GRETA
GREEN ENERGY TRANSITION ACTIONS

D4.1
Data integration and pre-processing

Prepared by: LUT
Contributions from: TEC, TNO, GESIS, CWD, UNIBO, FhG

Deliverable nature	Report
Dissemination level (Confidentiality)	Public (PU)
Delivery date	2022-10-31
Version	1.0
Total number of pages	42
Keywords	Data onboarding, data pre-processing, energy communities, energy citizenship
Cite as	Kuronen, T., et al. (2022) Data Integration and Pre-Processing. D4.1 of the Horizon 2020 project GRETA, EC grant agreement no 101022317, Lappeenranta, Finland
Project contact	Toni Kuronen, email: toni.kuronen@lut.fi

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101022317.

Disclaimer and acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101022317.

Executive summary

This data integration and pre-processing document is prepared in the context of WP4 – Data processing and Explicit modelling (Task 4.1, data integration and pre-processing) - of the GRETA project under Grant Agreement No. 101022317.

The overall objective of T4.1 is to (1) frame and identify the available data, (2) assess the relevancy of the data in terms of incorporating the emergence of energy citizens in the models, and (3) integrate, onboard, and pre-process the relevant data.

Chapter one of this document introduces the topic in the context of the GRETA project, WP4, and T4.1. Chapter two considers the data gathering, data needs, and data relevancy. This is achieved by linking the work of earlier work packages and using a WP4 related data workshop. Moreover, data from case studies, multinational survey and supplementary and secondary data are explained in the second chapter. Chapter three gives an outline for the data onboarding and processing practices including general data processing and specific data processing. Finally, chapter four concludes the document.

Data collection and processing in GRETA will follow the applicable national and EU legislation, especially, General Data Protection Regulation 679/2016 (GDPR) which is described in more detail in the D8.6 – Ethics and Privacy Management Plan. (Landeck J. , 2021) (European Commission, 2016)

Project information

Grant agreement No.	101022317
Acronym	GRETA
Full title	GReen Energy Transition Actions
H2020 Topic	H2020-LC-SC3-2020-NZE-RES-CC
Project URL	www.projectgreta.eu

Document information

	Number	Title
Deliverable	D4.1	Data integration and pre-processing
Work package	WP4	Data processing and Explicit modelling
Task	T4.1	Data integration and pre-processing

Delivery date	Contractual: M18, Actual: M18
Nature	☒Report ☐Other ☐ORDP
Dissemination level	☒Public ☐Confidential
Authors (partners)	Toni Kuronen (LUT), Nekane Hermoso (TEC), Anne Kantel (FhG), Martina Massari (UNIBO), Lurian Klein (CWD), Carlos Montalvo (TNO), and Lasse Lensu (LUT)
Reviewers (partners)	Lurian Klein (CWD)
Summary (for dissemination)	This report addresses the data integration and pre-processing of the GRETA project. Data collection and processing aims to support the work package 4 models that are an integral part of the GRETA project, which investigates the emergence of energy citizenship from the local level to the supranational level. The document contains information about: (1) data gathering, (3) relevant data identification, (4) data pre-processing, and (4) data onboarding, storing, and sharing.
Keywords	Data onboarding, pre-processing, energy communities, energy citizenship

Version	Date	Description
0.1	2022-02-28	The first draft (TOC)
0.2	2022-10-13	Second draft
0.3	2022-10-20	Third draft (for the internal review)

Table of contents

Disclaimer and acknowledgement **2**
Executive summary **3**
Project information **4**
Document information **5**
Table of contents **6**
List of Figures **8**
List of Tables **9**
Abbreviations and acronyms **10**
1 Introduction **11**
1.1 WP4 overview 11
1.1.1 T4.1 overview 11
1.2 Deliverable overview 11
2 Data supporting the emergence of energy citizenship **13**
2.1 WP4 workshop data 14
2.2 Data for developing the non-energy related models. 18
2.3 Data for developing energy-related models 19
2.4 Case studies data 20
2.4.1 CS1 (Renewable energy district – Bologna Pilastro-Roveri, Italy): 20
2.4.2 CS2 (Natural gas-free neighbourhoods, The Netherlands): 23
2.4.3 CS3 (Coopérnico – Renewable energy-driven cooperative, Portugal): 25
2.4.4 CS4 (UR BEROA – Energy efficiency-driven cooperative, Spain): 26
2.4.5 CS5 (Earnest App Case - A virtual community for sustainable mobility in Darmstadt, Germany): 30
2.4.6 CS6 (Electric autonomous and connected mobility network): 31
2.4.7 Case studies data summary 31
2.5 Multinational survey data 31
2.6 Supplementary/Secondary data 32
3 Data onboarding and processing **34**
3.1 General data pre-processing 34
3.1.1 Data Cleaning 34
3.1.2 Data Integration 35
3.1.3 Data transformation & reduction 35
3.1.4 Data format & processing 35
3.1.5 Geolocation processing and anonymization 36
3.1.6 Data organisation and Exploration 36
3.1.7 Data storage, sharing & onboarding 36
3.2 Specific data processing with energy related models 37
3.2.1 Data format 37

4 Conclusions 40
References..... 41

List of Figures

Figure 1 Identified data aims from the T4.1 Workshop 15

Figure 2 Identified data and modelling solutions (T4.1 WS) 16

Figure 3 Data and modelling challenges and barriers that were identified (T4.1 WS) 17

Figure 4 Future ideas and possible actions to take with data and models. (T4.1 WS) 18

Figure 5 Example of the unified geometries vs. the divided ones 21

Figure 6 Manual definition of the construction years by zone 22

Figure 7 Buildings that are part of the heritage (in orange) 22

Figure 8 Example of the parameters included in the final layer 23

Figure 9 Example of data per building unit 27

Figure 10 Example of data per unit/dwelling 27

Figure 11 BES data available from the EPC (red) and the municipal register (yellow) 28

Figure 12 Average household income of the 3 census tracts included in the study area 29

Figure 13 Example of hourly heating consumption profiles obtained for various DH substations. 30

Figure 14 data pre-processing 34

Figure 15 List of available formats for exporting results 39

List of Tables

Table 1 GRETA data overview 14

Table 2 Example of the most relevant input data for the BSEM generation 20

Table 3 List of the input data used for the energy model generation 37

Table 4 Main energy-related parameters at the building level for the baseline scenario 38

Abbreviations and acronyms

BSEM	Building stock energy model
CCAM	Connected and cooperative automated mobility
CLI	Community level indicator
CS#	Case study (#)
CTP	Community transition pathway
DHW	Domestic hot water
EU	European Union
FhG ISI	Fraunhofer institute for systems and innovation research ISI
GIS	Geographic information system
GDPR	General data protection regulation
GRETA	Green energy transition actions
OPCE	GRETA open portfolio for civic energy empowerment
PV	Photovoltaic
T#.#	Task (#.#)
UNIBO	Università di Bologna, Italy
WP#	Work package (#)
WS	Workshop

1 Introduction

GRETA intends to examine the global energy transition and especially the social side of it. Frameworks and models developed within GRETA aim to reveal the factors that influence the emergence of energy citizenship and energy transition. Those will be developed based on six case studies and a multinational survey. Data collection also forms the base for creating an EU survey database that contains European-wide data on energy citizenship practices. Moreover, the data should be processed so that it would allow the improved utilisation of energy-related data and allow better energy communications and knowledge building.

1.1 WP4 overview

WP4 aims to provide properly founded energy and non-energy models to form a solid foundation for the GRETA project analysis. This is achieved using diverse types of data, originating from surveys, interviews and existing datasets, and computational models. The models can be used as components in systems and services for guiding individuals and communities in their current energy citizen behaviours and predict which actions are the most prominent to achieve their decarbonisation and personal goals. Collaboration with WP2 deepens this connection between individuals and models by including the aspects of sensemaking and energy informatics/energy literacy.

1.1.1 T4.1 overview

This task ensures that the gathered data are integrated and pre-processed in a way that is used as inputs to the different models developed in WP4. This includes onboarding and storing all the data used in WP4 models. T4.1 also addresses the relevancy of data regarding the incorporation of the emergence of energy citizens in the models. This was achieved by establishing links with WP1 and WP2. Moreover, data gathered within GRETA is complemented with external data sources to cover all the data requirements within the diverse types of developed WP4 models. Lastly, to meet all the format requirements, these data are pre-processed so that they can be used in the developed models.

1.2 Deliverable overview

The overall aims of the deliverable are to provide guidelines for pre-processing and integration of the data gathered in other work packages and/or from other supplementary data sources. These data will then be used as inputs to the different WP4 models. D4.1 objectives are:

- To frame and identify the available data
- To assess the relevancy of the data in terms of incorporating the emergence of energy citizens in the models
- Integrate, onboard, and pre-process the relevant data
- Define the data storage

2 Data supporting the emergence of energy citizenship

Data for the GRETA project will be obtained from multiple sources. Data was collected from a multinational survey and from six GRETA case studies, using local surveys, workshops, interviews, as well as existing databases and documents. In the context of WP4, this data will be used to conduct analyses and process information regarding attitudes, social norms and engagement, to determine the factors behind energy citizens' engagement, aid the development of models that predict possible trends and envision different potential scenarios at the local improve energy informatics, and to develop behavioural models to assess possible scenarios and conditions to determine the emergence of energy citizenship. Table 1 provides an overview of the data collected from each case study (X indicates that the data is available).

The purpose of WP4 and GRETA is to gather and produce relevant data in sufficient amounts to support appropriate decision-making – i.e., data that allows defining the factors that support decarbonisation and lower emission behaviour, and data that helps to understand what influences stakeholders' decisions.

The data-gathering process is guided by the following definitions in the GRETA glossary:

"Energy Citizen

An energy citizen is an individual who participates individually or collectively in the transition of energy systems in a particular geographical area. Energy citizens use, consume, produce and/or store energy in an improved or reduced manner. Energy citizens' activities and behaviours affect the decarbonisation of current energy systems in the long run. Their energy-related knowledge, when shared, allow energy citizens to have also an advocacy role. The effects can be positive (e.g supporting the clean energy transition, investing in energy-efficient appliances or participating in a local energy initiative), negative (e.g. public resistance to new forms of renewable energy) or neutral. " (Project GRETA, 2022)"

"Energy Citizenship

Energy citizenship is about the active participation of citizens in energy systems in a particular geographical area. Active participation can both be socially and politically, either as individuals (e.g., through energy efficiency measures in households) or in larger groups (e.g., through engagement with energy policy in climate activist groups or energy communities). The effects of energy citizenship can be positive (e.g supporting the clean energy transition, investing in energy-efficient appliances or participating in a local energy initiative), negative (e.g. public resistance to new forms of renewable energy) or neutral. " (Project GRETA, 2022)"

Table 1 GRETA data overview

	Building	Interview	Workshop	Survey
CS1	X	X	X	
CS2	X	X	X	X
CS3	Not applicable	X	X	X
CS4	X	X	X	X
CS5	Not applicable	X	X	X
CS6	Not applicable	X	X	X

2.1 WP4 workshop data

A workshop was conducted on the 28th of September 2022 among WP4 participants (but open to all the consortium) to align views about data usage, solution, barriers/challenges, and future ideas within WP4. The objective of the workshop was to have a complete idea about the data availability, needs, and usage considering WP4 and different developed models. One of the goals was to define what kind of data has been used and could be used in the future, i.e., where one uses data and what kind, and which are the data requirements.

The data insights deriving from the workshop can be seen in Figures 1-4. Figure 1 presents insights from the first part of the workshop, which considered data aims - i.e., what participants expect to find and determine based on the available data. According to the results, the data should: be relevant; support the emergence of energy citizenship; allow individuals and energy communities to determine where they are related to others; be of good quality; and meet the needs of the developed WP4 models.

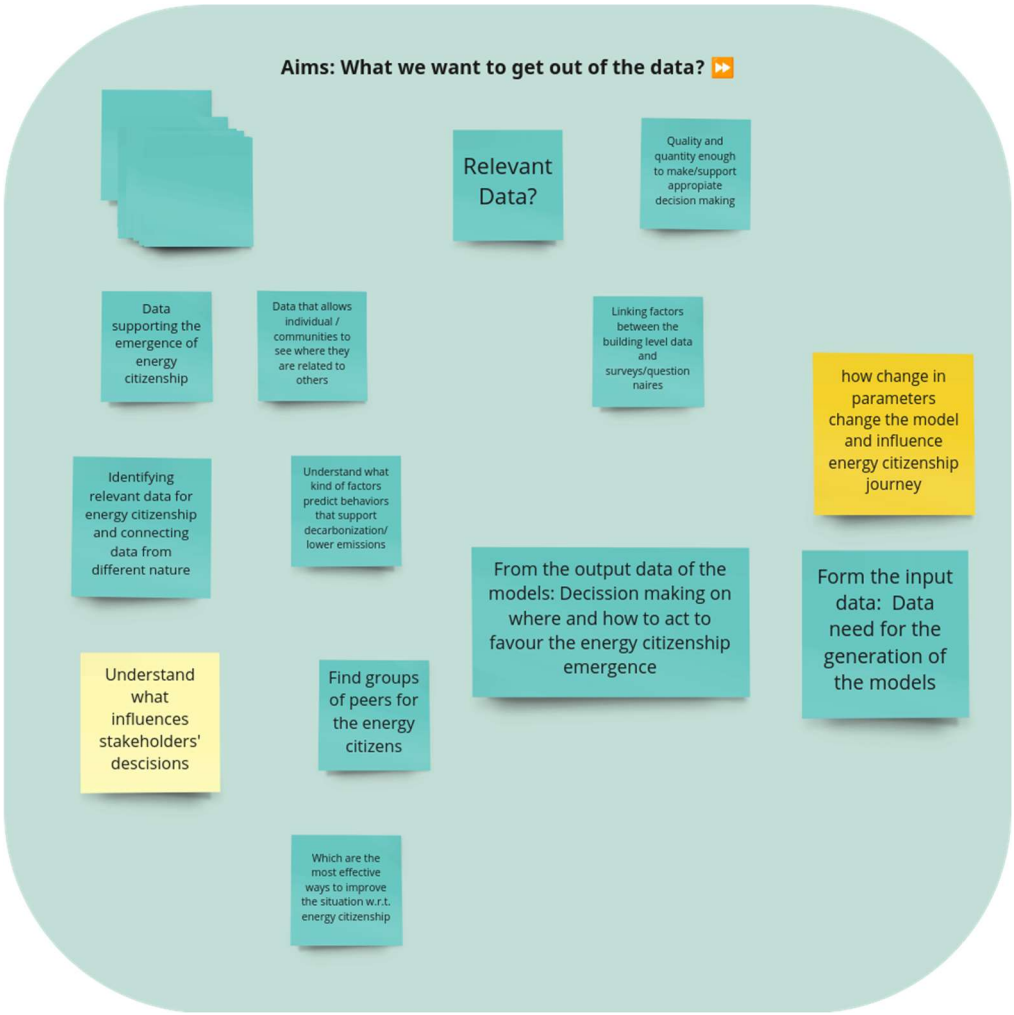

Figure 1 Identified data aims from the T4.1 Workshop

Figure 2 presents insights from the second part of the workshop, which considered data and modelling solutions to achieve the data aims presented in Figure 1. The main insights uncovered were: the better use of currently available data; data prediction; and identification of new data and ways to obtain it. Moreover, data uncertainty; people answering based on ideology instead of their views; answering based on public pressure; not paying attention to questions; or skipping questions that require thinking should also be considered, especially in the case of behaviour modelling and qualitative data. Furthermore, the solutions should enable the determination of similar energy citizens in an area and the identification of possible actions based on the current

situation and behaviour.

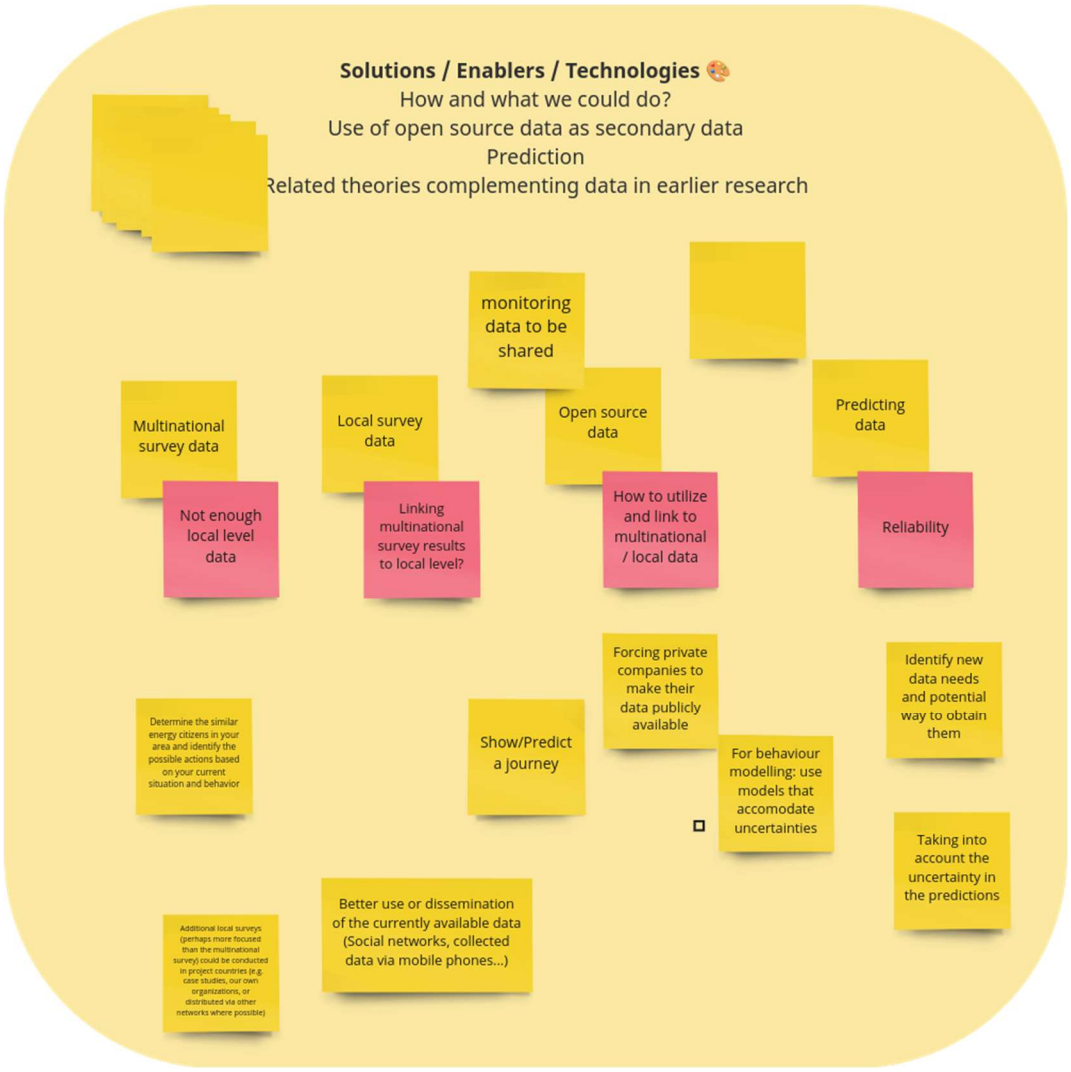

Figure 2 Identified data and modelling solutions (T4.1 WS)

Figure 3 presents insights on the challenges and barriers with data and models. Here the issues of locality, trust, uncertainty, data integration, missing data, and amount of available data were identified. This highlights the issues where the data integration and onboarding should concentrate. The data integration should consider, for example, differing spatial and temporal resolution, missing data, reliability/uncertainty in the data, and ensure correct encoding of the data for analysis. Furthermore, it was identified that the ongoing energy crisis might affect people's decisions.

Figure 3 Data and modelling challenges and barriers that were identified (T4.1 WS)

Figure 4 presents insights on possible actions and future ideas to address the challenges and data issues. Here the main results indicate that there should be more advantages from digitization and regulations promoting public data that would allow better knowledge sharing and transparency and promote data literacy and energy informatics.

Figure 4 Future ideas and possible actions to take with data and models. (T4.1 WS)

2.2 Data for developing the non-energy related models.

Data for developing the non-energy related models comes mostly from the surveys, questionnaires, workshops, and interviews conducted within the GRETA project, hereafter GRETA data. However, secondary data sources can be used to complement the gathered data. The data gathering is guided by the work of WP1, WP2, and WP3 and related deliverables. D1.1 explains the different types of energy citizens and which are the determinants per each type (Montalvo, 2021). D1.2 explains what are the factors of moving between different energy citizen engagement levels from unaware to advocate, including the drivers and barriers associated with each movement/level (Ruggieri, 2021). GRETA T1.3 provides a step-by-step method and heuristics to assess the conditions upon which the emergence of energy citizenship might arise for specific situations and actors. WP2 provides inputs more from the energy informatics and

energy literacy side. Moreover, WP2 provides information about CLIs that can aid in the determination of the engagement level of communities and individuals.

The data-driven approach, i.e., making decisions based on data analysis and interpretation, for clustering and finding patterns in data considers the collected GRETA data and where needed secondary data sources to complement the GRETA data. The aim is to cluster and profile individuals and groups based on the available GRETA data, i.e., creating the model based on the data and not creating the model first and fitting the data into the model.

The multinational survey covers a three-stage model for energy citizenship emergence and relevant themes defined in the Grant Agreement such as respondents' equipment/appliances and their use of energy data. The design of the question set related to the three-stage model has been filled with content with the results of the case study interviews. To avoid overburdening the survey respondents, each group of questions is randomly stratified into three or four (behaviour four only in the Netherlands and Germany since these countries have a special interest in this) sub-groups responding to questions related to one specific behaviour. These behaviours represent the targeted activities in the GRETA case studies; (1) cooperative self-generation of renewable energy; (2) use of sustainable transport (e.g., walking, using a bike, public transport or an electric vehicle); (3) use (buying, renting or leasing) of an electric vehicle with an autonomous and connected capacity and (4) activities to replace the use of gas in domestic appliances (cooking and/or heating).

The data analyses regarding the three-stage model have the following goals; (1) demonstrate and validate the structure and content of the model; (2) assess and rank the factors affecting the citizens' engagement in GRETA; (3) report the levels of engagement and distribution of responses describing the trends of factors affecting engagement positively or negatively; (4) estimate and simulate the propensity of citizens (and other actors) to engage in GRETA; (5) generate indicators/parameters that can be used in scenario exercises (e.g., simulating the conditions that optimize citizens engagement in GRETA).

2.3 Data for developing energy-related models

To develop the energy models, a certain set of mandatory data is necessary that must be provided by the case study leaders. These mandatory data are required to generate the model, see Table 2, and other additional data are required to generate a more complete model, although they are not mandatory.

The energy model generated is a geo-referenced Building Stock Energy Model (BSEM), which provides an overview of the energy demand of an area of interest, which can be either a district or a larger area, such as an entire city. These models should be at a local level and focus only on the building sector.

This section presents the information and processing necessary for generating the BSEM with the information in the required format and level of detail. The objective of this preliminary processing is to obtain a georeferenced layer with a detailed geometry at the building level, as well as the largest number of available attributes associated with that geometry, which will serve as input for the tool used.

The basic data for developing the model are in most cases publicly accessible data obtained from the cadastre, such as the building geometry in a Geographical Information Systems (GIS) layer, the use, height, and age.

Additional information can include data on the energy systems and fuels used, which makes it possible to calculate consumption and CO2 emissions, as well as information on existing thermal or photovoltaic systems, or the degree of protection of the buildings.

Table 2 Example of the most relevant input data for the BSEM generation

Parameter	Mandatory / Optional	Parameter	Mandatory / Optional
Building ID	Mandatory	Construction year	Mandatory
Footprint area	Mandatory	Energy system	Optional
Building Height	Mandatory / Optional	Boiler type	Optional
Number of floors	Optional*	Building use	Mandatory

2.4 Case studies data

2.4.1 CS1 (Renewable energy district – Bologna Pilastro-Roveri, Italy):

Interviews/surveys/workshops
The case study uses a qualitative and participatory approach. These allow for the analysis of needs, visions, and aims, and actions towards new energy behaviours, especially in the context of renewable energy communities.

Participants – citizens, students, academics and researchers, legal experts, policy, and decision-makers – have been involved in a series of meetings with experts and workshops to build a shared path of what energy citizenship is and how to engage in different forms of active participation in energy systems. A total of three preparatory meetings and two final CLI workshops have been conducted in the Roveri-Pilastro district.

Additionally, the UNIBO team conducted 16 qualitative interviews structured and semi-structured (WP1.3) with citizens, policymakers, businesses, associations and academics in Bologna at the local and regional levels. Interviews with citizens were

conducted in person, and the rest of the interviews were conducted online. Further short video interviews of about three minutes were conducted to understand the perception and knowledge about energy citizenship, behaviours, and policies.

Data for the energy model

To generate the basic model, cadastre information provided by the case study leaders from a direct request to the city council has been used. The information was reported in two different layers, one with the uses and one with the heights.

To develop the energy model, the buildings represented in a single geometry are needed, so, as there is no common identifier, it has been necessary to unify them manually. This is visualized in Figure 5.

Figure 5 Example of the unified geometries vs. the divided ones

Heights were defined in the small geometries, as some buildings have several different levels, so it is necessary to calculate an average height for each of the unified geometries using multiple geoprocesses.

Year of construction data was not available, so collaboration with the lead CSs has been necessary to complete the required information manually, assigning years of construction by zones visualized in Figure 6. As in the previous case, the attributes are joined by location using QGIS's geoprocessing tools.

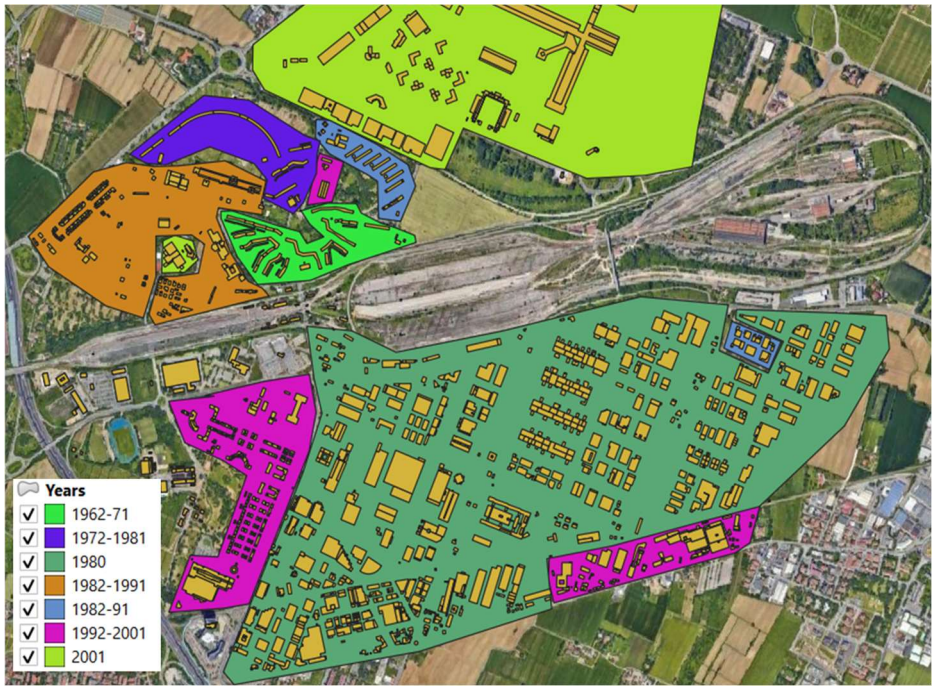

Figure 6 Manual definition of the construction years by zone

As additional information, a layer with heritage buildings, Figure 7, has been provided and included as "Protected buildings" in the final layer.

Figure 7 Buildings that are part of the heritage (in orange)

Finally, a single layer is obtained with one geometry per building and all the available data associated to it, Figure 8.

ID	Use	Area	Height	Year	Capacity k	Property	Energy Con	Protection_TIPO
1	Industrial	1232,173	11,3048976895...	1980	0	NULL	NULL	NULL
2	Other	20,859	4,09999990463...	1985	0	Full ownership	NULL	Piena Proprieta
3	Other	339,134	4,09999990463...	1985	0	Full ownership	NULL	Piena Proprieta
4	Other	46,042	4,09999990463...	1985	0	Full ownership	NULL	Piena Proprieta
5	Industrial	17803,628	12,6000003814...	1980	0	NULL	NULL	NULL
6	Industrial	3227,087	12,6000003814...	1980	0	NULL	NULL	NULL
7	Industrial	299,332	12,6000003814...	1980	0	NULL	NULL	NULL
8	Industrial	156,985	3,90000009536...	1980	0	NULL	NULL	NULL
9	Industrial	303,619	7,40000009536...	1980	0	NULL	NULL	NULL
10	Industrial	307,178	7,40000009536...	1980	0	NULL	NULL	NULL
11	Industrial	307,462	7,40000009536...	1980	10,35	NULL	NULL	NULL
12	Industrial	315,768	7,40000009536...	1980	0	NULL	NULL	NULL
13	Industrial	2318,595	7,80000019073...	1980	0	NULL	NULL	NULL
14	Industrial	199,765	7,80000019073...	1980	0	NULL	NULL	NULL
15	Industrial	845,643	8,5	1980	0	NULL	NULL	NULL

Figure 8 Example of the parameters included in the final layer

2.4.2 CS2 (Natural gas-free neighbourhoods, The Netherlands):

This case study considers the 46 Dutch neighbourhoods that are making the shift to being gas-free. The Dutch government chose to gradually restrict the extraction of natural gas from the Groningen gas fields, which has repeatedly resulted in damage and injuries due to subsidence accompanied by earthquakes, therefore the national programme on natural gas-free neighbourhoods began in 2018.

Interviews/surveys/workshops

This case study was conducted by performing interviews with policy-makers, and consulting experts and using previous research conducted by TNO and PAW. We consulted six experts at TNO who executed several research projects on different neighbourhoods involved. We interviewed them, asked them for input or let them review our reports to make sure we provided correct information. Below you find an overview of the research methods used for the different deliverables or work packages.

For D1.3 we interviewed four different experts with knowledge of the perceptions of homeowners, citizen initiatives, municipalities and suppliers. For the D1.3, WP2, and WP5 we used data from the following TNO research projects:

- de Koning, N., Kooger, R., Hermans, L., Tigchelaar, C. (2020). *Natural Gas-Free Homes: Drivers and Barriers for Residents*; Report P12006; TNO: The Hague, The Netherlands. This research entails fieldwork that was conducted in two comparable neighbourhoods in the Netherlands: Overwhere-Zuid in Purmerend and Wijk 03 Noord in Zwijndrecht. Various research methods have been used: 7 interviews were held with employees of the municipalities, 79 street interviews were conducted with residents, and 12 more extensive interviews were conducted with residents.
- Klösters, M., de Koning, N., Kort, J., Kooger, R. (2020). *De kracht van het collectief. Report* P12079. TNO: The Hague, The Netherlands. This research entails desk

research and interviews with ten participants from different collectives, distributed over 7 provinces in the Netherlands.

For the WP2 we used the following additional data from the TNO research projects:

- Eerland, D., de Koning, N., Kort, J., Paradies, G.L., van der Weerdt, C. (2020). *Samen Lokaal in Beweging. Werkboek.* Retrieved in August, 2022 from: Samen Lokaal in Beweging. Werkboek - Energy.nl . The result has been developed together with local energy initiatives in Nij Beets and Leeuwarden (Zuiderburen-Hempens). There were also two knowledge-sharing sessions organized with locally operating organisations, initiatives and other parties such as municipalities.

Moreover, we also used the following research reports from external parties:

- MarketResponse (2021). *Bewonerstevredenheid Proeftuinen Aardgasvrije Wijken*. Retrieved in August, 2022 from: Bewonerstevredenheid Proeftuinen Aardgasvrije Wijken | Rapport | Rijksoverheid.nl
- PAW (2022). *Voortgangsrapportage PAW Monitor 2021: Voortgang & Leerervaringen 1e en 2e ronde proeftuinen*. Retrieved in August, 2022 from: Programma Aardgasvrije Wijken - Programma Aardgasvrije Wijken

For WP 5 we used the following additional data:

- Klösters, M., De Koning, N., Dreijerink, L., Tigchelaar, C., Bijvoet, J., & Kooger, R. (2020). *Samenwerken in de wijkaanpak: ervaringen met bewonerparticipatie in acht proeftuinen aardgasvrije wijken.* Report available online in pdf format: Samen werkt het beter: bewonersparticipatie in de wijkaanpak | TNO
- PAW (2022)

For T6.1, policies were analysed for each case study including an evaluation of their impact and relevance for the case studies. This was done through desk research and three interviews. For the natural gas-free neighbourhoods, three policymakers that work at three different municipalities, namely Venlo, Voerendaal and Sittard-Geleen, were interviewed.

Energy model data

For the Natural gas-free neighbourhoods case study, the required information is available and complete. However, a BSEM has not been developed as there is already a geo-referenced model [1] with consumption data for the whole country, so it is

[1] https://dego.vng.nl/?tab=sociaal&label=topo&layer=layer21#11.53/52.0845/4.3875

considered that a model, such as the one developed for the other case studies, does not provide additional information.

In addition to these consumption data, they also report additional information such as occupancy, number of gas connections or income, which allows for additional analyses related, for example, to energy poverty.

2.4.3 CS3 (Coopérnico – Renewable energy-driven cooperative, Portugal):

Interviews/surveys/workshops
In the context of WP1 (T1.3), interviews were carried out with different stakeholders (i.e., (i) Coopérnico's members; (ii) Coopérnico as an entity; (iii) the Portuguese legislator; (iv) the Portuguese regulator; and (v) a business) for the development of the 3-stage model that also informed the design of the multinational survey – totalling 23 qualitative interviews.

In the context of WP2 (T2.1 and T2.3), an interview was conducted with a representative of Coopérnico for understanding the impact of digitalization and social media (from the perspectives of energy informatics and energy literacy) on energy citizenship (T2.1). Also, a virtual workshop was conducted with several Coopérnico's members for the co-design of personalized Community-Level Indicators (CLIs) (T2.3). A living working document was created and shared among Coopérnico's members for further co-design work on the proposed CLIs.

In the context of WP3, an interview was conducted with a representative of Coopérnico for the development of the Background Research for Case Studies, which aimed to:

- Explore likely enabling effects of new technologies, market conditions and institutional arrangements on energy citizenship. We are unlikely to catch this level of detail in the survey.
- Scope cases in the cases of interest and identify in more detail the stakeholders involved
- Provide a background to compare the general wisdom concerning the drivers and barriers in the sector with the empirical results of WP3 – i.e., What citizens and decision-makers perceive across the cases of interest;
- Place into context proposed policy options generated in WP4 (i.e., case study policy design workshops).
- Facilitate the analyses and lessons learned from what is the general wisdom and what else we learn in the case studies. Enable the across analysis of all cases in WP3 to WP6.

In the context of WP5 (T5.2 - which explores social science issues related to the emergence of energy cooperatives, focusing on geographic analyses and socio-spatial patterns), Coopérnico was an ideal CS to be explored due to its size and geographical scope. The data shared under the scope of this task related to the individual financial supporters/investors for two selected projects of Coopérnico: 1. Lar S. Silvestre

(https://www.coopernico.org/projects/14-lar-s-silvestre) and 2. Escola JG Zarco (https://www.coopernico.org/projects/17-escola-jg-zarco). Specifically, the following anonymized information and data were shared:

a. Number of financial supporters of the project
b. Location of the financial supporters (as detailed as possible, e.g. ZIP code level)
c. Amount of the individual investment
d. If available: socio-demographic data such as age, gender etc.
e. Date of investment.

Energy model data
In the context of WP4, Coopérnico was not selected for the development of the BSEM as Coopérnico is not a local case study and does not have all the required information available for its development.

2.4.4 CS4 (UR BEROA – Energy efficiency-driven cooperative, Spain):

Interviews / surveys / workshops
Based on the context of an energy community like Ur Beroa, the case study's design includes qualitative methods and a participatory approach that enables the analysis of Ur Beroa's and its members' energy citizenship level along with the analysis of the aims and actions towards more active participation in the energy transition.

Several members of the cooperative along with the management board of the cooperative have participated in meetings and in a workshop with academic and researchers to make a shift towards a higher level of decarbonisation and energy efficiency as well as increase the member base of the cooperative. In addition to building this shared vision, the cooperative defined a set of concrete actions to reach the vision. This direction-setting workshop was also used to reflect upon Community Level Indicators.

In addition, the TECNALIA team conducted 14 qualitative structure and semi-structured interviews (T1.3) with members of Ur Beroa, businesses and policymakers at the local, regional and national levels. All interviews were conducted either online or by telephone.

Energy model data
In the case of Ur Beroa, a combination of information from different sources has been used.

For generating the basic model, information from the cadastre has been used, which is publicly available and accessible to anyone. This information includes a layer of geometries with information on height. To obtain the data on the year of construction and use, it is necessary to work with other files in Excel format, which can also be downloaded from the cadastre website, and afterwards combine all the information.

The year of construction is defined at the building level, which is related to the geometry layer by the cadastral reference and the building unit number.

Mun	Referen	Superfic.	UC	U	C	T	FechaFiO
69	8297003	205	1	1	2	2	19000101
69	8297003	205	2	1	2	2	19500101
69	8297003	205	3	1	2	2	19000101

Figure 9 Example of data per building unit

From this file, it is possible to obtain the main use of the building, Figure 9, but if more precise data is required, there is another file with specific data at the dwelling level, Figure 10, from which the number of dwellings in each building can be obtained, as well as the real surface area for each end use (residential, commercial, storage rooms, garages...).

Mun	Referen	N.Fijo	Nuc	Cvia	Descripción Vía	Npor	Pl	Man	Use	Superfic.
69	8297003	139893	2	1910	GENERAL JAUREGI	12	6	DR	V	70
69	8297003	139894	2	1910	GENERAL JAUREGI	12	6	IZ	V	50
69	8297003	139895	1	1910	GENERAL JAUREGI	12	5	DR	V	62
69	8297003	139896	1	1910	GENERAL JAUREGI	12	4	DR	V	70
69	8297003	139897	1	1910	GENERAL JAUREGI	12	3	IZ	V	61,73
69	8297003	139898	1	1910	GENERAL JAUREGI	12	1	DR	V	70
69	8297003	139899	1	1910	GENERAL JAUREGI	12	1	IZ	V	70
69	8297003	139901	1	1910	GENERAL JAUREGI	12	2	IZ	V	70
69	8297003	139902	1	1910	GENERAL JAUREGI	12	2	DR	V	70
69	8297003	139903	1	1910	GENERAL JAUREGI	12	4	IZ	V	60
69	8297003	139904	1	1910	GENERAL JAUREGI	12	5	IZ	V	70
69	8297003	139905	1	1910	GENERAL JAUREGI	12	3	DR	V	70
69	8297003	6049615	3	1910	GENERAL JAUREGI	12	-1		C	174
69	8297003	6049615	3	1910	GENERAL JAUREGI	12	0		C	10,57
69	8297003	6049616	1	1910	GENERAL JAUREGI	12	0		C	152

Figure 10 Example of data per unit/dwelling

This basic information has been complemented with data on the energy systems used from the energy certificates, which are obtained by direct request to the municipality, as they are not publicly available Figure 11. Additionally, the municipality has the centralised boiler registers, information that has been included in the final layer, after processing.

Figure 11 BES data available from the EPC (red) and the municipal register (yellow)

Additionally, socio-economic data are also included, some of it public, such as educational attainment or average income at the census tract level, and some of it private, such as census data at the portal level. As income data are available at the census tract level, all buildings are assigned the same value.

As can be seen in Figure 12, the buildings in the study area correspond to three different census tracts in which, for example, the average rent data per household are quite distant.

Figure 12 Average household income of the 3 census tracts included in the study area

A layer with protected urban parcels is also available for the entire city, however, these do not apply to the study area.

The case study has hourly consumption data from the DH network at the substation level, an example is visualized in Figure 13, which is used to adjust the usage profiles of the tool's database and validate its results. These hourly profiles would be part of the information contained in the tool's database mentioned in section 2.6**Error! Reference source not found.**.

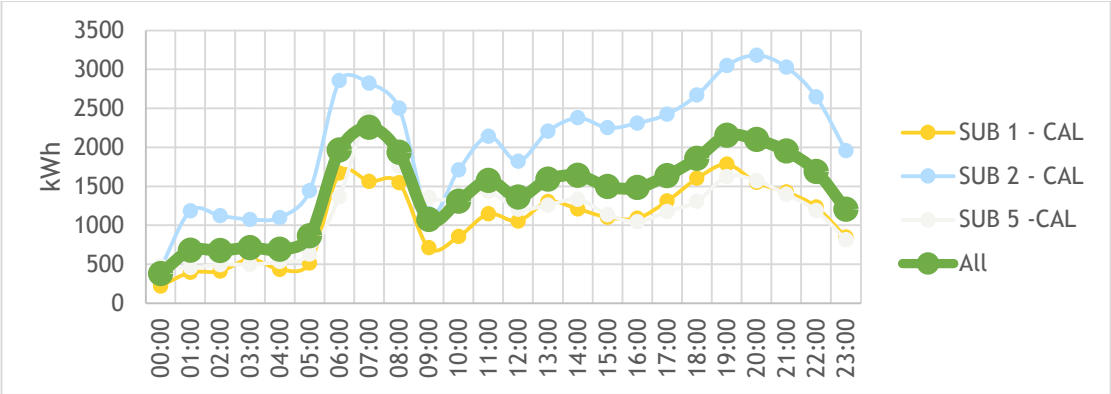

Figure 13 Example of hourly heating consumption profiles obtained for various DH substations.

2.4.5 CS5 (Earnest App Case - A virtual community for sustainable mobility in Darmstadt, Germany):

The case study explores whether game-like approaches, including a mobile application, can foster and influence sustainable mobility behaviour among members of a virtual community. The case study focuses specifically on sustainable mobility behaviour: On the one hand citizens' awareness and planned use of e-mobilities, public transport, and the reduction of long-distance travel; on the other hand, potential positive spill-over effects of changes within the mobility sector to other aspects of a sustainable lifestyle are explored.

Interviews/surveys/workshops

The case study design applies a mixed-method approach, allowing for the analysis of not only if and when people change their behaviour but also why and how. Participants responded to an online quantitative pre-and post-survey (before and after an eight-week phase of using the app) to monitor their behaviour and potential behavioural changes concerning mobility and travel-related activities (pre-survey had an n=50, post-survey still ongoing). Additionally, an online forum and a virtual qualitative discussion group (i.e., focus group) provided qualitative data assessing how and why participants changed levels of energy citizenship engagement. To contextualize the data, a policy and actor analysis was conducted (based on a policy document analysis and qualitative interviews with policymakers and citizens).

In addition to the core case study design, the FhG ISI Team conducted 16 qualitative interviews (WP1.3) with citizens, policymakers, civil society representatives and business leaders in and around the city of Darmstadt (and on the local, regional and national policy level). All the interviews were conducted online. Finally, two in-person CLI workshops (one of them including activities from the Community Transition Pathways) have been implemented up to this point.

Energy model data
The case study is not building-based, so it is impossible to develop a BSEM

2.4.6 CS6 (Electric autonomous and connected mobility network):

In contrast with the other case studies, this case addresses the conditions for the emergence of citizens' engagement in the use of a new form of transport that renders the current drivers into passengers. CCAM has five levels of automation and the most advanced reaches level four where the driver (now passenger) still has some driver's functions and responsibilities. In CCAM level five there are only passengers in the car, the car is the driver, and this development remains in the future. Up to recently, citizens (the common user) have not been involved in the development and deployment of advanced CCAM (level 3). Now there is an urgency to include citizens in the process of deployment. Thus, the case study aims to understand why or why not citizens engage in the use of CCAM and assess the influence of other actors in such engagement.

Interviews/surveys/workshops
To gain a first understanding of this and to contribute to the development of the survey questionnaire twelve interviews were conducted. The interviews included citizens (potential users of autonomous cars), policymakers at the EU level, the automotive industry, and CCAM experts. As the technology is not widely deployed and this is a European initiative no Community workshop was possible to conduct. A workshop will be conducted with policymakers and the industry in the first quarter of 2023. The case study aims as well to make a small survey complementary to the European survey. The additional survey will target participants in the European CCAM partnership.

Energy model data
The case study is not building-based, so it is impossible to develop a BSEM

2.4.7 Case studies data summary

Case studies data consists of interviews, surveys, workshop results and building-level data. Interviews with case study members, citizens, businesses and policymakers, were conducted for the development of the 3-stage-model, to inform the design of the multinational survey, and to aid the development of the WP4 models and allow linking factors to the multinational survey. Moreover, workshops were arranged to co-design personalized CLIs. The gathered data included: (1) investor information: investments, dates, etc.; (2) socio-economic data: age, gender, education, income, etc.; (3) building data: ID, footprint, height, construction year, etc; (4) CLIs; (5) interview results; and (6) survey results.

2.5 Multinational survey data

To better understand the emergence of energy citizenship and behaviour, an EU-wide survey was envisioned. Consequently, the GRETA consortium designed a

multinational survey that targets 16 EU Member States including the 10 largest EU countries by population. The multinational survey process and questions are explained in detail in the multinational citizen consultation results database document (Annala, 2022).

The aim of the multinational survey was to get 650 responses per country: 586 citizens (90%), 32 for both businesses (5%) and policymakers (5%). The multinational survey included questions related to the following areas:

- General background information, demographics and engagement level
- Energy literacy and energy information
- Social cohesion and capacity
- Energy use and characteristics of buildings
- Emergence model: attitudes, perceptions, norms, agency, influence
- Transversal integration of energy justice issues.

The areas that are mostly considered within WP4 are energy use and characteristics of the building, the emergence model, general background information, demographics and engagement level, social cohesion and capacity, and energy literacy and energy information.

General background information contains questions, for example, on age, gender, location, and income. Engagement level addresses, for example, the use of sustainable transport and participation in activities that support decarbonisation. Energy literacy and information use questions to determine the energy information sources and their trust in them. Respondents' neighbourhood vs larger area opinion is addressed in social cohesion and capacity area. For example, heating and electricity cost types questions are addressed in the energy use and characteristics of building questions. Finally, the emergence model contains questions related to the 3-stage energy citizenship emergence model that is elaborated in D1.1 (Montalvo, 2021).

A review of survey demographic distribution against known demographics will be conducted when the information becomes available.

2.6 Supplementary/Secondary data

Supplementary data include, for example, national/regional information about demographics/socio-economic situation, weather/climate data, geodatabases, and earth observation data, and cadastral and building information. Supplementary data are used to augment the models with the necessary data that are not exclusively collected during the GRETA project.

To support the development of data analysis methods for the surveys, an energy behaviour survey was conducted as a part of a master's thesis project at LUT which

was directed at LUT staff members and students. The survey consisted of 53 questions targeting various aspects including, basic information and living arrangements, energy-saving actions, environmental actions, travelling, beliefs and attitudes regarding climate and energy, and information gathering. It received 149 responses and resulted in mixed-type data. The most meaningful questions regarding the clustering process considered age, the use of a dishwasher, and the ownership of a car. The least meaningful included switching lights in unused rooms and other energy use questions (Pekkola, 2022).

To identify the secondary data input for the project, extensive desk research was performed. The identified datasets were categorized according to a pre-defined analytical framework and structured accordingly. The resulting database was saved in an excel-file and will be deposited in the project repository on Zenodo. Moreover, the GIS-based tool will include additional data from existing European surveys to complement the multinational survey, draw conclusions over time, and map national and EU-wide patterns of citizens' energy behaviour and attitudes. We have identified relevant survey projects and suitable items within these surveys based on comprehensive desk research. Furthermore, we have classified these surveys according to their general theme and year of implementation, methodology, the number and type of respondents, and geographical scope and level of georeferencing. A more detailed view of the aforementioned secondary data sources is available in D5.2 (Abel, 2022).

The energy models have an internal database generated from public data at the European level as well as data found in the literature. In the case of specific data, it is possible to modify this information to suit the particularities of the case study. This information includes U-values of building envelope elements, temperatures, usage patterns or data on energy systems. U-value information is available, together with other data, at the European level in the EU Buildings Database (European Commission, 2022)) In the case of climate files, they are obtained from the Energy+ climate database (Weather Data, 2022)

3 Data onboarding and processing

3.1 General data pre-processing

The general data pre-processing is visualised in Figure 14 and it contains data cleaning, data integration, data transformation, and data reduction. These are explained in more detail in the following sections.

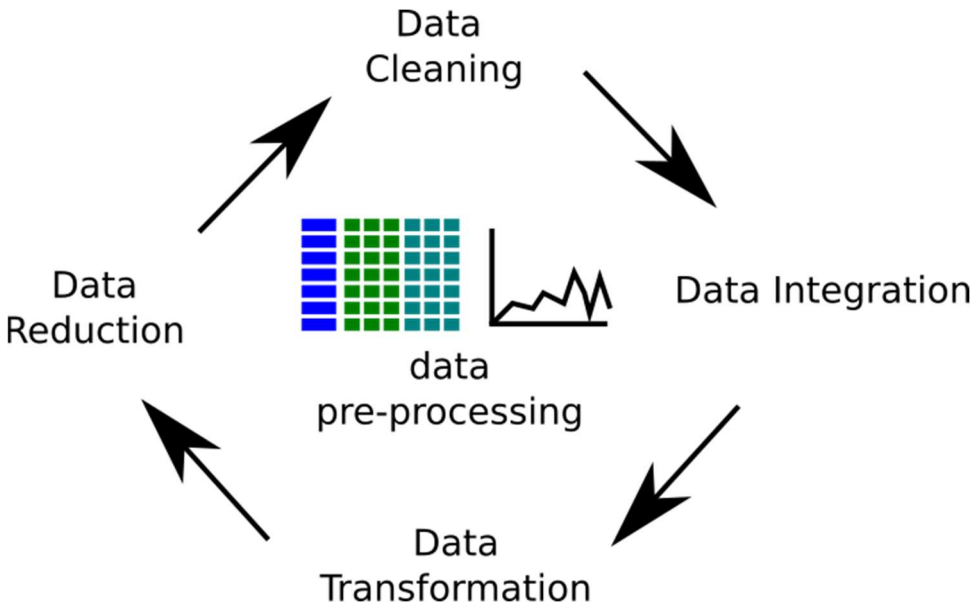

Figure 14 data pre-processing

3.1.1 Data Cleaning

Data originating from surveys commonly contain missing values. In such cases, the following strategies can be used: (1) ignore, (2) remove, (3) fill manually, and (3) fill automatically. Usually, removing entries with missing values leads to sparse data. Thus, either manually filled-in values or automatically filled values are generally used. In the case of bigger datasets, manually filling data becomes too burdensome and automatic filling with a strategy is used. This strategy can be for example filling all the missing values with zeros (0) or with the mean value of that specific variable. However, more advanced methods can be used if needed, for example, model-based data augmentation or machine learning-based predictions. The finalized strategy will be selected based on the experiments with the data.

Noisy data handling is in general easier with quantitative data than with qualitative data. However, the same strategy can be used to analyse and detect outliers and noisy entries in both. Data can be filtered and discretised, or the noisy data could be removed. However, usually filtering or finding the correct value by binning or based on the other variables the noisy values can be set to the most probable correct values.

3.1.2 Data Integration

Data integration is the process of combining data from multiple sources into a single database. It includes the integration of metadata and resolving value conflicts, for example, different units, representation, etc. It also manages redundancy, i.e., performing correlation analysis to ensure good data quality (Doan, 2012). Moreover, D5.2 proposed to follow the "GISualization" framework proposed by (Adelfio, 2019) which was developed to adaptively and iteratively manage complex data integration tasks (Abel, 2022). Furthermore, the specific data processing with energy related models is explained in more detail in Section 3.2.

3.1.3 Data transformation & reduction

Feature scaling is needed in the case of quantitative data with varying value ranges. It is a process where the values are transformed into a standardised representation. Normalisation is used when there is a need for a bounded value range, [0, 1] or [-1,1]. Standardization is another scaling method where the values get transformed to have a zero mean and a unit standard deviation. It is not as severely affected by data with outliers as normalisation is and it is normally chosen for normally distributed data (Doan, 2012).

If needed data generalisation can be performed. For example, this could include discretization of values, i.e., with age data instead of having a continuous age range from 20 to 80, age groups could be used. Similarly, regions could be combined to create larger data areas. This could also apply to forming concept hierarchies, for example, replacing municipalities, with regions, nations, or even multinational areas.

Other data reduction techniques could entail data summarisation, dimensionality reduction, compression, and discretization. Overall, the aim is to have necessary data that are detailed enough without repeated information (Doan, 2012).

3.1.4 Data format & processing

Survey results come in as mixed-type data, i.e., have both categorical and numerical values. While categorical data can be either nominal or ordinal, ordinal does have an order but the differences between the values cannot be determined, for example, a range of values strongly disagree to strongly agree, while nominal data are categories that cannot be ordered, for example, truth values, city, and gender. Whereas numerical data can be interval or ratio type both of which have ordered values. These types of data can be encoded in various ways to suit the used methods. Encoding methods vary from basic one-hot encoding to more advanced machine learning/deep learning methods (Seger, 2018). However, for example, one hot encoding increases the dimensionality of the data. One way to encode and decrease the dimensionality is to combine the sub-questions and the ranking questions into one attribute. For example, when exporting a question with six sub-questions and four choices it would generate a vector, 103424, with a length of six and values ranging from 0 to 4, 0 meaning that the answer has been skipped. The final encoding type and format will be selected based on the experiment results.

3.1.5 Geolocation processing and anonymization

Data geolocation processing includes, for example, the conversion of indirect spatial identifiers, e.g., a city/municipality, to geocoordinates. The aim is to get harmonized georeferencing information to, for example, link different datasets. The same georeferencing and linking procedures apply here as in D5.2 where the process is explained in more detail (Abel, 2022).

In some cases, for example in case study interviews and with the building data, it has been necessary to aggregate and anonymize data to comply with the GDPR. Moreover, consideration and measures are being taken that the GRETA data is compliant with the GDPR and GRETA DMP (Landeck J. A., 2021).

3.1.6 Data organisation and Exploration

Data quality assessment is performed to check the data for completeness, accuracy/reliability, consistency, validity, and redundancy. Various methods can be used here, just from basic methods to check for incompletes, outliers, or missing values to correlation analysis and principal component analysis to determine if some data should be excluded from the models.

3.1.7 Data storage, sharing & onboarding

Smaller data and documents within WP4 are used and shared through a Microsoft Teams repository. However, when sharing larger files services, such as EUDAT (EUDAT, 2022) can be used to share the files.

GRETA will adopt licenses that allow the re-use of the data and datasets. Moreover, data will be made usable by other parties after the end of the projects by using open formats, standards, and appropriate metadata, such as Dublin Core Metadata Initiative (DCMI) (DublinCore, 2022)

The GRETA project participates in the Open Research Data Pilot (ORDP) initiative and provides open access to research data and scientific publications, whenever possible. Moreover, the Zenodo repository will be used to ensure the maximum dissemination of the generated information. However, the GRETA data management plan (D8.5) lists the following data sharing restrictions; (1) Access to data and reports will be limited to members of the GRETA project until publications and presentations are submitted; (2) Some data and reports may be held back temporarily until the results are reviewed and validated by the appropriate intellectual property office at LUT or the partner institutions; (3) some data might not be made available if deemed sensitive personal information by the supervising committee; (4) Proprietary information (if any) provided by commercial firms under a confidentiality agreement will not be made available without the consent and approval of the firms (Landeck J. A., 2021).

3.2 Specific data processing with energy related models

To ensure that the results obtained from the energy simulation represent reality as accurately as possible, it is necessary to adapt, clean and organise the input data, as the accuracy of the results depends to a large extent on the level of detail and the veracity of the data. All data must be in a GIS layer with building-level information, represented by a single geometry.

Much of the processing is done using QGIS, especially that related to building geometries. In this case, the modelling requires a single geometry for each building, so it is necessary to combine all geometries that correspond to a single building, as well as to eliminate overlaps and duplicate geometries.

It is also necessary to complete the data required by the tool. The information can be completed by combining the GIS layer with another GIS layer joining attributes by location or by combining it with an Excel file using a common identifier.

3.2.1 Data format

Input
The input file required by the tool used for generating the models supports specific parameters in a specific format, which are shown in Table 3. As can be seen, some are mandatory, but others are optional.

Table 3 List of the input data used for the energy model generation

Parameter		Format	Nature
1	Building ID	str	Mandatory
2	Footprint Area	float	Mandatory
3	Total Height	float	Mandatory*
4	Year of Construction	int	Mandatory
5	Use	str	Mandatory
6	Protection Degree	str	Optional
7	Gross Floor Area	float	Optional
8	Number of Floors	int	Mandatory*
9	Roof Area	float	Optional
10	Wall Refurbishment Year	int	Optional
11	Roof Refurbishment Year	int	Optional
12	Window Refurbishment Year	int	Optional
13	Heating Boiler	str	Optional

14	Heating Boiler Configuration	str	Optional
15	Number of Boilers	int	Optional
16	DHW System	str	Optional
17	DHW System Configuration	str	Optional
18	Cooling System	str	Optional
19	Cooling System Configuration	str	Optional
20	Number of Dwellings	int	Optional
21	Solar PV installed capacity	float	Optional
22	Solar thermal installed capacity	float	Optional
23	Solar Effective Surface Perc	float	Optional
24	Solar Total Irradiance Sqm	float	Optional

In addition to these data used by the tool for internal calculations, any other parameter of interest can be included and then exported together with the energy results of the model and worked with, but always at the building level.

Output

The used tool quantifies and stores the georeferenced thermal energy (heating, cooling and Domestic Hot Water) and electricity (lighting and appliances) demand results for each building in the area under study, which will lead to the calculation of the energy consumption, CO_2 emissions, and cost of the consumed energy, in addition to the production of energy using solar panels, either thermal or photovoltaic.

The main energy-related parameters obtained from the baseline scenario simulation are listed in Table 4:

Table 4 Main energy-related parameters at the building level for the baseline scenario

Parameter	Unit	Parameter	Unit
Heating demand	kWh and kWh/m^2	Cooling consumption	kWh and kWh/m^2
Cooling demand	kWh and kWh/m^2	DHW consumption	kWh and kWh/m^2
DHW demand	kWh and kWh/m^2	Lighting consumption	kWh and kWh/m^2
Lighting demand	kWh and kWh/m^2	Equipment consumption	kWh and kWh/m^2
Equipment demand	kWh and kWh/m^2	CO_2 emissions	CO_2 and CO_2/m^2
Heating consumption	kWh and kWh/m^2	Cost	€ and €/m^2
Fuel		Energy production	kWh and kWh/m^2

This information together with the rest of the parameters related to the building geometry added to any parameters that were included in the input file are provided in various formats at the building level and in CSV format grouped at the district or city level.

The different formats of the output results, Figure 15, allow the user to visualise the results graphically using any GIS tool or numerically in CSV or Excel files.

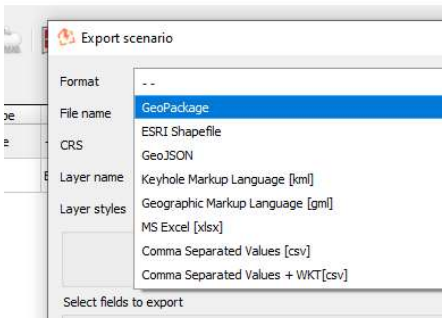

Figure 15 List of available formats for exporting results

4 Conclusions

This report details the data processes in WP4 of the GRETA project. It describes various aspects related to data gathering, data management, data processing, and data storing and sharing.

D4.1 summary:

- Describes the data gathering and identification process
- The main methods for data pre-processing with energy and non-energy data
- The determination of data integration and onboarding processes and related recommendations
- The protocols for data storing and sharing in WP4 while following the D8.5 Data Management Plan

References

Abel, D. J. (2022). *Interrelations among different types of citizens in different geographic contexts. D5.2 of the Horizon 2020 project GRETA, EC grant agreement no 101022317.* Cologne, Germany.

Adelfio, M. K.-H. (2019). GISualization: Visualized integration of multiple types of data for knowledge co-production. *Geografisk Tidsskrift-Danish Journal of Geography, 119(2)*, 163–184. doi:https://doi.org/10.1080/00167223.2019.1605301

Annala, S. M. (2022). *Multinational citizen consultation results database. D3.7 of the Horizon 2020 project GRETA, EC grant agreement no 101022317.* Lappeenranta/Lahti, Finland.

Doan, A. H. (2012). *Principles of data integration.* Elsevier.

DublinCore. (2022). *The Dublin Core Metadata Initiative*. Retrieved from DCMI: https://www.dublincore.org/

EUDAT. (2022). *EUDAT*. Retrieved from Collaborative Data Infrastructure: https://eudat.eu/

European Commission. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR)*. Retrieved from http://data.europa.eu/eli/reg/2016/679/oj

European Commission. (2022). Retrieved from EU Buildings Database: https://ec.europa.eu/energy/eu-buildings-database_en

Landeck, J. (2021). *Ethics and Privacy Management Plan. D8.6 of the Horizon 2020 project GRETA, EC grant agreement no 101022317.* Coimbra, Portugal.

Landeck, J. A. (2021). *Data Management Plan. D8.5 of the Horizon 2020 project GRETA, EC grant agreement no 101022317.* Coimbra, Portugal.

Montalvo, C. S. (2021). *Framework for research on energy citizenship emergence structure and dynamics D1.1 of the Horizon 2020 project GRETA, EC grant agreement no 101022317.* The Hague, The Netherlands.

Pekkola, T. (2022). *Machine learning techniques applied to energy behavior profiling.* Lappeenranta-Lahti University of Technology LUT: Master's thesis.

Project GRETA. (2022). *Glossary*. Retrieved from Project GRETA: https://projectgreta.eu/glossary/

Ruggieri, B. C. (2021). *Vision document on energy citizenship-based Energy Union (persons, essays, scenarios, winners and losers of energy transitions). D1.2 of the Horizon 2020 project GRETA, EC grant agreement no. 101022317.* Bologna, Italy.

Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.*

Weather Data. (2022). Retrieved from EnergyPlus: https://energyplus.net/weather