DW Akademie

Federal Foreign Office

Webinar 2: How to get data

Once you have your data story idea, you'll have to find data. But where to start?

Types of data sources

	Deliberate disclosure	Unintended disclosure
Active acquisition	**FOIA**	**Scraping**
Passive acquisitio	**Open Data**	**Leak**

Open data

Great, because: easy to get, freely available

DATA FROM AUTHORITIES

Exists on all levels:

- **international**
- **national**
- **regional**
- **local**

INTERNATIONAL DATA SOURCES

- United Nations: UNdata
- World Health Organization: WHO Data
- International Labour Organization: ILO Data
- World Bank: World Bank Open Data
- Organisation for Economic Co-operation and Development: OECD Data
- Eurostat: Database

and many more

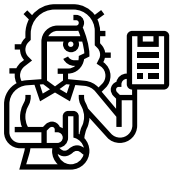
NGOS

Our World in Data
Gapminder

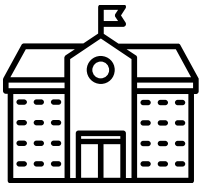
Collections of Open Data pages around the world:

- *Open Data Soft: Open Data Inception*
- *Open Knowledge Foundation (OKFN): Data Portals*

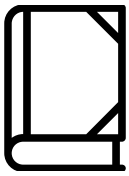
ABOUT NERDS AND SCIENTISTS



Ask scientists!

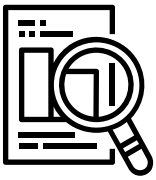


Cooperate with universities



find relevant studies

ask for the research data



ask local research labs

Freedom of Information

Many countries have Freedom of Information (FOI) laws that grant access to governmental data

- **Tip:**
  Ask nicely first, but know your rights
- **Pro:** potential for exclusive stories
- **Contra:** time-consuming of authorities don't cooperate

RESOURCES:

- Overview: Freedom of Information Laws — GIJN
- I've Sent Out 1,018 Open Records Requests, and This Is What I've Learned – ProPublica
- Data Journalism Handbook 1.0: Wobbing Works. Use it! – DataJournalism.com

Scraping

Extract data from websites if they don't have a download button

- **Example:** Automatically save a table from a Wikipedia page
- **Pro:** potential for exclusive stories
- **Con:** basic knowledge of HTML & CSS

RESOURCES:

- A web scraping toolkit for journalists – Journocode

BROWSER ADD-ONS

▶ Table Capture (Chrome): Scrape tables from websites

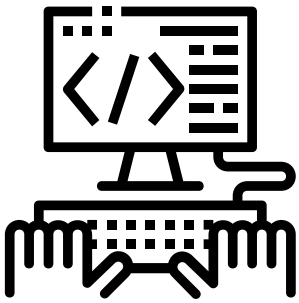
▶▶ Scraper (Chrome): Scrape any content from one page at a time.

▶▶▶ Web scraper (Chrome): Scrape MORE complex content, multiple pages at once

TOOLS WITHOUT PROGRAMMING

Often expensive!

- Octoparse ..... **Freemium**
- Luminati ..... Not free
- Scrapinghub - Scrapy Cloud ..... **Freemium**
- Dexi.io ..... Not free



Alternatively:
Learn to code!

Leaks

Whistleblowers may leak data to the public or the press

- **Pro:** potential for exclusive stories
- **Con:** requires investigative rapport, hard to plan for

Tips for smarter research

As with all research, finding data involves asking:

WHO MIGHT KNOW THIS?

- Demographic data ..... → *Statistical offices, ministries, cities and municipalities*
- Economic data ..... → *OECD, WorldBank, ...*
- Health data ..... → *WHO, ...*
- Satellite data ..... → *NASA, ESA, ...*
- …

Search engines

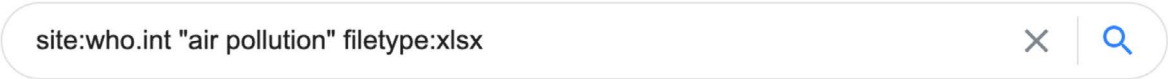
Tip: Go beyond Google! There are other search engines like Yahoo, Bing, DuckDuckGo, Yandex, …

If you do google, google smarter with search operators:

Google search operators

`"search term"`	search for exactly this expression
`-"search term"`	exclude results that contain this expression
`site:website.com`	only search on this specific website
`filetype:pdf`	search for specific file types
`jobs And gates`	search for pages with both terms
`jobs OR gates`	search for pages with either term

Example:



Suche auf der Website der WHO nach Excel-Tabellen mit dem Begriff “air pollution”.

Wenn du deine Daten findest...
...behandle sie wie jede andere Quelle

Frage dich:

- **Wer** hat diese Daten erhoben?
- **Zu welchem Zweck** wurden die Daten erhoben?
- **Wie** wurden die Daten erhoben?
- **Wie aktuell** sind die Daten?
- Kann eine **zweite Quelle** die Daten bestätigen?