

# How to use open data for your research?

Unin. europn academy

25 October 2024 10:00 – 11:30 CEST

# Rules of the game



The webinar will be recorded and published on the data.europa academy



For questions, please use the ClickMeeting chat.



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

# Introduction



Flora Kopelou

Data.europa.eu

**Publications Office** 

of the EU

Jim Rovekamp, Senior Consultant Data Strategy, Capgemini Invent



Calum Inverarity Senior researcher, Open Data Institute



Neil Majithia Researcher, Open Data Institute



Maria Ioanna Maratsi PhD Researcher, University of the Aegean



Mohsan Ali PhD Researcher, University of the Aegean





# Agenda

academy	
11.20 - 11.30	Q&A and closing remarks – <i>Flora Kopelou</i>
11.00 – 11:20	Open Data for Research & Research for Open Data – <i>Maria Ioa<mark>nna Marats</mark>i &amp;</i> <i>Mohsan Ali</i>
10.45 – 11.00	Using open data for genuinely interesting research projects - Neil Majithia
10.40 - 10.45	The State of UK Open Data: from big bang open data to responsibly stewarding data with a purpose – <i>Calum Inverarity</i>
10.20 - 10.40	A case with open data from data.europa.eu – <i>Jim Rovekamp</i>
10.10 - 10.20	Search data.europa.eu like a pro – <i>Flora Kopelou</i>
10.00 - 10.10	Opening and introduction – <i>Flora Kopelou</i>

# Search data.europa.eu like a pro





## What is open data

6



### 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





## What is data.europa.eu

8



A platform providing central point of access to **European open data** from international, European Union, national, regional, local and geodata portals.



The platform was set up by the European Commission to implement EU open data and reuse policies. It is managed by the Publications Office of the EU.



### 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



# 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
- **Export citation** in several formats



Paquets Arome Outre-Mer - Guyane - Résolution 0,025°

# Data.europa.eu as information hub





# Information around open data

Data stories Studies News Newsletter Events Podcast



# 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.
- Soon: Working on creating learning paths for our users



	European Union	Log in	S	earch		Search
	data.europa.eu The official portal	for European				
	Home Data 👻 EU Open Data Days	Academy Community	/ ~ Publications		entation C	
	Home > Academy					
						_
E		Search results	(11)	Sort by	Published on	_
	Target audience		E-learning   Reading   Tools Introducing open dat Are you looking to understar course is for you! Our e-lear Show more	Videos ta nd the concept an ming modules will	Course d the benefits of open data? This introduce you to the concept of	
	Theme Select an item	~	Content: Theme:	5 lessons	YOLICY QUALITY	
	Level Select an item	~	Audience:	Academia ( governmen	Civil servants Journalists Non- tal organisations Private sector	
	Format Select an item		Level: E-learning   Reading   Tools	Videos	Course	
	Apply	E A	Understanding the le Are you looking to understan publication and reuse of ope Show more ~	gal side of op nd how legislation en data? This cour	and regulations can impact the se is for you! Our e-learning	
			Content:	4 lessons		
			Theme:	LEGAL	DFICA	
			Audience:	Academia ( Private sec	Civil servants Data providers for	
			Level:	Beginner		

# Case with open data

Electricity prices by user type





# **Case study**

Imagine yourself as part of a team of data analysts working closely with the EU to evaluate electricity prices by user type over the period from 2013-2023. 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 electricity prices in the EU for the given period. 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. How did electricity prices for medium sized household in Germany develop from 2013-2020?
- 2. Which country had the highest average electricity price for medium sized households in 2023?
- 3. Is there a correlation between the average electricity price and a countries' GDP?

### doto. europc academy



:

Feel free to ask your questions in the chat



# Task 1: Find the Dataset

				Dataset Electricity pric	es by type of user		
icial website of the European Union How	do you know? 🗸		i.	Eurostat	Publisher: Eurostat	U	odated: 09.10.2024
European Union		💄 Login 🛛 🔀 English		Dataset Quality Similar da	itasets	Dataset feed	Linked data - Cite - Embed
European data data.europa.eu The official pe Home Data Y t <sup>2</sup> U Open data d Home > Datasets	ortal for European data ays Academy Community v Publications v Do	cumentation 🕑		This indicator presents electricity prices cha non-household consumers are defined as f kWh without taxes applicable for the first se industrial consumers (Consumption Band la 2000 MWh). Electricity prices for household national price in Euro per kWh including tax semester of each year for medium size hou with annual consumption between 2500 an	arged to final consumers. Electricity prices for ollows: Average national price in Euro per imester of each year for medium size c with annual consumption between 500 and d consumers are defined as follows: Average tes and levies applicable for the first isehold consumers (Consumption Band Dc d 5000 kWh).	or Created: J Updated: Landing Page Publisher:	11.09.2012 09.10.2024 : https://ec.europa.eu/eurostat/dat owser/product/page/ten00117 Name: Eurostat
HVDs only	electricity prices by type of user	⑦ Datasets v Q	:	Distributions (5)		Show More	•
igh-Value-Dataset category ⑦	Datasets found (1 469 647)	Sort by: Relevance		Link to the data	Format	Updated	Actions
			3	Bownload dataset in SDMX 2.1 format Show more v	it (CSV)	UNKNOWN	Preview Access V Linked data V Valio
				Download dataset in TSV format	TSV	UNKNOWN	Access ✓ Linked data ✓ Vali

### doto. europo academy



# 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 Electric	ity prices by type of user	
Eurostat	Publisher: Eurostat	Updated: 09.10.2024
Dataset Quality	Similar datasets	Dataset feed Linked data ▼ Cite ▼ Embed

This indicator presents electricity prices charged to final consumers. Electricity prices for non-household consumers are defined as follows: Average national price in Euro per kWh without taxes applicable for the first semester of each year for medium size industrial consumers (Consumption Band Ic with annual consumption between 500 and 2000 MWh). Electricity prices for household consumers are defined as follows: Average national price in Euro per kWh including taxes and levies applicable for the first semester of each year of the first semester of each year for medium size household consumers (Consumption Band Dc with annual consumption between 2500 and 5000 kWh).

**Distributions (5)** 

Download dataset in SDMX 2.1 format

Download dataset in SDMX-CSV format

Download dataset in TSV format

Link to the data

Show more V

Show more V

Show more V

final consumers. Electricity prices fo Average national price in Euro per	Created:	1	11.09.2012				
of each year for medium size nnual consumption between 500 and	Updated:	(	09.10.2024				
ners are defined as follows: Average levies applicable for the first consumers (Consumption Band Dc	Landing P	age: h	https://ec.europa.eu/eurostat/databr owser/product/page/ten00117				
kWh).	Publisher:		Name: Eurostat				
	Show Mor	Show More 🗸					
Format	Updated	Actions					
XML	UNKNOWN		Access 🗸	Linked data 🗸	Validate		
(CSV)	UNKNOWN	Preview	Access 🗸	Linked data 🗸	Validate		
TSV			Access 🗸	Linked data 🗸	Validate		

# Task 2: Understand the Data

#### ten00117

Currency	
[currency]	(1/1)
Energy indicator	
[indic_en]	(2/2)
Geopolitical entity (reporting)	
[geo]	(45/45)
Products	
[product]	(1/1)
Time	
[time]	(12/12)
Time frequency	
[freq]	(1/1)
Unit of measure	
[unit]	(1/1)

Search by code and label Type to filter ( special filter with ? or * )				
<pre>[MSHH]</pre>	Medium size households			
[MSIND]	Non-household, medium size consumers			

# Task 3: Answer the Questions

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

Statement 1: I have the information to identify how electricity prices for medium sized household in Germany developed from 2013-2023?

- 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 identify which country had the highest average electricity price for medium sized households in 2023?

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

Statement 3: I have the information to identify if there a correlation between the average electricity price and a countries' GDP?

- 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 CSV file.

# How did electricity prices for medium sized household in Germany develop from 2013-2023?



Electricity prices were relatively stable until 2020, then the prices rose quickly between 2020-2023

# Which country had the highest average electricity price for medium sized households in 2023?



Liechtenstein had the highest average electricity price in 2023 for medium sized households, do you see any surprises?

# Is there a correlation between the average electricity price and a countries' GDP?



We are missing information to calculate any correlation between electricity prices and GDP. Therefore, we would have to enrich the data to draw meaningful conclusions.

# Is there a correlation between the average electricity price and a countries' GDP?



If we were to combine the two datasets (getting the GDP data from Eurostat), we can get insights into the relationship between GDP and electricity prices.

In this example we see a positive correlation of 0.42. We would consider this moderately positive

# The State of UK Open Data: from big bang open data to responsibly stewarding data with a purpose

Calum Inverarity 25-10-2024

### Agenda

- 1. Introduction: the ODI and our model for open data
- 2. Part 1: Open government data
- 3. Part 2: Public sector / regulatory-driven data sharing
- 4. Part 3: Private-sector driven coalitions and data institutions
- 5. The future of open data





### The Data Spectrum

Small / Medium / Big data

Personal / Commercial / Government data

	Internal access	Named access	Group-based access	Public access	Anyone
	Employment contract + policies	Explicitly assigned by contract	Via authentication	Licence that limits use	Open licence
/	Sales reports	Driving licences	Medical research	Twitter feed	Bus timetable



Shared



theodi.org/data-spectrum



### Why open data matters?

The \$3tn per year **valuation of the open data market** by McKinsey in 2013 centred on the value of combining open government data with shared data held by businesses

In 2014, Lateral Economics estimated that the **value of open data to the G20** would be around \$2.6tn a year, contributing to the group's cumulative gross domestic product (GDP) of around 1.1% from 2014–2019

In 2020, the European Data Portal estimated that the **value of open data for the EU28+** was €184bn in 2019, and forecast it to reach between €199.51 and €334.21bn by 2025

#### TfL's free open data boosts London's economy

"This new research from Deloitte

backs our strong belief that

transparent and free-to-access

can be massively beneficial for

both London and the wider

Managing Director of Customers, Comm

economy"

Vernon Everitt

and Technology at TfL

providing data in an open,

13 October 2017

Research by Deloitte shows that the release of open data by TfL is generating annual economic benefits and savings of up to  $\pm$ 130m a year

- Media
- Press releases

- Customers, road users, London and TfL itself all benefit
- More than 80 data feeds now available for developers through the free unified API, which ensures accurate real-time data is available from one system for over I3,000 developers

The provision of free, accurate and real-time open data by TfL is helping London's economy by up to £I30m a year, new research reveals.

The research, commissioned by TfL and conducted by Deloitte, shows that by providing of

### Open data: Unlocking innovation and performance with liquid information

By James Manyika, Michael Chui, Diana Farrell, Steve Van Kuiken, Peter Groves, and Elizabeth Almasi Doshi Share

Economic Impact of Open Data

Open data—public information and shared data from privat —can help create \$3 trillion a year of value in seven areas c global economy.

pen data—machine-readable information, particularly government data, that's made available to others—has generated a great deal of excitement



### Agenda

- 1. Introduction: the ODI and our model for open data
- 2. Part 1: Open government data
- 3. Part 2: Public sector / regulatory-driven data sharing
- 4. Part 3: Private-sector driven coalitions and data institutions
- 5. The future of open data





### **Retrospective: Open data in the UK since 2010**



provide a link to this licence;
### Annual economic benefits and savings upto:

# £130 Million

# The cost? £1 Million

https://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf

Assessing the value of TfL's open data and digital partnerships

July 2017

Deloitte.



# Over 700 Apps now rely on open transport

- City Mapper (Millions of users)
- Google Maps (Millions of users)
- Real time trains (500,000 users)
- Station master (10,000's users)





### Agenda

- 1. Introduction: the ODI and our model for open data
- 2. Part 1: Open government data
- 3. Part 2: Public sector / regulatory-driven data sharing
- 4. Part 3: Private-sector driven coalitions and data institutions
- 5. The future of open data





### **Open Banking: driven by regulation**

- Open data and data-sharing ecosystem
- Focused on secure sharing of current account transaction data with trusted 3rd parties
- Provides standards for open data publishing from financial institutions





### Agenda

- 1. Introduction: the ODI and our model for open data
- 2. Part 1: Open government data
- 3. Part 2: Public sector / regulatory-driven data sharing
- 4. Part 3: Private-sector driven coalitions and data institutions
- 5. The future of open data





### **Data Institutions**

- hold data on behalf of an organisation or person, or group of them, and share it with some of those who want to use it
- combine or link data from different sources and provide insights and other services back to those that have contributed data

# • maintain common data infrastructure

for a sector or domain, such as by registering identifiers or publishing open standards Organisations whose purpose involves **stewarding data on behalf of others** to create economic and social value





### Agenda

- 1. Introduction: the ODI and our model for open data
- 2. Part 1: Open government data
- 3. Part 2: Public sector / regulatory-driven data sharing
- 4. Part 3: Private-sector driven coalitions and data institutions
- 5. The future of open data





"The emergence of the latest generative AI, LLMs (large language models) and FMs (foundation models) – indeed all of modern machine learning – depends on vast amounts of data. Without data, there would be no AI."

Sir Nigel Shadbolt



Sir Nigel Shadbolt Executive Chair and Cofounder of the ODI



Thank you!



# Using open data for genuinely interesting research projects - my experience

Neil Majithia 25-10-2024

### Contents

### Context

### **Motivations**

### The data and the research

How I linked two open data sources to get insights

The insights and outcomes



### Context

- This presentation is based on an article I wrote for the ODI's Medium page, Canvas
- I'm a researcher at the ODI with an allround skillset, but my background is more in quantitative and computational methods
- All of this research was done in Python using relatively basic libraries, albeit not commonly used in data science
- This research is actually ongoing, so any feedback is appreciated

### Medium Q Search



### Open Data Science To Be Proud Of: Geographic Healthcare Analysis with Healthsites.io and WorldPop



Neil Majithia Published in Canvas · 12 min read · Feb 22, 2024

### 🏐 81 🛛 📿

⊷ ڭ ∿



Right: The Medium piece I wrote, accessible at



### **Motivations**

- 1. I was tired of my research portfolio looking dry
- But, generally speaking, this was because the open data sources I was using for my projects were dry in the first place
- 3. I wanted to change things, and do something
  - interesting with interesting open data



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	ТАХ	PTRATIO	в	LSTAT
C	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
5	0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21
e	0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43
7	0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15
8	0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93
9	0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10





# Part 1 - finding interesting data; healthsites.io

- I used to work with a professor of Catastrophe Risk and Disaster Reduction, and under another who worked in digital public health
- So I decided to look for data in the humanitarian sector using the UN Humanitarian Data Exchange
- Here, I found healthsites.io
  - Healthsites.io is the result of the Global Healthsites Mapping Project, which is a participatory program aiming to put every hospital, clinic, or other medical site onto a global map
  - The resulting dataset is a list of all health sites everywhere, with some added information about them
  - On the right here is the dataset for South Africa









# Part 2 - finding something interesting to do with the data

- A long time ago, I studied network science, which is an entire field of study dedicated to points on a map
- I could use Dijkstra's algorithm to find the most efficient paths for suppliers or blood bikers to take between hospitals
- Or I could do some cluster analysis to build an analysis of healthcare coverage
- But I decided to perform Voronoi tessellation, using the healthsites locations to create 'Voronoi regions' on the map which indicated the geographical 'catchment area' of each hospital in South Africa





# But really, is this interesting?





# Part 3 - finding interesting data for linkage; WorldPop

- I went back to the Humanitarian Data Exchange, looking to find geospatial datasets that could enhance my Voronoi tessellation by making the polygons mean more
- Now, I found Worldpop
  - The WorldPop project's mission is to collect high spatial resolution data on human population distributions using censuses and satellite imagery.
  - A main output is gridded population estimates, which are datasets that split a country into 1km x 1km squares (or 100m x 100m) and provide the estimated population within each square.
  - On the right here is the dataset for South Africa







WorldPop (www.worldpop.org School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project ... Funded by the Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00671

©2020 This work is licensed under a Creative Commons Attribution 4.0 International License

53

### Part 4 - linkage and results

- I could use Worldpop to have an understanding of the population within each Voronoi region
- That is to say that my diagram now provides insight on both the geographical area a hospital has to serve as well as the population burden it has
- So analysing the diagram will provide insights on the effectiveness of healthcare planning in a subject country





# Insights and Outcomes



### **Final results - with Senegal**

- I'm working with the creator of healthsites.io to build an implementation of this methodology for Senegal
- This is to be put in front of the Ministry of Health there, and has already been seen by the World Bank transport team, whose findings line up with findings here
- Note that healthcare planning is very good in the north of the country, in the St. Louis region, but on the inner, sub-Saharan part, hospitals are few and far between
- Improving access to these health sites and reducing their burden is therefore mandatory, especially given climate change causing increased incidence of Malaria and other disease

### Voronoi Diagram: Hospitals and Clinics Service Areas in Senegal



**Open data** enabled me to do some genuinely interesting and impactful research

Voronoi Diagram: Hospitals and Clinics Service Areas in Senegal





### How might you do the same?

- 1. Use open data portals like data.europa and HDX to find interesting open data
- Then look at the fields of study you enjoy, examine the things you feel are most important, talk to friends and colleagues, to find something interesting to do with it



Thank you!







# ODECO

# Open Data for Research & Research for Open Data

Maria Ioanna Maratsi Mohsan Ali

University of the Aegean, Greece

### 25-10-2024





# **ODECO: Towards a Sustainable OD Ecosystem**

- 4-year H2020-MSCA-ITN-2020 project: <u>https://odeco-research.eu</u>
- <u>Vision</u>: Address current and future challenges in the creation of **user driven**, **circular** and **inclusive** open data ecosystem



### **Envisioned situation**







## **ODECO Network - Beneficiaries**









# **ODECO Network - Partners**









# **ODECO Research Projects**

### These topics are the focus of 15 PhDs in ODECO









## **Open Data**

- Open data is data freely available to use, reuse, and distribute.
- Various sources of open data (OGD, geospatial data, citizen-contributed data, Scientific NGOs, International Organizations, etc.).







Icons from Freepik.com



# **Open Data for Scientific Research**

## **Open Data for Research** Research for Open Data

Research for Open Data  $\Leftrightarrow$  Open Knowledge for Research

• Interoperable and Findable data













# **Open Data for Research (1) - Case**

## Case-study: Greek Economic crisis (2009-2014) and Open Data

- Greek statistical Authority collected data from firms during Economic crisis
- Released OD on request (Under specific terms and use









# **Open Data for Research (2) - Case**

- In market-based economies often appear significant decreases of economic activity, which lead to recessionary economic crises.
- Economic crises have negative consequences for firms, as they lead to significant decrease of sales revenues:
- Firms respond by **decreasing**:
  - their production, general operational activities and expenses, personnel employment and materials' procurement, and
  - their investments in production equipment, **digital technologies**, etc., which leads to technological obsolescence.
- The reduction of investments, (especially in **digital technologies)** can have negative impact on their future competitiveness.









# **Open Data for Research (3) - Case**

- These negative consequences differ significantly among firms:
  - Some exhibit a **lower vulnerability** to the crisis, so they have fewer negative consequences,
  - Other firms exhibit a **higher vulnerability**, and have more negative consequences;
  - The competitive position of the former is significantly strengthened with respect to the latter, and finally the former are the **'winners'** of the crisis, while the latter are the **'losers'**.









# **Open Data for Research (4) - Case**







Icons from Freepik.com





# Conclusions

- Implications for research and practice, mainly to 2 research streams:
  - The growing research stream concerning the use of AI in government, by developing a novel approach for a highly beneficial use of AI/ML for support and enhancement of government critical activity.
  - The OGD research stream, by providing an approach for increasing the economic/social value generation from OGD through advanced AI/ML.

### **SPRINGER LINK**








## **Open Data for Scientific Research**

### **Open Data for Research \Leftrightarrow Research for Open Data**

Research for Open Data  $\Leftrightarrow$  Open Knowledge for Research

• Interoperable and Findable data













## The Role of Good Research Data Management

- Good research data management is essential for research collaborations. [ERUA, 2024]
- Research data should be published to encourage collaboration among varying scientific domains, while re-using existing research data.
- Data Re-use: research data discoverability identified as a big issue. *[ERUA, 2024]* Several researchers not aware of the advantages of reusing existing data.
- All universities should work on conveying the benefits and possibilities of sharing research data. The value of sharing and re-using research data needs to be made clearer.

#### **Open science practices become pivotal.**







This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569.



## **Semantically Interoperable and Linked Data**

"Semantically interoperable schemas allow information to be automatically exchanged by sharing common meanings through the use of universally accepted standards."

#### **Key aspects of semantic interoperability:**

- Standard schemas
- Domain-specific knowledge representation
- Controlled Vocabularies
- Good metadata









## Linked Sources for Scientific Data Discovery

- Linked data important for semantic search capabilities within large data collections.
- A lot of scientific research remains undiscovered due to poor linking to its respective scientific domain.

Research articles can be retrieved by keyword searches of the respective domain; difficult when:

i) Domain not explicitly mentioned in the article's metadata
ii) Multidisciplinary research
iii) Use of shared scientific methodological tools of more than one domain.







This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569.

Icons from Freepik.com



## **Case: Scientific Data Discovery on** Wikidata



- Experiment with the Greek Open Technologies Alliance (GFOSS)
- Employing ChatGPT and GPT-4, on how to facilitate the process of identifying and enriching data relationships to make scientific data discoverable.

#### An initial analysis revealed several issues:

- Inability to retrieve all (or most) relevant results using only simple SPARQL queries.
- Scholarly publications contain knowledge not captured sufficiently by the existing Wikidata codes.
- The search was made more difficult due to the lack of intermediate and semantically meaningful Wikidata codes.



### **Technology for Linked Knowledge and Open Science (1)**

#### How can Large Language Models facilitate the process of linking knowledge?

- How to improve open data provision and enhance open data repositories? How can scientific data become easier to find (discoverable) and have higher semantic value?
- A proposed technologically-facilitated methodology taking the best of machine intelligence and human expertise to improve knowledge linking in open data repositories.



The future of semantics?







This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569.



## **Technology for Linked Knowledge and Open Science (2)**





## Conclusions

- Findable scientific information improved conceptual and hierarchical structure of represented knowledge.
- Improve retrievability of available scientific work, independent of discipline, allowing for cross-domain discoverability and the identification of common areas and linking points of different scientific disciplines.
- Interlinked knowledge sources are essential for the promotion of social development, scientific research, and innovation.











# Questions?

mohsan@aegean.gr

ioanna.m@aegean.gr







## Q&A





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

Jim Rovekamp, Senior Consultant Data Strategy, Capgemini Invent





Neil Majithia Researcher, Open Data Institute



Maria Ioanna Maratsi PhD Researcher, University of the Aegean



Mohsan Ali PhD Researcher, University of the Aegean





# Stay up-to-date on our 2024 activities!

Unin. CUropo academy

#### **WEBINAR**

Data spaces: experience from the European Health and Common Energy data spaces

doto. europo academy

22 November 2024 10.00 – 11.30 CET

doto. europo academy

# Your opinion is important to us



doto. europo academy

## Thank you!

