BuiltHub

D4.2 – Matrix, describing the available datasets and main data sources and type

Final version

Project acronym	BuiltHub
Full title	Dynamic EU building stock knowledge hub
GA no	957026
WP, Deliverable #	4.2
Version	1.0
Date	29.09.2023
Dissemination Level	Public
Deliverable lead	TUW
Author(s)	Carla Rodríguez Alonso, CARTIF Víctor Iván Serna González, CARTIF Iná Eugenio Noronha Maia, TUW
Reviewer(s)	Mikael Mangold, RISE
Keywords	Dataset, matrix, thematic area, entity, granularity

Disclaimer

This project has received funding from the EU's Horizon 2020 programme under grant agreement no 957026.

BuiltHub

Table of contents

Figures

Tables

Abbreviations and acronyms

Acronym	Description
AC	Air conditioning
BSO	Building Stock Observatory
CC	Creative Commons
CHP	Combined Heat and Power
CO2	Carbon Dioxide
DHW	Domestic Hot Water
EC	European Commission
EDGAR	Emissions Database for Global Atmospheric Research
EEA	Energy Environment Agency
EPC	Energy Performance Certificate
EU	European Union
FEC	Final Energy Consumption
FP7	Seventh Framework Programme for Research
GDP	Gross Domestic Product
GHG	Greenhouse Gas
GLA	Gross Leasable Area
H2020	Horizon 2020 programme

Acronym	Description
IEA	International Energy Agency
IEE	Intelligent Energy Europe programme
IEQ	Indoor Environmental Quality
IRENA	International Renewable Energy Agency
JRC	Joint Research Centre
LAU	Local Administrative Units
nZEB	Nearly Zero-Energy Building
NUTS	Nomenclature of Territorial Units for Statistics
OECD	Organisation for Economic Co-operation and Development
PEC	Primary Energy Consumption
PPS	Purchasing Power Standards
RES	Renewable Energy Source(s)
SDG	Sustainable Development Goal
URL	Uniform Resource Locator
WP	Work Package

Executive summary

For the making of a datahub, as the BuiltHub platform is about the European building stock, a protocol to produce reliable figures representing the EU building stock is needed. In the present deliverable an investigation and classification of the datasets identified to produce relevant indicators is performed, in terms of related thematic areas and cross-cutting entities, time resolution, spatial resolution, source, availability, license and usability.

This classification is done in a matrix form (https://builthub.eu/resource?uid=637), in a homogenised and organised way, which is key for the inclusion of the datasets in the BuiltHub platform, at the same time as it provides a relevant metadata information on each of them.

The analysis starts from the identified datasets in the project, and the aspects for the analysis included in the matrix come from the project research, as well as from the thematic areas and entities defined in the project for producing indicators.

This analysis is beneficial when making the datasets accessible through the platform, as well as being a proper protocol in a coherent and structure manner that ease the integration of data in the platform, and thus the production of indicators, visualisation diagrams, maps and tables from different data sources.

The matrix analysis of datasets allows also the production of analytics to assess each of the fields analysed, providing an overview of the information they contain.

1. Introduction

The objective of **WP4** "Data processing and analytics" is to process, present and visualise comprehensive information and knowledge related to existing building stock and possible transformation scenarios, using machine learning and other data manipulation methods, which allow certain flexibility on processing different types of data. The main result of this WP is a data-post processing workflow to be IT implemented in WP5, with the main objective of providing focused, target-oriented and easy access to knowledge.

Task 4.2 "Protocol to produce reliable figures representing EU building stock" will support the achievement of WP4 objectives through the investigation and classification the datasets in terms of related areas, resolution, source, availability and usability. This classification of the different datasets leads to an identification of possible interlinks and overlaps between the different data, and to prepare the ground for the activities in Task 4.3 "Data organization and structure", where data is organised and structured, considering all metadata from T4.2, in a repository format.

The investigation of the datasets identified in Task 3.1 "Building stock inventory setting" followed a matrix approach, where the available datasets are identified and analysed and classified according to a set of aspects, including time and spatial resolution, main source, availability and format, license and usability.

The main outcome of T4.2 is the present **Deliverable 4.2** "Matrix, describing the available datasets and main data sources and type", which is of type "other", that is, the D4.2 matrix itself is the "**BuiltHub_D4.2MATRIX_Final_v1.0.xlsx**" (**https://builthub.eu/resource?uid=637**), and the present report should be considered as complementary to the Excel Spreadsheet main document.

This complementary document is structured into 6 main sections:

- **Section 1**: (current section) introduces the document and its context under BuiltHub project.
- **Section 2**: recaps the organisation and structure of the D4.2 matrix Excel Spreadsheet.
- **Section 0**: presents the overview of the datasets included in this analysis.
- **Section 4**: delves into the aspects under which the datasets are analysed, informing on the different options and considerations for each of them.
- **Section 5**: provides conclusions and final remarks from the overall matrix analysis.
- **Section 6**: includes the references used in the document.

2. Excel Spreadsheet organisation

The D4.2 Excel Spreadsheet is organised into three tabs:

- The first tab "**D4.2 introduction**" includes the presentation of the deliverable and a brief presentation of the content in the other tabs.
- The second tab "**Aspects analysed**" includes the detailed information on the aspects analysed in the matrix for each dataset. These aspects come from the Task 4.2 description, as well as from the developments in Task 3.1 and own criteria with the common work and approach for WP4 and WP5. In this tab, the titles of the aspects analysed are included in the rows 2 and 3, in the same order as in the matrix in the following tab. Below these rows, for each aspect the specific items considered for the datasets classification can be found, as well as some further explanations in the following cell, inlcuding schemes in some cases to make it more understandable.
- The third tab "**Datasets Matrix**" corresponds to the matrix itself. It is structured with the aspects analysed in rows 2 and 3, the datasets list in "C" column, and the subsequent analysis and classification of each dataset in its corresponding row. The datasets analysed are taken from the Task 3.1 list (and ID), and are: **1, 2, 3, 4, 5, 7, 11, 12, 13, 14, 15, 17, 19, 21, 22, 23, 24, 25, 27, 28, 34, 36, 37, 39** and **40**. Additionally, in some of the cells there is an information icon (**i**) which indicates that the there is more information or explanation for that field when user clicks on the cell.

3. Datasets included in the analysis

The identification of datasets to be considered for the BuiltHub platform was done under Task 3.1 “Building stock inventory setting”, and reported in D3.1 “Inventory structure and main feature and datasets”, so that list of 30 datasets is the starting point for the dataset selection. From task 3.1 a complementary work was performed to extend that initial list until 45 datasets. The extended list (Table 1) was also investigated for the work in the present T4.2.

Table 1: Extended list of datasets (from Task 3.1)

ID	Name
1	Horizon 2020 HotMaps project: Building stock analysis
2	IEE TABULA project: Typology Approach for Building Stock Energy Assessment
3	IEE EPISCOPE project: Focus of building stock monitoring
4	IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020
5	IEE ENTRANZE project: Policies to Enforce the TRAnsition to Nearly Zero Energy buildings in the EU27
6	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland.
7	FP7 CommONEnergy Project: building stock
8	JRC IDEES 2015
9	SET-Nav - Strategic Energy Roadmap
10	H2020 ExcEED Project: building stock data
11	FP7 iNSPiRe project: building stock analysis
12	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT
13	Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - FINAL REPORT
14	EUROSTAT: Final energy consumption in households
15	EUROSTAT: Final energy consumption in households by fuel
16	EUROSTAT: Disaggregated final energy consumption in households
17	ZENSUS 2011
18	DPE - Diagnostic de Performance Energetique
19	Towards a sustainable Northern European housing stock - Sustainable Urban Areas 22
20	DEEP - De-risking Energy Efficiency Platform

ID	Name
21	Energy consumption and efficiency technology measures in European non-residential buildings
22	Dataset of the publication: Europe's Building Stock and Its Energy Demand: A Comparison Between Austria and Italy
23	National Housing Census: European statistical System
24	Energy prices in 2019 - Household energy prices in the EU
25	EUROSTAT: GDP per capita in PPS
26	EUROSTAT: Population on 1 January by age, sex and NUTS 2 region
27	EUROSTAT - Cooling and heating degree days
28	EDGAR (Emissions Database for Global Atmospheric Research) CO2 Emissions
29	CORDEX - Regional climate model data on single levels for Europe
30	PVGIS - Photovoltaic Geographical Information System
31	Concerted Action EPBD
32	Global Buildings Performance Network (GBPN)
33	BuildingRating.org
34	IEA World energy balances database
35	ENERFUND: Building Retrofit Potential
36	Statistical pocketbook 2017
37	IRENA
38	Air Quality e-Reporting (AQ e-Reporting)
39	Approximated estimates for the share of gross final consumption of renewable energy sources in 2019 (EEA 2019 RES share proxies)
40	Approximated estimates for the primary and final consumption of energy in 2019 (EEA 2019 proxies on primary and final energy consumption)
41	Copernicus Land Monitoring Service - Urban Atlas
42	INSPIRE Buildings Theme (Annex 3) Datasets
43	EOCD Data
44	MRS ESPON
45	ESPON DB

The final list of analysed datasets is: **1, 2, 3, 4, 5, 7, 11, 12, 13, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27, 28, 34, 36, 37, 39** and **40**.

The ones that have been left out of the analysis have been for the following reasons:

- Data could not be downloadable or webpage is not available anymore: **6**, **10**, **32**, **35**.
- Datasets that contain a huge amount of data or indicators that would need previous selection or filtering to precise what to download (or would require of API connection to be incorporated to the BuiltHub platform) or are more service-data: **8**, **30**, **41**, **42**, **45.**
- Data in report and PFD format, from which data should be searched in tables or graphs within the PDF: **18**, **20**, **43**.
- No actual or usable data (e.g. data is more literature-focused, methodological, or not related with any BuiltHub thematic areas) found in the identified dataset: **9**, **33**, **38**, **44.**
- Datasets that require the selection of certain parameters to proceed with the download of data (or would require of API connection, service data): **29**.
- Data disaggregated (e.g. per country) and in PDF report format: **31**.

4. Aspects analysed for each dataset

The aspects defined to analyse each datasets can be seen all at once in the following Figure 1.

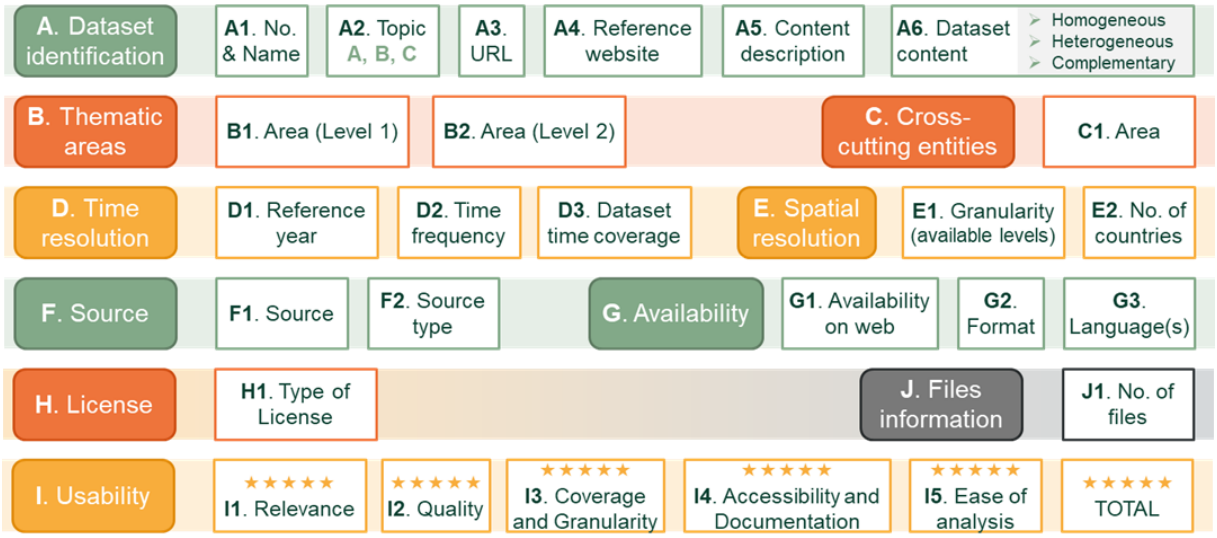

Figure 1: Overview of the aspects analysed of each dataset, including the sub-aspects for each of them

The **dataset identification** field contains the **number and name** of the dataset for identification and reference purposes, being the number the ID from Task 3.1. The **topic** is to identify the type of dataset in general terms, with the following options: A) Building stock related datasets; B) Socio-economic datasets; and C) Climatic datasets. **URL** is also included to locate the dataset in the website, if available, as well as the **Reference website** to the general one where the specific dataset belongs, if available (e.g. the project website, the EU website, etc.). The identification field contains a **content description** as well, including a brief description of the data contained in the dataset, and how it is structure in there. Finally, **dataset content** is included to know if the dataset contains single-topic focus information or if they contain a lot of information from different topics. It is assessed in relation to the different thematic areas identified for the information contained in the datasets. The three options for this sub-field are: homogeneous dataset, for those with data or information on a single thematic area; heterogeneous dataset, for datasets with data or information on more than one thematic area; and complementary dataset, for those datasets with data or information only on cross-cutting entities.

The **thematic areas** and cross-cutting entities fields come from outcomes of Task 4.1 “Specification of flexible indicators and platform information”, where a classification of areas for the indicators content was defined, taking the BSO classification as basis of this work. The thematic areas have **two levels**, and are depicted in the Figure 2, the areas in the left side are those coming from the BSO classification, while those with (*) included in the right side are additional ones added to complement the BSO areas.

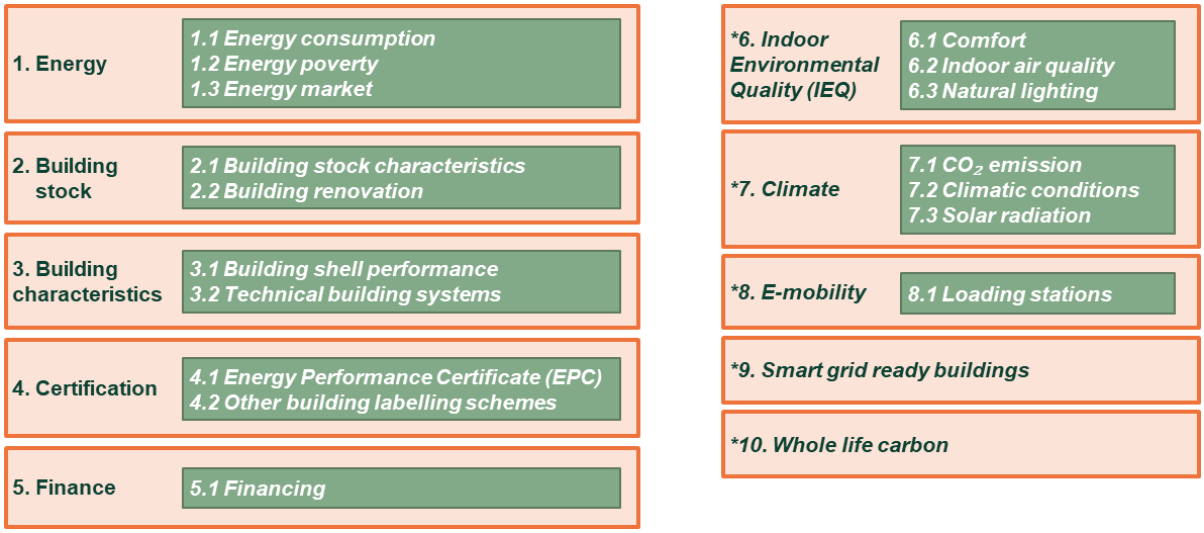

Figure 2: Thematic areas (level 1 and 2) for the analysis of datasets in the matrix

The **cross-cutting entities** are also related to T1.4 work, and are configured as the following list of options, taking into account that a single dataset can have all of them: 0.1) Building type, if it includes information on residential and/or non-residential buildings; 0.2) Demography, if it includes data on the degree of urbanisation and population density; 0.3) Energy carrier, if it includes information on the primary energy source (both renewable and non-renewable) and/or on secondary energy source; 0.4) Housing occupation, if it includes data on the number of personas and number of rooms of buildings; 0.5) Tenure status, if it includes information related to if it is owner-occupied, if it is rented at market price, if it is rented at reduced price (social), or if it is free housing; 0.6) Country, if it includes separate data for the EU Member States, Norway, Switzerland, and the UK; 0.7) Construction year class, for datasets that includes yearly information: before 1920, 1921-1945, 1946-1969, 1970-1994, 1995-2008, 2009-now

The **time resolution** is analysed according to the **reference year**, which is the year or range of years where the data is available in the dataset. The **time frequency**, to depict the frequency of the data that is available, with the following pre-defined options: (1) accumulated data (total) in the reference year, (2) data at daily basis, (3) weekly basis, (4) monthly basis, (5) bi-monthly basis, (6) quarterly basis, (7) bi-yearly basis, (8) yearly basis, or (9) other. In the time resolution field, the **dataset time coverage** is also included to explain the time coverage of the data, and to indicate if data is presented in an accumulated manner. It includes the following options: (1) between 0-1 year data, (2) between 1-2 years data, (3) between 2-3 years data, (4) between 3-4 years data, (5) between 5-10 years data, (6) between 11-20 years data, (7) more than 21 years data.

The **spatial resolution** is assessed through the **granularity**, referred to the available levels, according to NUTS[1] and LAU[2] classification: (1) NUTS 0, (2) NUTS 1, (3) NUTS 2, (3) NUTS

[1] https://www.europarl.europa.eu/factsheets/en/sheet/99/la-nomenclatura-comun-de-unidades-territoriales-estadisticas-nuts

[2] https://ec.europa.eu/eurostat/web/nuts/local-administrative-units

3, (4) LAU 1, (5) LAU 2, (6) other resolution. The **number of countries** is also informed, to include the number of countries covered by the dataset information.

The **source** is assessed through the name of the **source** where the dataset belongs to, and the **source type**, to include the type of source according to a specific classification (taken from the BSO report [1] for the classification for quality assessment of the data sources): (1) Official statistics, for official European/national/regional statistics, (2) EU project, for statistics developed in research EU project, (3) Literature, if the dataset comes from literature research such as reports with data based on multiple sources and/or own analyses, (4) Model output/calculation, if the dataset comes from results of modelling based on statistical data and experts' assumptions, (5) Assumptions/ Estimations, if dataset comes from expert assumption, and no analytical modelling is involved.

The **availability** field is analysed taking into account the **availability of the web**, selecting the availability of the dataset according to this specific classification: (1) Viewable but not downloadable, (2) Open – downloadable, (3) Registration required and then downloadable, (4) Downloadable only selected fields/indicators (not all data at once), (5) Registration required and downloadable only selected fields/indicators (not all data at once), (6) Web reserved to EU universities and research centres for non-commercial uses, (7) Registration required and data only available for EU universities and research centres. The availability analysis also includes the **format**, to indicate the format in which the data is available: (1) xlsx, (2) csv, (3) xml, (4) pdf, (5) nc, (6) Esri ASCII Grid, (7) json, (8) flat, (9) epub, (10) tsv. And finally, the **language(s)**, to inform on the languages in which the data is available, for which it includes all European languages: EN (English), BG (Bulgarian), CS (Czech), DA (Danish), NL (Dutch), FI (Finnish), FR (French), DE (German), EL (Greek), HU (Hungarian), GA (Irish), IT (Italian), LV (Latvian), LT (Lithuanian), NO (Norwegian), PL (Polish), PT (Portuguese), RO (Romanian), RU (Russian), SR (Serbian), SL (Slovenian), ES (Spanish), SV (Swedish), TR (Turkish), UK (Ukrainian), or other.

For the **license** field, the **type of license** is assessed according to the following list: CC0 1.0 Universal (Public Domain Dedication), CC Attribution 4.0 International (CC BY 4.0), CC Attribution-ShareAlike 4.0 International (CC BY-SA 4.0), CC Attribution-NonCommercial 4.0 International (CC BY-NC 4.0), CC Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0), CC Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0), CC Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), Data licence Germany - attribution - v2.0 (Data license BY 2.0).

The usability field was introduced to analyse the data according to the five dimensions of data usability defined in Data Usability and Analysis. Transportation Performance Management (TPM) Guidebook [2]. It intends to provide an assessment of the usability or usefulness of the dataset according to the BuiltHub purposes. Although it is difficult to provide such assessment in an objective manner, it is structured according to the five dimensions and relating them with some defined parameters from the previous aspects analysed, to obtain a final score from 1 to 5 in each of them, and obtaining a final one as the mean of all five dimensions. The dimension of relevance is related to the thematic areas field, while the quality dimension is assessed through the source type field (Table 2). The dimension of coverage and granularity is related to the time resolution and spatial resolution fields, focusing in the data at country level for all EU countries, which is interesting but not much relevant the data at other NUTS levels, since for most of the datasets with these conditions, data is not available for all countries (only few of them); and dimension of accessibility and documentation is assessed through the availability on the web field (

Table 3). Finally, the dimension of **ease of analysis** is assessed through the link with format and language(s) fields (

Table 4).

Table 2: Relevance and Quality dimensions of the usability field, with the scoring assessment descriptions

Score	I1. Relevance	I2. Quality
	*Linked to **B**. Thematic areas field*	*Linked to **F2**. Source type sub-field*
★	No Area (Level 1) covered	Assumption/ Estimations
★★	1 Area (Level 1) covered	Model output/ Calculations
★★★	2 Areas (Level 1) & 2 Areas (Level 2) covered	Literature
★★★★	2 Areas (Level 1) & ≥3 Areas (Level 2) covered	EU project
★★★★★	3 or more Areas (Level 1) covered	Official statistics

Table 3: Coverage and Granularity and Accessibility and Documentation dimensions of the usability field, with the scoring assessment descriptions

Score	I3. Coverage and Granularity	I4. Accessibility and Documentation
	*Linked to **D**. Time resolution and **E**. Spatial resolution*	*Linked to **G1**. Availability of the web*
★	NUTS0 only for few countries (less than 20)	Viewable but not downloadable data
★★	NUTS0 for EU countries & accumulated data	Registration required and then only donwloadable selected fields/indicators
★★★	NUTS0 for EU countries & yearly data (or more) for 2 - 5 years	Open but downloadable only selected fields or indicators
★★★★	NUTS0 for EU countries & yearly data (or more) for 5 - 20 years	Registration required but then downloadable all data
★★★★★	NUTS0 for EU countries & yearly data (or more) for > 21 years	Open - downloadable all data

Table 4: Ease of analysis dimension of the usability field, with the scoring assessment descriptions

Score	I5. Ease of analysis
	*Linked to **G2**. Format and **G3**. Language(s)*
★	Not downloadable data - only visualisation (picture)
★★	Data in a format other than xlsx, csv and xml, and not in EN language
★★★	Data in a format other than xlsx, csv and xml, but in EN language
★★★★	Data available in xlsx, csv or xml format, but not in EN language
★★★★★	Data available in xlsx, csv or xml format, and in EN language

For quick checking, the information of the stars scoring for each dimension on the usability is added as information box in the corresponding cell in the matrix tab.

The last field of the matrix, the **files information**, is about the **number of files** where the dataset information used in BuiltHub project can be found.

5. Conclusions

The current analysis of datasets in the matrix form in a homogenised and organised way is key for their inclusion in the BuiltHub platform. It serves as metadata information for the platform. This is also beneficial when making the datasets accessible, linking them to indicators, including them in the dashboard and creating relevant visualisations.

This proper protocol to analyse datasets in a coherent and structured manner is needed for the making of the datahub. This makes the integration of data in the platform easier and more coherent, and the production of visualisation diagrams, maps and comparison tables from different data sources more dynamic and integrated. It is also key to link the data from different datasets across the EU with the defined indicators in which the data is also made available through the Platform.

The matrix allows also to perform analytics on each of the fields analysed for the datasets list, more specifically for the fields in which the options are pre-defined to classify them.

In the dataset identification, the topic to which the datasets are related can be assessed (**¡Error! No se encuentra el origen de la referencia.**), and it can be extracted that most of t he datasets are related to the building stock, which makes sense as for the purpose of the Project, while the others should remain as complementary data to build the indicators. It can also be observed together with the classification according to the content, which is quite balanced between the homogeneous datasets (focus on a concrete thematic area) and the heterogeneous ones (with data on more than a thematic area).

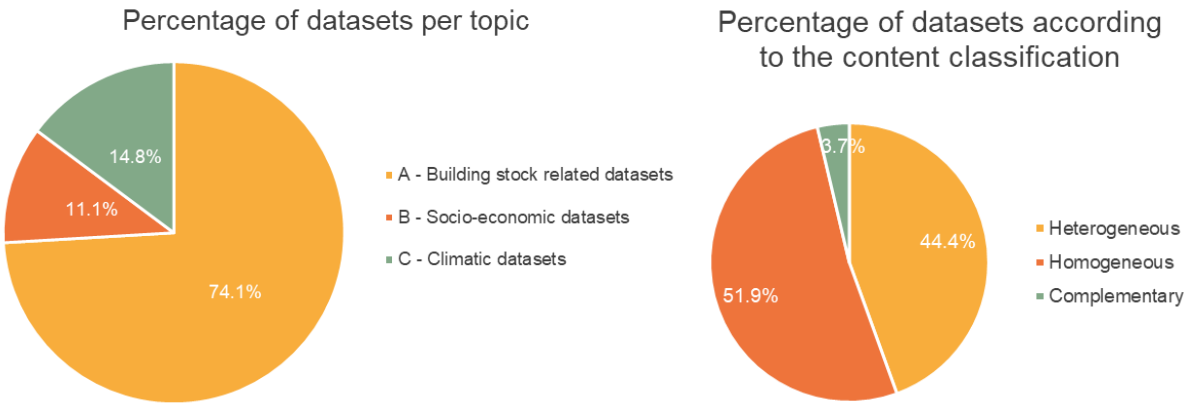

Figure 3: Percentage of datasets per topic and according to their content classification

With respect to the thematic areas, the analytic information (Figure 4) presents that the thematic area with most information from datasets is Energy, followed by building characteristics, building stock and climate. It can also be highlighted that there are some defined thematic areas where no datasets are included in the matrix with information on that. With respect to the analysis at datasets level (Figure 5), it is remarkable that most of the datasets (76%) include information in the Energy thematic area, followed in the same proportion as the previous graph (Figure 4) by the other thematic areas.

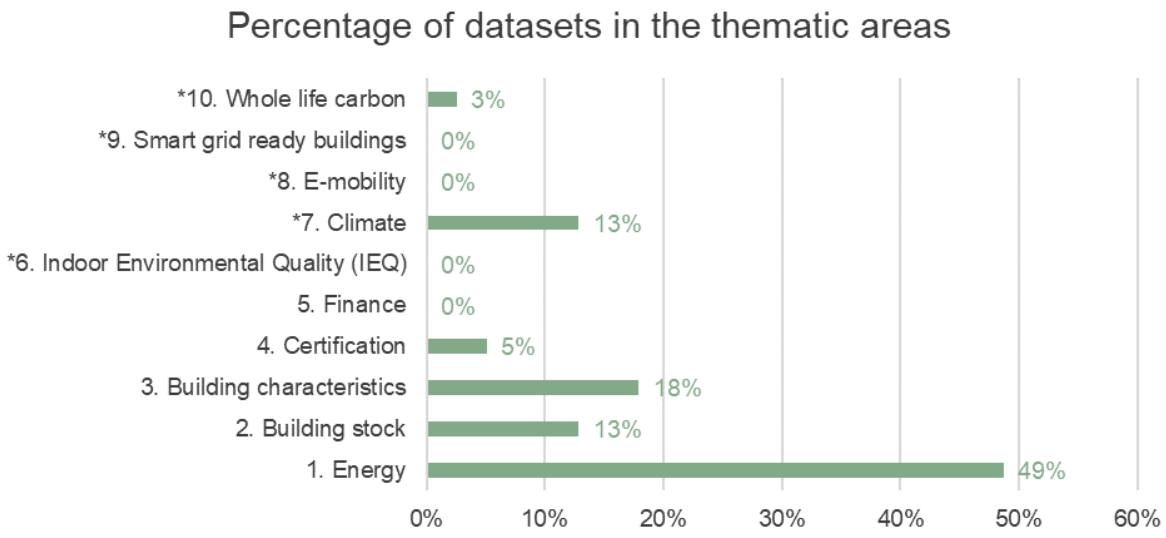

Figure 4: Percentage of datasets that include the different per thematic areas out of all the thematic areas that appear across all of them (taking into account that some datasets had more than one – until three in the matrix)

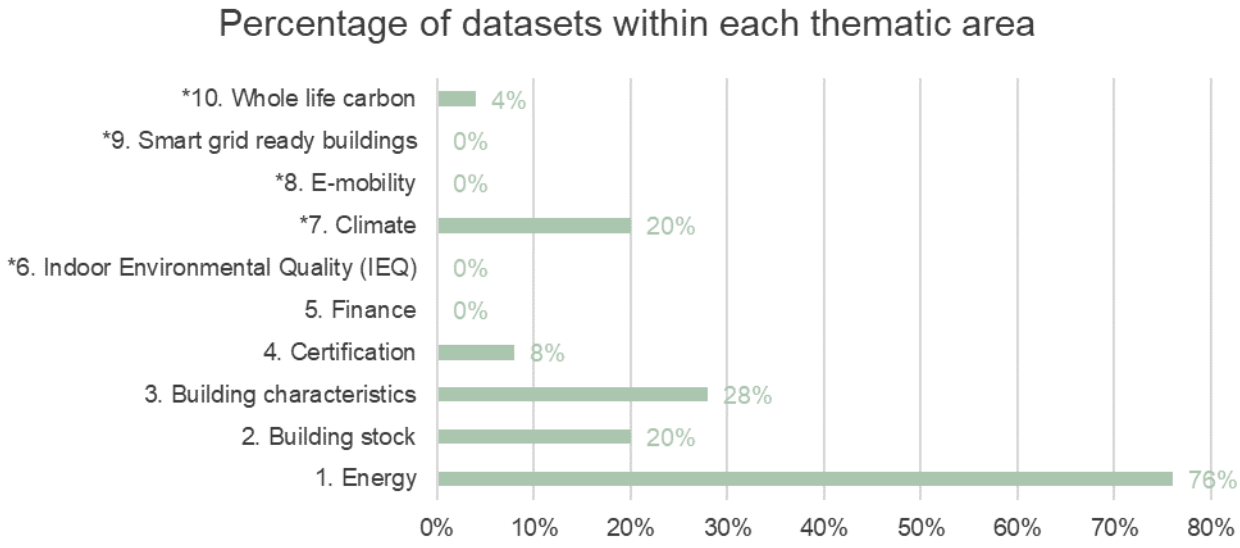

Figure 5: Percentage of datasets within each thematic area from the total of datasets

In the cross-cutting entities analysis, it is remarkable that all datasets include information at country level (Figure 7), being the cross-cutting entity with more information (Figure 6). For the others, it is followed by building type, energy carrier, construction year class and tenure status.

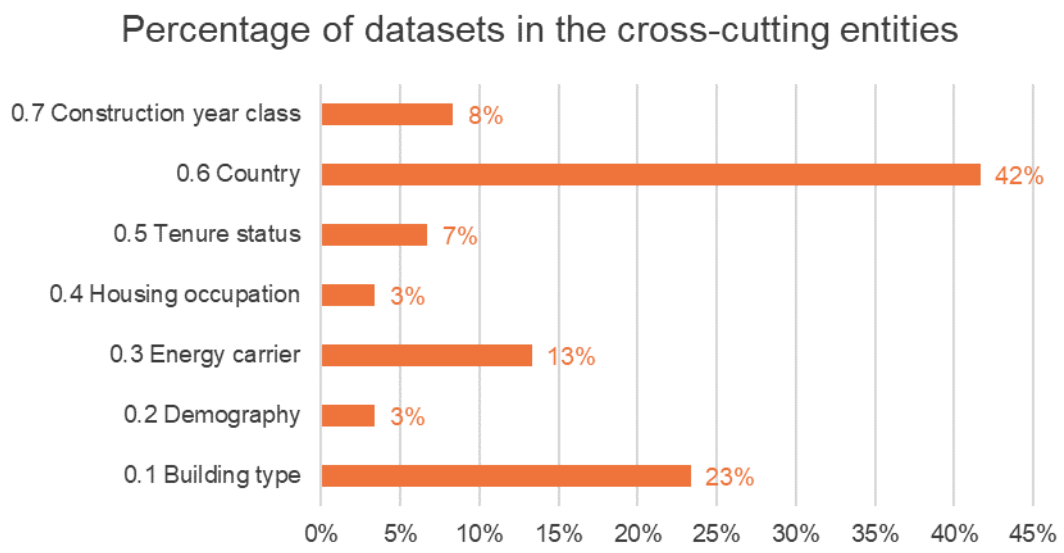

Figure 6: Percentage of datasets that include the cross-cutting entities out of all the cross-cutting entities that appear across all datasets (taking into account that some datasets have more than one, until all of them)

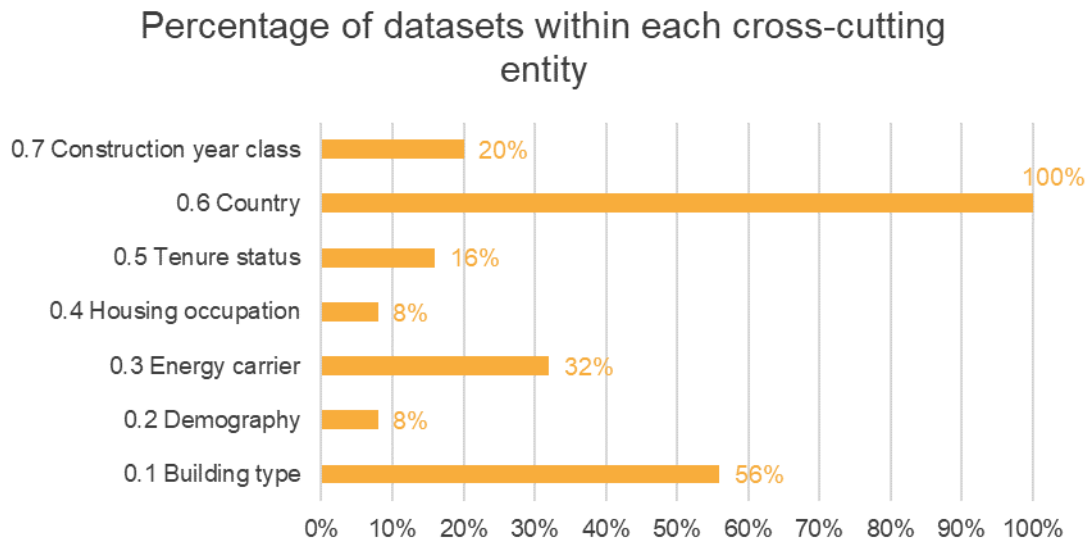

Figure 7: Percentage of datasets within each cross-cutting from the total of datasets

With respect to the source type (Figure 8), there are only three of the predefined list of types that appear in the datasets analysed. Most of them come from official statistics (around 67%), and almost 26% come from EU projects.

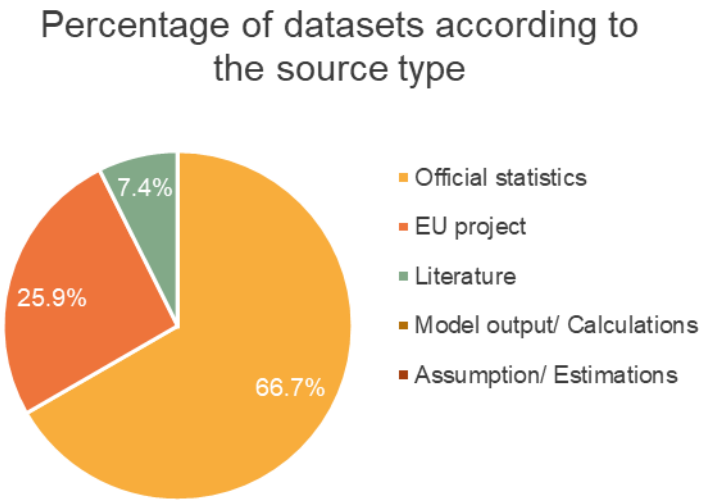

Figure 8: Percentage of datasets according to the source type

With respect to the availability on the web (Figure 9), there are only three of the predefined list of types that appear in the datasets analysed. Most of them are open-downloadable (around 74%), while other 22% are downloadable but only selected fields or indicators, and not all data at once. The same for the other almost 4% of the datasets, which are downloadable by selected fields or indicators, and prior registration is required.

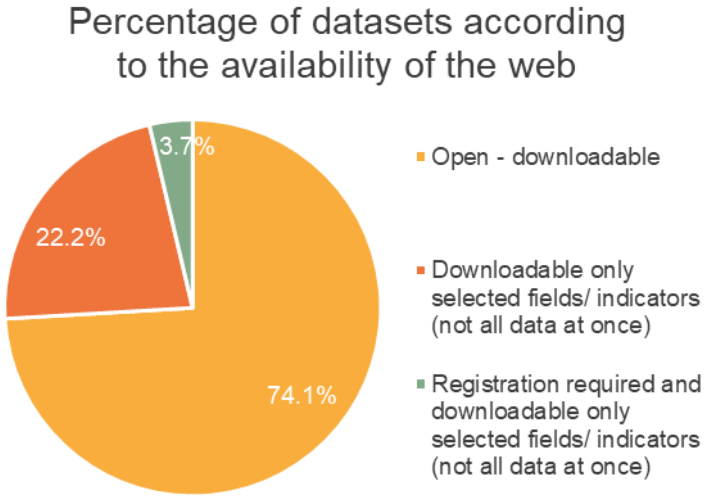

Figure 9: Percentage of datasets according to the availability on the web

For the usability field, the analytic (Figure 10) shows that the total score for all datasets is only between three and four stars, being the 85% of the datasets of four stars score. In the relevance dimension, most of the datasets (52%) obtain two stars, which means that only one thematic area is covered. In the quality dimension, the 67% of datasets obtain a five-star score, which means that the source type is official statistics. For the coverage and granularity dimension, 44% of datasets obtain 2 stars, which corresponds to NUTS0 for EU countries and accumulated data (in time resolution); while 30% of datasets obtain 4 stars, meaning the NTUS0 for EU countries and yearly data (or more frecuent) for 5-20 years. With respect to the accessiblity and documentation dimension, it turns out that most of the datasets analysed

(74%) obtain a five-star score, which corresponds to open-downloadable data (related to the availability on the web field). Finally, in the ease of analysis dimension, five-star is also the score for most of datasets (78%), which means that data is available in xlsx, csv or xml format, and in English language.

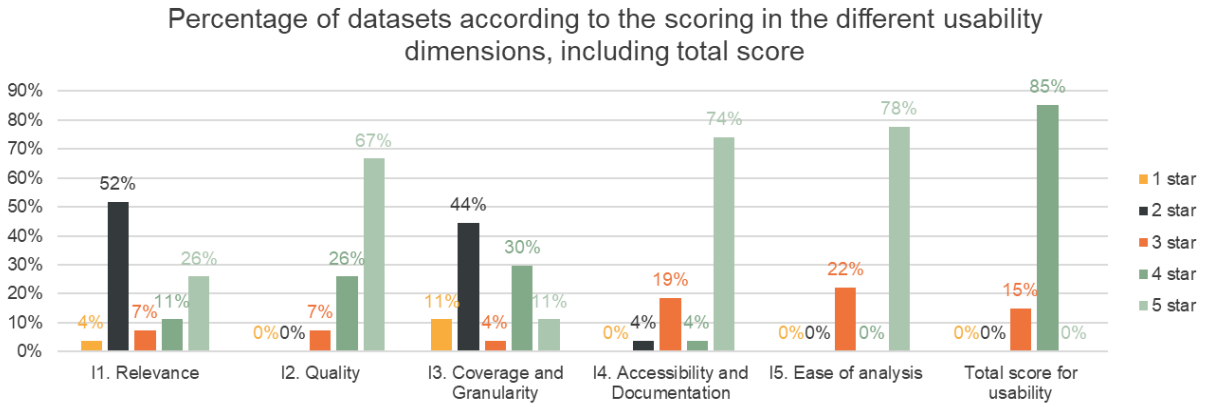

Figure 10: Percentage of datasets according to the scoring in the different usability dimensions, including total score for usability

6. References

[1] Directorate-General for Energy (2017). Support for setting up an observatory of the building stock and related policies. https://energy.ec.europa.eu/publications/support-setting-observatory-building-stock-and-related-policies_en

[2] U.S. Department of Transportation, Federal Highway Administration, Component D: Data Usability and Analysis. Transportation Performance Management (TPM) Guidebook. www.tpmtools.org/wp-content/uploads/2016/09/guidebook-component-d.pdf

BuiltHub

builthub.eu