PROJECT	
Project number:	101188188
Project acronym:	SoGreen
Project name:	Social Aspects of the Green Transition (SoGreen)

Work Package	7
Deliverable Number	7.3
Deliverable Name	Paper on Methodologies and Provenance in Data Integration
Deliverable Type	R
Dissemination Status	PU
Lead Beneficiary	Sikt
Authors	Benjamin Beuster; Hilde Orten
Date Submitted	26 June 2026

PROJECT	
Project number:	101188188
Project acronym:	SoGreen
Project name:	Social Aspects of the Green Transition (SoGreen)

Work Package	7
Deliverable Number	D7.3
Deliverable Name	Paper on Methodologies and Provenance in Data Integration
Deliverable Type	R — Document, report
Dissemination Status	PU
Lead Beneficiary	Sikt
Authors	Benjamin Beuster; Hilde Orten
Date Submitted	25 June 2026

Abstract

This paper documents the methods and provenance developed in Work Package 7 of the SoGreen project for integrating climate and air quality contextual data with three European social-survey infrastructures: the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the Generations and Gender Programme (GGP). It describes the data sources used, the raw variables retrieved, the derived variables produced for the integrated dataset, and the provenance approach that links every derived variable to its source, version and processing step. The work extends the methodology developed in the EOSC Future Science Project "Climate Neutral and Smart Cities" (Orten et al. 2023) and operates within the pseudonymity-preserving geo-imputation framework defined for SoGreen in deliverable D7.1 (Fenton 2025). New to the SoGreen design are health-relevant derived variables based on the Universal Thermal Climate Index (UTCI) and the World Health Organization (WHO 2021) air quality guidelines, alongside the established European Air Quality Index of the European Environment Agency. The paper is intended both as a formal Horizon Europe deliverable and as a reusable methodological reference for cross-domain integration of survey and environmental data.

1. Introduction

SoGreen is a Horizon Europe research and innovation action that brings together three of the largest European social-survey infrastructures — ESS, SHARE and GGP — to investigate how people across Europe experience and respond to climate change and the green transition. The three infrastructures field a harmonised survey module on social attitudes to sustainability, the perceived effects of green-transition policies, and behaviours related to consumption, mobility and employment. To allow these responses to be analysed in the context of the conditions in which respondents actually live, Work Package 7 — the project's conceptual and planning phase — develops the methodology for a set of contextual variables describing local climate and air quality and for linking them to individual survey cases, while the integration itself and its provenance documentation are carried out in Work Package 8.

2. Background

Personal experience of weather extremes and air pollution has long been thought to shape attitudes and behaviours related to climate change and environmental policy. Self-report studies have shown that respondents who are more concerned about climate change are also more likely to report recalling unusual weather events in their local area (Leiserowitz et al. 2012). In a related design, Whitmarsh (2008) found that respondents' self-reported experience of air pollution was significantly associated with concern about anthropogenic climate change, although experience of flooding showed little such relationship; either way, this design cannot separate perception bias from genuine exposure. Linking survey responses to independently measured environmental data offers a more objective route. Hamilton and Stampone (2013) found that beliefs about climate change among self-declared independent voters in the United States were predicted by temperature anomalies on the day and day before interview. Bergquist et al. (2019) reported that state-wide annual temperature anomalies were associated with climate concern in U.S. opinion data. Hoffmann et al. (2022) extended the analysis to a Europe-wide setting, constructing indices for monthly temperature anomalies, UTCI-based heat episodes and SPEI-based dry and wet periods, and reported that an extended period of anomalous weather, rather than a single anomalous month, is required to influence environmental concern.

The EOSC Future Science Project "Climate Neutral and Smart Cities" (Orten et al. 2023) demonstrated that climate and air quality indicators can be integrated with ESS data at the NUTS region level for a selection of large European urban regions, using ECMWF ERA5 reanalysis and European Environment Agency observation-station data. SoGreen WP7 extends that work in three ways. First, the integration is generalised from ESS to all three SoGreen survey infrastructures (ESS, SHARE and GGP), each of which has its own data-management environment, spatial-identifier scheme and dissemination platform. Second, the indicator catalogue is extended with health-relevant indicators based on the WHO air quality guidelines (WHO 2021) and on UTCI thermal stress thresholds (Bröde et al. 2012; European Environment Agency 2025). Third, the spatial linkage is performed under the geo-imputation framework set out in D7.1, which avoids transferring or holding exact respondent locations while also enabling linkage at a finer spatial resolution than the NUTS region level.

As described in Deliverable D3.1, the SoGreen inventory consists of 30 questions (23 of them fielded by all surveys, with an extension of 7 questions fielded only by the ESS and GGS), covering concepts describing perception, attitudes, and behaviours related to climate change and the effects of the green transition, among them:

- Exposure to extreme weather events and pollution,
- perceived environmental risks,
- climate change believes,
- behavioral intentions, and
- policy impacts.

In the SoGreen inventory the above concepts are operationalized into questions, measuring

- concern about, exposure to, and effects of air quality indices on health and wellbeing,
- concerns about climate change and exposure to extreme weather events, and
- attitudes related to environmental protection and climate change policies and responsibilities

When selecting air quality and climate related variables for integration with the survey data, focus has been put on capturing observable extreme events within a one-year window, as well as on selecting indicators that matches the questions in the SoGreen inventory.

The raw input variables are described in Section 5, while Section 6 describes the derived variables that will be integrated with the survey data in WP8.

3. Data sources

The integrated dataset draws on three types of inputs: survey data from the SoGreen partner infrastructures, historical weather data, and air quality data. The combinations of provider, modelling approach and resolution are summarised in Table 1.

3.1 Survey data (ESS, SHARE, GGP)

The SoGreen survey module is fielded across three infrastructures. For ESS the module is implemented within the CRONOS3 panel (Cross-National Online Survey, third iteration), a probability-based online panel of ESS respondents. The CRONOS3 panel began fielding earlier with other modules; the SoGreen module is carried in rounds 6, 7 and 8, fielded in 2025–2026 across eleven participating countries (Rød and Beuster 2026). SHARE and GGP field the same module within their respective fieldwork rounds. All three infrastructures release public scientific-use datasets that include a date of interview and a regional identifier at NUTS levels 1, 2 or 3, but no finer-grained spatial reference. The temporal detail in these public files varies — for GGS, for example, the public file reports only the interview month — whereas linkage to environmental data requires the interview date at daily resolution. For the integration, Sikt therefore works from a restricted, non-public version of the data that contains the exact day of interview, together with the regional identifier from which a plausible point location can be drawn under the geo-imputation method (Section 4).

3.2 Weather data (Open-Meteo Historical Weather API)

Historical weather variables are retrieved from the Open-Meteo Historical Weather API (Zippenfenig 2023), using the ERA5-Seamless reanalysis: air temperature and humidity come from ERA5-Land (Muñoz-Sabater et al. 2021) on a 0.1° grid (about 9 km), and wind and solar radiation from ERA5 (Hersbach et al. 2020) at about 0.25° (25 km). Daily values (mean, maximum and minimum) are retrieved for the European area of interest (25°W–45°E, 30°N–72°N). Full details of the API and the available models are documented at https://open-meteo.com/en/docs/historical-weather-api.

3.3 Air quality data (Open-Meteo Air Quality API, CAMS-backed)

Air quality variables are retrieved from the Open-Meteo Air Quality API (Zippenfenig 2023), which serves the Copernicus Atmosphere Monitoring Service (CAMS) European air-quality reanalysis, which builds on the MACC-II multi-model ensemble system (Marécal et al. 2015; Copernicus Atmosphere Monitoring Service 2024). This product provides hourly concentrations on a 0.1° grid (about 10 km) over Europe from 2013 onwards, covering the SoGreen fieldwork period, and is built from an eleven-model ensemble that assimilates European Environment Agency (EEA) in-situ observations. The five pollutants used in SoGreen (PM10, PM2.5, NO2, SO2 and O3) are all regularly validated against in-situ data. The reanalysis is preferred to raw monitoring-station series because it gives gap-free spatial coverage that suits the geo-imputation approach (Section 4) while still drawing on the same observations. Full details are documented at https://open-meteo.com/en/docs/air-quality-api.

Dataset	Provider	Underlying source	Resolution (space / time)	Coverage
Historical Weather API	Open-Meteo (open-meteo.com)	ECMWF ERA5-Seamless reanalysis: ERA5-Land + ERA5	0.1° temp. & humidity, 0.25° wind & radiation (ERA5-Seamless); hourly, daily aggregates	Global; SoGreen uses Europe 25°W–45°E, 30°N–72°N
Air Quality API	Open-Meteo (open-meteo.com)	CAMS European air-quality reanalysis: 11-model ensemble median, EEA-assimilated	0.1° / hourly	Europe 25°W–45°E, 30°N–72°N
ESS / CRONOS3	ESS ERIC, Centerdata, Sikt	CAWI/CAPI fieldwork 2025–26	NUTS 1–3 / interview date	11 SoGreen countries: AT, BE, CZ, FI, FR, GB, HU, IS, PL, PT, SI
SHARE	SHARE ERIC, Centerdata	Self-completion web & paper-and-pencil fieldwork, Jun–Jul 2025	NUTS 3 (also postcodes for contextual data) / interview date	6 SHARE countries (DE, EL, IT, NL, PL, SE)
GGP	GGP consortium	CAWI fieldwork (possibly some CAPI in Poland, TBC)	NUTS 2 (postcodes used for linkage) / interview date	Croatia, Poland (HR, PL)

Table 1. Data sources used in the SoGreen integrated dataset.

4. Sampling and spatial linkage

Linking gridded environmental data to individual survey responses requires a spatial reference for each respondent. The SoGreen partner infrastructures release only regional identifiers, typically at NUTS 3, NUTS 2 or NUTS 1 level. Direct linkage at the address or coordinate level is not used. Instead, the project applies the pseudonymity-preserving geo-imputation method developed in deliverable D7.1 (Fenton 2025), summarised here in compact form.

For each survey case, the regional identifier (and, where available, a coarser additional attribute such as urban-versus-rural setting) is passed to a sampling-point generation service. The service draws one or more random points from within the region, with the draw weighted by high-resolution gridded population counts from WorldPop (Stevens et al. 2015). The drawn point represents a plausible residential location for the case: it lies near settled areas and not in uninhabited terrain, but does not correspond to any published spatial unit and cannot be reverse-inferred from the released contextual data. Environmental values at the drawn point are then extracted from the Open-Meteo daily product and applied to the case. Only the regional identifier and a single-use throwaway identifier are transferred between organisations; no exact residential coordinates are held, transferred or released. The approach is consistent with the principle of data minimisation set out in Article 5(1)(c) and Article 89 of the GDPR.

Within SoGreen the current intention is to generate one imputed location per case and to use that single point to extract values from all contextual datasets, so that all derived variables for a given respondent share a common spatial reference. This is methodologically simpler than independent draws per derived variable and avoids the introduction of within-respondent spatial variation that would not correspond to any real exposure pattern. A test of error and bias in the

geo-imputation approach is planned within Work Package 7 using specimen address data and will be reported in the survey documentation (Fenton 2025).

5. Raw input variables

Raw input variables are retrieved from the Open-Meteo daily products listed in Table 2. Weather variables are obtained at daily resolution (mean, maximum and minimum where available); air-quality concentrations are obtained at hourly resolution and aggregated as required by the derived variable definitions in Section 6. All variables are retrieved at the imputed point coordinates for each survey case (Section 4) and merged on the interview date.

Variable (source name)	Label	Unit	Source product
temperature_2m_mean	Mean daily air temperature at 2 meters above ground	°C	Open-Meteo Historical Weather API
temperature_2m_max	Maximum daily air temperature at 2 meters above ground	°C	Open-Meteo Historical Weather API
temperature_2m_min	Minimum daily air temperature at 2 meters above ground	°C	Open-Meteo Historical Weather API
apparent_temperature_mean	Mean daily apparent temperature	°C	Open-Meteo Historical Weather API
apparent_temperature_max	Maximum daily apparent temperature	°C	Open-Meteo Historical Weather API
apparent_temperature_min	Minimum daily apparent temperature	°C	Open-Meteo Historical Weather API
precipitation_sum	Sum of daily rain	mm	Open-Meteo Historical Weather API
wind_gusts_10m_max	Maximum wind gusts on a day	m/s	Open-Meteo Historical Weather API
pm10	Particulate matter with diameter smaller than 10 µm (PM10) close to surface (10 meter above ground)	µg/m³	Open-Meteo Air Quality API (CAMS)
pm2_5	Particulate matter with diameter smaller than 2.5 µm (PM2.5) close to surface (10 meter above ground)	µg/m³	Open-Meteo Air Quality API (CAMS)
nitrogen_dioxide	Atmospheric gas close to surface (10 meter above ground)	µg/m³	Open-Meteo Air Quality API (CAMS)
sulphur_dioxide	Atmospheric gases close to surface (10 meter above ground)	µg/m³	Open-Meteo Air Quality API (CAMS)
ozone	Atmospheric gases close to surface (10 meter above ground)	µg/m³	Open-Meteo Air Quality API (CAMS)

Table 2. Raw input variables retrieved from Open-Meteo APIs.

In addition to the environmental inputs, the integrated dataset retains a set of survey variables from the SoGreen module and the ESS main questionnaire, used both as analytical covariates and to link the contextual derived variables to respondent characteristics, attitudes and reported experiences. These variables are listed in Table 3, grouped by their source module within the questionnaire; the variable names shown are the ESS variable names.

Variable	Label
From the ESS main questionnaire	
gndr	Gender
eduyrs	Years of full-time education completed
eisced	Highest level of education (ES-ISCED)
netusoft	Internet use, how often
mnactic	Main activity, last 7 days (post-coded; all respondents)
hincfel	Feeling about household's income nowadays
hinctnta	Household's total net income, all sources
region	Region
ctzcntr	Citizen of country
vote	Voted in last national election
mode	Mode of data collection
yrbrn	Year of birth
From the well-being module	
w6wq8	Feel safe and secure in your life
From the SoGreen core and extended module	
w6sgq2	Concern about air pollution in local area
w6sgq3	How often experience health issues related to air pollution
w6sgq11	How worried about climate change
w6sgq12	Extent of personal responsibility to reduce climate change
w6sgq13	Trust in [country] government to address climate change
w6sgq14	Priority preference between environmental protection and economic growth
w6sgq15	Familiarity with climate change policies in [country]
w6sgq16	Confidence that [country's] climate change policies consider everyone's views
w6sgq17	Confidence in fair outcomes of [country's] climate change policies
w6seq1_2	Experienced extreme weather in the past 5 years: drought
w6seq1_3	Experienced extreme weather in the past 5 years: wildfire
w6seq1_4	Experienced extreme weather in the past 5 years: heavy storm

Variable	Label
w6seq1_5	Experienced extreme weather in the past 5 years: extended extreme heat
w6seq1_6	Experienced extreme weather in the past 5 years: none
w6seq1_7	Experienced extreme weather in the past 5 years: no answer
w6seq2	Worry about extreme weather events in the next 5 years where you live
w6seq3	Had to move or consider moving due to extreme weather risks
w6seq4	Extent [country] government is prepared for extreme weather events
w6seq5	Air pollution impact on ability or willingness to go outside
w6seq6_1	Air-pollution issue experienced: temporary ban on outdoor burning activities
w6seq6_2	Air-pollution issue experienced: restrictions on entering certain city areas
w6seq6_3	Air-pollution issue experienced: advised to reduce or avoid fossil-fuel vehicles
w6seq6_4	Air-pollution issue experienced: limited outdoor recreation opportunities
w6seq6_5	Air-pollution issue experienced: respiratory issues
w6seq6_6	Air-pollution issue experienced: reduced visibility
w6seq6_7	Air-pollution issue experienced: closure or limited access to public spaces
w6seq6_8	Air-pollution issue experienced: increased health warnings from authorities
w6seq6_9	Air-pollution issue experienced: other
w6seq6_10	Air-pollution issue experienced: none of the above
w6seq6_11	Air-pollution issue experienced: no answer
w6seq7	Had to move or consider moving due to air pollution

Table 3. Related ESS survey variables retained for analysis and linkage.

6. Derived variables

The derived variables in the integrated SoGreen dataset are organised in four groups: climate derived variables (Section 6.1), thermal-stress derived variables based on UTCI (Section 6.2), air quality derived variables based on the European Air Quality Index (Section 6.3), and health-based air quality derived variables based on the WHO 2021 guidelines (Section 6.4). All derived variables share three methodological conventions. First, they are referenced to one of three temporal anchors: the interview date itself, the calendar month containing the interview date, or the 365-day period ending on the interview date (in line with the design choices that emerged from the working group following the review of the EOSC Future derived variable catalogue). Second, anomalies and standardised measures are computed against the 1991–2020 baseline period, consistent with current World Meteorological Organization practice. This baseline is computed from the same ERA5-Seamless daily archive, extracted at the same imputed point coordinates as the interview-period values, so that each anomaly contrasts like with like; the 1991–2020 normal period lies well within the archive coverage (ERA5 from 1940, ERA5-Land from 1950). Third, all spatially-resolved values are extracted at the imputed point coordinates for the case (Section 4).

6.1 Climate derived variables

The climate derived variables retained from the EOSC Future catalogue describe daily and monthly temperature, precipitation and wind-gust conditions experienced around the time of interview. Indicators referring to the calendar month of interview are constructed once the full month has elapsed, so that the same definition can be applied to every case in a wave. Anomalies are expressed in the native unit of the underlying variable (e.g. °C for temperature) and are computed as the value for the reference period minus the corresponding 1991–2020 mean for the same calendar month and location. The full set of climate derived variables is given in Table 4.

Variable	Label	Unit	Type
tmpdca	Temperature in degrees Celsius, interview date average	°C	decimal
tmpdcmx	Temperature in degrees Celsius, interview date maximum	°C	decimal
tmpdcmn	Temperature in degrees Celsius, interview date minimum	°C	decimal
tmpdcacm	Temperature in degrees Celsius, calendar month of interview average	°C	decimal
tmpdcamb	Temperature in degrees Celsius, multi-year calendar month of interview averages, baseline 1991–2020	°C	decimal
tmpanod	Temperature anomaly, interview date	°C	decimal
tmpanocm	Temperature anomaly, calendar month of interview	°C	decimal
pacct	Total precipitation, interview date	mm	decimal
pacctcm	Total precipitation, calendar month of interview	mm	decimal
pacctmb	Total precipitation, multi-year calendar month of interview averages, baseline 1991–2020	mm	decimal
paccdcm	Total precipitation, calendar month of interview relative to normal	mm	decimal
iwg10mx	Instantaneous 10 metre maximum wind gust, interview date	m/s	decimal
iwg10mxamb	Instantaneous 10 metre maximum wind gust maximum, multi-year calendar month of interview averages, baseline 1991–2020	m/s	decimal
iwg10anocm	Instantaneous 10 metre maximum wind gust maximum, calendar month of interview average relative to normal	m/s	decimal

Table 4. Climate derived variables in the integrated dataset.

6.2 Thermal-stress derived variables (UTCI-based)

The Universal Thermal Climate Index (UTCI) combines air temperature, humidity, wind speed and radiation into a single equivalent temperature representing the human physiological response to outdoor thermal conditions (Bröde et al. 2012). UTCI categories are widely used to flag thermally stressful days; in particular, daily maximum UTCI values at or above 32 °C correspond to strong, very strong or extreme heat stress, and daily minimum UTCI values at or below −13 °C correspond to strong, very strong or extreme cold stress (European Environment Agency 2025). For SoGreen WP7 the derived variables in Table 5 are constructed from UTCI values computed in the Open-Meteo product. Indicators marked "applies" / "does not apply" are coded 0/1; count derived variables are integers. The "three-day period" derived variables count consecutive runs of at least three days meeting the stress threshold and capture sustained stress episodes, which the literature suggests are more behaviourally relevant than isolated extreme days (Hoffmann et al. 2022).

Variable	Label	Unit	Type
utcishd	Strong thermal heat stress, interview date	0/1	integer
utciscd	Strong thermal cold stress, interview date	0/1	integer
utcishndy	Number of days with strong heat stress, year before interview date	count	integer
utciscndy	Number of days with strong cold stress, year before interview date	count	integer
utcishn3dy	Number of three-day periods with strong heat stress, year before interview date	count	integer
utciscn3dy	Number of three-day periods with strong cold stress, year before interview date	count	integer

Table 5. Thermal-stress derived variables based on the Universal Thermal Climate Index (UTCI).

Type/Unit codes: 0 = does not apply, 1 = applies; count = number of days/periods (integer).

6.3 Air quality derived variables (EEA AQI-based)

The European Air Quality Index (AQI) developed by the European Environment Agency (European Environment Agency 2024) assigns daily concentration values for each pollutant to one of six categories — Good (0), Fair (1), Moderate (2), Poor (3), Very poor (4) and Extremely poor (5) — based on health-impact thresholds. For SoGreen WP7 the AQI level is computed at the imputed point coordinates for each respondent on the interview date, for each of the five pollutants individually and as the worst level across all pollutants. To capture exposure over longer windows, counts of "poor or worse" days are constructed for the calendar month of interview and for the 365 days ending on the interview date. The derived variable set is shown in Table 6.

Variable	Label	Unit	Type
aqiwdpm10	Worst European Air Quality Index level PM10, interview date	0–5	integer
aqiwdpm2_5	Worst European Air Quality Index level PM2.5, interview date	0–5	integer
aqiwdso2	Worst European Air Quality Index level SO2, interview date	0–5	integer
aqiwdno2	Worst European Air Quality Index level NO2, interview date	0–5	integer
aqiwdo3	Worst European Air Quality Index level O3, interview date	0–5	integer
aqiwd	Worst European Air Quality Index level across pollutants, interview date	0–5	integer
ndyprm	Number of days with 'poor' European Air Quality Index level or worse on one or more pollutants, calendar month of interview date	count	integer
ndypry	Number of days with 'poor' European Air Quality Index level or worse on one or more pollutants, year before interview date	count	integer

Table 6. Air quality derived variables based on the European Air Quality Index.

AQI level (0–5): 0 = Good, 1 = Fair, 2 = Moderate, 3 = Poor, 4 = Very poor, 5 = Extremely poor; count = number of days (integer).

6.4 Health-based air quality derived variables (WHO 2021 guidelines)

The WHO global air quality guidelines (WHO 2021) define annual and short-term concentration thresholds for the most health-relevant pollutants. For SoGreen WP7, four health-based derived variables are constructed for PM2.5 and NO_2, the two pollutants most strongly linked to mortality and respiratory disease in the European context (Schraufnagel et al. 2019; WHO 2021). The

annual derived variables flag whether the mean concentration of the pollutant over the 365 days ending on the interview date equals or exceeds the WHO annual guideline (5 µg/m³ for PM2.5; 10 µg/m³ for NO_2). The short-term derived variables count three-or-more-day consecutive episodes in the same 365-day window during which the daily-mean concentration equals or exceeds the WHO 24-hour guideline (15 µg/m³ for PM2.5; 25 µg/m³ for NO_2). The four derived variables are listed in Table 7.

Variable	Label	Unit	Type
whohypm2_5	Harmful air quality for PM2.5, year before interview date, WHO guidelines	0/1	integer
whohyno2	Harmful air quality for NO2, year before interview date, WHO guidelines	0/1	integer
whoh3dypm2_5	Number of three-day periods with harmful air quality for PM2.5, year before interview date, WHO guidelines	count	integer
whoh3dyno2	Number of three-day periods with harmful air quality for NO2, year before interview date, WHO guidelines	count	integer

Table 7. Health-based air quality derived variables based on the WHO 2021 guidelines.

Type/Unit codes: 0 = does not apply, 1 = applies; count = number of three-day periods (integer).

7. Provenance and reproducibility

Each derived variable in the integrated SoGreen dataset comes with a provenance record that traces it back to its origins: the source dataset and version the raw values come from, where they were retrieved (the API endpoint or download channel), the resolution of the raw input, the rule used to turn raw values into the derived variable (formula, threshold, aggregation, baseline), and the processing code and its version. This record is kept in two forms — a human-readable row in the derived variable catalogue (Table 8), and a machine-readable description following the DDI Cross-Domain Integration (DDI-CDI) process model piloted in the EOSC Future Science Project (Orten et al. 2023; DDI Alliance 2023).

The DDI-CDI model lets us describe the workflow at two levels. General tasks — such as retrieving data from an API or running a data-processing routine — are recorded as activities, described in plain text and linked to the data they use and produce. The more demanding tasks, above all the computation of each derived variable, are recorded as steps that list the input variables, the output variable, and the script that does the work, so the path from raw input to derived value can be followed from end to end (an API call can also be written up as a step where more detail is useful). For SoGreen the variable computations are the priority and are documented in the greatest detail. A worked example from the EOSC Future predecessor serves as the template, and a tool to generate and visualise these provenance descriptions will be built in Work Package 8.

Provenance element	Description	DDI-CDI class / property	Example
Activity (overall operation)	A higher-level processing operation described in prose — e.g. an API retrieval or a data-processing routine — without embedded code; it groups the detailed steps and may nest sub-activities.	cdi:Activity → cdi:Activity-name, cdi:Activity-description, cdi:Activity-standardModelMapping; cdi:Activity_has_Step; cdi:Activity_hasSubActivity_Activity	“EEA / CAMS data processing”: retrieve and aggregate air-quality data to regions
Activity entities	The data products an activity	cdi:Activity-entityUsed /	used: CAMS

Provenance element	Description	DDI-CDI class / property	Example
used / produced	consumes and generates, referenced by URL rather than as individual variables.	cdi:Activity-entityProduced → cdi:Reference (URI)	reanalysis product (URL); produced: regional air-quality file (URL)
Step (variable computation)	Computation of one derived (output) variable from its input variables; holds the computation rule in prose and carries the executable code.	cdi:Step → cdi:Activity-name, cdi:Activity-description, cdi:Activity-standardModelMapping	Compute 'iwg10mxam' from 'iwg10mx', 'region_id', 'date'
Step inputs	Variables the step consumes — raw retrieved variables or variables produced by earlier steps	cdi:Step_receives_Parameter → cdi:Parameter (cdi:InstanceVariable)	iwg10mx, region_id, date
Step output	The derived variable the step produces	cdi:Step_produces_Parameter → cdi:Parameter (cdi:InstanceVariable)	iwg10mxam
Script and language	Executable code implementing the step and the language used (steps only; activities carry no code attribute).	cdi:Step-script → cdi:CommandCode / cdi:CommandFile; cdi:Step-scriptingLanguage	Python; GitHub sikt-no/sogreen-wp7, commit <sha>
Variable definition	Concept, value domain and code list defining each variable	cdi:InstanceVariable / cdi:Concept / cdi:ValueDomain / cdi:CodeList	AQI levels 0–5 (Good…Extremely poor)
Source and version	Upstream product and version from which raw variables are retrieved	cdi:Reference / cdi:Identifier	ERA5-Seamless; CAMS European reanalysis
Step order	Position of the step within the processing sequence	cdi:Sequence / cdi:SequencePosition	step n of N
Environment and agent	Software environment/version and the organisation responsible	cdi:ProductionEnvironment; cdi:Agent / cdi:Organization	Python <version>; ESS ERIC / Sikt

Table 8. Provenance record for each derived variable and its DDI-CDI representation.

Provenance documentation is designed as a structural feature of the integration workflow rather than a separable post-hoc activity. Work Package 7 (the project's first eighteen months) is mainly planning and conceptual work, including this deliverable's methodology; the climate-data integration and its provenance documentation are carried out in Work Package 8. There, code, derived variable definitions and dataset releases are versioned together, so that any change in a transformation rule or source product propagates to the derived variable and its provenance record — consistent with the FAIR principles (Wilkinson et al. 2016), in particular Reusability, and with the ESS/CRONOS3 processing standards in D9.3 (Rød and Beuster 2026).

8. Limitations and disclosure risk

The integration approach described in this paper has several known limitations. Spatially, the geo-imputation method introduces deliberate randomisation in the assignment of contextual values to cases (Fenton 2025). For point-level variables this means that the value linked to a case is a plausible local value rather than a value measured at the respondent's residence; the magnitude of the resulting error cannot be estimated without reference to exact respondent

coordinates, and a dedicated bias-and-error test is therefore being conducted in Work Package 7 using specimen address data. Temporally, derived variables referenced to the interview date are sensitive to the time of day on which the interview took place: the daily-aggregated environmental value can be influenced by hours of the day that the respondent had not yet experienced when responding. This is partially mitigated by the calendar-month and 365-day derived variables, which dominate the literature on attitudinal effects (Hoffmann et al. 2022).

Air-quality values exhibit substantial within-region variability, particularly for primary pollutants such as NO_2 where concentrations near major roads can differ markedly from urban background levels. The CAMS reanalysis used for SoGreen provides a smoother estimate than dense in-situ networks and is therefore conservative for studying exposure at the imputed point. In analyses that combine the released SoGreen contextual variables with regional survey identifiers, strong within-region variation of pollutant levels may in principle increase disclosure risk; analyses of statistical variation of within-region air-quality values will be performed during preparation of the released datasets so that any residual disclosure risk can be detected and mitigated (Fenton 2025). The same logic applies to weather variables, though their typical within-region variability is lower than for air quality.

A further limitation is that the SoGreen integrated dataset captures a snapshot of environmental conditions linked to the interview date and a backward-looking 365-day window. It does not attempt to reconstruct respondents' lifetime or recent exposure histories, nor their movement patterns. The derived variables are designed for analyses of how recent local environmental conditions relate to attitudes and behaviours captured in the interview, and should not be interpreted as proxies for chronic exposure.

9. Conclusion and next steps

This paper has documented the methods and provenance developed in SoGreen Work Package 7 for integrating climate and air quality data with three European social-survey infrastructures. The approach builds on the EOSC Future Science Project predecessor (Orten et al. 2023) and extends it in three respects: a multi-infrastructure design covering ESS, SHARE and GGP; an derived variable catalogue enriched with UTCI-based thermal-stress derived variables and WHO-based health derived variables alongside the established European Air Quality Index; and a pseudonymity-preserving geo-imputation method that supports point-level extraction of contextual values without ever transferring or holding exact respondent locations (Fenton 2025). Each derived variable is accompanied by a full provenance record linking it to its upstream source, version, transformation rule and processing code.

The next steps within SoGreen move this methodology into production in Work Package 8: (i) extracting the interview date for each respondent in the ESS, SHARE and GGP surveys; (ii) obtaining the imputed point coordinates for each case from the geo-imputation service (Fenton 2025); and (iii) calling the Open-Meteo weather and air-quality APIs at those coordinates and dates to collect the contextual data, from which the derived variables are computed and released through each infrastructure's standard channels under FAIR-aligned governance (Wilkinson et al. 2016; Rød and Beuster 2026).

Acknowledgements

This deliverable was produced within Work Package 7 of the SoGreen project. SoGreen has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101188188. The authors thank colleagues at ESS ERIC, SHARE-ERIC, the GGP consortium, SGH Warsaw School of Economics, and Sikt for

collaboration on the methodology described here, and conferred with the Climate and Environmental Research Institute (NILU); the authors also acknowledge the EOSC Future Science Project "Climate Neutral and Smart Cities" (Horizon 2020 grant 101017536) on which the derived variable approach builds. The authors are also indebted to expert advice received during the EOSC Future project from the Norwegian Meteorological Institute (MET Norway).

Preparation of this document was supported in part by AI-assisted tools (Claude, Anthropic) for drafting, structuring and language refinement; the authors remain fully responsible for all content.

References

Bergquist, Parrish, and Christopher Warshaw. 2019. "Does Global Warming Increase Public Concern about Climate Change?" The Journal of Politics 81 (2): 686–691. https://doi.org/10.1086/701766.

Bröde, Peter, Dusan Fiala, Krzysztof Błażejczyk, Ingvar Holmér, Gerd Jendritzky, Bernhard Kampmann, Birgit Tinz, and George Havenith. 2012. "Deriving the Operational Procedure for the Universal Thermal Climate Index (UTCI)." International Journal of Biometeorology 56 (3): 481–494. https://doi.org/10.1007/s00484-011-0454-1.

Copernicus Atmosphere Monitoring Service (CAMS). 2024. CAMS European Air Quality Reanalyses. Copernicus Atmosphere Data Store (ADS). https://ads.atmosphere.copernicus.eu/datasets/cams-europe-air-quality-reanalyses.

DDI Alliance. 2023. DDI–CDI: Integrating Data for Better Science. Public Review 1. DDI Alliance. https://ddialliance.org/ddi-cdi.

European Environment Agency. 2024. European Air Quality Index. Copenhagen: EEA. https://airindex.eea.europa.eu/.

European Environment Agency. 2025. "High UTCI Days." Indicator metadata, Climate-ADAPT. https://climate-adapt.eea.europa.eu/en/metadata/indicators/high-utci-days.

Fenton, Alex. 2025. "Geo-Imputation: Pseudonymous Linkage of Spatial Contextual Data." SoGreen Deliverable D7.1, Horizon Europe grant agreement No. 101188188. 23 December 2025. https://sogreen-project.eu/wp-content/uploads/2026/01/D.7.1.-Regional-Data-Concept.pdf.

Hamilton, Lawrence C., and Mary D. Stampone. 2013. "Blowin' in the Wind: Short-Term Weather and Belief in Anthropogenic Climate Change." Weather, Climate, and Society 5 (2): 112–119. https://doi.org/10.1175/WCAS-D-12-00048.1.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, et al. 2020. "The ERA5 Global Reanalysis." Quarterly Journal of the Royal Meteorological Society 146 (730): 1999–2049. https://doi.org/10.1002/qj.3803.

Hoffmann, Roman, Raya Muttarak, Jonas Peisker, and Piero Stanig. 2022. "Climate Change Experiences Raise Environmental Concerns and Promote Green Voting." Nature Climate Change 12 (2): 148–155. https://doi.org/10.1038/s41558-021-01263-8.

Leiserowitz, Anthony, Edward Maibach, Connie Roser-Renouf, and Jay D. Hmielowski. 2012. Extreme Weather, Climate and Preparedness in the American Mind. New Haven, CT: Yale Project on Climate Change Communication.

Marécal, Virginie, Vincent-Henri Peuch, Christian Andersson, Stephen Andersson, Joaquim Arteta, Mathieu Beekmann, Anna Benedictow, et al. 2015. "A Regional Air Quality Forecasting System over Europe: The MACC-II Daily Ensemble Production."

Geoscientific Model Development 8 (9): 2777–2813. https://doi.org/10.5194/gmd-8-2777-2015.

Muñoz-Sabater, Joaquín, Emanuel Dutra, Anna Agustí-Panareda, et al. 2021. "ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications." Earth System Science Data 13 (9): 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.

Orten, Hilde, David Rayner, Eirik Stavestrand, Iris Alfredsson, Ilze Lace, Irena Vipavc Brvar, Maja Dolinar, Joachim Wackerow, and Hannah Clark. 2023. "Climate and Air Quality Indices for the European Social Survey." Research Article, RIO/ARPHA Preprints. https://doi.org/10.3897/arphapreprints.e114675.

Rød, Linn-Merethe, and Benjamin Beuster. 2026. "Codebook and Processing Guidelines." SoGreen Deliverable D9.3, Horizon Europe grant agreement No. 101188188. 31 March 2026.

Rød, Linn-Merethe, and Benjamin Beuster. 2026. "Preliminary CRONOS 1 Data." SoGreen Deliverable D9.4, Horizon Europe grant agreement No. 101188188. 10 May 2026. https://sogreen-project.eu/wp-content/uploads/2026/06/D9.4.pdf.

Schraufnagel, Dean E., John R. Balmes, Clayton T. Cowl, Sara De Matteis, Soon-Hee Jung, Kevin Mortimer, Rogelio Perez-Padilla, et al. 2019. "Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 2." Chest 155 (2): 417–426. https://doi.org/10.1016/j.chest.2018.10.041.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." PLOS ONE 10 (2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

Whitmarsh, Lorraine. 2008. "Are Flood Victims More Concerned about Climate Change than Other People? The Role of Direct Experience in Risk Perception and Behavioural Response." Journal of Risk Research 11 (3): 351–374. https://doi.org/10.1080/13669870701552235.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18.

World Health Organization. 2021. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. Geneva: WHO. https://www.who.int/publications/i/item/9789240034228.

Zippenfenig, Patrick. 2023. Open-Meteo.com Weather API. Zenodo. https://doi.org/10.5281/zenodo.7970649.