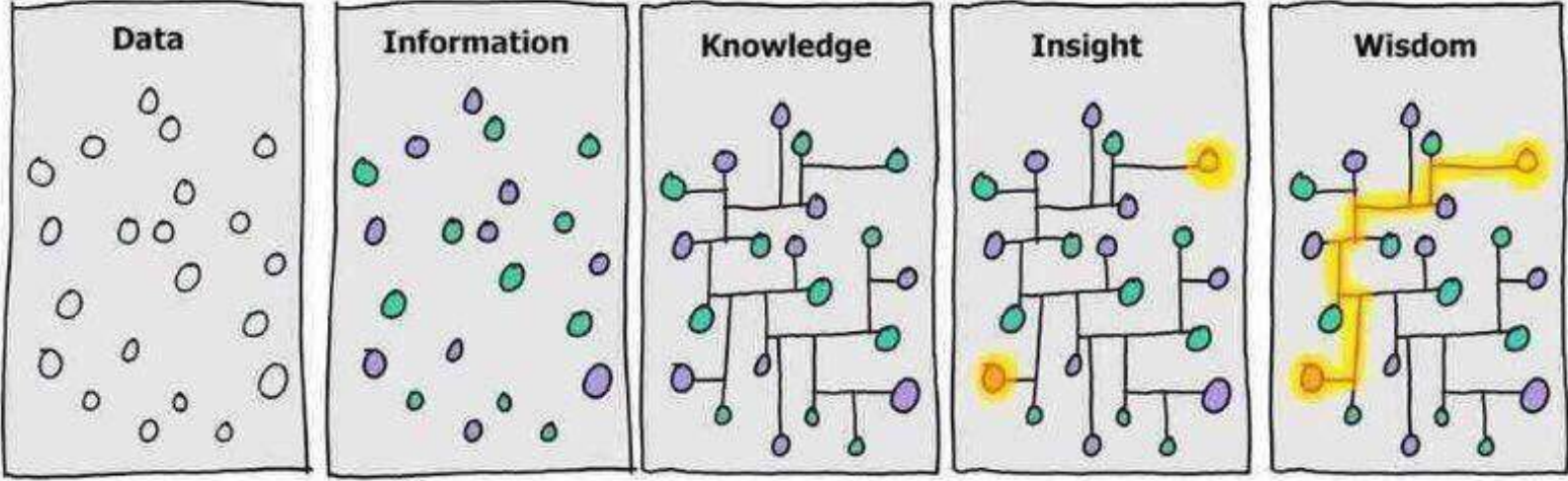
«[DIKW Pyramids and Car Crashes](#)» by Shôn Ellerton, October 16, 2019 on Medium

Data as base ingredients



by Mark Johnstone, originally published at epicgraphic.com [[archived copy](#)]

Data as base ingredients



By [@gapingvoid](#), via [@thinkmariya's tweet at 2017-04-30 at 11:23](#)

Opportunities

- Open Data
- Big Data
- Personal Data

3. Open Data Ecosystems

Open Data Producers

- International bodies (OECD, World Bank, United Nations, EU)
- Government/Public Administration
- Communities
- Corporations:
 - Open Data 500
 - Kaggle

World Bank Open Data

Free and open access to global development data

Search for a country or indicator

Browse by [Country](#) or [Indicator](#)

MOST RECENT

Making risk data work harder for adaptation in small island states

By [Christine L. Gaudin](#) (Paris, 10/2019)
[View](#), Feb 04, 2020

10 transformative charts from the past decade of development (part 1)

By [Helen Fokunang](#) (Jan 30, 2020)

World Bank data infrastructure: shortening the path from data to insights

By [T. Ng](#), Jan 28, 2020

[View all news](#)

[View all blogs](#)

WHAT YOU CAN LEARN WITH OPEN DATA



[Data Download Here](#)

Extreme Poverty

The proportion of the world's population living in extreme poverty has dropped significantly.

INTERNATIONAL DEBT STATISTICS

2020



International Debt Statistics 2020

Jan 02, 2020



RECENTLY UPDATED DATASETS

[Afghanistan - Transmission Network](#)

Oct 01, 2019

[Vietnam - Solar Radiation Measurement Data](#)



Search Datasets

Enter keywords...

Search Q

SPARQL Search



Browse Datasets by Categories



Agriculture, Fisheries,
Forestry & Foods



Energy



Regions & Cities



Transport



Economy & Finance



International Issues



Government & Public
Sector



Justice, Legal System &
Public Safety



Environment



Education, Culture &
Sport



Health



Population & Society



Science & Technology



Catalogues



All data

Latest News



Open Data Maturity
Report 2019
09/02/2020



Optimise water irrigation
and promote
sustainability
21/01/2020



Release of a new study
into high-value datasets
30/01/2020



Save the date: Open
Data Day 2020
26/01/2020

RSS

More news



The European Union Open Data Portal (EU ODP) gives you access to open data published by EU institutions and bodies. All the data you can find via this catalogue are free to use and reuse for commercial or non-commercial purposes.

Show results with
 all of these words | any of these words | the exact phrase

Search for metadata using our SPARQL endpoint query editor or access the API.

Discover our datasets [View datasets by subject](#) [View all datasets](#) [View all publishers](#)

Focus on



Antibiotic resistance in Europe

> European Centre for Disease Prevention and Control

Twitter



EU Open Data
@EU_odp_en

From #EUDataViz to visual thinking. Learning over lunch with visual thinkers at @EULawDataPubs today! #dataviz #GraphicDesign

Open Data for All New Yorkers

Open Data is free public data published by New York City agencies and other partners. **Sign up for the NYC Open Data mailing list** to learn about training opportunities and upcoming events.

Or join us in building **Open Data Week 2020!**

Search Open Data for things like 311, Buildings, Crime

The Next Decade of Open Data

2019 Open Data for All Report



Learn about the next decade of NYC Open Data, and read our 2019 Report

CHE TEMPO FA, OGGI, IN TRENTINO

tutti i dataset sul meteo:

le previsioni giornaliere, le previsioni in montagna, le probabilità di pioggia

Tags popolari

servizi dipendenti organi politici regolamenti
politici aree tematiche albi ed elenchi avvisi
moduli piani e progetti

Cerca dati

Cerca i dati, e ottieni gli aggiornamenti per i dataset e le categorie a cui sei interessato.

Ultime modifiche



Botteghe storiche del Trentino

Elenco delle botteghe di interesse storico del Trentino, disciplinate dall'art. 63 della legge provinciale 17/2010 "Disciplina dell'attività commerciale". Sono considerate...

[CSW](#)

Tweet

di @DatiTrentino



Ritwittato da dati.trentino.it



Alfonso Fuggetta

@AlfonsoFuggetta

Un mio commento al dibattito di queste ore su SPID: "A proposito di SPID e di ciò che serve realmente al Paese"

Open government

Home

What is it

Transparency

Open data

Public participation

Procedures

News

Contact

Home

Opendata

Opendata



Search the open data catalogue

Search

Most popular datasets

- Contractació de Catalunya
- Dades d'informació dels punts de mesurament de la taxa de Vigilància i Prestació de la Contractació Administrativa
- Contractació de la Generalitat de Catalunya. Contractes Menors

Newly added datasets

- Premi Ramon Margalef d'Ecologia
- Recerca capital per a projectes i altres tipus d'activitats relacionats amb la R+D+i a les universitats
- Personal Docent Investigador a les universitats catalanes

Iniciativa de **datos abiertos** del Gobierno de España

Destacados.



Encuentro Aporta 2019

Seriedad y datos abiertos

Desafío Aporta 2019

Guía COTEC

Predictibilidad judicial

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).

GET STARTED

SEARCH OVER 230,221 DATASETS



BROWSE TOPICS



Agriculture



Climate



Consumer



Ecosystems



Education



Energy



Finance





WIKIPEDIA

The Free Encyclopedia

<https://dumps.wikimedia.org>

Datasets

Documentation

New Dataset

Help the community by creating and solving Tasks on datasets!



Search 23,279 datasets

Feedback Filter

PUBLIC

Sort by: Hottest



Caselaw Dataset (Illinois)

Caselaw Access Project

1 year 829 MB 81 2 Files (CSV) 1 Task

68



FIFA 20 complete player dataset

Statista Limited

4 months 15 MB 67 7 Files (CSV)

106



eCommerce behavior data from multi category store

Mahesh Electronics

2 months 4 GB 100 27 Files (CSV) 1 Task

82



Cost of Living

DL

25 days 22 MB 64 1 File (CSV) 1 Task

93



Screen Actors Guild Awards, 1994 - 2020

Richard Forbis

12 days 51

13

Open Tasks

Add Average Data Scientist Salary

2 Submissions - in Cost of Living

What to watch on Netflix?

2 Submissions - in Netflix Movies and TV Shows

Top 50 Spotify Songs - previous years

1 Submission - in Top 50 Spotify Songs - 2019

Visualize US Accidents Dataset

3 Submissions - in US Accidents (I 95) Road

Background analysis

0 Submissions - in Credit Card Fraud Detect



The OD500 Global Network

The OD500 Global Network is an international network of organizations that seek to study the use and impact of open data. Coordinated by the Governance Lab (GovLab) the OD500 Global Network enables participating organizations to analyze open data in their country in a manner that is both globally comparative and domestically specific. The OD500 Global Network starts from the assumption that only by mapping the use of open data within and across countries, can new approaches for understanding the economic and social impact of open government data be generated.



Australia



México



United States



Italy



Korea



Canada

Data Portals

A Comprehensive List of [Open Data](#) Portals from Around the World

[588 Data Portals listed](#) »

This service is run by [Open Knowledge Foundation](#) | [Source Code](#) | [Download Data \(CSV\)](#) | [Download Data \(JSON\)](#) | [Data License \(Public Domain\)](#) | [Privacy Policy](#)



A WORLD WHERE
KNOWLEDGE
CREATES POWER
FOR THE MANY,
NOT THE FEW.

THIS IS THE WORLD WE CHOOSE.

FRICTIONLESS DATA TOOL FUND

A halfway point update



PASSIONATE TEAM

Passionate about openness. Using advocacy, technology and training to unlock information and enable people to create and share knowledge.



GLOBAL NETWORK

Meet, campaign, learn, innovate, share, train, create, support, explore; some of the ways you can help open up knowledge for everyone. Join us.

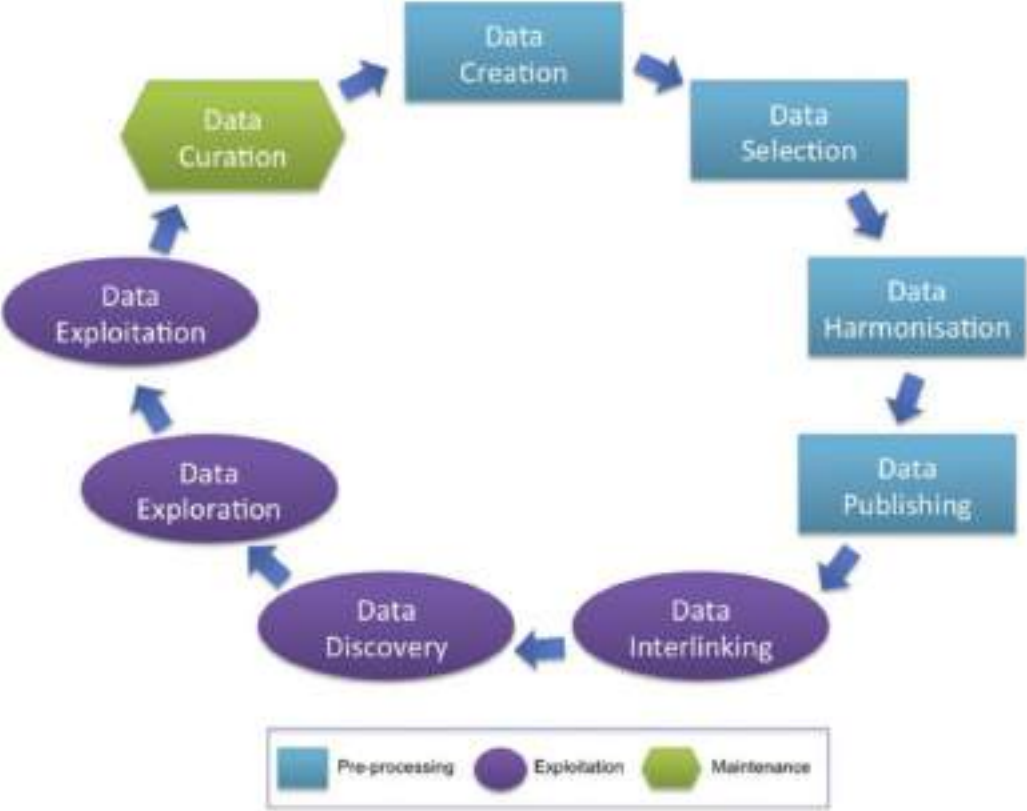


DIVERSE PROJECTS

Through our projects, research and collaborations, we explore niche areas of data, and ways in which it can be used to empower people around the world.

3. How to make Open Data?

Integrated Open Data Cycle



Open data can be:

- **structured or unstructured;**
- **available in a given format (e. g. CSV, JSON, XML, etc.);**
- **organized following a schema**
- **Shared with different technologies/procedures (e. g. using APIs, download, SPARQL endpoints, etc.)**
- **released under a free license**

Disclaimer

"I AM NOT A LAWYER"

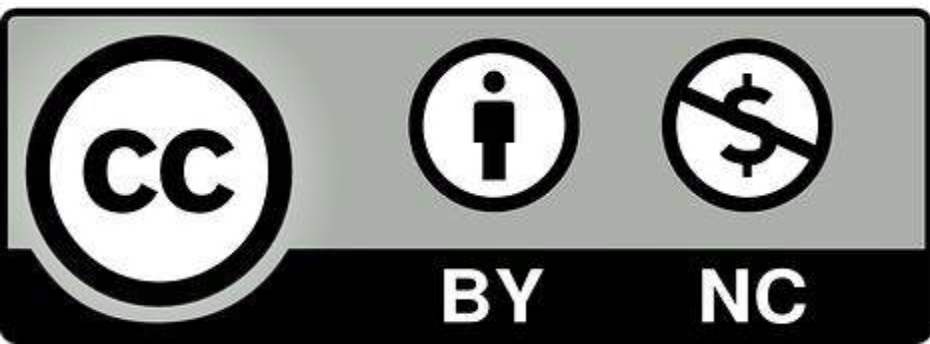
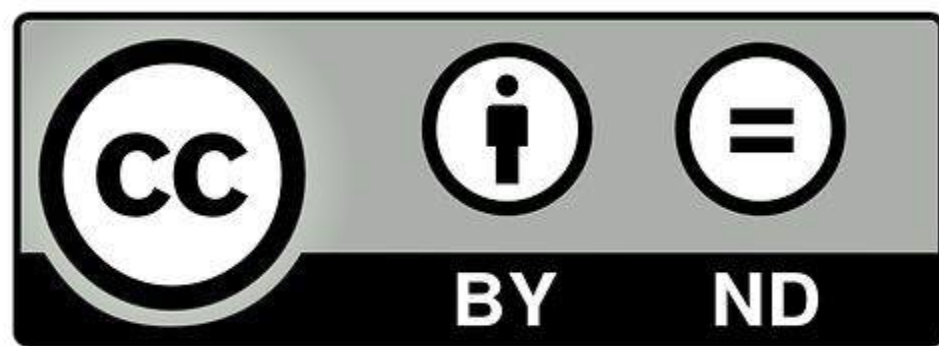
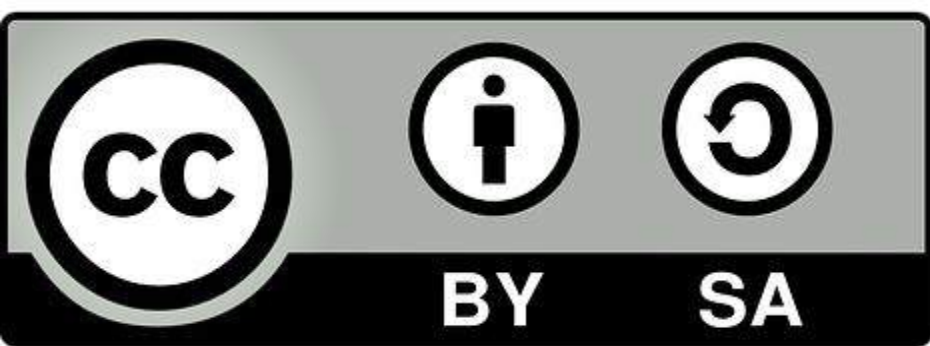


**I CAN SHORTEN IT TO IANAL!
EVERYONE WILL KNOW
EXACTLY WHAT I MEAN!**

- A **contract** authorize a some use (copying software, translating a book, etc.) to a licensee
- Their foundations is **copyright law**, but may touch also patents, trademarks, personal rights and privacy, etc.
- Different licenses are suited for different kind of works:
 - Content (text, images, sounds): CC, GFDL, ...
 - Software: GPL, LGPL, MIT, ...
 - Data: CC0, ODbL

Creative Commons licenses

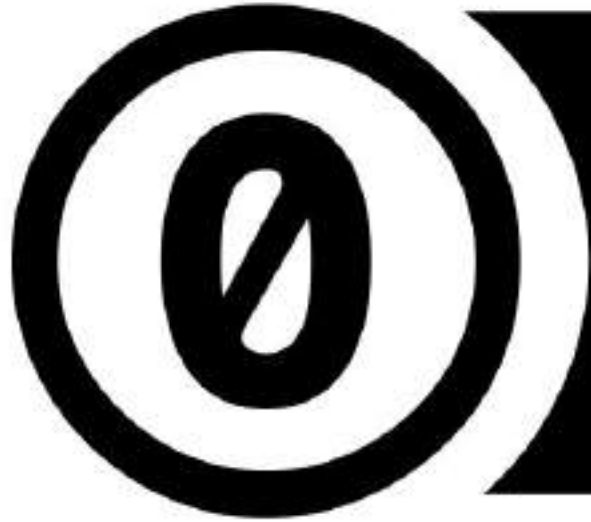
- **BY (attribution)**: cite the original author of the work;
- **NC (*non-commercial*)**: you can redistribute this work as long as you don't do it for commercial purposes;
- **ND (*non-derivative*)**: the work can be redistributed but it must be unchanged;
- **SA (*share-alike*)**: you can create derivative works, but you must license the new creation under identical terms of the original work, i.e. **copyleft or virality**;



Public Domain “licenses”

- **CC0:** a “waiver” by the author that grants all uses permitted by the law.
- **Public Domain Mark,** for works whose copyright terms have ended

CC0



**PUBLIC
DOMAIN**

- CC0
- Open Data Commons - Open Database License:
 - Database \neq Content
 - Additional IP rights
 - Can distinguish content and the DB
 - Derivative works
 - Licensing terms themselves can be a problem



ODbL

Public Domain “licenses”

- LGPL
- GPL
- AGPL
- MIT/BSD



Data Formats

- Non structured:
 - PDF
 - Word (.doc)
- Tabular data:
 - Excel (.xls)
 - CSV, TSV
- Dictionaries and trees:
 - XML
 - JSON
- Linked data:
 - RDF

Tabular Data: CSV, TSV

CSV

Name, Surname, Age, Street, City, Postal_code, Phone

John, Doe, 25, "Carrer de Bilbao,72", Barcelona, 08005, 602030230

TSV

Name Surname Age Street City Postal_code Phone

John Doe 25 Carrer de Bilbao,72 Barcelona 08005 602030230

Dictionaries and Trees: XML, JSON

- Data in “*key–value*” form (or *name–value pair*, *attribute–value*).
- Accepts values in many formats: strings, dates, numbers, booleans
- Allows for lists and *nested structures*
- They are flexible
- They can have a schema

Tree-like Data: XML

```
<person>
  <name>John</name>
  <surname>Doe</surname>
  <age>25</age>
  <address>
    <street>Carrer de Bilbao, 72</street>
    <city>Barcelona</city>
    <postal_code>08005</postal_code>
  </address>
  <contact>
    <type>home</type>
    <number>059234565</number>
  </contact>
  <contact>
    <type>mobile</type>
    <number>602030230</number>
  </contact>
</person>
```

Dictionary: JSON

```
{
  "name": "John",
  "surname": "Doe",
  "age": 25,
  "address": {
    "street": "Carrer de Bilbao, 72",
    "city": "Barcelona",
    "postal_code": "08005"
  },
  "phone": [
    {
      "type": "home",
      "number": "059234565"
    },
    {
      "type": "mobile",
      "number": "602030230"
    }
  ],
}
```

- **Descriptive:** to identify and retrieve digital resources (e. g. *title, author*)
- **Administrative:** to manage, preserve and administer rights (e. g. *format, license*)
- **Structural:** used to describe the internal structure of digital resources and manage relations between its components (e. g. *chapter, page*)

schema.org is an initiative heralded by Google, Bing, and Yahoo that started in 2011 to “*create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.*”

Example: book

- <https://schema.org/Book>
- <https://openlibrary.org/books/OL1102372M.json>

Schema.org: example

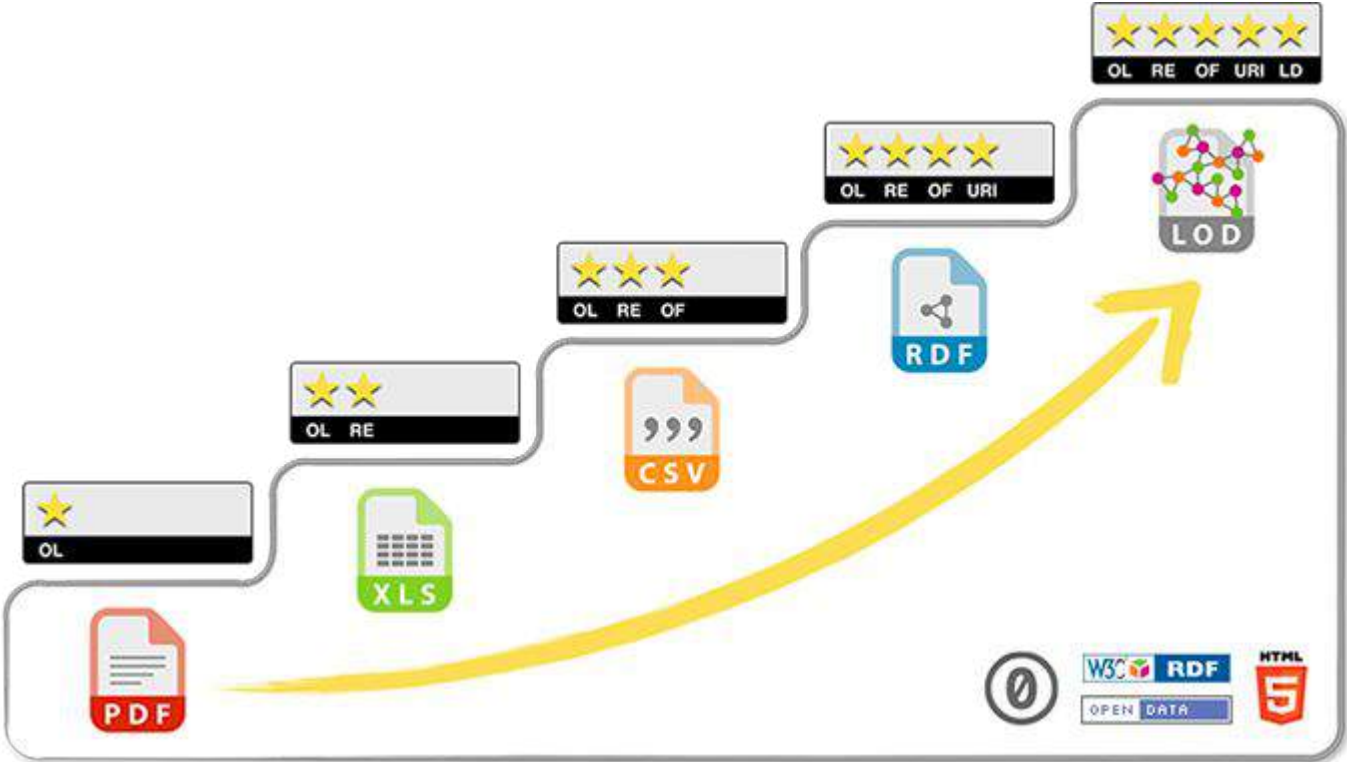
```
<div itemscope itemtype="http://schema.org/Book">
  <h1 itemprop="name">Ficciones</h1>
  <div itemprop="author" itemscope itemtype="http://schema.org/Person">
    Author: <span itemprop="name">Jorge Luis Borges</span>
    (born on <time itemprop="birthDate" datetime="1899-08-24">August, 24th
1899</time>)
  </div>
  Genre: <span itemprop="genre">Fiction</span>
  Publisher: <span itemprop="name">Adelphi</span>
</div>
```

Making Open Data: a guide in 3 steps

1. Use very liberal *open licenses* such as **CC0** to obtain a complete **liberation of data**
2. Use open API and **standards** (e. g. REST API; open data formats such as JSON, CSV).
3. Share your code on the internet with a free software license, you can use platforms such as **GitHub**^(*) (<https://github.com>)

(*) although being very popular and free-of-charge for open source projects the platform itself is not a free software platform. Furthermore, the company that owns the platform - GitHub, Inc - is based in the United States, which could be problematic for some projects.

5-star Open Data



5-star Open Data: description

★ Make your stuff available on the Web (whatever format) under an open license

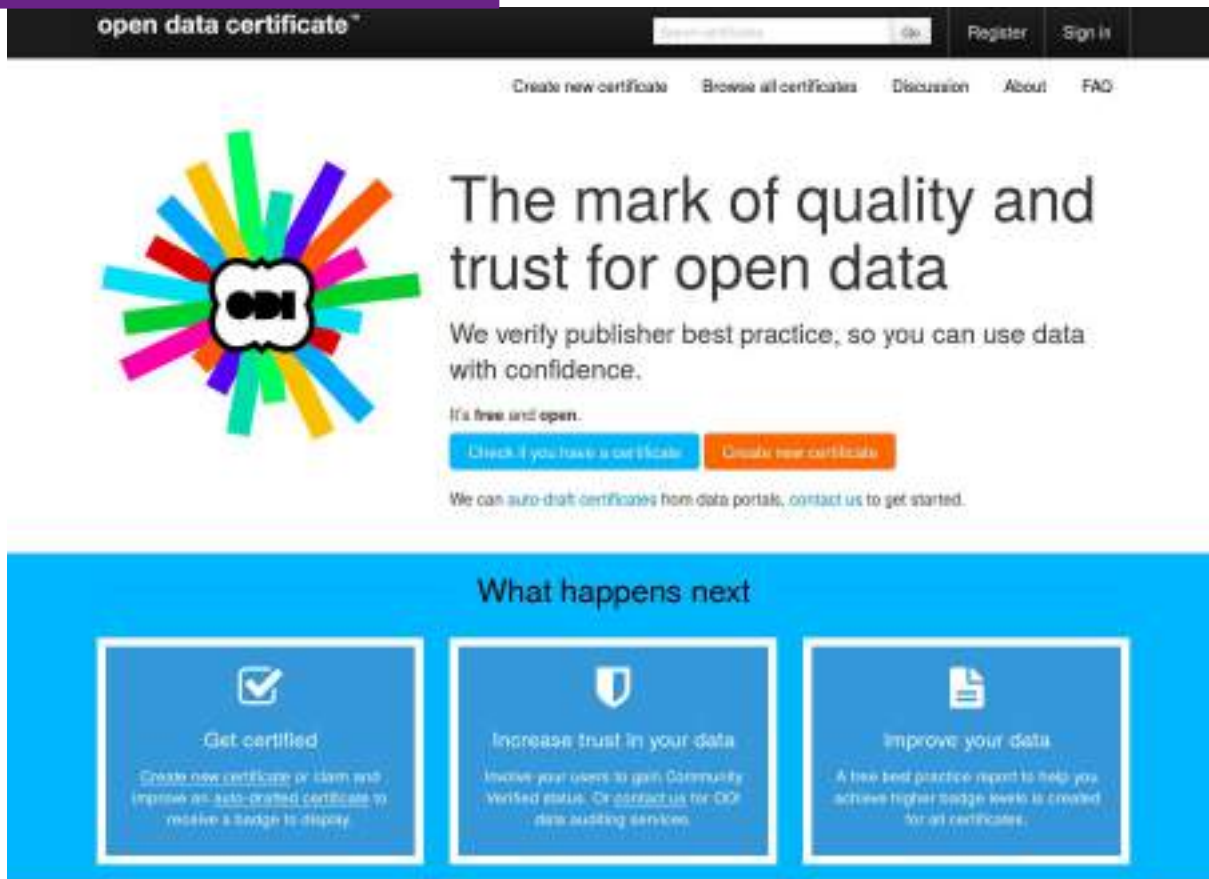
★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ Make it available in a non-proprietary open format (e.g., CSV instead of Excel)

★★★★ Use URIs to denote things, so that people can point at your stuff

★★★★★ Link your data to other data to provide context

ODI Open Data Certificates




The screenshot shows the homepage of the ODI Open Data Certificates website. At the top, there is a navigation bar with the logo "open data certificate™", a search bar, and links for "Register" and "Sign In". Below the navigation bar, there are links for "Create new certificate", "Browse all certificates", "Discussion", "About", and "FAQ". The main content area features a large, colorful logo on the left consisting of many sticks radiating from a central circle with "ODI" inside. To the right of the logo, the text reads "The mark of quality and trust for open data" followed by "We verify publisher best practice, so you can use data with confidence." Below this, it says "It's free and open." and provides two buttons: "Check if you have a certificate" and "Create new certificate". A note at the bottom of this section states "We can auto-draft certificates from data portals, contact us to get started." The bottom section, titled "What happens next", contains three blue boxes with icons and text: 1. "Get certified" with a checkmark icon, describing the process of creating or improving a certificate to receive a badge. 2. "Increase trust in your data" with a shield icon, describing how to gain community-verified status or contact ODI for auditing services. 3. "Improve your data" with a document icon, describing how a best practice report can help achieve higher badge levels.

open data certificate™

Register Sign In

Create new certificate Browse all certificates Discussion About FAQ



The mark of quality and trust for open data




We verify publisher best practice, so you can use data with confidence.

It's free and open.

[Check if you have a certificate](#) [Create new certificate](#)

We can auto-draft certificates from data portals, contact us to get started.

What happens next

- 
Get certified
Create new certificate or claim and improve an auto-drafted certificate to receive a badge to display.
- 
Increase trust in your data
Involve your users to gain Community Verified status. Or contact us for ODI data auditing services.
- 
Improve your data
A free best practice report to help you achieve higher badge levels is created for all certificates.

<https://certificates.theodi.org/en/>

How to make Open Data

- *Legal level:*
 - open license (e. g. **CC0**)
- Practical level:
 - Accessible over the web
- *Technical level:*
 - Open format: **JSON, CSV, XML**
 - Metadata schema: **Dublin Core, Schema.org**
 - Open procedures: **API, SPARQL**
- *Social level:*
 - Documentation
 - Contact information

3. Open Access

Open Access: definition

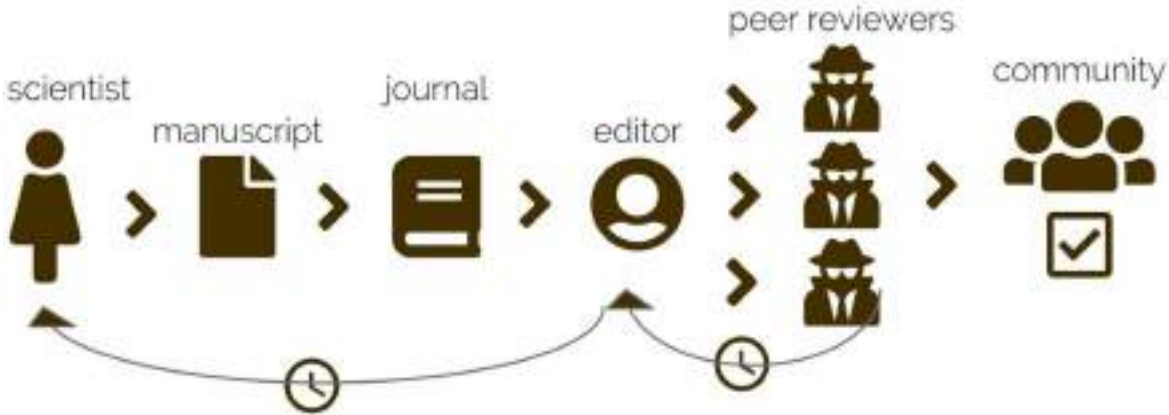
Open Access is a movement is focused around free access to peer-reviewed research literature and a **set of mechanisms** by which research outputs are distributed online, **free of cost or other access barriers**.

- Green open access (*self-archiving*)
- Gold open access



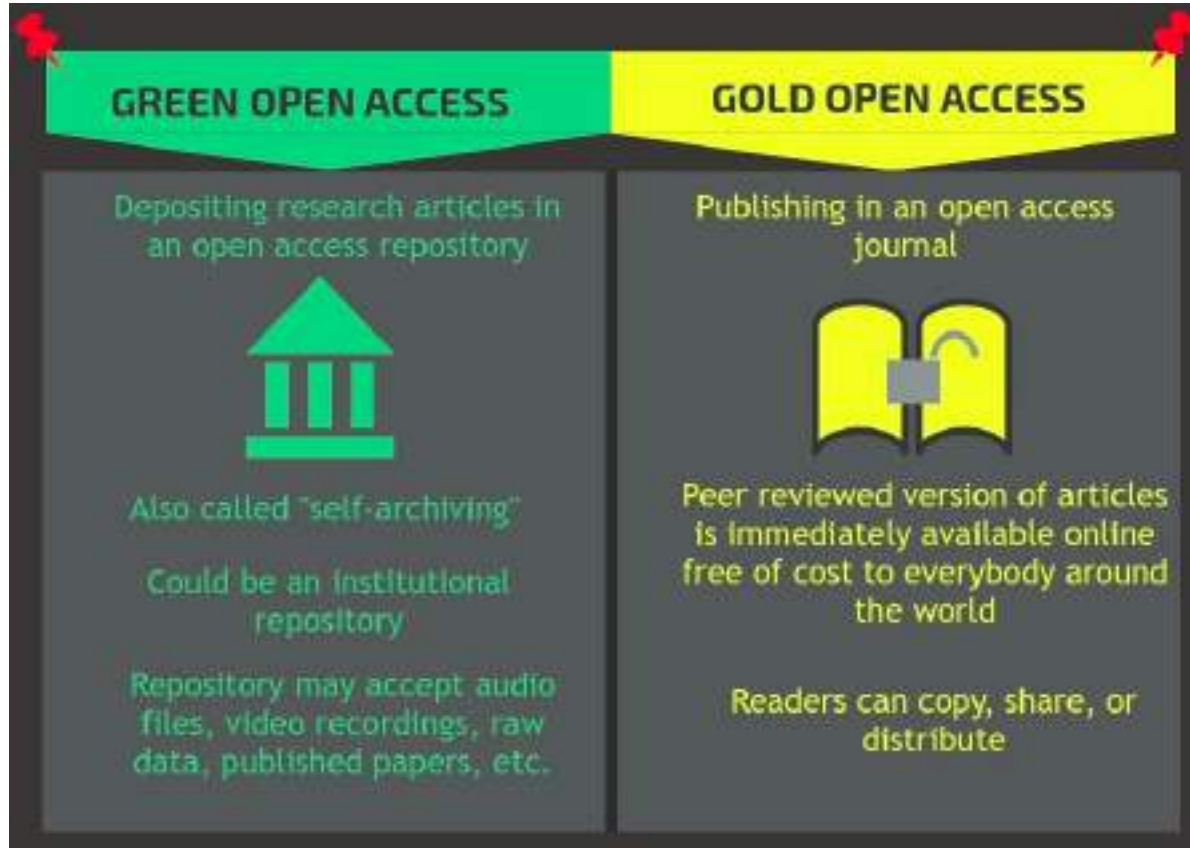
Open Access logo by [Public Library of Science](#)

How **scholarly communication** works



«Ten Hot Topics around Scholarly Publishing» Publications 2019, 7(2), 34;
<https://doi.org/10.3390/publications7020034>

Open Access: two-path system



Open Access Guerrilla Manifesto

«Information is power. But like all power, there are those who want to keep it for Themselves. [...]

Those with access to these resources – students, librarians, scientists – you have been given a privilege. You get to feed at this banquet of knowledge while the rest of the world is locked out. [...]

But all of this action goes on in the dark, hidden underground. It's called stealing or piracy, as if sharing a wealth of knowledge were the moral equivalent of plundering a ship and murdering its crew. **But sharing isn't immoral – it's a moral imperative. Only those blinded by greed would refuse to let a friend make a copy.»**



Aaron Swartz at Boston Wikipedia Meetup, 2009-08-18 by [Sage Ross](#) via [Wikimedia Commons](#) [CC BY-SA 2.0]

4. Free and Open Source Software

Source code

```
#include <stdio.h>

int main() {
    printf("Hello, world!\n");
    return 0;
}
```



```
7f45 4c46 0201 0100 0000 0000 0000 0000
0200 3e00 0100 0000 3004 4000 0000 0000
4000 0000 0000 0000 d819 0000 0000 0000
0000 0000 4000 3800 0900 4000 1f00 1c00
0600 0000 0500 0000 4000 0000 0000 0000
4000 4000 0000 0000 4000 4000 0000 0000
f801 0000 0000 0000 f801 0000 0000 0000
0800 0000 0000 0000 0300 0000 0400 0000
3802 0000 0000 0000 3802 4000 0000 0000
3802 4000 0000 0000 1c00 0000 0000 0000
```

```
$ gcc hello.c -o hello
$ ./hello
Hello, world!
$
```

Software Freedoms

Four essential software freedoms:

0. The freedom to **run** the program as you wish, for any purpose
1. The freedom to **study how** the program works, and **change** it so it does your computing as you wish. Access to the source code is a precondition for this
2. The freedom to **redistribute copies** so you can help others
3. The freedom to **distribute copies** of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

Free Software and Open Source **history**

- 1983: GNU Manifesto written
- 1989: First version of GNU GPL released
- 1991: Linux release
- 1994: Mosaic (then Netscape) navigator started
- 1998: Mozilla project started
- 1998: Open Source Initiative started



API: Application Programming Interface

An **API** is a communication protocol to interact, in real time, with the data of a website or service.

Example

`https://en.wikipedia.org/w/api.php` → base URL

`?action=query`
`&list=search`
`&srsearch=Jorge%20Luis%20Borges`
`&format=json`

parameters



API: example results

JSON Raw Data Headers

Save Copy Collapse All Expand All Filter JSON

```
batchcomplete: ""
continue:
  sroffset: 19
  continue: "-||"
query:
  searchinfo:
    totalhits: 2227
  search:
    0:
      ns: 0
      title: "Jorge Luis Borges"
      pageid: 15781
      size: 100900
      wordcount: 12578
      snippet: "offspring also included the painter Norah <span class=\"searchmatch\">Borges</span>, sister of <span class=\"searchmatch\">Jorge</span> <span class=\"searchmatch\">Luis</span> <span class=\"searchmatch\">Borges</span>. At age nine, <span class=\"searchmatch\">Jorge</span> <span class=\"searchmatch\">Luis</span> <span class=\"searchmatch\">Borges</span> translated Oscar Wilde's The Happy Prince"
      timestamp: "2020-02-08T11:52:44Z"
    1:
      ns: 0
      title: "Jorge Luis Borges bibliography"
      pageid: 1557427
```

5. Data Papers

Data paper: definition

«A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data. Data papers provide recognition for this effort by means of a scholarly article.» [Data paper](#) via Global Biodiversity Information Facility

- Nature Scientific Data
<https://www.nature.com/sdata/>
- Data in Brief
<https://www.journals.elsevier.com/data-in-brief>
- ICWSM, International Conference on Web and Social Media
<https://www.icwsm.org/>

Example

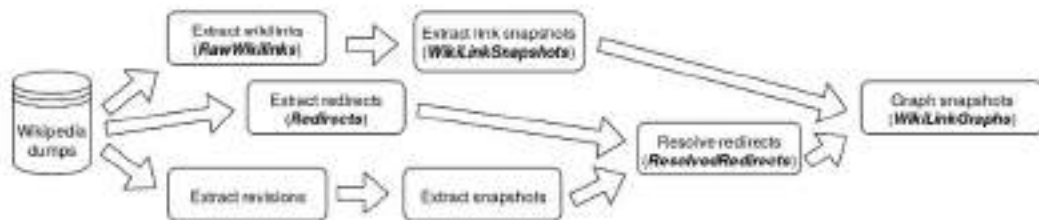


Figure 2: The process to produce the WIKILINKGRAPHS dataset from the Wikipedia dumps. In bold and italics the name of the intermediate datasets produced.

Redirects and Link Resolution A redirect in Media-Wiki is a page that automatically sends users to another page. For example, when clicking on a *wiki link* `[[NYC]]`, the user is taken to the article *New York City* with a note at the top of the page saying: “(Redirected from NYC)”. The page *NYC*⁴ contains special Wikitext: `#REDIRECT [[New York City]]` which defines it as a redirect page and indicates the target article. It is also possible to redirect to a specific section of the target page. Different language editions of Wikipedia use different words¹⁹, which are listed in Table 2.

| lang | words |
|-----------------|-----------------------------|
| de | #WEITERLEITUNG |
| en | #REDIRECT |
| es | #REDIRECCIÓN, #REDIRECCION |
| fr | #REDIRECTION |
| it | #RINVIA, #RINVIO, #RIMANDO |
| nl | #DOORVERWIJZEN |
| pl | #PATRZ, #PRZEKIERUJ, #TAM |
| ru [†] | #PERENAPRAVLENIE, #PERENAPR |
| sv | #OMDIRIGERING |

creation of a snapshot for a given year entails the following process:

1. we list all *revisions* with their timestamps from the dumps;
2. we filter the list of revisions keeping only those that existed on March 1st, i.e. the last revision for each page created before March 1st;
3. we resolve the redirects by comparing each page with the list of redirects obtained as described above;

At the end of this process, we obtain a list of the pages that existed in Wikipedia on March, 1st of each year, together with their target, if they are redirects. We call this dataset **RESOLVEDREDIRECTS**.

It should be noted that even if we resolve redirects, we do not eliminate the corresponding pages: in fact, redirects are still valid pages belonging to the namespace 0 and thus they still appear in our snapshots as nodes with one outgoing link, and no incoming links.

Link Snapshots We then process the **RAWWIKILINKS** dataset and we are able, for each revision of each page, to establish whether a wiki link in a page was pointing to an existing page or not. We add this characteristics to the **RAWWIKI-**

6. Tech



numpy / numpy

Sponsor

Watch 484

Star 12.8k

Fork 4.2k

Code

Issues 1,779

Pull requests 219

Actions

Projects 3

Wiki

Security

Insights

The fundamental package for scientific computing with Python. <https://www.numpy.org/>

numpy python

22,335 commits

20 branches

0 packages

166 releases

862 contributors

BSD-3-Clause

Branch: master

New pull request

Find file

Clone or download

merge Merge pull request #15430 from sethroisnel/helgguide_warning Latest commit atbeee 10 hours ago

.circleci DOC, BLD: sphinx 2.2.0 -> 2.3.1 to get rid of deprecation warning 15 days ago

.dependabot MAINT: Add "MAINT" tag to dependabot commit msg 4 months ago

.github Update FUNDING.yml 3 months ago

benchmarks BENCH: adding benchmarks for np.maximum 7 days ago

branding/icons add .gitattributes and fix line endings 9 years ago

doc Merge pull request #15427 from aeberg/deprecate-unused-c-api 2 days ago

numpy Merge pull request #15500 from eric-wieser/type__reduce__breaks 11 hours ago

tools Merge pull request #15430 from sethroisnel/helgguide_warning 10 hours ago

.codecov.yml DOC: Removing mentions of appveyor (#14711) 4 months ago

.coveragerc MAINT: TST: remove test-installed numpy cv 7 months ago

.ctags.d 14 months ago

<https://github.com/>



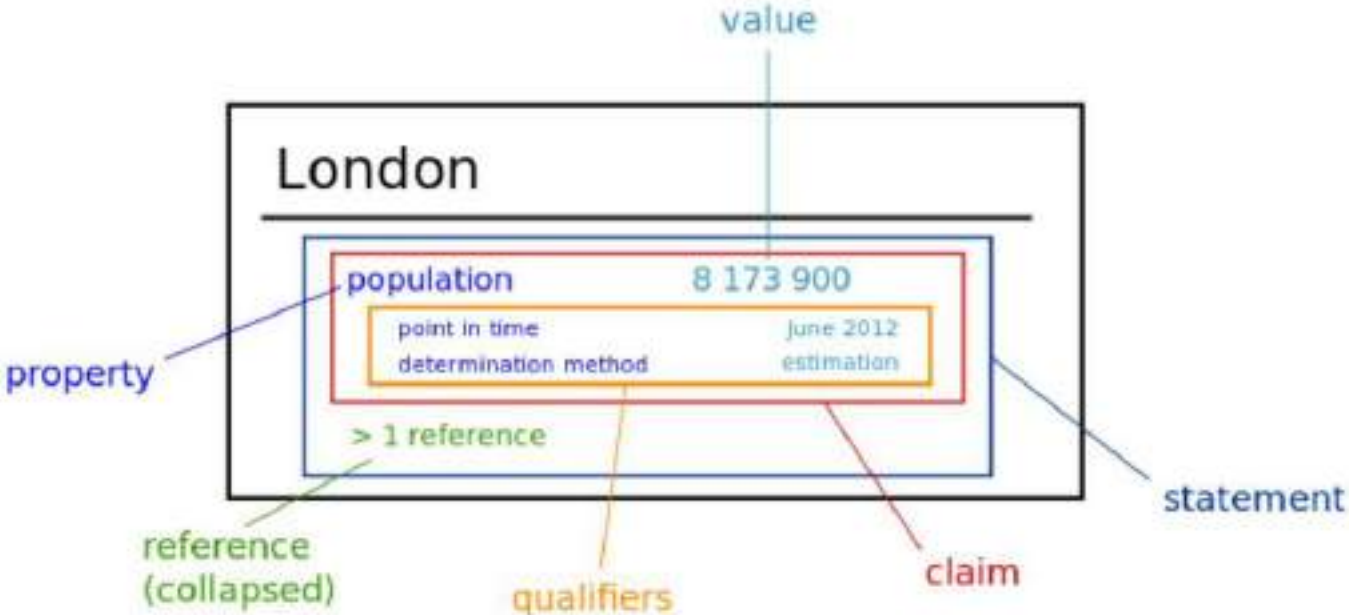
Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

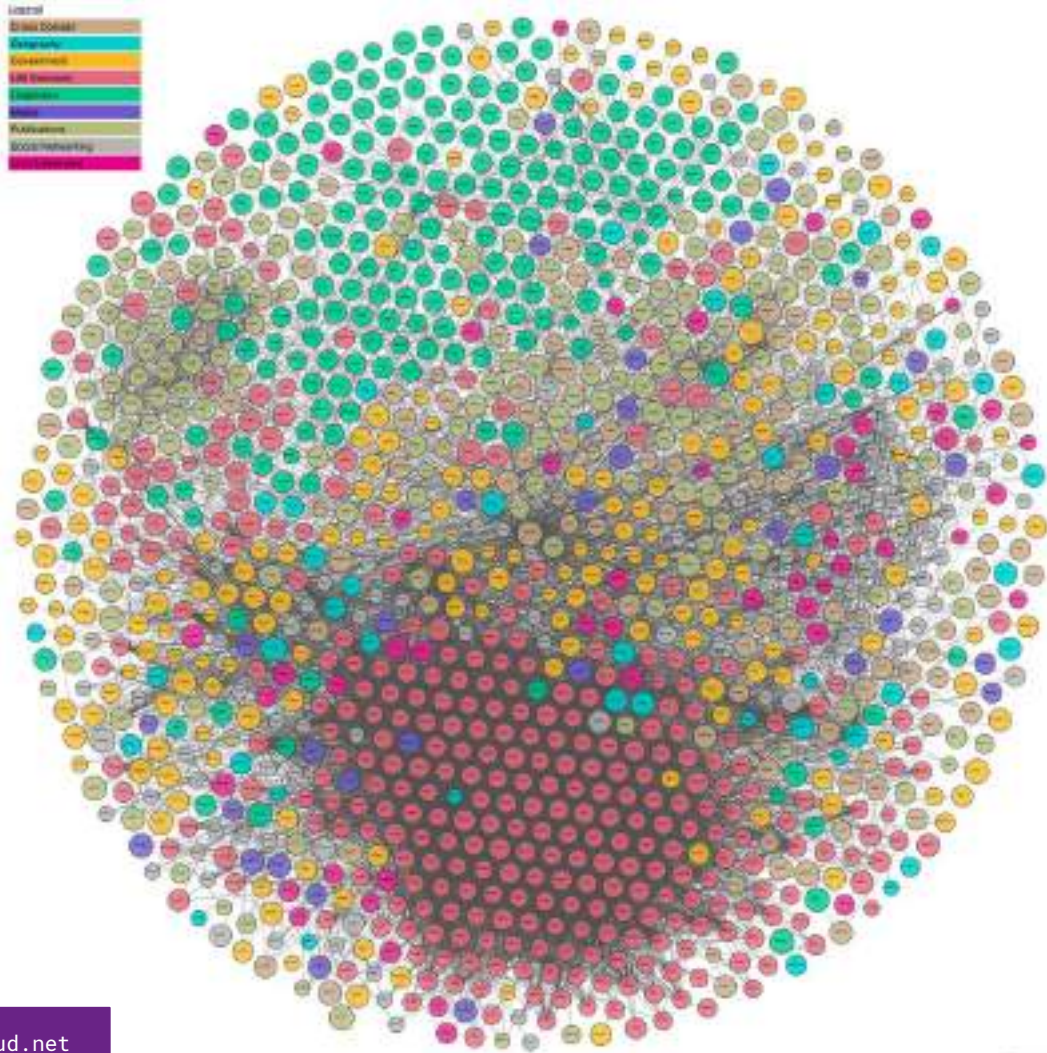
«A Collaborative Knowledge Base (i. e. “Wikipedia for data”) that makes data accessible to Wikipedia in all languages. Data are also available as Linked Open Data and accessible to external services»

Characteristics:

- Wikidata is centralized, one instance for all languages;
- Wikidata exposes *Linked Open Data*: you can also download the data in various formats (e. g. RDF)
- Wikidata hosts, maintains and curates all data available for all Wikipedia versions, all Wikimedia projects and other;
- Wikidata is a *central hub* of the linked open data web

Wikidata: example item





Questions?

cristian.consonni@eurecat.org



Credits and Acknowledgements and Licensing

- This presentation is released under the **Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0)** license. Specific parts of the presentation - such as images - may have a different licences.
- This presentation is based on the presentation: “*Bibliotecari digitali: dati, digitalizzazione, visualizzazione*” by Andrea Zanni (<https://twitter.com/aubreymcfato>)



Backup

Smart Cities and the Data Deluge



Source: <https://www.zonamovilidad.es/fotos/2/smart-cities-iot-lora-alliance.jpg>

The Memex: Knowledge as links

(“As We May Think” by V. Bush, The Atlantic, 1945)

- *«This is the essential feature of the memex. The process of tying two items together is the important thing. Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified»*
[“As We May Think”](#), V. Bush, The Atlantic, 1945



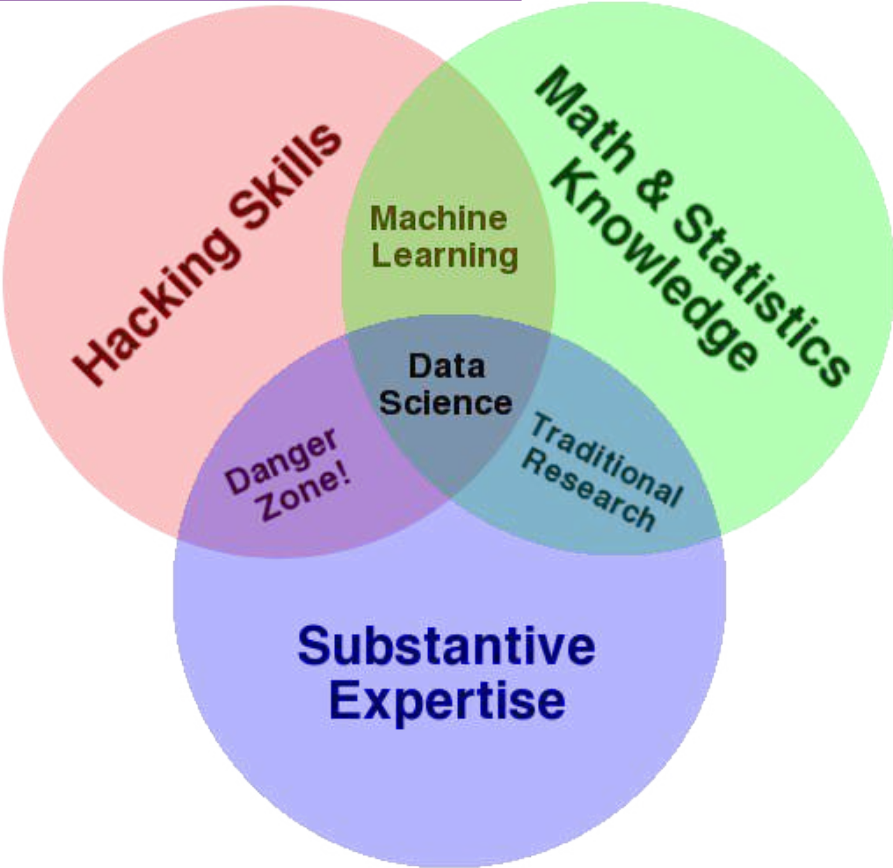
[The Memex](#), via Wikimedia Commons ([CC BY 2.0](#))

Data as base ingredients

The four V's

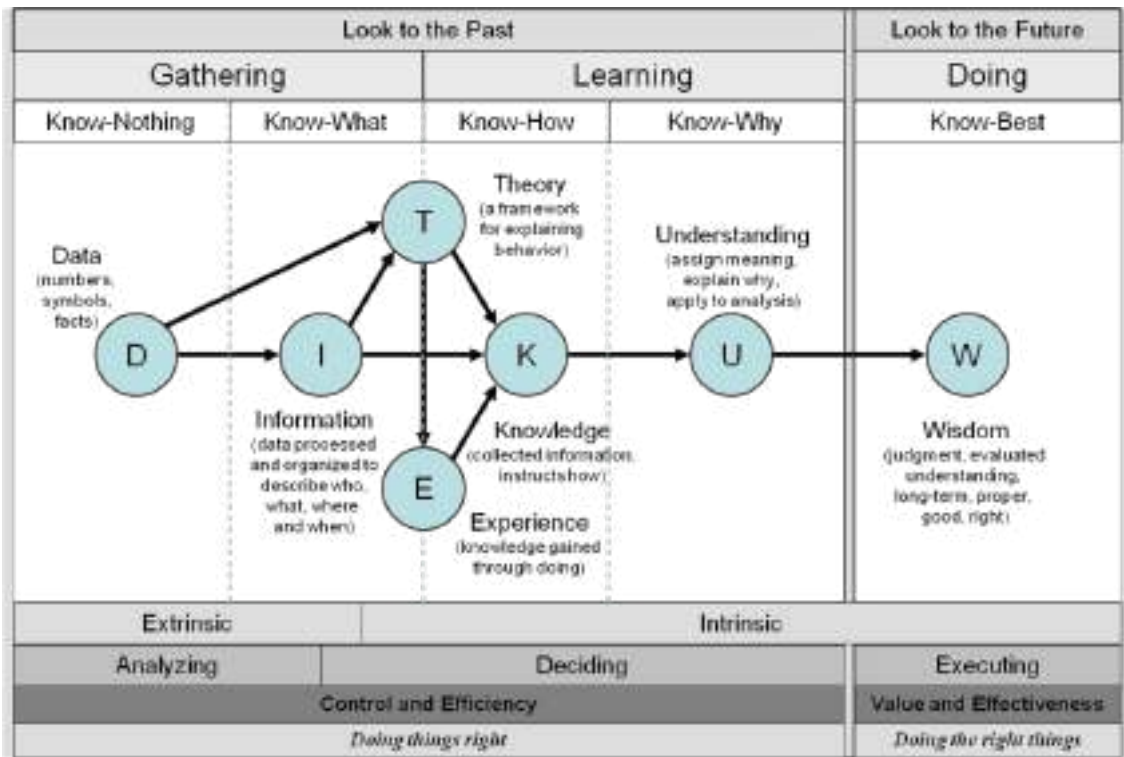


From the presentation "Big Data, Apache Spark e Apache Kafka" by Clayton Kossoski
[<http://silverio.net.br/heitor/disciplinas/eeica/aulas/aula8a-BD.pdf>]



DIKW pyramid

(it doesn't have to be a pyramid)



Ervick, Michael (2012)
DIKW Perspective

ODI Certificates: badge levels

- **Bronze:** data is openly licensed, available with no restrictions, accessible and legally reusable.
- **Silver:** satisfies the Bronze requirements, the data is documented in a machine readable format, reliable and offers ongoing support from the publisher via a dedicated communication channel.
- **Gold:** satisfies the Silver requirements, is published in an open standard machine readable format, has guaranteed regular updates, offers greater support, documentation, and includes a machine readable rights statement.
- **Platinum:** satisfies the Gold requirements, has machine readable provenance documentation, uses unique identifiers in the data, the publisher has a communications team offering support. This is an exceptional example of an information infrastructure.