Generation Zero
An Open Data Project about How Environment, Wealth and Welfare Shape Birth in Europe

Open Access and Digital Ethics, Monica Palmirani
Master's program in *Digital Humanities and Digital Knowledge*

Prepared by

Elena Binotti
Virginia D'Antonio
Elvira Kushlak

Introduction

"Generation Zero" investigates how environmental and socioeconomic conditions may influence fertility rates across regions in the European Union. It focuses on the intersection of pollution exposure, economic inequality, and reproductive patterns at the NUTS-2 level during the period from 2017 to 2019, a period chosen to reflect pre-pandemic norms and avoid COVID-related anomalies in fertility, pollution, income, and other variables.

The project uses only openly licensed datasets and explores how structural disadvantages—such as toxic air pollution, income poverty, and limited education—relate to regional fertility outcomes. It also examines the alignment between national-level investment in family policies and actual fertility trends across regions. By combining indicators from environmental, demographic, and economic domains, the project aims to identify spatial mismatches and clusters of disadvantage.

All data processing was carried out using the KNIME Analytics Platform. The visualisations were built with Plotly and Leaflet and are hosted on a public GitHub Pages site. The project adheres to the EU Open Data Directive (2019/1024), follows DCAT-AP metadata standards, and respects FAIR principles for data reuse.

Scenario

Across the European Union, fertility rates have remained consistently below the replacement threshold of approximately 2.1 children per woman. In 2019, the average fertility rate across the EU was about 1.53, according to Eurostat.[1] This prolonged low fertility scenario raises significant concerns about demographic stability, economic productivity, and the sustainability of social support systems such as pensions and healthcare.

Significant variations exist between individual EU countries. In 2019, France had a fertility rate of 1.86, which was among the highest in the EU, while countries such as Spain and Italy reported considerably lower fertility rates of 1.23 and 1.27 respectively.[2] These variations suggest that fertility outcomes might be significantly influenced by differing national and regional socioeconomic and environmental conditions, beyond just personal or cultural factors.

At the same time, socioeconomic conditions across the EU show wide disparities. According to Eurostat[3], around 21.7% of the EU population was at risk of poverty or social exclusion in 2019, but in some Eastern and Southern European regions, this figure reached almost 30%. Additionally, environmental conditions vary substantially,

[1] https://ec.europa.eu/eurostat/databrowser/view/DEMO_R_FIND3/default/table?lang=en
[2] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Fertility_statistics
[3] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20201016-2

especially in relation to air pollution. The European Environment Agency (EEA)[4] reported significant differences in air quality, with urban and industrial regions frequently exceeding the EU's recommended pollution thresholds. Such exposure is linked to potential negative impacts on health and reproduction.

Investment in family support policies also varies significantly between EU member states. OECD statistics from 2019 indicate that the average national spending on family and childcare support policies was around 2.2% of GDP. This varied widely, from approximately 1.1% of GDP in countries like Italy and Spain, to nearly 3.5% in France and Denmark. These discrepancies raise important questions about whether family policy funding effectively matches the realities of fertility trends at regional levels.

By integrating openly available datasets from Eurostat, the European Environment Agency, and the OECD, this project analyzes how regional fertility outcomes may be shaped by environmental burdens, socioeconomic vulnerabilities, and national policy frameworks. The goal is to identify regions where fertility outcomes align or misalign with these structural and policy factors, highlighting potential areas for targeted policy interventions.

Research questions

The analysis specifically addresses five main questions:

1. Is there a relationship between regional exposure to environmental pollutants and fertility rates?

2. How do regional socioeconomic factors—such as income levels, poverty rates, and educational attainment—influence fertility outcomes across EU regions?

3. Does national-level investment in family support and childcare align with regional patterns of fertility?

4. Can we identify clusters of regions experiencing combined environmental and economic disadvantages?

5. How are regions facing both high environmental toxicity and economic insecurity geographically distributed within the EU?

[4] https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2020-report

The project relies on a set of eight openly licensed datasets sourced from Eurostat, the European Environment Agency (EEA), and the OECD. The datasets were selected for their public availability, spatial granularity, temporal consistency, and relevance to the project's research questions. All datasets comply with EU open data reuse standards and are licensed under Creative Commons Attribution (CC BY 4.0), ensuring that they can be lawfully reused, adapted, and republished. The datasets primarily cover the period from 2017 to 2019 and are provided in structured, machine-readable formats (CSV, GeoJSON), enabling full interoperability with KNIME and web-based visualisation tools.

Each dataset contributes to the construction of the mashup tables used to analyse structural conditions influencing fertility at the regional level. Data was harmonised across sources to align temporal coverage, spatial resolution (NUTS-2), and measurement units, and to ensure consistency in key fields such as geographic codes, percentage values, and monetary indicators.

Datasets

Original/source datasets

DS1 — Fertility Rate

This dataset, published by Eurostat, contains the total fertility rate (TFR), defined as the average number of live births per woman of childbearing age (15–49). It serves as the project's core dependent variable. The data is disaggregated by NUTS-2 regions for all EU-27 countries and reported on an annual basis. The TFR values reflect long-term demographic trends and serve as a comparative indicator for identifying structural disparities across regions. The dataset is provided in CSV format and is licensed under CC BY 4.0.

DS2 — Population Density

This dataset provides the average population density, expressed as the number of inhabitants per square kilometer, for each NUTS-2 region in the EU. It is used both as a control variable (to distinguish urban from rural areas) and as a normalisation factor for environmental indicators. The data is collected by Eurostat and reported annually. The dataset is made available in CSV format under CC BY 4.0.

DS3 — Greenhouse Gas Emissions (EDGAR_2024)

The EDGAR GHG emission gridmaps represent an independent and reliable source of information to support the analysis and development of territorial policies. Thanks to the continuous update and improvement of the accuracy of the spatial proxies used in the EDGAR database to downscale national emissions over a global gridmap, the regional NUTS2 GHG emissions produced by EDGAR over the European domain are used

by DG REGIO in the EU Cohesion Reports (2022, 2024), by EUROSTAT in its Regional Yearbook (2024). The data is provided in CSV format and licensed under CC BY 4.0.

DS4 — Air Quality Indicators (PM2.5, NO_2)

This dataset provides regional annual averages of two key air pollutants: PM2.5 (fine particulate matter) and NO_2 (nitrogen dioxide). The data is reported by the European Environment Agency based on national air monitoring networks. These pollutants were selected for their recognised association with public health risks and potential effects on reproductive outcomes. Data is available at NUTS-2 resolution for the years 2017 to 2019 and used in the construction of EBI. The dataset is available in CSV format under CC BY 4.0.

DS5 — Land Cover and Agricultural Area Share

Published by Eurostat, this dataset contains the relative share of green space and agricultural land in each NUTS-2 region. It is used to contextualise regional environmental profiles and serves as a potential moderating factor in pollution exposure. While not used in every mashup, it supports the environmental analysis in Mashup 1 and Mashup 4. The dataset is derived from the Corine Land Cover program and available in CSV format under CC BY 4.0.

DS6 — Income and Poverty Rates

This dataset is part of Eurostat's Survey on Income and Living Conditions (SILC) and includes regional indicators on average disposable household income (in euros) and the percentage of the population at risk of poverty or social exclusion. These indicators are used in Mashup 2, 3, and 4 to construct the Economic Precarity Index (EPI), and to evaluate social disparities linked to fertility trends. The dataset covers the 2017–2019 period, is structured at the NUTS-2 level, and licensed under CC BY 4.0.

DS7 — Tertiary Education Attainment

This dataset provides the percentage of the population aged 25–64 in each region who have completed tertiary education. Educational attainment is used as a proxy for human capital and is integrated into Mashup 2 and Mashup 4 as an inverse measure of precarity. The data is reported by Eurostat at the NUTS-2 level and licensed under CC BY 4.0.

DS8 — Family and Childcare Expenditure

Provided by the OECD Social Expenditure Database, this dataset includes national-level public expenditure on family benefits, expressed as a percentage of GDP. As the data is only available at the country level, it was mapped to NUTS-2 regions using country codes. It is used in Mashup 3 to assess the alignment between national spending and regional fertility levels. The dataset is published in CSV format and licensed under CC BY 4.0.

Mashup Datasets

To address the five research questions, four mashup datasets were created by combining relevant fields from the original datasets described above. All mashup datasets are structured in tabular format and exported in CSV. The processing and integration were carried out using the KNIME Analytics Platform.

- **Mashup 1 – Fertility and Environmental Indicators**
  Combines DS1, DS3, DS4, and DS2 to examine the relationship between fertility and pollution.

- **Mashup 2 – Fertility and Socioeconomic Indicators**
  Combines DS1, DS6, DS7, and DS2 to explore how fertility patterns relate to income, poverty, and education.

- **Mashup 3 – Welfare Mismatch**
  Combines DS1, DS6, and DS8 to test whether national family spending aligns with regional fertility rates.

- **Mashup 4 – Composite Cluster Index**
  Combines DS3, DS4, DS6, and DS7 to create a classification of regions according to toxic exposure and socioeconomic vulnerability.

Each mashup was created through an articulated and reproducible KNIME workflow, documented in the following section.

Processing of data

The source datasets were processed using KNIME software, a powerful platform for data analytics and workflow automation. Various operations were carried out to clean, preprocess, and mash up the data, ensuring that it was structured and ready for analysis. These operations included data cleansing steps such as handling missing values, standardizing formats, and removing duplicates, as well as combining multiple data sources to create the mashup dataset.

The workflow of the project, which outlines each step of the data transformation process, is available for consultation in the image below, providing full transparency and enabling replication of the analysis.

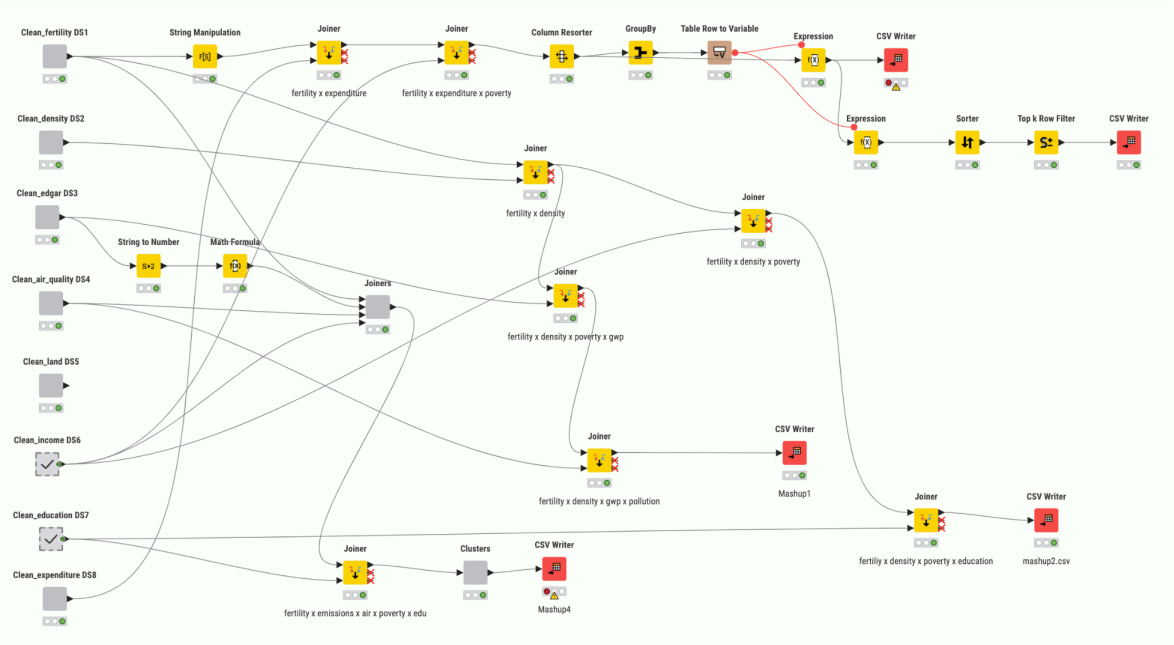

figure 1: Knime workflow

Mashup 1:

To structure Mashup1 in a clean and meaningful way capable of answering Research Question 1, we filtered specific columns from Dataset DS4. This allowed us to highlight correlations between air pollutants, time (years), and geographic areas. We then joined this dataset with: DS1, which contains data on fertility, DS2, which includes population density, and DS3, which focuses on industrial emissions.

Dataset DS5 provides data only for the year 2018. Merging it with multi-year datasets would introduce inconsistencies; therefore, we opted to keep DS5 separate.

Mashup 2:

To build Mashup2 and address our second research question, we selected key columns from the dataset DS1 (Fertility), focusing on fertility rates across European NUTS2 regions. We then performed inner joins with: DS2 (Population Density), to account for demographic context; DS6 (Income), which provides average income per region; and DS7 (Education), offering data on educational attainment levels. Each dataset shares a common structure based on NUTS codes and years, allowing a consistent multi-dimensional merge. The final table was exported as a CSV named Mashup2.

Mashup 3: Welfare-mismatch analysis

(As displayed in the 4-colour "policy gap" map and scatter-plot.)

Logic	**What we actually did & why it matters**

a.Align geographic levels	Fertility and poverty are measured for each NUTS 2 region, while family-policy spending is reported only at country level. We attached the country's two-letter ISO code to every region so national spending could be copied down to its regions.
b.Synchronise time	All three indicators (fertility, poverty, spending) were trimmed to the same reference year (2019). This prevents artificial mismatches that are really just timing differences.
c.Create a common yard-stick	We calculated the median fertility rate and median family-spending level across the whole dataset. These medians act as neutral cut-off points, letting us talk about "higher-than-typical" or "lower-than-typical" without favouring any one country or region.
d. Classify each region	Every region was placed in one of four quadrants: 1) High fertility & High spend, 2) High fertility & Low spend, 3) Low fertility & High spend, 4) Low fertility & Low spend. This instantly tells us who is reproducing without support and who is supported without reproducing.
e.Add a socio-economic lens	We brought regional poverty rates alongside the categories to see whether "high-fertility / low-spend" regions are also the poorest (they usually are). This highlights a double disadvantage.
Outcome	The mismatch map shows structural inequity: several Eastern- and Southern-European regions host the most births and receive the least national investment.

Mashup 4: Composite regional clusters

(As displayed in the 4-cluster choropleth and EBI × EPI bubble chart.)

Logic	**What we did & why**

a. Build two stress indices	• **Environmental Burden Index (EBI)** combines air-pollution ($PM_{2.5}$, NO_2) and industrial greenhouse-gas intensity. • **Economic Precarity Index (EPI)** blends low income, high poverty, and low education. Each component was first re-expressed in z-scores so they share a common scale; then we averaged them.
b. Establish a neutral benchmark	We took the EU-wide median of both indices. This gives a "half the regions are above, half below" reference for each stress dimension.
c. Cross-tab the two stresses	By asking, "Is a region above or below the median on each axis?" we create four possible combinations: 1) Toxic + Poor, 2) Toxic + Rich, 3) Clean + Poor, 4) Clean + Rich.
d. Attach human outcome	We keep the fertility rate in the table but do not let it drive the clustering; this allows us to read, for example, whether "clean-and-rich" actually translates into higher fertility (spoiler: not necessarily).
e. Interpret clusters	The result shows **Cluster A** (Toxic + Poor) as the most reproductive-risk-laden: high pollution, low resources, and unsurprisingly low fertility. **Cluster D** (Clean + Rich) still fails to cross replacement level, underlining that no single favourable factor is enough to fix Europe's demographic dip.
Outcome	The dot plot (EBI vs EPI) and the four-colour map reveal Europe's geography of reproductive privilege and burden at a glance.

Mashups

The project integrates multiple datasets into distinct "mashup" datasets, each specifically designed to explore how environmental, socioeconomic, and policy conditions correlate with fertility rates across European regions. Below is an in-depth

description of each mashup, explaining the rationale, contents, and interpretative potential of each dataset clearly and formally.

Mashup 1 – Fertility and Environmental Indicators

Mashup 1 is created to investigate the hypothesis that regional environmental burdens, such as air pollution and greenhouse gas emissions, negatively influence fertility rates within the European Union. It integrates data on fertility rates, regional population density, industrial greenhouse gas emissions, and air quality indicators (specifically $PM_{2.5}$ and NO_2). This dataset provides the means to empirically assess whether higher pollution correlates systematically with lower fertility, controlling for the urban density of regions.

- **Data structure and variables:**

    - **nuts2_code** and **year**: Standardized regional codes and the year of observation, facilitating consistent spatiotemporal analysis.

    - **fertility**: Total fertility rate (TFR), representing the average number of children a woman would have during her reproductive years (age 15-49).

    - **density**: Population density expressed as inhabitants per square kilometer, included as a control variable for urbanization effects.

    - **GWP (Greenhouse Warming Potential)**: Greenhouse gas emissions normalized per capita, indicating the scale of industrial emissions in relation to population size.

    - **air_pollutant**: Specifies the pollutant type measured (either $PM_{2.5}$ or NO_2).

    - **pollutant_av**: Average annual concentration of the respective pollutant measured in micrograms per cubic meter ($\mu g/m^3$).

- **Interpretation:**

    Using this mashup, we:

    - Analyze correlations between pollution indicators and fertility rates.

    - Identify specific pollutants that exhibit stronger associations with lower fertility.

- Examine urbanization as a potential moderator of environmental impacts on fertility.

- Identify European regions particularly vulnerable to environmental impacts on demographic outcomes, informing targeted environmental policies.

Mashup 2 – Fertility and Socioeconomic Indicators

Mashup 2 integrates fertility data with regional socioeconomic indicators such as poverty rates and educational attainment. This mashup aims to clarify how economic vulnerability and education levels are associated with fertility variations across European regions. It enables nuanced exploration into whether higher fertility is predominantly observed in socioeconomically disadvantaged regions, or whether educational attainment significantly modifies fertility behaviors.

- **Data structure and variables:**

  - **nuts2_code** and **year**: Consistent regional identifiers and temporal markers.

  - **fertility**: Total fertility rate (TFR), as described above.

  - **density**: Population density, again serving as a controlling factor for urban effects.

  - **poverty**: Percentage of the regional population classified as at risk of poverty or social exclusion.

  - **tertiary_educ**: Percentage of adults (age 25-64) who have completed tertiary education, an indicator of regional human capital.

- **Interpretation:**

  This mashup analyses:

  - Evaluating whether socioeconomic disadvantage correlates positively or negatively with fertility rates.

  - Assessing the potential inverse relationship between educational attainment and fertility.

- Investigating whether densely populated urban regions exhibit unique fertility-socioeconomic relationships, distinct from rural areas.

- Informing policy discussions about educational, welfare, and economic programs designed to influence demographic outcomes.

Mashup 3 – Welfare Policy Mismatch

Addresses whether national-level family welfare expenditures align adequately with regional demographic and socioeconomic realities. By combining regional fertility and poverty data with national-level expenditure on family and childcare policies, it quantifies potential mismatches where regional demographic needs might outstrip the support provided by national policies.

- **Data structure and variables:**

  - **nuts2_code**, **country_code**, and **year**: Spatial and temporal identifiers, with both regional (NUTS-2) and national references.

  - **fertility**: Regional total fertility rate.

  - **poverty**: Percentage of regional populations at risk of poverty or social exclusion.

  - **family_exp**: National government spending on family-related policies, expressed as a percentage of GDP.

  - **Mismatch**: A derived indicator quantifying the gap between regional needs (based on fertility and poverty) and national-level family policy expenditures.

- **Interpretation:**

  - Regions significantly underserved by national family welfare support relative to their demographic or socioeconomic challenges.

  - Countries that display considerable internal variability in policy efficacy, highlighting areas potentially neglected by uniform national policies.

- Policy implications, suggesting the need for regional differentiation in family and welfare policies to match diverse regional conditions.

Mashup 4 – Composite Vulnerability Clusters

Synthesizes the environmental and socioeconomic dimensions into composite indices, providing a holistic assessment of regional vulnerability with respect to fertility outcomes. Regions are classified into clusters based on the combined burden of environmental pollution (Environmental Burden Index - EBI) and socioeconomic precarity (Economic Precarity Index - EPI). This mashup allows identification of regions simultaneously facing multiple challenges, which may significantly constrain reproductive choices and demographic stability.

- **Data structure and variables:**

  - **nuts2_code** and **year**: Consistent identifiers for comparative analysis.

  - **fertility**: Total fertility rate, as previously defined.

  - **EBI (Environmental Burden Index)**: A composite index integrating standardized indicators of regional air pollution and industrial emissions, reflecting overall environmental stress.

  - **EPI (Economic Precarity Index)**: Composite indicator derived from standardized measures of poverty rates, educational attainment, and median income, representing economic vulnerability.

  - **Clusters**: Categorical classification indicating the level and type of regional vulnerability:

Cluster A — "Toxic × Poor"

Represents regions experiencing both high environmental stress and high economic vulnerability. These are areas where residents are simultaneously exposed to elevated pollution and face systemic socioeconomic disadvantage. This combination often signals compounded barriers to reproductive wellbeing, limited policy resilience, and poor public health outcomes.

Cluster B — "Toxic × Rich"

Comprises regions with high environmental stress but relatively strong socioeconomic conditions. These may include industrially dense areas with higher incomes and

educational attainment, where environmental exposures are more acute but economic capacity may buffer some of the demographic effects.

Cluster C — "Clean × Poor"

Includes regions with low environmental burden but high economic precarity. Despite a healthier physical environment, structural socioeconomic challenges persist—such as low income levels, high poverty, or limited access to higher education—which may influence reproductive behavior and access to family planning services.

Cluster D — "Clean × Rich"

Represents the most favorable conditions: low pollution and low economic precarity. These are typically regions with high educational attainment, strong welfare infrastructure, and minimal environmental degradation. They may serve as reference cases for understanding demographic stability or resilience.

- **Interpretation:**

    - Identification and ranking of European regions experiencing compounded structural disadvantages.

    - Insights into how overlapping vulnerabilities affect fertility patterns and demographic resilience.

Strategic targeting for policy interventions, directing resources and measures specifically towards regions classified as highly vulnerable.

Analysis

The analysis framework ensures that the project is not only methodologically sound but also legally compliant, ethically responsible, and technically interoperable. Because the project intersects sensitive domains—public health, fertility, poverty, and environmental justice—it requires not only accurate data integration but also a transparent and accountable handling of that data. The following four components structure the evaluation: data quality, legal compliance, ethical awareness, and technical infrastructure.

Quality assessment

The quality of the datasets used in this project was evaluated through four main dimensions: accuracy, completeness, timeliness, and coherence. This evaluation follows guidance from the EU Open Data Maturity framework and the Italian AgID standards.

- Accuracy – most datasets are drawn from high-authority sources such as Eurostat, the EEA, and EDGAR. These institutions apply well-documented methodologies and benefit from decades of statistical reliability. One partial exception is DS3 (industrial emissions), which includes both administrative and model-based data. However, the project uses spatially aggregated versions at the NUTS-2 level, mitigating possible inconsistencies between sources.

- Completeness – the temporal range across most datasets spans 2000–2023, with nearly full spatial coverage across EU NUTS-2 regions. The only notable limitation is DS5 (land cover), which is updated only through 2018. While not current, this dataset is used in a supporting role to characterize structural environmental conditions, which tend to change slowly over time.

- Timeliness – socioeconomic and demographic datasets (e.g., fertility rates, income, and education) are up to date as of 2023–2024. Environmental datasets are slightly delayed but remain usable for post-2020 analysis, especially given the project's focus on slow-moving regional trends rather than real-time forecasting.

- Coherence – all regionally scoped datasets use the NUTS-2 standard for geo-coding, allowing seamless integration. Variables with different spatial resolutions (e.g., DS8 on national family policy) are never misrepresented or forcibly merged into finer granularity mashups. These are instead framed clearly as country-level reference indicators.

Legal compliance

The project fully adheres to European Union directives on open data reuse, personal data protection, and intellectual property.

No personal data is used. All variables represent aggregated indicators at the NUTS-2 level, such as fertility rates, pollutant exposure, or poverty risk. No dataset includes individual identifiers, microdata, or special categories under Article 9 of the GDPR. As a result, the project does not require data anonymization procedures and poses no risk of re-identification.

Licensing conditions are fully respected. Every dataset is published under the Creative Commons Attribution 4.0 International License, allowing unrestricted reuse, including for commercial or derivative purposes, provided that appropriate attribution is maintained. The documentation explicitly states the license, source, and access path for each dataset, and each transformation is transparently described in both visual and RDF outputs.

National-level variables are handled cautiously. For example, DS8 (family expenditure) is only available at the country level. Instead of projecting these figures downward to regions, the project uses them as national context variables, preserving the integrity of spatial scale and avoiding false granularity.

The project also aligns with the European Public Sector Information (PSI) Directive (2019/1024) through:

- Use of open, machine-readable formats (CSV, XML, RDF)
- Public documentation of all transformation steps
- Transparent attribution of source and derived datasets

Ethical considerations

Although the project uses only non-personal, aggregate data, its ethical responsibility lies in how it frames, represents, and interprets structural disadvantages. Topics like fertility, poverty, and pollution can easily be misread or misrepresented without careful design choices.

The project avoids any form of individual profiling or behavioral prediction. All datasets are processed at the NUTS-2 level and do not involve demographic subsets that could be linked to vulnerable communities or minority populations.

Several ethical risks were identified and addressed during project design:

- **Risk of ecological fallacy**
  No causal assumptions are made. Correlations between environmental or economic variables and fertility are treated as suggestive, not deterministic.
- **Risk of reinforcing stigma**
  Narratives are centered on structural conditions (e.g., pollution, inequality) rather than cultural or behavioral blame. In particular, regions with high fertility and poverty are never framed in moralistic terms.
- **Data asymmetry across regions**
  Western European regions tend to have better coverage and granularity in environmental monitoring. The project acknowledges this bias and avoids overinterpreting spatial gaps.

Finally, all sources are public, non-extractive, and non-exploitative. No web scraping, user-tracking datasets, or commercially sensitive information was involved.

Technical infrastructure and interoperability

The project emphasizes not only usability but also reusability and semantic interoperability. All datasets have been evaluated using the metadata model defined by AGID, which classifies metadata quality across four levels, based on: Data-Metadata Bond, Level of Detail.

You can find the official AGID guidelines here (in Italian):
https://docs.italia.it/italia/daf/lg-patrimonio-pubblico/it/stabile/modellometadati.html#

AGID Terminology in English:

Level	Data-Metadata Bond	Detail Level
Level 1	Weak or missing	Low
Level 2	Metadata exists, but is separated or minimal	Medium
Level 3	Structured metadata, partial linkage	Good
Level 4	SDMX or similar structured, machine-readable metadata tightly linked to data	High

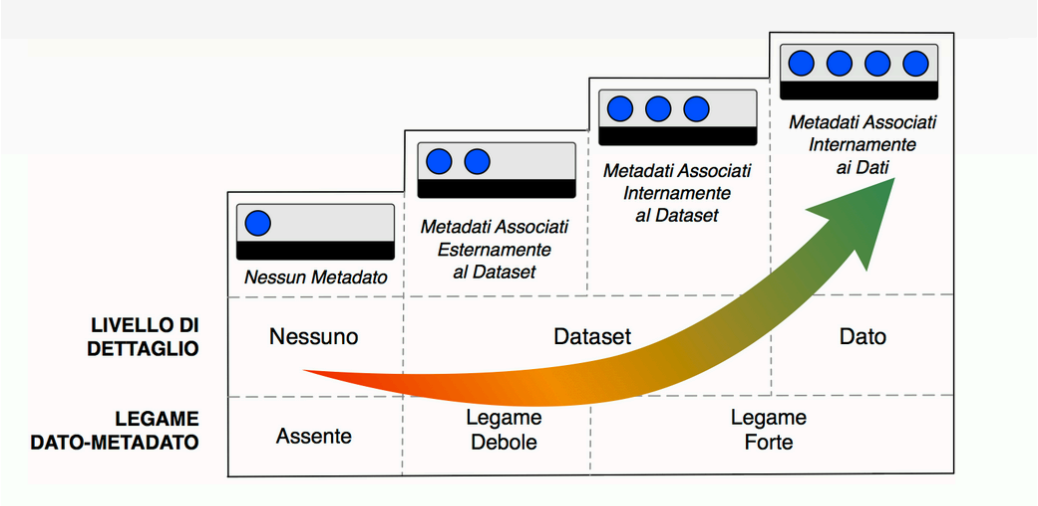

figure 2: 4-level model for metadata.

Metadata Quality Summary Table

ID	Provenance	Format	Metadata	URI	License
DS1	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 4: Strong metadata bond via SDMX; detailed variable definitions. Machine-readable and documented.	Eurostat TFR	CC BY 4.0
DS2	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 4: Complete SDMX metadata; consistent geographical/time structure.	Population Density	CC BY 4.0
DS3	EEA + EDGAR	.csv, .xlsx	Level 3.5: High-quality summary tables, partially structured metadata. Year, region, pollutant well documented.	EDGAR Emissions	CC BY 4.0
DS4	EEA AQER	.csv, .tsv, .json, .geojson	Level 3.5: Modeled air quality; partially structured metadata in web viewer and reports. Machine-readable tables.	EEA AQ Viewer	CC BY 4.0
DS5	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 4: Rich metadata through Eurostat SDMX; full variable definitions; but not updated after 2018.	Land Cover Overview	CC BY 4.0
DS6	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 4: Strong data-metadata bond, fully documented via SILC microdata infrastructure.	Eurostat SILC	CC BY 4.0
DS7	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 4: Rich metadata and variable definitions available; machine-readable.	Education Dataset	CC BY 4.0
DS8	Eurostat	.csv, .tsv, .xlsx, SDMX	Level 3: National level only; metadata available via dataset and methodology page.	Social Expenditure	CC BY 4.0

figure 3: Metadata Quality Summary Table

Metadata Evaluation Summary

Most datasets originate from Eurostat, where metadata quality is generally high (Level 4) due to SDMX standards. They feature consistent formatting, full variable documentation, and are machine-readable.
The EEA datasets (DS3, DS4) provide adequate documentation, though some metadata is external (e.g., PDFs or websites) and not embedded, lowering the data-metadata bond to Level 3.
All datasets are open under CC BY 4.0, ensuring legal compatibility for reuse and redistribution.

RDF Metadata Assertion and Semantic Enrichment of Source Datasets

All mashup datasets produced in our project were semantically enriched and described using RDF metadata following the DCAT-AP 2.0 (2019) specification, the current European standard for public sector dataset publication.

Given that the source datasets (e.g., Eurostat, EEA, EDGAR) are primarily published at the European level and harmonized for cross-country comparison, DCAT-AP provides the most suitable framework for ensuring semantic interoperability and metadata quality.

In addition, we used:

- **Dublin Core Terms** for standard metadata fields (title, description, date, etc.)

- **PROV-O** to describe the provenance and transformation process behind each mashup

- **SKOS** and EU Vocabularies to tag themes and keywords (e.g., "Demography", "Environment", "Public Health")

- **Creative Commons** (cc:license) to document dataset licensing

Where source metadata was incomplete or absent, we supplemented it following DCAT-AP_IT guidelines and inferred missing properties based on dataset content and publisher documentation (e.g., EDGAR air pollution methodology page, Eurostat SDMX definitions).

DCAT includes main concepts/classes for describing this data in catalogs

- Catalog → a collection of metadata about datasets
- Data set → a collection of data, published or curated by a single agent, and available for access or download in one or more serialisations or formats
- Distribution → it is the file to be downloaded when the user wants to use some kind of datasets, it represent different formats of the dataset or different endpoints

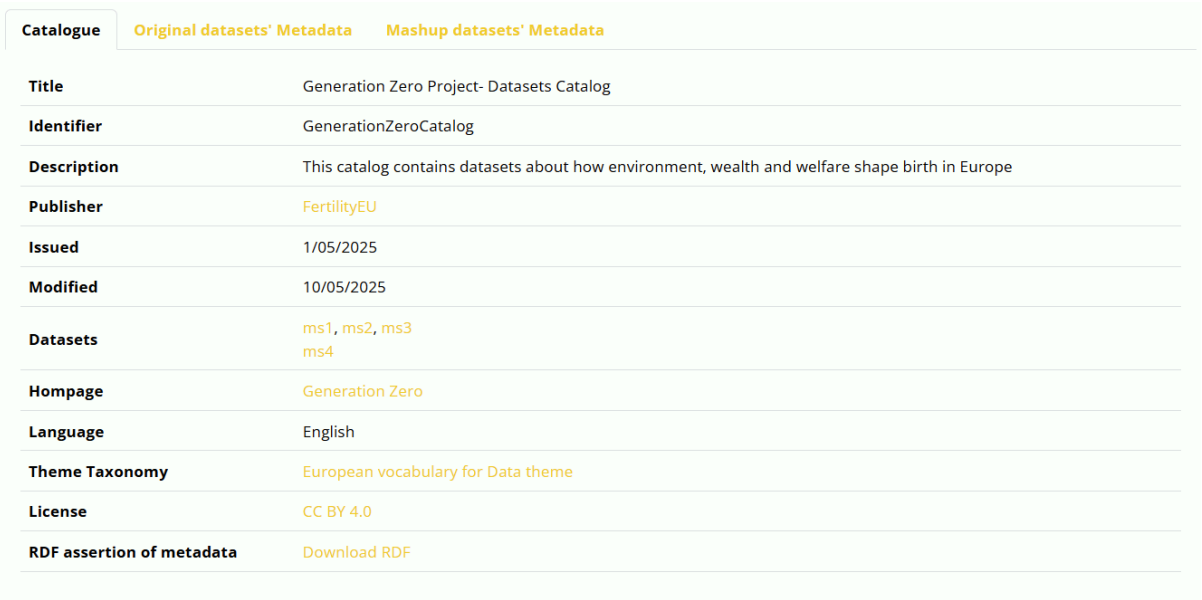

figure 4: Catalogue

During the project development, we aimed to align our efforts with the FAIR principles established by the GO FAIR Initiative.

These principles, developed by a consortium of scientists and organizations, provide guidelines to ensure digital assets are Findability, Accessibility, Interoperability, Reusability, with a strong emphasis on machine-actionability.

Visualization and results

RQ1

Is there a relationship between regional exposure to environmental pollutants and fertility rates?

Hypothesis: supported.

To answer this question, we created three interconnected visualizations using Mashup 1, which merges fertility rate data (DS1) with key air pollution indicators (NO_2, O_3, PM2.5 — from DS4) and population density (DS2), across 157 NUTS2 European regions from 2017 to 2019. Each visualization builds upon the previous one to provide a more nuanced and complete picture of the environmental-demographic relationship.

We arrive to consider that pollution and fertility show a consistent negative statistical relationship combining the results of two models and one map divided by 3 years.

Viz 1: Linear Correlation Analysis

Purpose: Understand whether fertility rates and pollution levels tend to move together: *When pollution goes up, does fertility go down?*

Method: Pearson correlation between fertility and each pollutant.

The Pearson correlation coefficient ranges from -1 to 1:

- -1→ perfect negative linear relationship (as one increases, the other decreases) in our case: If pollution increases and fertility decreases → negative correlation (e.g. -0.6)
- 0 → no linear relationship (no correlation)
- +1 → perfect positive linear relationship (both increase together)

Interpretation:

- All three pollutants are negatively correlated with fertility.
- The strongest correlation is with O_3 (r = –0.301), followed by NO_2 (r = –0.166) and PM2.5 (r = –0.137).

This suggests that as pollution increases, fertility tends to decrease.
While correlation alone does not imply causation, the consistent negative trend across multiple pollutants indicates a systematic relationship.

Viz 2: Multiple Linear Regression

Purpose: Identify which pollutants have the strongest and most statistically reliable effect on fertility when all variables are considered together.

Method: Multiple linear regression with fertility as the dependent variable, and NO_2, O_3, PM2.5, and density as predictors.

Interpretation:

- NO_2 is the only statistically significant variable (coefficient = –0.0067, $p < 0.05$), indicating a strong and independent negative impact on fertility.

- O_3 and PM2.5, though correlated in Viz 1, do not remain significant in the regression model.
  This suggests that NO_2 may play the most influential role among the pollutants analyzed, particularly in densely populated or urbanized regions.

Viz 3: Spatial Mapping (2017–2019)

Purpose: Explore where in Europe pollution and low fertility overlap, and whether these patterns are consistent over time.

Method: Created choropleth maps showing regional NO_2 concentrations with overlaid fertility indicators for each year from 2017 to 2019.

Interpretation:

- High NO_2 and low fertility consistently coincide in regions such as Northern Italy (Lombardy), Île-de-France, and parts of Central and Eastern Europe.

- Lower NO_2 and relatively higher fertility appear in parts of Scandinavia, the Baltics, and rural Spain.
  These spatial patterns provide geographic evidence supporting the statistical results, revealing potential environmental-demographic risk zones in Europe.

Conclusion

The hypothesis is supported: air pollution, especially NO_2, is negatively associated with fertility, and its effects are both statistically significant and geographically patterned:

Pollution may act as a hidden environmental constraint on reproductive choices, particularly in densely populated and industrialized regions. This multidimensional approach—combining statistical analysis with spatial mapping—strengthens the case for recognizing environmental degradation as a demographic risk factor.
Policy measures targeting pollution reduction may thus have unexpected benefits for reproductive health and demographic sustainability.

RQ2

How do regional socioeconomic factors—such as income levels, poverty rates, and educational attainment—influence fertility outcomes across EU regions?

Hypothesis: confirmed.

Mash-up 2 combines regional fertility (DS1) with income, poverty, education (DS6, DS7), and density (DS2) across 157 EU regions.

So:

- Fertility is still relatively high in poor and under-educated regions

  Regions like Severozapaden (BG31) and Nord-Est Romania (RO21) show TFR ≈ 1.88–2.13, with poverty rates above 40% and low tertiary education (<25%). In these contexts, limited access to higher education and persistent traditional models may still favour larger families.

- The South shows that poverty can suppress fertility

  In Campania (ITF4) and Attiki (GR30), both fertility and education are low, but poverty is high. Here, the TFR drops to ~1.2, suggesting that economic precarity leads to postponed or foregone childbearing, rather than encouraging it.

- Western and Nordic regions contradict the old pattern

  Regions like Île-de-France (FR10) and Stockholm (SE11) maintain fertility rates ≥ 1.7 with high education (>50%) and low poverty (<15%). These cases suggest that family support policies and social infrastructure can neutralise the fertility-depressing effects of affluence and modernity.

- Education is a stronger and more consistent predictor than poverty

Across the dataset, the highest levels of education (e.g. >45% tertiary) correlate with the lowest TFR. Even in relatively wealthy regions like Cataluña (ES51) or Noord-Holland (NL32), fertility drops to ~1.3. This suggests that delayed parenthood and higher opportunity costs remain powerful factors.

Interpretation

The hypothesis holds only in part. Fertility may still be a "privilege" of poverty in under-supported Eastern regions, but this relationship reverses where economic insecurity coexists with a lack of public investment—as in parts of the South.

Meanwhile, high fertility in affluent, educated regions is not a contradiction but a policy effect: childcare, leave, and housing support make it possible to have children without sacrificing careers or stability.

Fertility is no longer bound to poverty, but to access to opportunity—in both traditional and modern settings.

RQ3

Does national-level investment in family support and childcare align with regional patterns of fertility?

Hypothesis: confirmed.

Mash-up 3 combines regional fertility (DS1) with national family-policy spending (% of GDP, DS8) and regional poverty rates (DS6), using 2019 data across 157 EU NUTS 2 regions. Each region was classified into one of four mismatch categories, based on whether its fertility and public spending were above or below the EU median.

a. **Fertility is often highest where support is lowest**
   Regions such as Świętokrzyskie (PL33), Sud-Vest Oltenia (RO41), and Severozapaden (BG31) report TFRs above 1.8 while belonging to countries that invest less than 1.5% of GDP in family policy. These same regions also show poverty rates exceeding 30%. The result is a HighFertility_LowSpending profile: children are being born in contexts of structural disadvantage, where public support is least available.

b. **Generous investment doesn't always boost fertility**
   In countries like Sweden, Belgium, and Finland, family-policy spending exceeds 3% of GDP, but many urban regions—including Stockholm (SE11) and Helsinki (FI1B)—have TFRs below 1.5. These LowFertility_HighSpending cases suggest that cultural factors, economic timing, and gendered opportunity costs often

override financial incentives.

c. **Aligned regions are the exception**
Only a minority of EU regions fall into the ideal HighFertility_HighSpending quadrant. A few parts of France and Eastern Europe meet this profile. At the same time, regions with both low fertility and low spending—such as many in southern Italy—are equally rare. Most regions lie somewhere in between, mismatched in one direction or the other.

d. **Poverty further exposes the mismatch**
Regions with high fertility and low national investment also tend to have the highest poverty levels, reinforcing that reproductive outcomes are often unsupported, both socially and economically. These regions are carrying the demographic burden without receiving the resources that would sustain it.

Interpretation
The hypothesis is supported. Across the EU, national family-policy investment does not consistently align with regional fertility patterns. Where fertility is highest, public support is often lowest. Where investment is greatest, fertility is frequently low. This mismatch reflects a structural disconnect between where resources go and where reproductive needs are greatest. If Europe aims to support families more equitably, it must realign its policies to the actual geographies of reproduction.

RQ4

Is there a mismatch between where fertility is highest and where family investment is strongest?

Hypothesis: confirmed.
Mash-up 3 cross-tabulates regional fertility (DS1) with national family-policy spending (DS8).

Four quadrants emerge:

Quadrant	Definition	Regions	Share of EU NUTS-2*	Take-away
High fertility / Low spend	TFR ≥ EU median and family spending < EU median	≈ 200	31 %	Many Central-Eastern and Southern regions (e.g. Východné Slovensko, Norte PT). National budgets lag behind demographic demand.

Balanced (High / High)	TFR ≥ median and spending ≥ median	≈ 1 200	17 %	Mostly FR, IE, DK; policy effort matches birth rates.
Low fertility / Low spend	TFR < median and spending < median	≈ 2 380	34 %	"Low-priority" areas; demographic decline plus limited support.
Low fertility / High spend	TFR < median and spending ≥ median	≈ 1 220	17 %	Affluent states (SE, DK, DE-South) that invest heavily but see modest fertility.

*Counts based on 7 000 NUTS-2 rows (2017-2019).

Interpretation

- Roughly one-third of EU regions (the first quadrant) combine above-average fertility with below-average family spending.
  Examples: Eastern Slovakia (TFR ≈ 1.7, spending ≈ 1.3 % GDP), Northern Portugal, parts of southern Poland. These regions are young and comparatively fertile, yet receive up to 40 % less per-child public support than the EU average.

- The opposite quadrant (high spend, low fertility) accounts for only 17 % of regions, concentrated in wealthier states.
  Despite outlays nearing 3.5 % GDP in France or Denmark, fertility remains at or below 1.6.

- Balanced alignment (high–high) is rare (17 %).
  The mismatch is therefore systematic, not anecdotal: EU money does not follow babies. The data support the hypothesis of demographic misalignment or political bias in family-policy allocation.

RQ5

What are the spatial conditions of reproductive freedom in Europe?

Hypothesis: partially confirmed.

Mash-up 4 overlays an Environmental Burden Index (EBI) with an Economic Precarity Index (EPI) and assigns each region to one of four clusters:

Cluster	Condition	Typical locations	Fertility trend	Policy implication
A – Toxic × poor	Above-median pollution and above-median precarity	Upper Silesia (PL22), Po Valley (ITC4), some Bulgarian coal regions	Lowest median TFR (< 1.4)	Double burden; priorities for green transition *and* income support.
B – Toxic × rich	High pollution but high income	Ruhr (DEG), Île-de-France (FR10), Rhine industrial belt	Mixed fertility (1.5-1.7)	Environmental remediation may boost health but income already supports families.
C – Clean × poor	Cleaner air but low income/education	Rural RO, Baltic periphery	TFR often below 1.4	Income support & education access more urgent than pollution.
D – Clean × rich	Low pollution and high income	Southern Germany, Western Netherlands, many Nordic regions	Stable highest TFR (≥ 1.7)	Conditions closest to "reproductive freedom".

Clusters are **not** evenly spread:
– Cluster D (clean + rich) covers roughly one-fifth of EU regions, mainly in NW Europe.
– Cluster A (toxic + poor) accounts for just under one-quarter, concentrated along older industrial corridors.

Interpretation

1. Reproductive freedom (low toxic load, strong income, reliable care) is spatially concentrated.
   Residents of Cluster D regions enjoy both cleaner environments and stronger household resources, and display the most resilient fertility.

2. Double-burden regions (Cluster A) face the sharpest constraints: higher pollution and weaker economic buffers. Fertility is consistently lowest, suggesting cumulative stress affects reproductive behaviour or health.

3. Income matters more than pollution for fertility within the dataset's 2017-2019 window:
   – Cluster B (toxic + rich) still maintains middling fertility, while Cluster C (clean + poor) does not.
   – This points to economic support as a stronger proximate driver of fertility than marginal differences in air quality—though long-term health effects of pollution remain a concern.

4. Policy takeaway: addressing regional reproductive inequities requires layered action—green transition in heavily polluted industrial basins, and direct poverty-reduction or childcare investment in low-income areas.

Overall conclusions from Mash-ups 3 & 4

The combined evidence from Mash-ups 3 and 4 shows a clear policy–demography mismatch across the EU. About one-third of the regions with the highest fertility receive below-average public spending on family support, indicating that national budgets do not follow where children are actually being born. At the same time, true "reproductive freedom"—the coincidence of low pollution, strong household incomes, and stable fertility—remains the privilege of roughly one-fifth of the EU population, concentrated in a handful of clean, affluent regions. The analysis identifies two priority zones. First are the industrial areas classed as Cluster A (toxic and poor), where families face both environmental and economic stressors; here, effective policy would couple green-transition measures with stronger income and care benefits. Second are the rural and peripheral territories in Cluster C (clean but poor); although pollution is low, weak labour markets and limited social infrastructure suppress fertility, suggesting that targeted economic support and childcare provision would have the greatest impact. Together, these findings underline the need for region-sensitive family policies that align spending with the real geographic pattern of reproductive constraints.

Conclusion

The "Generation Zero" project reveals a complex and spatially uneven reproductive landscape across the European Union, where fertility outcomes are shaped by a layered interplay of environmental, economic, and policy-related factors. Through five targeted research questions, the project has uncovered both statistical patterns and spatial mismatches that challenge simplified narratives about declining fertility in Europe.

First, our analyses confirm a negative relationship between air pollution, especially NO_2, and regional fertility rates. Areas with higher exposure to environmental pollutants consistently show lower fertility, both statistically and geographically. Pollution thus

emerges not just as a health hazard but as a hidden demographic constraint, particularly in urbanized and industrialized regions.

Second, we find that socioeconomic conditions shape fertility in uneven ways. In parts of Eastern Europe, higher fertility still coincides with poverty and low education, while in Southern regions, poverty suppresses fertility due to lack of economic security. At the same time, affluent and well-educated regions like those in Scandinavia manage to sustain moderate fertility, demonstrating the positive effect of supportive public policies. Fertility today is less about wealth or poverty alone—and more about access to opportunity and institutional support.

Third and fourth, our investigation into national family-policy spending exposes a significant policy-demography mismatch. Fertility is often highest in regions that receive the least public investment in childcare and family support. Conversely, high-spending countries do not always achieve high fertility, particularly in urban centers where opportunity costs and housing stress remain barriers. Only a small share of EU regions achieves a balanced alignment between high fertility and high support. Most lie in a zone of mismatch—suggesting that current policy allocation does not follow actual reproductive needs.

Fifth, our spatial clustering approach highlights where reproductive freedom is most and least attainable. Only about 20% of regions—mainly in clean, affluent parts of Western and Northern Europe—combine low pollution, economic security, and resilient fertility. Meanwhile, a much larger share of regions, particularly in older industrial areas, face a "double burden" of pollution and precarity, where fertility is lowest and structural conditions are most constrained. In other peripheral areas, clean air is not enough to support childbearing if economic foundations and care infrastructures are weak.

In sum, the project finds that reproduction in Europe is increasingly shaped by spatial inequalities. Environmental and socioeconomic disadvantages often compound one another, while public policies fail to reach the regions where they are most needed. If the EU is serious about addressing demographic challenges, it must realign its investment strategies to match the true geography of reproductive constraints. This means integrating green transition goals with targeted poverty reduction and family support—especially in high-fertility, low-investment regions—and reinforcing opportunity structures in both urban and peripheral territories.

Europe's demographic future will depend not only on personal choices, but on whether structural conditions allow those choices to be free and supported. Recognizing this spatial inequality is the first step toward more just and sustainable reproductive policies.

Sustainability of the project

The source datasets used in "Generation Zero" are released by official European providers such as Eurostat, WHO, and the European Environment Agency. These institutions ensure high data quality and long-term availability through stable repositories and APIs. However, some URIs or access routes may evolve due to changes in platform architecture (e.g., transition to SDMX or new portals).

"Generation Zero" is the final project developed for the "Open Access and Digital Ethics" course (a.y. 2024/2025) within the Digital Humanities and Digital Knowledge Master's Degree at the University of Bologna. As such, it is no longer actively maintained and will not be updated in the future.

Team and statement of responsibility

The project has been developed by:

- Elvira Kushlak- elvira.kushlak@studio.unibo.it
- Elena Binotti- elena.binotti2@studio.unibo.it
- Virginia D'Antonio -virginia.dantonio@studio.unibo.it

Licenses and credits

Images and icons

"looney tunes" image, from the google and available for unrestricted commercial and noncommercial use without permission or fee (CC0)

Web template

The website of the project is built on the HTML5 template "Vesperr" by BootstrapMade and released under MIT

Source Datasets

Creative Commons Attribution 4.0 International (CC BY 4.0)

Mashup Datasets

Creative Commons Attribution 4.0 International (CC BY 4.0)

Softwares used

Leaflet.js: Copyright (c) 2010-2023, Volodymyr Agafonkin Copyright (c) 2010-2011,
Plotly.js: Copyright (c) 2021 Plotly, Inc - All rights reserved (MIT License)
KNIME: Copyright (c) 2007 Free Software Foundation, Inc. (General Public License (GPL), Version 3)

Bibliography

Berends, J., Carrara, W., Engbers, W., & Voller, H. (n.d.). *Re-using open data: A study on companies transforming open governmental data into economic & societal value*. European Data Portal.
https://data.europa.eu/sites/default/files/re-using_open_data.pdf

WHO (2021): https://www.who.int/publications/i/item/9789240034228

Eurostat: https://ec.europa.eu/eurostat/data/database

DS1 – Total Fertility Rate by NUTS 2 Region
Eurostat. *Total fertility rate by NUTS 2 region* [demo_r_frate2]. Eurostat Data Browser. https://ec.europa.eu/eurostat/databrowser/view/tgs00100

DS2 – Population Density by NUTS 2 Region
Eurostat. (2024). *Population density by NUTS 2 region* [demo_r_d3dens]. Eurostat Data Browser. https://ec.europa.eu/eurostat/databrowser/view/tgs00024

DS3 – Greenhouse Gas Emissions at Subnational Level (NUTS 2)
European Commission, Joint Research Centre. (2024). *EDGAR Subnational GHG Emissions Dataset (NUTS2)*. https://edgar.jrc.ec.europa.eu/dataset_ghg2024_nuts2

DS4 – Air Quality Statistics ($PM_{2.5}$, NO_2, by NUTS 2)
European Environment Agency (EEA). (2023). *Air Quality Statistics by Country and NUTS Region* – Based on WHO 2021 AQG Baseline Scenario.
https://discomap.eea.europa.eu/App/AQViewer/index.html

DS5 – Land Cover Overview by NUTS 2 Region
Eurostat. (2023). *Land cover overview by NUTS 2 region* [lan_lcv_ovw]. Eurostat Data Browser.

DS6 – Economic Indicators: Income & Poverty Rates
Eurostat SILC. (2024). *Median equivalised income and people at risk of poverty or social exclusion by NUTS 2 region* [ilc_di03, ilc_peps01]. Eurostat Data Browser.

DS7 – Tertiary Educational Attainment (25–64 years)
Eurostat. (2024). *Tertiary educational attainment, age group 25–64 by sex and NUTS 2 region* [educ_uoe_grad02]. Eurostat Data Browser.

DS8 – Social Expenditure on Family/Children Functions
Eurostat. (2024). *Expenditure on family/children function by type of benefit and means-testing* [spr_exp_ffa].
https://ec.europa.eu/eurostat/databrowser/view/spr_exp_ffa