

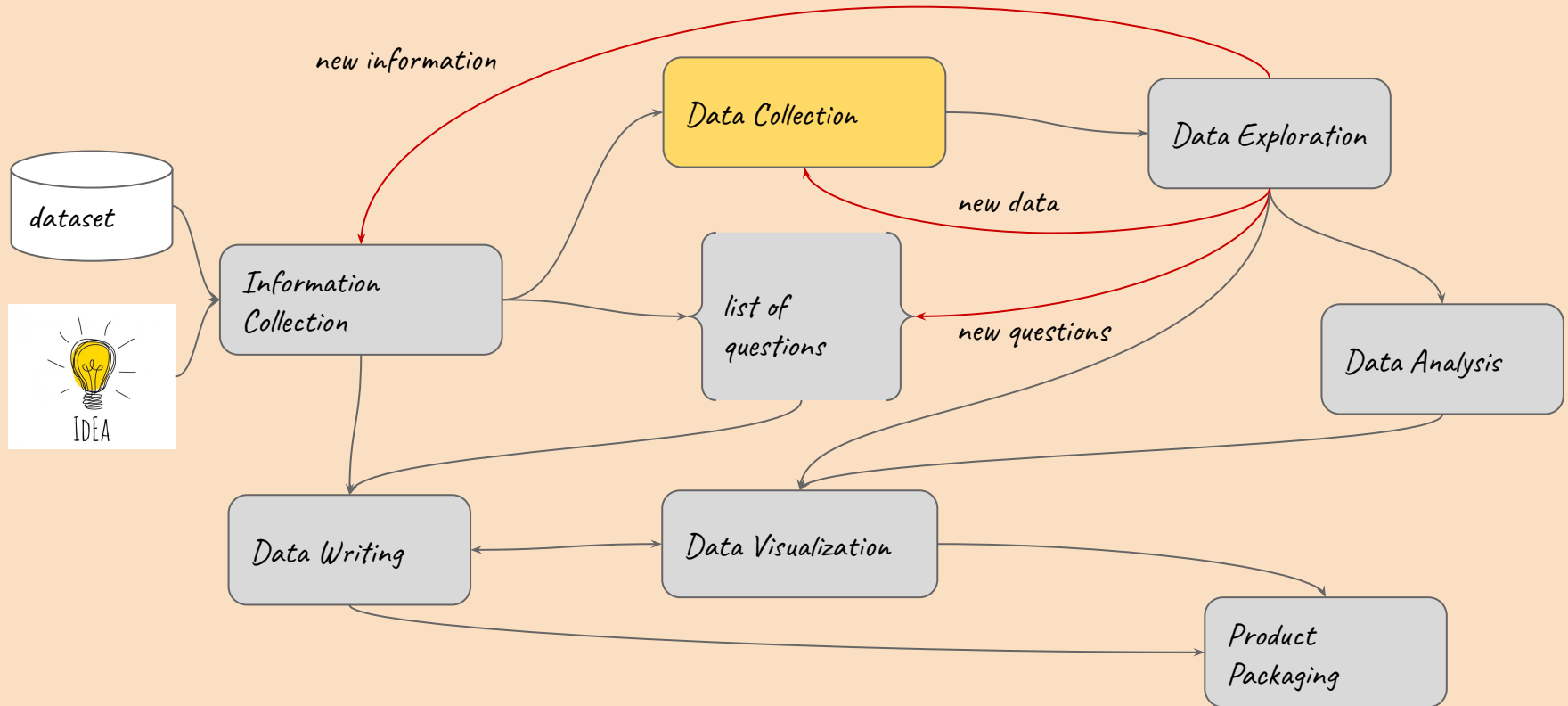


Data Journalism

Data Collection

InfoUma 2021-22 *Andrea Marchetti*

Data Collection



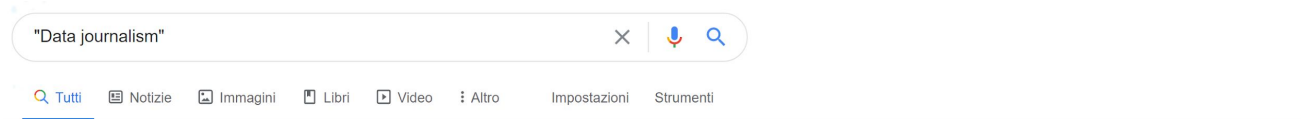
Data Collection Sources

1. Open Dataset
2. Web Scraping
3. Web Crawling through API
4. Scraping from PDF

Finding data via Google

First tool for Data Journalism is Google

<https://google.com/search?q=data+journalism>



Circa 2.310.000 risultati (0,48 secondi)

it.wikipedia.org › wiki › Data_journalism ▾

Data journalism - Wikipedia

- 1 Per **data journalism** o giornalismo dei dati s'intendono le inchieste o i reportage realizzati con gli strumenti della matematica, della statistica e delle scienze ...
Le origini · Evoluzione storica · Inchieste di data journalism · Note

datajournalism.com ▾ Traduci questa pagina

Homepage | DataJournalism.com

The world's largest **data journalism** learning community. Featuring free video courses, long reads, resources and a discussion platform.

www.classup.it › blog › data-journalism-scopriamo-cos... ▾

Data Journalism: scopriamo cos'è e chi è il Data Journalist ...

Facciamo un po' di chiarezza su cos'è il **Data Journalism** e perché si è diffuso a macchia d'olio, e approfondiamo la conseguente professione: il Data Journalist.

www.datajournalism.it ▾

datajournalism.it – dati, non (solo) parole

31 dic 2020 — "Se il più grande esercito del mondo, quello americano, ha solo quattro tipologie di navi da guerra e solo un tipo di carro armato, non ha molto ...



Knowledge Graph Node
Altre Immagini
Wikipedia Page



Le persone hanno chiesto anche

What does a data journalist do? ▼

Why is data journalism important? ▼

How do journalists learn Data? ▼

What is open data journalism? ▼

Feedback

Altre ricerche fatte,
come domande

www.mdpi.com › pdf ▼ PDF

Data Journalism Practices Globally: Skills, Education ... - MDPI

di BR Heravi · 2020 · [Articoli correlati](#)

22 ott 2020 — **Data journalism**, also known as data-driven journalism, is an emerging discipline that brings together knowledge from several disciplines, ...

www.lsd.i.it › assets › Andrea-Fama-Data-Journalism ▼ PDF

primo ebook sul data journalism in Italia - LSDI

Andrea Fama: Open Data - **Data Journalism**. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. h p://lod-cloud.net/. As of September ...

www.rivisteweb.it › download › article ▼

Data journalism «Made in Italy» - Rivisteweb

di M Antenore · 2016 · [Articoli correlati](#)

SAGGI. **Data journalism** «Made in. Italy». Un'analisi preliminare del giornalismo basato sui dati nella stampa quotidiana di Marzia Antenore, Elisabetta Trinca.

www.studiabo.it › pdf › tesi › Galanti_Data-journalism-... ▼ PDF

Comunicazione Pubblica e d'Impresa Il data journalism come ...

Il **data journalism** come nuova forma di informazione. Dai dati alla narrazione: un'applicazione pratica. Tesi di laurea in. Sociologia della comunicazione ...

5

6

7

8

9

<https://www.ascuoladiopencoesione.it> › lezione-2 › cos'!... ▼

STEP 8 – COS'E' IL DATA JOURNALISM | A Scuola di ...

In questa pillola vengono descritti i metodi e gli strumenti di lavoro di un **data journalist** e le tecniche necessarie per realizzare un buon prodotto di giornalismo ...

10

<https://medium.com> › il-fallimento-del-data-journalism-... ▼

Il fallimento del data journalism italiano: quali sono le cause ...

What conditions are the **data journalists** there working under? 5) Story telling: I've put this in twice for a simple reason — we want to encourage the best journalism ...

Ricerche correlate

🔍 data journalism **unipi**

🔍 data journalism **corso**

🔍 **cos'è** il data journalism

🔍 **esempi** di data journalism

🔍 data journalism **esempi**

🔍 data journalism **italiano**

🔍 data journalism **italia**

🔍 data journalism **wikipedia**

Ricerche Correlate



1 2 3 4 5 6 7 8 9 10

Avanti

Oltre la barriera dei
10 risultati

Tips and tricks for Google Search

1. Exact search (e.g. [open data](#) vs [“open data”](#))
2. Search with a specific extension (pdf, csv, xlsx) :
 - [Data journalism filetype:pdf](#)
3. Search inside a site
 - [covid-19 site:ru.wikipedia.org](#)



Find the right keyword

Prigioni, Carceri, Istituti penitenziari

Definition

A **prison**^[a] (also known as a **jail**^[b] or **gaol** (dated,^[c] *British* and *Australian English*), **penitentiary** (*American English*), **detention center** (or *centre* if outside the US),^[d] **correction center** (*American English*), **correctional facility**, **lock-up**^[e] or **remand center**) is a facility in which *inmates* (or prisoners) are *forcibly* confined and denied a variety of freedoms under the authority of the *state*.

Google Trends

Esplora cosa una regione sta cercando
Confronto tra ricerche

[Joe Biden vs Donald Trump in USA](#)

Quanto è popolare una ricerca?



Open Dataset

Open Data Sources

- **User generated data:** [Wikipedia](#), [DBPedia](#), [Wikidata](#)
- **Open Government Data:** [data.gov](#), [data.gov.uk](#), [data.gov.it](#)
- **Statistics Institutes:** [istat](#), [eurostat](#), [worldbank](#), [oecd](#)
- **Open Data aggregators:** [awesome dataset](#), [cooldataset](#),
[Kaggle](#)
- **Open Data search engines:** [Google Data Search](#), [Google trends](#)

User Generated Data

[Medaglie vinte dall'Italia alle Olimpiadi](#)

[Pandemie nella storia](#)

Carceri in Italia

- [Carcere, Prigione](#)
- https://www.penitenziaria.it/indirizzi_carceri.asp
- <https://www.poliziapenitenziaria.it/wp-content/uploads/2019/05/elenco-carceri-italiane.pdf>
- https://www.giustizia.it/giustizia/it/mg_2_3_2.page

Open (Government) Data

data.gov

data.gov.uk

data.gov.it

datiopen.it (sistemi territoriali)

dati.lazio.it dati.toscana.it ... (open data regionali)

Statistics Institutions

Istat

Eurostat

U.S. Census Bureau

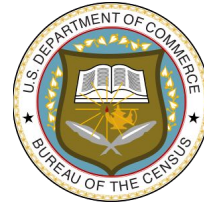
World Bank (ONU)

Worldometer (World Population Clock)

Oecd (Organizzazione per la cooperazione economica dei paesi sviluppati)



EUROPEAN
STATISTICAL
SYSTEM



Web Scraping

Web Scraping

Il **web scraping** (detto anche **web harvesting** o **web data extraction**) è una tecnica di estrazione di dati da un sito web per mezzo di programmi software che appartengono alla famiglia dei **bot**.

Un esempio di web scraping è strettamente correlato **all'indicizzazione** dei siti Internet effettuato dai motori di ricerca - crawler

Il web scraping si concentra nell'estrarre **dati non strutturati** presenti nella pagine HTML e immagazzinarli in **database**

Principali siti attaccati

- agenzie immobiliari (immobiliare.it)
- agenzie di viaggio (booking.com)
- commercio elettronico (amazon.it)
- motori di ricerca (google.com)
- siti di scommesse (bet.com)

...

Eterna lotta tra web scraper e web master

Metodi per prevenire il web scraping	Metodi per continuare a fare web scraping
Utilizzare Robots Exclusion Standard (Googlebot è un esempio) per bloccare i bot che dichiarano la loro identità (a volte lo fanno usando stringhe degli user agent)	Impostare un user agent accettato
Monitorare l'eccesso di traffico.	Inserire dei ritardi magari random nel codice
Utilizzare tool come CAPTCHA	Utilizzare servizi di terze parti che manualmente risolvono i quiz esposti
Aggiungere piccole variazioni di HTML/CSS per circondare dati importanti ed elementi di navigazione	Usare espressioni regolari per bypassare l'HTML
Monitorare il traffico da un particolare IP	Usare dei servizi proxy

Aspetti legali ed economici

Ubiquity and danger: The web scraping economy

(31/08/2016)

“If your content can be viewed on the web, it can be scraped,”
said Rami Essaid, CEO of Distil Networks

58.000\$ annui la paga media di un web scraper ma può raggiungere i 128.000\$

Metodi di Web Scraping

- Alcuni siti web sono popolati con chiamate web api a Database e i risultati sono wrapped in html
 - Sfruttare le web api e ottenere i dati in json o xml
 - Tecniche di de-wrapping
- **Dom Parsing**
- **Text Pattern matching**
- **Computer Vision + Machine Learning**

Web Scraping Tools Bibliography

[The 10 Best Data Scraping Tools and Web Scraping Tools](#)
(31/12/2019)

[Top 7 Python Web Scraping Tolls for Data Scientists](#)
(12/11/2019)

[5 Tasty Python Web Scraping Libraries](#)

[Selenium Vs Scrapy](#) (15/12/2018)

Web Scraping Tools

Librerie

[Beautifulsoap](#) (python)

[Scrapy](#) (python)

[Selenium](#) (python)

[Puppeteer](#) (javascript)

Browser Extensions (semplici)

[Web Scraper](#)

[Grepsr](#)

Servizi a pagamento (puntano sulla semplicità a sull'efficienza)

[Scraper api](#)

[ScrapeSimple](#)

[Octoparse](#)

[ParseHub](#)

Web API

What

Molti **social media** mettono a disposizione degli utenti registrati delle funzioni per accedere ai loro dati

- Twitter (post)
- Flickr (immagini)

Altri non lo fanno

- Facebook (post)
- Instagram (immagini)

Why

Sono un termometro delle inclinazioni della massa

PDF Scraping

Why

In modo molto anacronistico molte PA continuano a fornire i propri dati come file PDF