

How to use open data for your research?

ປດໄດ. europo academy 24 May 2024 10.00 – 11.30 CET

Rules of the game



The workshop will be recorded



Please reserve 3 min after the workshop to help us improve by filling in our feedback form



For questions, please use the ClickMeeting chat



Introduction







Flora Kopelou Data.europa.eu Publications Office of the EU

Hyo Mee Duerinck Senior Consultant in Enterprise Data & Analytics at Capgemini Invent Daniel Garijo Researcher at Universidad Politécnica de Madrid





	10.00 - 10.05	Opening and introduction – Flora Kopelou						
	10.05 - 10.20	Search data.europa.eu like a pro – <i>Flora Kopelou</i>						
	10.20 - 10.40	A case with open data from data.europa.eu – <i>Hyo Mee</i> <i>Duerinck</i>						
	10.40 - 11.00	Open Data and the FAIR principles: A use case in Citizen Science – Daniel Garijo						
	11.00 - 11.15	Q&A						
	11.15 – 11.20	Closing remarks – <i>Flora Kopelou</i>						
doto. europo academy								

Search data.europa.eu like a pro





What is open data

5



What is open data?

Data that can be **freely accessed**, **used**, **re-used**, and **shared** by anyone for any purpose, without restrictions from copyright, patents, or other mechanisms of control.



Availability and access

Reuse and distribution

Universal participation

6

doto. europo academy

What is data.europa.eu

7



- The official portal for European data
- Managed by the Publications Office of the European Union, European Commission
- Supporting data supply and data reuse across Europe and beyond



data.europa.eu has been set up by the European Commission to implement EU open data and reuse policies under the legal acts adopted by the EU institutions and EU Member States.



The portal is a bridge between the data providers and data users

Data providers

Data users



Our services in a nutshell



Data

Providing access to free public data resources across Europe via a single platform (the portal).

Academy

Improving data literacy with thematic courses designed according to users' needs.

Community

Organising conferences; communicating via social media and newsletters.

Publications

Assessing open data maturity in Europe; providing reports, studies and training via data.europa academy.

data.europa.eu connects:





Data.europa.eu as data hub





Data.europa.eu as data hub

Jata.

academy



Data.europa.eu as data hub

- More than 1.7 million datasets, from 183 data providers
- Navigate or search to get to the data you are looking for
- Benefit from many filters
- Metadata translations in all EU languages

data

academy

CUI

• **Export citation** in several formats

European Union			-	Login	🕀 English				
European data data.europa.eu The official portal for European data									
Home Data Y Academy Comm	unity Y Publications V Documentation 년								
Home > Datasets									
Datasets									
Publisher ©	Search	0	Data	sets	୰				
- Any -	Datasets found (1 671 091)	Sor	rt by:	Relevance	~				
- Any - History of the political groups of the National Assembly (since 2012) - Information and statistics									
- Any - This dataset presents all the political groups of the French National Assembly since the 14th legislature (2012). The data comes from open data from the National Assembly. This new game brings together data on politica									
Categories [®]	CSV								
- Any -	Plate	eforme ou	uverte de	s données publi	ques françaises				
Keywords ©	History of deputies of the National Assembly Information and statistics	/ (sir	ice 2	.002) —					

Data.europa.eu as information hub





Information around open data

Data stories Podcasts Studies News Newsletter Events



doto. europo academy

Data.europa.eu as knowledge hub





Data.europa academy

your open data knowledge hub

- Improving data literacy with 11 thematic courses (legal, technical, business, data visualisation...)
- Designing according to users' needs
- Offering training material for public sector, private companies, students, researchers, etc.





Case with open data from data.europa.eu





Case study

Imagine yourself as part of a team of data analysts working closely with the EU to evaluate the midwife and nurse coverage in the EU. Your role is to utilize datasets from Eurostat, the official statistical office of the EU, to perform a thorough data analysis and provide actionable insights.

Task 1: Finding the dataset

Your first task will be to source the necessary dataset using <u>data.europa.eu</u>. Specifically, you need to locate the dataset that contains information on the inhabitants per <u>Nursing and caring professionals</u>, from the 80's until 2021. Use the <u>access</u> link to download the dataset in <u>csv format</u>. Then convert the csv to Excel.

Task 2: Understanding the data

Now that you have the dataset, let's see if it contains the data you expected for our analysis. To check this, you need to understand what data is inside the dataset and how you can interpret data. Use the landing page link to guide you to the source website to find metadata.

Task 3: Answering the questions

Now that you understand the data you have at hand, you have been given the task to answer three questions. Check if the dataset contains what you need to start your analysis for the following questions.

- 1. What are the top 3 and bottom 3 EU countries in terms of midwife coverage in 2020?
- 2. How did the midwife coverage in Belgium change over the last 40 years?
- 3. Can a similar trend be found for nurse coverage as was found for midwife coverage in Belgium?

doto. europe academy





Feel free to ask your questions in the chat



Task 1: Find the dataset

European Union		L	ogin 🌐 English			Dataset Nursing and caring profe	essionals - hi	istorical data	a (1980-
European data						Eurostat Publisher	: Eurostat		Updated: 12.07.2022
data.europa.eu The official po	rtal for European data					Dataset Quality Similar datasets		Dataset feed	Linked data • Cite • Embed
Home Data 🗸 Academy Com	munity \checkmark Publications \checkmark Documentation \mathbb{C}					Nursing and caring professionals - historical data (1980-2021		Created:	05.03.2014
Home > Datasets								Updated:	12.07.2022
								Landing Pa	ge: https://ec.europa.eu/eurostat/databr owser/product/page/hlth_rs_prsns
Datasets								Publisher:	Name: Eurostat
Data scope ③	Nursing and caring professionals	⑦ Datasets	s 🗸 २	2				Show More	×]
- Any - 🗸 🗸		Sort but Do				Distributions (5)			_
Provenance ©	Datasets found (882 968)	Soft by.				Link to the data	Format	Updated	Actions
- Any -	Nursing and caring professionals - historica	al data (1980)-2021)			Download dataset in SDMX 2.1 format Show more V	XML	UNKNOWN	Access ✓ Linked data ✓ Validate
Publisher [®]	Nursing and caring professionals - historical data (1980-2021)			3	4	Download dataset in SDMX-CSV format	CSV	UNKNOWN	Preview Access v Linked data v Validate
- Any -	XML TSV HIML CSV					Snow more 🗸			
Catalogues (?)			Eurostat	1		Download dataset in TSV format Show more v	TSV	UNKNOWN	Access ✓ Linked data ✓ Validate



YouTube: <u>'How to Extract tar.gz File In PC/Laptop Using 7zip [EASY]'</u> How to extract a GZ file \longrightarrow

How to convert a CSV file to Excel — YouTube: <u>'How to Convert CSV to Excel (Simple and Quick)'</u>



Task 2: Understand the data

Task 2: Understanding the Data

Now that you have the dataset, let's see if it contains the data you expected for our analysis. To check this, you need to understand what data is inside the dataset and how you can interpret data. Use the <u>landing page</u> link to guide you to the source website to find metadata.

doto. europo academy

Dataset Nursing and caring professionals - historical data (1980- 2021)							
Eurostat	Publisher: Eurost	at		Updated:	12.07.2022		
Dataset Quality Sim	ilar datasets	Dataset feed Linked data - Cite - Embed					
Nursing and caring professionals - h	storical data (1980-2021)		Created:	C	5.03.2014		
			Updated:	1	2.07.2022		
			Landing Pa	ge: h	ittps://ec.eu wser/produ	ropa.eu/euros ct/page/hlth_r	tat/databr s_prsns
			Publisher:	١	lame: Euro	stat	
			Show More	~			
Distributions (5)							
Link to the data		Format	Updated	Actions			
Download dataset in SDMX 2.1 f Show more v	ormat	XML	UNKNOWN		Access 🗸	Linked data 🗸	Validate
Download dataset in SDMX-CSV	format	CSV	UNKNOWN	Preview	Access 🗸	Linked data 🗸	Validate
Download dataset in TSV format	:	TSV	UNKNOWN		Access 🗸	Linked data 🗸	Validate

Task 2: Understand the data

Dataset Nursing and 2021)	caring professionals - I	nistorical da	ita (1980-	eurostat	Data	a Browser	
Eurostat	Publisher: Eurostat		Updated: 12.07.2022	ALL DATA RECENTLY UPD	ATED DOWN	ILOADS	
Dataset Quality Simila	r datasets	Dataset fe	eed Linked data - Cite - Embed	Is this your first visit to the 'Data brow	vser'? Let us brie	fly introduce it and its functionalities	to you. View presentation Don't ask me again Ask me again at the next visit
ursing and caring professionals - histo	vrical data (1980-2021)	Created:	05.03.2014	Survive data and caring	professior	nals - historical data (1	980-2021)
		Updated	: 12.07.2022	Source of data:	hith_rs_prsn Eurostat	IS	
		Landing	Page: https://ec.europa.eu/eurostat/databr owser/product/page/hlth_rs_prsns	Last data update: Last structure update:	12/07/2022 03/01/2024	11:00 (1 year ago) 23:00	
		Publishe	r: Name: Eurostat	Overall data coverage: Number of values:	1980 - 2021 34 986		
		Chan M		Data navigation tree location	Population a	nd social conditions > Health > Hea	alth care > Health care resources > Health care resources - historical data
Distributions (5)		Show M			DI	EFAULT	FULL DATASET (hl
Link to the data	Format	Updated	Actions	Activity and employment statu [wstatus]	s 🔛 (3/3)	Search by code and label Type to filter (special filt	ter with ? or *)
Download dataset in SDMX 2.1 form	mat XML	UNKNOWN	Access ✓ Linked data ✓ Validate	Geopolitical entity (reporting) [geo]	(38 / 38)	<pre>[0C222_322_5321_5322]</pre>	Nurses, midewives, health care assistants and home-based personal care workers
Download dataset in SDMX-CSV fo	rmat CSV	UNKNOWN	Preview Access → Linked data → Validate	International Standard Classifi [isco08]	cation 🔡 (9/9)	[OC222_322]	Nurses and midwives
Download dataset in TSV format	TSV	UNKNOWN	Access ~ Linked data ~ Validate	Time [time]	(42/42)	[0C222_3222]	Nursing professionals and midwives
Snow more 🗸				Time frequency		[0C2221]	Nursing professionals
loto				Unit of measure	25	[OC2222_3222]	Midwives
JUICI.				[unit]	(3/3)	[OC3221]	Nursing associate professionals
2Uron	0					[OC5321_5322]	Health care assistants and home-based personal care workers
cadomy						[0C5321]	Health care assistants
caueiny							



Task 3: Answer the questions

Based on the data provided in the attached Excel file, please determine if you can answer the following research statements. For each statement, indicate whether the information in the Excel file <u>is sufficient</u> to answer the question.

Statement 1: I have the information to identify the top 3 and bottom 3 EU countries in terms of midwife coverage in 2020.

- Yes, I have the information to answer this question
- □ No, I do not have the information to answer this question

Statement 2: I have the information to research how midwife coverage in Belgium has changed between 1980 and 2021.

- Yes, I have the information to answer this question
- □ No, I do not have the information to answer this question

Statement 3: I can research whether a similar trend is found for nurse coverage as was found for midwife coverage in Belgium.

- Yes, I have the information to answer this question
- □ No, I do not have the information to answer this question

Please review the data carefully and provide your answers based on the information available in the Excel file.

doto. europo academy

What are the top 3 and bottom 3 EU countries in terms of midwife coverage in 2020?



europa

academy

- Belgium, Serbia and Turkey has the highest midwife coverage
- Spain, Italy and Luxembourg has the lowest coverage
- The top 3 countries have about 6 time less inhabitants per midwife than to bottom 3 countries

How did the midwife coverage in Belgium change over the last 40 years?



- The coverage more than doubled over the last 40 years
- The doubling in coverage happened in the 10 years between 1998-2008

doto. europo academy

Can a similar trend be found for nurse coverage as was found for midwife coverage in Belgium?





- Nurse coverage doubled in 10 years over 2008 and 2018
- In 2019 nurse coverage data shows a halving in one year
- Availability of data on nurses does not go back in time as far as the midwife data

doto europo academy







Open Data and the FAIR principles: A use case in Citizen Science

Esteban González (OEG - UPM) Daniel Garijo (OEG - UPM) daniel.garijo@upm.es @dgarijov • This work is licensed under the license CC BY-NC-SA 4.0 International

- <u>http://purl.org/NET/rdflicense/cc-by-nc-sa4.0</u>
- You are free:
 - to Share to copy, distribute and transmit the work
 - to Remix to adapt the work
- Under the following conditions
 - Non-commercial You cannot use it for commercial purposes, nor for training inside a commercial company
 - Attribution You must attribute the work by inserting
 - "[source http://www.oeg-upm.net/]" at the footer of each reused slide
 - a credits slide stating: "These slides are partially based on "Open Data and the FAIR principles: A use case in Citizen Science" By Esteban Gonzalez and Daniel Garijo
 - Share-Alike





The importance of Open Data in Science





Retraction Watch Tracking retractions as a window

into the scientific process

Scientific publications are changing

Traditional Published Articles

Text: Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Modern Published Articles

Text: Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Data: Supplementary materials, pointers to data repositories

Source: The Scientific Paper of the Future. Yolanda Gil et. al. http://dx.doi.org/10.5281/zenodo.159206

Reproducible Articles

Text: Narrative of method, the data is in tables, figures/plots, and software mentioned

Data: Supplementary materials, pointers to data repositories

Software: Data preparation, data analysis, and visualization

Provenance and methods: Workflow/scripts describing dataflow, codes, and parameters

Ontology Engineer ingGroup

Ever heard about FAIR?



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

FINDABLE

- F1: Global PIDs
- F2: Use rich metadata
- F3: Metadata include data identifier
- F4: Metadata indexed in a searchable engine



Source: https://www.go-fair.org/fair-principles/

FINDABLE

ACCESSIBLE

- F1: Global PIDs
- F2: Use rich metadata
- F3: Metadata include data identifier
- F4: Metadata indexed in a searchable engine

A1: Standardize communications protocol

- A1.1: Open protocol
- A1.2: Protocol with authentication
- procedure
- A2: Metadata accessible

Terminal Archivo Editar Ver Buscar Terminal Ayuda jimmy@KEVIN:~\$ curl -I http://192.168.1.60 HTTP/1.1 200 OK Server: nginx/1.14.0 (Ubuntu) Date: Fri, 13 Jul 2018 16:28:23 GMT Content-Type: text/html Content-Length: 612 Last-Modified: Fri, 13 Jul 2018 15:17:19 GMT Connection: keep-alive ETag: "5b48c27f-264" Accept-Ranges: bytes

https + content negotiation

Source: https://www.go-fair.org/fair-principles/

FINDABLE

ACCESSIBLE

- F1: Global PIDs F2: Use rich metadata F3: Metadata include data identifier
- F4: Metadata indexed in a searchable engine
- A1: Standardize communications protocolA1.1: Open protocolA1.2: Protocol with authentication procedureA2: Metadata accessible

INTEROPERABLE

I1: Meta(data) use a formal language for knowledge representation
I2: Meta(data) use vocabularies that follow FAIR principles
I3: Meta(data) include references to other metadata



Source: https://www.go-fair.org/fair-principles/

FINDABLE

ACCESSIBLE

F1: Global PIDsF2: Use rich metadataF3: Metadata include data identifierF4: Metadata indexed in a searchable engine	A1: Standardize communications protocolA1.1: Open protocolA1.2: Protocol with authentication procedureA2: Metadata accessible			
INTEROPERABLE	REUSABLE			
I1: Meta(data) use a formal language for knowledge representationI2: Meta(data) use vocabularies that followFAIR principles	R1: Meta(data) are richly described R1.1: Meta(data) use a clear and accesible license R1.2: Meta(data) include detailed			

13: Meta(data) include references to other metadata

provenance R1.3: Meta(data) meet domain-relevant community standards

How data.europa.eu adopts the FAIR principles?

Example: https://data.europa.eu/data/datasets/244310-16?locale=en

European data	Metadata quality			
Home Data V Academy Community V Publications V Documentation ^[2]	on which metrics we use for indicator measurements, please l			
Home > Datasets > % 25-49 years old	Accessibility			
Dataset % 25-49 years old Image: data.gov.be Publisher: Waals Instituut voor Evaluatie, Prospectie en Statistiek (IWEPS) Updated: 04.07.2023	Download URL 100% Most frequent a			
Dataset Quality Similar datasets Dataset feed Linked data Cite 	Reusability			
Careful: FAIR is not always a	Access restrictions true License information			
But quality data must be FAIR	Publisher true			

Open Data and Citizen Science: Why is FAIR needed?

Participation of citizens in the research lifecycle

- Producing data
- Data Analysis

Observations are done by **amateur citizens**, not domain experts

- More data, but less quality

How does quality vary across use cases?

Human-based observations: Globe At Night



- Objective: raise public awareness of the impact of light pollution by inviting citizenscientists to measure their night sky brightness and submit their observations
- Citizens need astronomical expertise to identify the stars.
- More than 180,000 measurements have been contributed from people in 180 countries over the last 12 years
- Open Data & Open Access

Hybrid observations: Loss of the Night



- Loss of the Night App challenges citizen scientists to identify as many stars as they can in order to measure light pollution.
- Uses software to guide annotating stars
- Evolution of Globe At Night initiative
- Open data and Open Access

Semi-automated observations: Dark Sky Meter



- Contributes to a global map of nighttime light pollution using the lphone's camera.
- Uses an external weather forecast service.
- Open data is available in Globe At Night initiative.
- Paid application

Fully automated observations: Nixnox



Manual Process. Duration: 30 minutes (experts) and results with mistakes



Automatic process: 1 - 2 minutes



https://tess.stars4all.eu

Credits: Nixnox Project

Thank you to an amateur initiative. Based on collaboration!

Ontology Engineer ingGroup

Citizen Science Photometer Network



The role of FAIR in Citizen Science

FAIR:

- Each dataset has its own persistent identifier
- Data follow standard specifications
- Quality **metadata** (provenance)
 - Key for integration!

Citizen Science



- Gathering data
 - How to improve data quality
- Data analysis

FAIR + Citizen Science = Reproducible Science!

Side effects of FAIR and Open Data: Cities At Night



Astronauts from the ISS have produced thousand of images.

Objective: Citizens participation to classify images



Results were used in a thesis! and published in Zenodo

Credits: NASA & Cities At Night Project

Ontology Engineer ingGroup







Open Data and the FAIR principles: A use case in Citizen Science

Daniel Garijo (OEG - UPM) Esteban González (OEG - UPM)

Q&A





Flora Kopelou Data.europa.eu Publications Office of the EU

Hyo Mee Duerinck Senior Consultant in Enterprise Data & Analytics at Capgemini Invent



Daniel Garijo Researcher at Universidad Politécnica de Madrid

49



Please provide your feedback!







Stay up-to-date on our activities!

Unin. europo academy

Join our next webinars!



Exploring the potential of high-value datasets: connecting with real use cases



data. europo academy

14 June 2024

10.00 - 11.30 CET

Stories from the Use Case Observatory -Volume 2

data. europa academ



The European Drug Report: using an open data approach to improve data visualisation

data. europo





Thank you



