COLLECTiEF

D.1.3 DATA MANAGEMENT PLAN

Project acronym: COLLECTiEF

Project title: Collective Intelligence for Energy Flexibility

Call: H2020-LC-SC3-2018-2019-2020 (Building a low-carbon, climate resilient future: Secure, Clean and Efficient energy)

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

Disclaimer

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683. The only responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the European Commission. The European Commission is not responsible for any use that may be made of the information contained therein.

Project no.	101033683
Project acronym:	COLLECTiEF
Project title:	Collective Intelligence for Energy Flexibility
Call:	H2020-LC-SC3-2018-2019-2020 (Building a low-carbon, climate resilient future: Secure, Clean and Efficient energy)
Start date of project:	01.06.2021
Duration:	48 months
Deliverable title:	Data_Management_Plan.docx
Due date of deliverable:	30.11.2011
Actual date of submission:	30.11.2011
Deliverable Lead Partner:	Partner No. 10, European Research Centre for Technology, Design and Materials (CETMA)
Work Package:	1
No of Pages:	36
Keywords:	Data, DMP, Data Management Plan

Name	Organization
Antonio Luparelli	CETMA
Amedeo Ingrosso	CETMA

Dissemination level

PU	Public

History

Version	Date	Reason	Revised by
0.1	20/09/2021	Initial draft version shared with the partners	All Partner
0.2	22/10/2021	Second draft version shared with the partners	All Partner
0.3	05/11/2021	Revision of the draft by R2M partner	R2M
0.4	23/11/2021	Submission of document after final review for approval by project management board	NTNU
1.0	29/11/2021	Final version	

Executive Summary

In accordance with the H2020 data management guidelines[1], the COLLECTiEF data management plan defines the management policy for the data that will be generated in the project, specifically what types of data are to be generated in the project, whether and how they will be made open and accessible for verification and re-use. It also specifies how it will be curated, processed and stored, with details of ethical privacy and security issues. Applicable regulations on human participation (e.g., informed consent, data processing, data security) and relevant regulations such as GDPR or H2020 Ethics or fair, will be mentioned in this document. The project will support openness according to the EU FAIR approach and the principle "as open as possible, as closed as necessary".

This document is part of WP number 1 of the COLLECTiEF project (task 1.3) and is responsible for the elaboration of the project's data management requirements and policies. The general requirements of the project are directly related to the nature of the data collected from the pilot buildings located in Norway, France, Italy and Cyprus, which will be processed and analysed only by the project partners, with the aim of developing software and hardware packages to increase the energy flexibility and climate resilience of buildings in urban areas by controlling and managing the interactions between buildings and energy systems using collective intelligence (CI).

This version of the document is the first version due at the end of the sixth month of the project (M6). This document will evolve during the project lifecycle and will be updated at M24 and M36, as project implementation progresses and significant changes occur to include new information, new datasets and results.

[1] European Commission., Data management - H2020 Online Manual. Available at: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [Accessed 22 November 2021].

Table of contents

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

List of Acronyms

API	Application programming interface
BACS	Building Automation and Control System
BMS	Building Management System
CC	Creative Commons licenses
CI	Collective Intelligence
CO	Confidential
CVS	Comma-Separated Values
DB	Database
DBMS	Database Management System
DDI	Documentation Data Initiative
DMP	Data Management Plan
DOCX	Microsoft Word extension file
DOI	Digital object identifier
DPO	Data Protection Officer
FAIR	Findable, Accessible, Interoperable, Re-usable
GA	Grant Agreement
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
MD	Markdown Language
MQTT	Message Queue Telemetry Transport
ODF	Open Document Format for Office Applications
ODP	OpenDocument Presentation
ODS	OpenDocument Spreadsheet
ODT	OpenDocument Text
PDF	Portable Document Format
PID	Persistent Identifier
POE	Post Occupancy Evaluation
PPT	Microsoft PowerPoint extension file
PY	Python extension file
REST	Representational state transfer

List of figures

List of tables

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

1. Introduction

1.1 Purpose

This document, D1.3 - Data Management Plan (DMP) is a result of the COLLECTiEF project, funded by the European Union's Horizon 2020 programme under Grant Agreement 101033683. The main points covered by the general plan for the management of data generated and collected during the project are listed below:

1. The management of research data during and after the project.
2. What data will be collected, processed or generated.
3. What methodologies and standards will be applied.
4. Whether data will be shared / made open and how.
5. How the data will be processed, stored and protected.

As specified in the guidelines, the DMP is a document outlining how research data will be managed during a research project, and also after the project has been completed. It should describe what data will be collected, processed or generated, making explicit what methodology and standards will be used, whether and how this data will be shared and/or made open and how it will be curated and archived. As written above, equally important will be the ethical management plan, which also requires some technical restrictions associated with the security of the personal data that will be collected in the pilot sites. Moreover, this DMP is consistent with the exploitation and protection of the results that will be developed within the COLLECTiEF project.
The preparation of this deliverable aims at answering all these questions, to achieve this objective, document D1.3 is organised as follows:

- **Project overview** (Chapter 2): description of architecture and data flow
- **Data Description** (Chapter 3): data description and data management used in the COLLECTiEF project. Here we will describe the tables of datasets that will be used within the project and that will be updated throughout the duration of the project.
- **FAIR Data** (Chapter 4): description of policies and methodologies to make data adhere to FAIR Data principles (findable, openly accessible, interoperable, re-usable)
- **Allocation of resources** (Chapter 5): description of resources and costs to implement the plan.
- **Data access and security** (Chapter 6): description of storage, backup and security methods and procedures
- **Ethical aspect** (Chapter 7): description of applicable regulations on human participation, informed consent, data processing, data security and relevant regulations such as GDPR.
- **Other** (Chapter 8): description of national/fundamental/departmental procedures for data management
- **Responsibilities** (Chapter 9): description of partners' roles and responsibilities in the project

2. Project overview

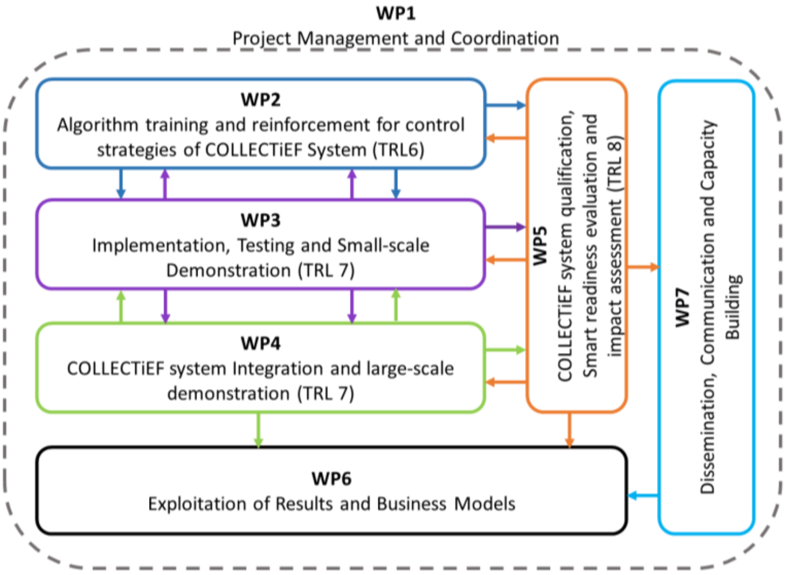
2.1 Project abstract

COLLECTiEF is an EU-funded H2020 project, running for 4 years (2021 to 2025). The vision of COLLECTiEF is to improve energy flexibility through the development of sustainable and resilient urban energy solutions. Concretely, the COLLECTiEF project will implement, test and qualify a Collective Intelligence (CI)-based interoperability and communication platform that enables easy and seamless integration of legacy equipment into a collaborative network for scalable energy management within existing buildings and neighbourhoods in larger urban systems with low installation costs, reduced data transfer and computing power, while increasing data security, energy flexibility and climate resilience. This collaborative network interacts to provide energy efficiency and flexibility, based on user preferences and requirements, and uses self-learning to maximise user comfort. This is done by developing software and hardware packages to install and make buildings and their legacy equipment intelligent on a large scale, while maintaining simple and robust communication with the energy network.

2.2 Project objectives

COLLECTIEF Data Management Plan (DMP) is developed to provide the necessary tools to outline the appropriate data management procedures. In relation to the Work Packages (WP) of the project, the development and evaluation of the COLLECTiEF algorithms and frameworks for Edge and Cluster nodes (WP2) will be adapted to the use case and will be tested in the G2ELab as a small-scale demonstrator (WP3), which will pave the way for the integration, demonstration activities of the large-scale system (pilot buildings) for capturing the baseline values with regard to energy consumption and other parameters (WP4). System qualification, smart readiness assessment and impact evaluation (WP5) will be based on data analysis. In conjunction with these activities, COLLECTiEF plans to develop a strategy for the effective development and exploitation of the project results, with a focus on replicability and upscaling across Europe (WP6), as well as a targeted set of dissemination and communication activities (WP7) that will promote the project results.

Figure 1 – Overview of Work Packages


WP1
Project Management and Coordination
WP2
Algorithm training and reinforcement for control strategies of COLLECTiEF System (TRL6)
WP3
Implementation, Testing and Small-scale Demonstration (TRL 7)
WP4
COLLECTiEF system Integration and large-scale demonstration (TRL 7)
WP5
COLLECTiEF system qualification, Smart readiness evaluation and impact assessment (TRL 8)
WP7
Dissemination, Communication and Capacity Building
WP6
Exploitation of Results and Business Models

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

The data generated, collected, processed and analysed during the project will contribute to the following objectives:

Table 1 – COLLECTiEF Projects Objectives

NO	Objectives	Description
1	**Development of control strategies, algorithms and structures for demand management**	In this phase, existing algorithms based on collective intelligence will be improved, considering different communication logics, indicators and targets, also based on external information such as dynamic prices, weather forecasts, imbalances between supply and demand, etc.). In addition, occupant-centred control algorithms will be developed, based on user needs and health requirements that have been identified beforehand (e.g., thermal, visual, acoustic, indoor air quality, etc.).
2	**Realisation of system components and adaptation of technical components to the use case**	This includes the adaptation and combination of Cluster and Edge Nodes and the integration of IoT devices relevant for smartness upgrades, including algorithms for smart plugs, smart thermostats and DALI dimmers. Corresponding hardware/software interfaces will also be developed. In addition, a user-friendly interface will be developed.
3	**Evaluation of COLLECTiEF algorithms and control strategies**	The partner CSTB has developed the DIMOSIM simulation platform in order to have a test bed for testing innovative control and optimisation algorithms and products. This platform makes it possible to connect either algorithms (software) to test them in a co-simulation or products (physical equipment with algorithms installed) to evaluate them through emulation techniques (hardware in the loop). This platform will be made available throughout the project in order to better develop and improve the algorithms and control strategies and subsequently to ensure the correct functioning of the overall solution (hence also the hardware) before implementing them in the pilot sites.
4	**Testing in a small-scale demonstration site**	The test will take place in the G2ELab, with access provided by the consortium partner CSTB. a part of the building is used as a "living laboratory" which will allow the evaluation of COLLECTiEF algorithms and control strategies in a real building. the living laboratory is equipped with a large number of meters and sensors which will allow the validation of COLLECTiEF measurement and metering products. The algorithms will be linked to the energy system in place (energy production for the building, energy delivery)

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

5	**Large-scale demonstration of the COLLECTiEF solutions in three countries**	Four countries in different climate zones, ranging from "cold" (Norway) to "warm" (Cyprus), were chosen for the demonstration. Comparing performance data from these different climate zones will help COLLECTiEF understand how COLLECTiEF should be adapted on the technical performance side to be universally applicable across Europe, but also how different users/occupants from these countries have different preferences that need to be taken into account.
6	**Development of business models and exploitation of the results**	While the large-scale demonstration will give indications on the technical feasibility of the COLLECTiEF system, its market potential and commercial viability will also need to be assessed and refined, especially with regard to replicability and market upgrade across Europe. An important pillar is the analysis of the regulatory framework and standardisation needs in order to adapt the system to many different markets in Europe.
7	**Capacity building and stakeholder engagement**	COLLECTiEF will also address the scientific community with journal articles and conference presentations, as academia is often a catalyst and multiplier for new and complex approaches. Furthermore, large building complexes owned by academic institutions would be ideal use cases for the COLLECTiEF system, making these institutions also potential customers for COLLECTiEF.

2.3 COLLECTiEF Conceptual architecture

The following is an architectural diagram for the data flow generated in the COLLECTiEF network system including Cluster Node, Edge Node, legacy supervisory systems (e.g., BMS, BEMS, SCADA), Human-Building Interface, fully integrated Dashboards, legacy home-building equipment, Sphensor™, and smart plugs (**Figure 3 – COLLECTiEF: Operational phase data flow schema**). Specifically, project data will be generated from:

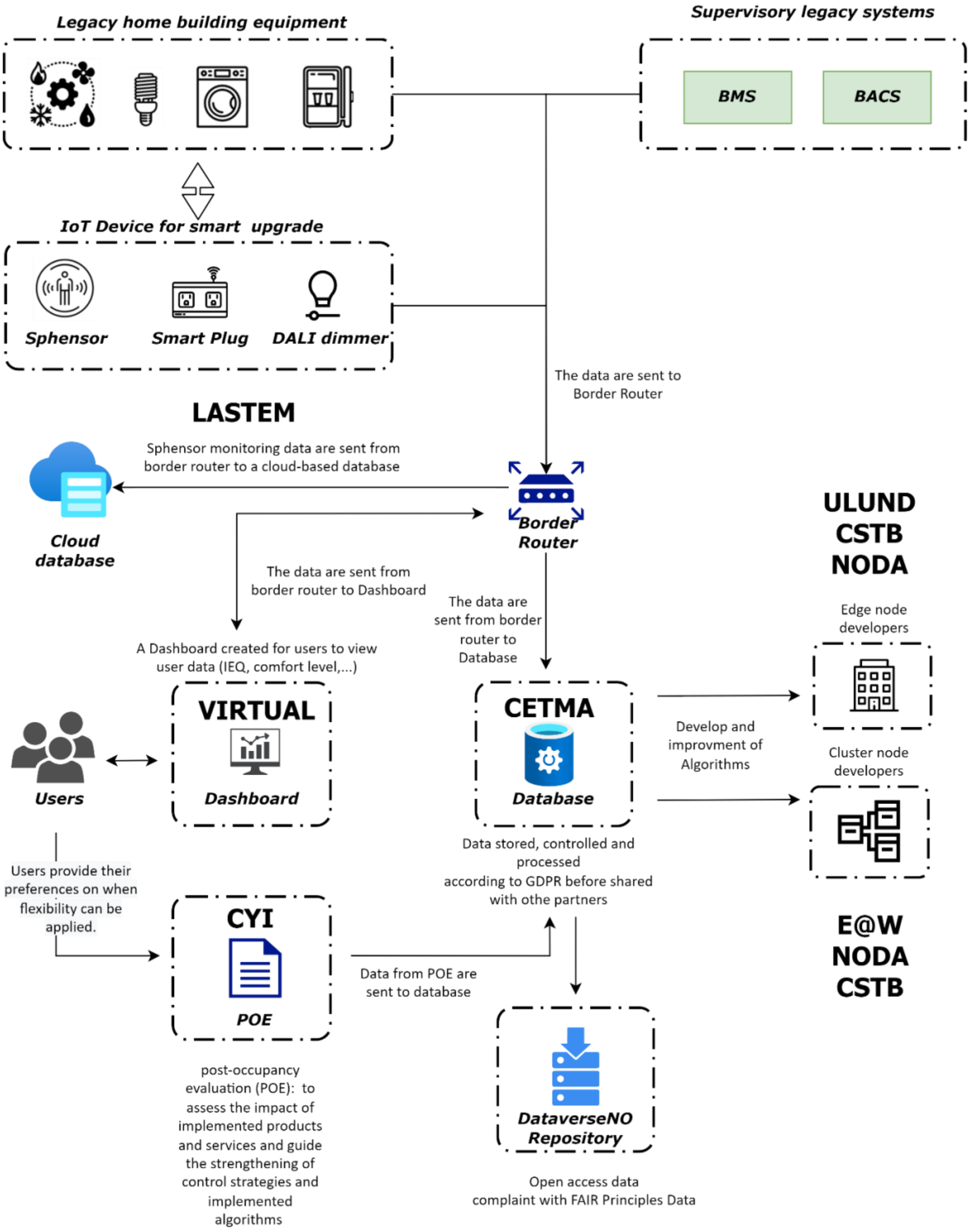
(i) existing legacy Hardware (HW) devices,

(ii) new IoT devices and systems required to encompass legacy barriers by interconnecting existing HW devices and systems.

Timewise, two implementation phases are planned:

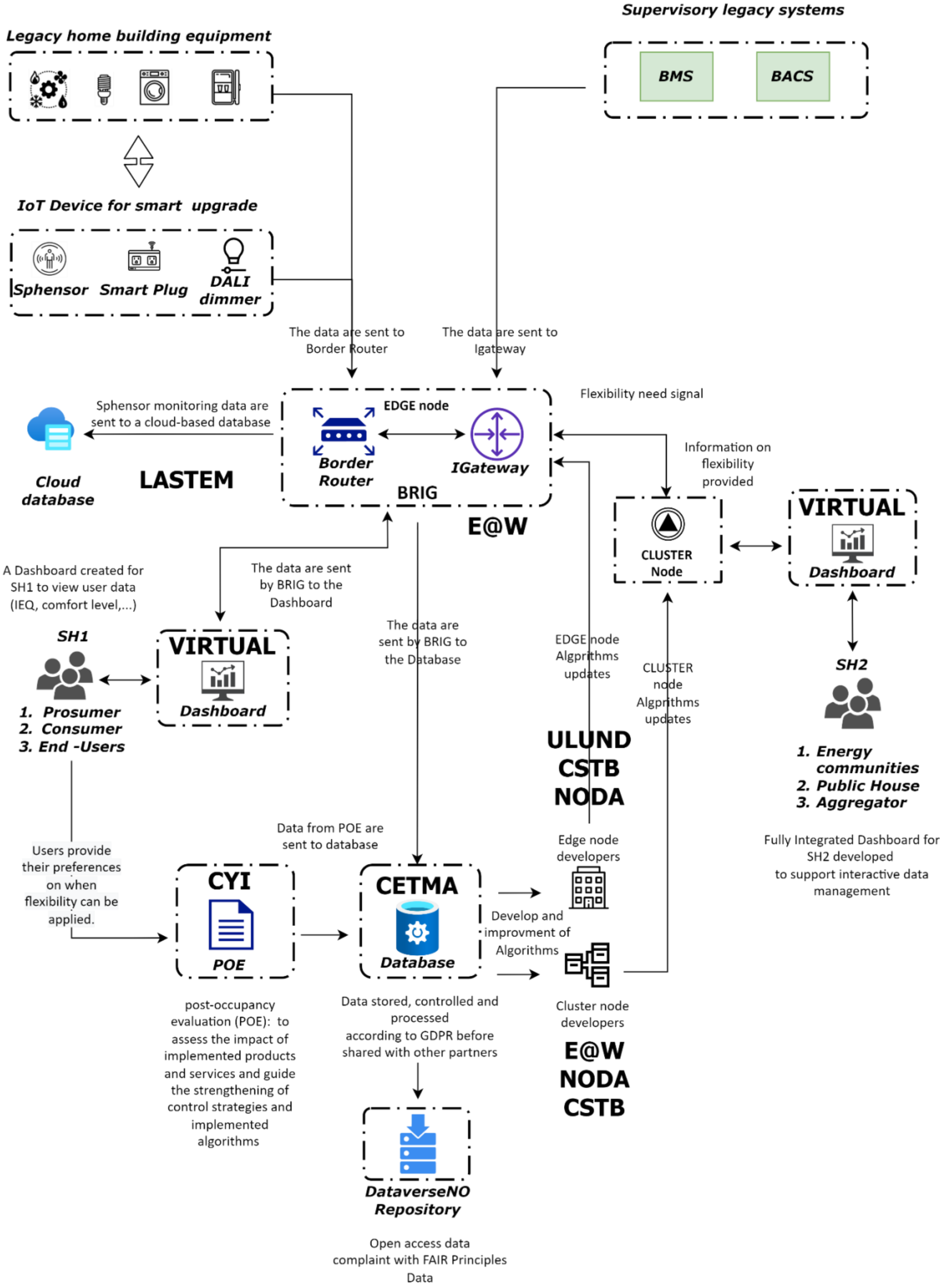
Monitoring phase (Figure 2): is between month M13-24 of the project and involves the acquisition of data generated by occupant-centric fusion sensor network (Sphensor™), with measurements related to Air temperature, Relative humidity, Illuminance, CO_2, VOC, PM 1, PM 2.5, PM 4, PM 10; data generated by supervisory systems (BACS, BMS) and energy consumption data from energy bills and/or smart meters.

Figure 2 - COLLECTiEF: Monitoring phase data flow schema


Legacy home building equipment
Supervisory legacy systems
BMS
BACS
IoT Device for smart upgrade
Sphensor
Smart Plug
DALI dimmer
The data are sent to Border Router
LASTEM
Sphensor monitoring data are sent from border router to a cloud-based database
Cloud database
Border Router
ULUND
CSTB
NODA
The data are sent from border router to Dashboard
The data are sent from border router to Database
Edge node developers
A Dashboard created for users to view user data (IEQ, comfort level,...)
VIRTUAL
Dashboard
CETMA
Database
Develop and improvment of Algorithms
Cluster node developers
Users
Users provide their preferences on when flexibility can be applied.
Data stored, controlled and processed according to GDPR before shared with othe partners
E@W
NODA
CSTB
CYI
POE
Data from POE are sent to database
post-occupancy evaluation (POE): to assess the impact of implemented products and services and guide the strengthening of control strategies and implemented algorithms
DataverseNO Repository
Open access data complaint with FAIR Principles Data

Operational Phase (Figure 3): the second is between month M 25-48 of the project, when COLLECTiEF solutions including edge and cluster nodes will be deployed in the pilot buildings. During this phase, the interoperability platform (iGateway) is introduced and apart from the data collected from Sphensor™, BMS and energy consumption, the addition of smart plugs and smart thermostats will generate data on existing HW devices, i.e., legacy home building equipment (heating, cooling, controlled, ventilation, lighting, etc.).

Figure 3 – COLLECTiEF: Operational phase data flow schema


Supervisory legacy systems
BMS
BACS
Legacy home building equipment
IoT Device for smart upgrade
Sphensor
Smart Plug
DALI dimmer
The data are sent to Border Router
The data are sent to Igateway
Sphensor monitoring data are sent to a cloud-based database
Cloud database
LASTEM
EDGE node
Border Router
IGateway
BRIG
E@W
Flexibility need signal
Information on flexibility provided
CLUSTER Node
VIRTUAL
Dashboard
A Dashboard created for SH1 to view user data (IEQ, comfort level,...)
The data are sent by BRIG to the Dashboard
SH1
1. Prosumer
2. Consumer
3. End -Users
VIRTUAL
Dashboard
The data are sent by BRIG to the Database
EDGE node Algprithms updates
CLUSTER node Algprithms updates
SH2
1. Energy communities
2. Public House
3. Aggregator
ULUND
CSTB
NODA
Users provide their preferences on when flexibility can be applied.
Data from POE are sent to database
CYI
POE
CETMA
Database
Edge node developers
Develop and improvment of Algorithms
Cluster node developers
E@W
NODA
CSTB
Fully Integrated Dashboard for SH2 developed to support interactive data management
post-occupancy evaluation (POE): to assess the impact of implemented products and services and guide the strengthening of control strategies and implemented algorithms
Data stored, controlled and processed according to GDPR before shared with other partners
DataverseNO Repository
Open access data complaint with FAIR Principles Data

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

a) Data between field devices, Border router and cloud database

The system consists of several existing sensors (Sphensor™) and HW devices connected in a radio network to a border router. The border router acts as a gateway for measurement data transmitted to the outside world through the radio network, where the cloud application platform operates. Monitoring data from the occupant-centric fusion sensor network, specifically the Sphensor™ system, will be collected by the border router and sent to a cloud-based database to be stored (containing no sensitive data) and used for diagnostic purposes during pilot deployments and to configure and control the entire system. With the aim to optimize the data communication and avoid any eventual redundancies and/or problems related to data congestion and transmission delays, within the edge node, iGateway and border router (BR) devices will be integrated into a single Hardware platform (BRiG) based on Raspberry 4 open hardware, enabling also a further level of flexibility in installations.

b) Data between BRiG and central server

Sensor monitoring and supervisory legacy systems data will be collected by BRiG and sent to a database where data generated by the research activity in the project will be stored. A persistent Data Storage and Management will be structured that will provide all the necessary tools for the management of the context history and data that can be stored using different relational DBMS technologies (such as MySQL, PostgreSQL, etc.). The logical-entity relations schema will be modelled, considering its three main layers:

- **internal (or physical) layer** for the representation of the logical schema in which the physical data storage structures will be defined
- **logical layer** for the description and representation of the database in terms of entities, entity attributes, and relationships between entity attributes
- **external layer** for the display and use of partial views made available for data access queries only to authorized users who have acquired permissions to access the repository.

This part of the task will be managed by CETMA. State-of-the-art methods and best practices on the concepts of privacy, security, and trust will be applied. Demonstration building data will be securely stored on a central server provided by NTNU - HUNT Cloud[2]. Data storage will have to observe compliance with certain standards set by the GDPR to ensure a certain level of logical and physical security systems (surveillance, monitoring, encrypted transmission, AES encryption with the ability to encrypt the content on the storage, etc.). The selected repository complies with Norwegian laws and regulations governing research, such as the EU Data Protection Directive (GDPR) and complies with the international quality management standard ISO 9001 and ISO 27001.

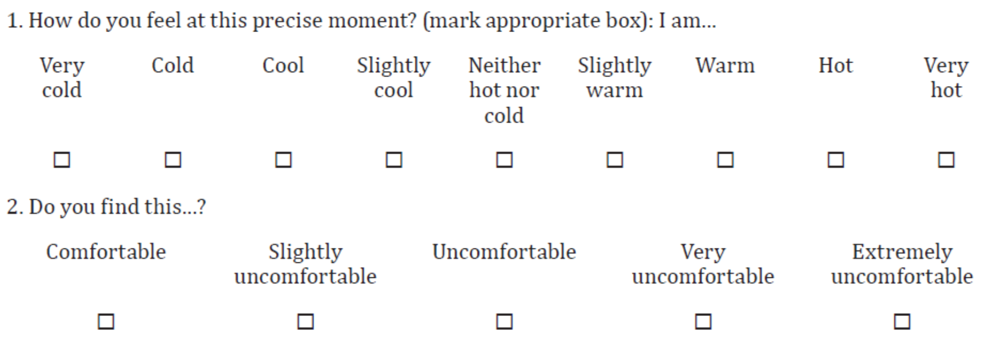
c) Data collected through post occupancy evaluation (POE) questionnaire

Data will also be generated from user satisfaction evaluation questionnaires. A post-occupancy evaluation (POE) platform will be developed by CyI for the systematic study of occupied buildings. The POE will assess various aspects of the operation and performance of occupied buildings from both a chemical/physical (indoor environmental quality, indoor air quality, thermal and lighting

[2] NTNU, HUNT Cloud: General information on cloud services, available at: https://www.ntnu.edu/mh/huntcloud [Accessed 22 November 2021].

environment) and subjective/interactional (space use, user satisfaction, etc.) perspective. These data will be collected in native digital format and will be sent to the central server.

Figure 4 – Example of POE questions from ISO 10551:2019 (Ergonomics of the physical environment – Subjective judgement scales for assessing physical environments).



1. How do you feel at this precise moment? (mark appropriate box): I am...

Very cold	Cold	Cool	Slightly cool	Neither hot nor cold	Slightly warm	Warm	Hot	Very hot
□	□	□	□	□	□	□	□	□

2. Do you find this...?

Comfortable	Slightly uncomfortable	Uncomfortable	Very uncomfortable	Extremely uncomfortable
□	□	□	□	□

d) Data between the BRiG and the user dashboard

A digital interface on top of the IoT operating system will be developed by Virtual to support interactive data management to support decision makers with relevant information at the right time. Graphical analytical interfaces which will basically consist of digital indicators, meters and diagrams to visualize information/knowledge; and responsive buttons, selectors and controllers to receive interaction responses. The digital dashboards at edge and cluster node will ensure that each stakeholder only sees and acts on relevant data. Security and privacy issues are considered to expose the correct information to the different parties.

e) Data between central server edge node and cluster node

The data generated by the COLLECTiEF project activities contained in the Data Storage and Management system developed by CETMA will be used for verification of building simulation models developed in WP2, and to develop the control algorithms based on CI at the building level (Edge Nodes) and for the improvement of the control algorithms available for COLLECTiEF at the Cluster Node level in WP3 and WP4. However, with the aim of facilitating software development, by fixing some technical details and promote a common framework it is possible to imagine situations where edge solutions use several local databases and share only some data with the central database. The choice and evaluation of the optimal architecture will be provided in the next updates of the plan.

f) Open data – DataverseNO

Finally, part of the data generated by the project will be made public and deposited in the DataverseNO repository[3].

[3] "*NTNU Open Research Data is an institutional repository for open data from all fields and disciplines. The archive is part of DataverseNO, which is operated by UiT The Arctic University. NTNU Open Research Data adheres to the guidelines and policy of DataverseNO, and all data sets will be curated before publication. DataverseNO is a Core Trust Certified repository and assigns DOIs (Digital Object Identifiers) to data sets. The standard license is CC0 (Creative Commons Zero), but other open licenses can be considered if needed*" NTNU Open Data - Wiki - innsida.ntnu.no. [Accessed 22 November 2021].

3. Data Description

3.1 Data description and management

A detailed description of each data category is provided in the following paragraphs. The following parameters - based on existing knowledge about the characteristics of the data - are examined:

- Purpose of data collection/generation in relation to project objectives.
- Types and formats of generated data collected.
- Reusability of the data.
- Data origin.
- Size of data.
- Use of data.

3.2 Collection and generation of data in COLLECTiEF related to the project objectives.

The purpose of data collection/generation in COLLECTiEF relates to the development of collective intelligence (CI)-based systems and processes that, as such, promote distributed self-organization and collaboration across the energy system. In COLLECTiEF, will be apply CI to a range of technological challenges with the goal of "making the existing European building stock smarter" for energy efficiency while ensuring user privacy and promoting broad access to data. The aim is to collect and harmonize data using a human-centric fit-for-purpose approach that integrates data related to, Energy consumption, Indoor physical environment, Ventilation systems, Outdoor environment, Information and data directed towards users.

The collection and generation of data in the COLLECTiEF project is linked to two-time patterns. The first is related to the monitoring phase which will take place between month 13 and month 24 of the project.

Project objectives during the Monitoring Phase (M 13-24): in this phase, the existing algorithms based on collective intelligence will be improved. The aim is to test the COLLECTiEF algorithms through co-simulation based on modelling and analysis of the building and energy system. Communication logics for the Edge and Cluster nodes of the COLLECTiEF system will also be implemented. In addition, occupant-centred control algorithms will be developed, based on user needs and health requirements that have been identified (e.g., thermal, visual, acoustic, indoor air quality, etc.). Existing solutions developed by the partners and new solutions will be tested and improved using the virtual test bench, DIMOSIM, and in a real environment at G2Elab. Demonstration Human-building Interfaces for prosumers, consumers, end-users will also be created to visualise/interact with the test data of the COLLECTiEF algorithm through co-simulation based on modelling and analysis of the building and energy system.

Project objectives during the operational phase (M 25-48): In this phase, the focus will be on the demonstration of the systems (Edge Node, Cluster Node, occupant-centric fusion sensor network, IoT operating system, human-building interface) in a large-scale operational environment, with a detailed description of the demonstration and monitoring performed, the main conclusions and recommendations on adaptation measures. Validation of the effectiveness of COLLECTiEF in different large-scale operational environments will be useful to provide data for further improvement of algorithms and control strategies based on test results. A fully integrated dashboard will be developed to support interactive data management for aggregators, energy communities, public

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

housing, building owners and managers, providing digital indicators, gauges and diagrams to visualise information/knowledge. Human-Building user-friendly interface requirements will be collected and digital dashboards designed for prosumers, consumers and end users will be improved/adapted to Human-Building user interface requirements.

Data collection will comply with all national and European ethical and legal requirements:

- Data and information management according to the General Data Protection Regulation (GDPR)[4].
- FAIR Data Principles: set of guiding principles for making data findable, accessible, interoperable and reusable (Wilkinson et al, 2016)[5].

The following standards will be used to ensure that data and information comply with the quality standards:

- EN 15603:2008 / ISO 52000:2017 (Energy performance of buildings - EPB global assessment).
- EN 16798:2019 (Energy performance of buildings - Ventilation for buildings).
- EN ISO 7726 (Ergonomics of the thermal environment - Instruments for measuring physical quantities).
- ISO 10551:2019 (Ergonomics of the physical environment - Subjective judgment scales for the assessment of physical environments).

3.3 Types and formats of data that the project will generate and collect

The data that will be managed in this DMP are informative (such data are related to e.g., reports, publications, dissemination activities, questionnaires, etc.), and technical (such data are related to e.g., measurements, simulations and datasets produced by partners, etc.). From this classification, COLLECTiEF will distinguish data derived from the project into the following key categories:

Table 2 – Key categories of data generated from the project the project

Key categories of data	Description
Underlying Research Data	Data produced by the research activities (including associated metadata), and used to validate the results presented (e.g., in scientific articles, dissemination activities, etc.). In line with the general principle 'as open as possible, as closed as necessary', the partnership will provide open access to research data and linkage to the respective publications, to enable the scientific community to review and validate results based on the underlying data.

[4] European Parliament and Council. *Directive 95/46/EC*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=EN [Accessed 22 November 2021].

[5] Wilkinson et al, M., *The FAIR Guiding Principles for scientific data management and stewardship*. Available at: https://www.nature.com/articles/sdata201618.pdf [Accessed 22 November 2021].

Operational and observational data (raw data, as building simulation models, laboratory tests. Etc.)	This category includes curated or raw data generated from technical and research activities, such as building simulation models, laboratory tests, operation of the pilot building (operational data), and data from qualitative activities, such as the post-assessment occupancy questionnaire (POE). The data in this section is mainly of a confidential data, only accessible to project partners. Data that will be made open will be published at the end of the project through other channels, such as dataverseNO.
Monitoring and evaluation data	This data will be captured to track KPIs of project performance in WP5. It will be assessed how it will be reported and published in the DataverseNO repository in a clearly defined and open way.
Reusable documentation, tools, and knowledge	These types of data relate to both general and project-specific documentation, including the methods, tools, software and underlying source code needed to replicate the results. This category also includes data that will be used for dissemination activities. All data that will be made public will be published in the DataverseNO repository and part of them, into Website of the project.

Pre-existing datasets, such Energy consumption data, energy bills, historical data (climate data, indoor environmental data from BMS, future weather files), owned by COLLECTiEF partners, will be used to providing inputs and boundary conditions data for COLLECTiEF network to be used in virtual test-bed. The partner who owns that dataset and makes it available to the project (unless otherwise stipulated in section 25 GA).

It is expected that the data generated by the research activities will be quantitative (the numerical data generated by field sensors and legacy home building equipment), and qualitative (such as text, images contained in reports, peer-reviewed publications). In general, the types of data expected include, but are not limited to:

- Integer
- Booleans
- Characters
- Floating-point numbers
- Alphanumeric strings

The management of data generated or collected in COLLECTiEF will be organised according to the three levels of accessibility defined below, i.e., data from private individuals, data confidential to project partners, and data made public:

- **Private data**: partners may decide to store these specific data on the servers of the institution/company to which they belong. It is specified that these types of data are outside the scope of the COLLECTiEF DMP.

- **Confidential Data, only accessible to project partners**: this type of data will be stored on the central server, which will be used for the duration of the project. It will also be agreed with the partners how the data stored on the server will be managed during and at the end of the project. This includes, for instance, raw data and software development activities (source codes and related documentation) which will be managed by the partners concerned using the GitLab repository. As part of the management of this category of data and in order to

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

facilitate software development activities, other repositories can be used by partners, as long as they offer similar accessibility resources and meet the requirements of the European Commission, especially where sensitive data are present.

- **Data made public:** the data that will be made public will be deposited on an open access repository FAIR aligned, which will be DataverseNO. This type of data includes all data used in the project that is not protected by a confidentiality or security clause, regulated by data protection, or any other clause that would conflict with the publication of the data. Other mechanisms for publishing public data will include the website and Open access to scientific publications data.

At the time of delivery, most Tasks have not yet fully defined the type and structure of data they need or will generate or can make available. Pending detailed descriptions, the following table shows the data management summary template to be used within the DMP and within Tasks for documentation.

Table 3 – Dataset description template

Data Field	Description
Data set reference and name	Identifies the data set to be produced
Data set description	This will describe the data generated or collected, their origin, nature, and whether they are used for scientific publications. Information will also be provided on the existence (or otherwise) of similar data and the possibilities for integration and re-use, and to whom they might be useful.
Data source, data ownership	Name of the partner that produced the data and explanation of the purpose of treatment
Metadata Standards, data formats, vocabularies	Reference will be made to appropriate existing metadata standards and vocabularies of the discipline governing data collection, aggregation, storage and sharing.
File Format	Description of file format used
Storage	Description of the procedures that will be put in place for data retention, including an indication of how long the data are to be retained, their approximate final volume and what the associated costs are and how they are expected to be covered.
Data sharing	Description of how the data will be shared. This section will also cover identify the repositories where the data will be stored, indicating the type of repository
Security & Privacy considerations	Should consider ethical issues, related to rules on personal data, but also implications on intellectual property, commercial, privacy and security aspects.
Dissemination Level	Detailing access procedures (e.g., whether widely open or restricted to specific groups), embargo periods (if any), dissemination and sharing methods (e.g., software required and other tools to enable re-use).
Stakeholders	Description of stakeholders interested in data reuse

Data will be all in digital format that can be interpreted by various software technologies (open and not). The data generated by the research activities in the COLLECTiEF project will be raw data, analysed, processed and published, the data analysis source code, documentation (metadata and software), result and conclusion. The following is a list of widely used 'open' data formats classified according to the nature of the data. However, depending on the specific needs of partners, they may be used interchangeably by specifying other data types and related specific formats. These will be specified in COLLECTiEF, and in case of significant changes, the DMP will take these implementations into account.

- Databases (.SQL etc.)
- tables and spreadsheets (CSV, ODS, TSV)
- text documents (TXT, ODT, PDF/A, XML)
- structured text (HTML, JSON, MARKDOWN, XML, RDF)
- source code (PY).
- images (JPG, PNG, etc.)
- video/movies (MPEG, WMV, MP4, etc.)
- audio (MP3, WAV, AIFF, etc.)

Table 4 – Type and format of data

Data	Description	Type	Format
Raw Data	Data obtained through sensor (Sphensor™), legacy home building equipment, supervisory legacy systems (BACS, BMS) and POE (post occupancy evaluation survey)	• Database • Tables and spreadsheet • Structured text	• *.SQL* • *.CSV* • *.JSON* • *.XML*
Post processed Data	Data after applying algorithms: the processing of the raw data in a more readable format and easy to understand (graphs, documents, etc.),	• Tables and spreadsheet • Text / Structured Text • Images	• *.CSV* • *.JSON* • *.ODT* • *.PDF/A* • *.JPEG,* • *.PNG*
Data analysis codes	Computer codes for analysing data	• Source code	• *.PY*
Results, presentation and dissemination	Data used for presenting research findings to both specific stakeholders (e.g., the relevant scientific community, In the form of research journal articles. scientific publications, technical reports, or peer-reviewed publications.) and a general audience for dissemination activities (e-learning materials, articles, dissemination materials,	• Text • Images • Video • Audio	• .PDF/A, • .ODT, • .JPEG, • .PNG • *MPEG, .AVI, .WMV, .MP4* • *MP3, .WAV*

	social media posts, local networks and platforms. etc.).		
Metadata documentation	Metadata along with every dataset about how the things were achieved.	• Structured text • Text	• .MD • .TXT
Documentation (requirements, architecture and design, technical, user)	Written text accompanying the description of the software, with the aim of explaining what functions it performs, how it is structured and implemented and how it is used. This section will also cover the Database development documentation relating to the description of the conceptual, logical and physical design elements (Entity-Relationship Scheme; Definition of entities and attributes; Relations and their attributes; Cardinality, Candidate Keys; Attributes and Tuples; Tables, Attributes, and Properties; Views; Primary Keys, Foreign Keys, etc.)	• Structured text • Text • Images	• .MD • .TXT • .PDF/A • .ODT • .JPEG, • .PNG

The decision to use these data formats is motivated by a preference for open formats, since they are widely accepted as standard by data centers and very widespread within the scientific community of reference. In general, standardized, interchangeable and open formats will be used for file storage and sharing. Although is a standard in business environments, Microsoft formats such as Excel, Power Point and Microsoft Word can hardly be considered neutral. The use of formats such as .DOCX and .PPT will be restricted in favour of ODF (Open Document Format) due to interoperability issues with Linux and MacOS operating systems.

Text-based documents such as scientific publications, articles, paper for dissemination etc. will be stored in mainly in .PDF/A format, but if necessary, also formats such as .TXT or .ODT are accepted. If formatting is required formats such as markdown format (.MD), combined with the original file will be used. Sets of related files will be packaged using standard open-source tools (e.g., zip). The format that will be used to store and export the raw data, tables and spreadsheet is relational database (SQL) and to facilitate data exchange, CSV, JSON and XML formats, will also be accepted. The same applies to post-processed data, where .JPEG, .PNG .PDF/A and ODT are also used. Visualizations of simulated data may be stored in standard graphics formats (.JPEG, .PNG). A suitable format for metadata and documentation (manuals, etc.,) is MarkDown (.MD) but .PDF/A or Plaintext TXT will also be used as required. Specific data formats are related to the type of ICT systems and tools developed by COLLECTiEF (Sphensor™ sensor network, etc.). As a standard data exchange format, JSON files will be preferred, but other formats will also be used.

3.4 Data management and Sharing plans

All data generated by the project (public and confidential for the project partners) will be stored on the central server provided by NTNU - HUNT Cloud and managed by CETMA, with the exception of strictly private data which might will be stored by the partners on the servers of the institution/company to which they belong and which are outside the scope of this data management plan.

Depending on their level of openness (public or partners only), and taking into account the ethical, commercial and confidentiality constraints of handling sensitive or closed data, the project considers the use of different sharing mechanisms, with the aim of making the research data of the COLLECTiEF project replicable, or at least reproducible or reusable. The main strategies to make the data generated by the project both open and FAIR compliant are described below.

For data whose nature will be public, the following publication mechanisms will be used to ensure the openness of the data:

A. **Website:** With reference to the objectives outlined in WP 7 the data produced by the research activities in the form of tangible and reusable knowledge (articles, publications, etc.) will be published in a publicly accessible form on the website (https://collectief-project.eu/).

The website will be mainly used for the opening of the results and conclusions for the general public and also for the expert public (researchers, energy experts, professionals, etc.). Information will be conveyed through e-learning materials, articles, dissemination materials, social media posts, local networks and platforms, etc.

With reference to the objectives outlined in WP 6 and 7, the partners R2M and Geonardo developed a template for the stakeholder database. The objectives of the template are the following, which are relevant for the market analysis of the project solutions

- to build a list of market segments/sectors useful for the market analysis in the exploitation process (WP6 task.6.1)
- to build a list of interested organisations (WP6 and WP7)
- building a list of contacts for dissemination activities (WP7 task.7.2)

Personal data (e.g., contact details of stakeholders/companies, newsletter subscriber data, etc.) will be treated confidentially and in accordance with legal regulations on the protection of sensitive data, and the information will be stored in accordance with the retention periods stipulated by law. Finally, confidential results that could harm the commercial interests of partners will not be published. The same applies to ethical issues (GDPR).

B. **Open Access to scientific publications:** as specified in the GA (Art. 29.2 and 29.3) each beneficiary must ensure open access (free online access for any user) to all peer-reviewed scientific publications related to its results. In addition, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications. Data underlying research activities, will be provided as support material for research articles published in journals, typically with the data files published by the publisher of the article. A large number of journals and publishers support the addition of supplementary material to research articles, including datasets.

C. **DataverseNO:** open data is data that can be freely shared and reused by anyone, FAIR data is data that follows a set of good practices for data sharing, respecting any ethical, legal or contractual restrictions (data may contain personal information, be subject to copyright, be protected by patents or trade secrets). In order to make open data in line with FAIR principles, the data produced by the research will be deposited in a repository designed to support the publication of research data. The identified repository is DataverseNO (https://dataverse.no) a curated and FAIR-aligned national generic repository for open research data from all academic disciplines.

For data that will be confidential, only accessible to project partners the following publication mechanisms will be used:

D. **GitLab**: source code and software documentation data (DIMOSIM) will be published on GitLAB https://gitlab.com/collectief_members/collectief. GitLab will be used to provide operational and observational data such as construction simulation models, as well as documentation, tools and reusable knowledge such as Dimosim documentation and example simulation projects.
   - Source codes: Python scripts that allow to run Dimosim simulations via its web interface
   - Raw data: Data needed to run simulations, such as profiles and building description files
   - Documentation: Includes Dimosim simulation software documentation and input data file formats as required.

It can also be used to receive the actual simulation projects used for the study, as well as their subsequent versions.
The repository is managed by the CSTB partner, and all members of the COLLECTiEF project will have read and write access to the Git repository. Each member is responsible for managing their account (setting a password, keeping it confidential, etc.). The data sent to GitLab is a snapshot of the work in progress with the possibility of seeing the evolution of each file. Any git client software can be used to obtain the data. The data can also be accessed through the Gitlab.com website, without installing any software. All project members who pull the project from the repository will have a copy of the entire repository on their local computer. The data will remain in the repository for the duration of the project and is subject to COLLECTiEF's confidentiality agreement. Data submitted to the repository are shared with all project partners but remain the intellectual and commercial property of their respective owners, unless otherwise stated. Data will be published at the end of the project through other channels, such as dataverseNO.
As mentioned before, with the goal of facilitating software development, fixing some technical details and promote a common framework, it is possible to imagine situations where edge node uses several local databases and share only some data with the central database. In this regard, the idea is to use the self-host developed by NODA (https://github.com/self-host/self-host) as a common platform for access control, rights management, time series database and for running the hooked software. NODA has released the open-source platform, https://github.com/self-host/self-host (under active development, still version < 1.0), which offers a possible solution for how to structure research and software development efforts in a way that allows co-simulation with the Dimosim platform as well as execution against a live environment. The open-source platform facilitates the management of time series data and the scheduling of containerized processes that act on metadata and corresponding data. Using the same platform for simulations, the cluster node software, the edge node software and the UI, it is possible to optimize the processes of communication, data storage and retrieval. This will make it easier to perform integration tests on the COLLECTiEF solutions, something that is crucial to succeed with the software development of the project.

3.4.1 Expected size of the data

With regard to the volume of data, this will be evaluated in the course of the project. The expected size depends on the extent and nature of the data made available. More details will be provided in future updates of the plan.

3.4.2 To whom they might be useful (usefulness of the data)

Depending on the area of expertise, the data generated under the COLLECTiEF project may be useful to:

- COLLECTiEF consortium;
- European Commission services and European Agencies;
- EU National Bodies;
- The general public including the broader scientific community;
- Building occupants
- Facility managers & owners
- ESCOs

4. FAIR Data

4.1 Making data findable, including provisions for metadata

In accordance with FAIR Data Principles guidelines: set of guiding principles for making data findable, accessible, interoperable, and reusable (Wilkinson et al, 2016)[6]. In the following, processes and mechanisms to make data adhere to the FAIR Data Principles will be made explicit. Starting with these guidelines, to ensure that data is **findable (F)** the following issues must be specified:

According to the **first principle (F1)**, to make data 'findable', **metadata must be assigned a unique and persistent global identifier (DOI)**. To meet this requirement, the dataverseNO repository uses two PID systems, DOI and Handle[7]. Therefore, each dataset that is deposited and made accessible will have a specific and persistent identifier called a Digital Object Identifier (DOI) at both the dataset and individual file level. This helps to ensure proper citation of research and to show which version of the data has been used.

According to the **second principle (F2)**, to make data 'findable', **data must be described with rich metadata**. To meet this requirement a specific metadata model, as an extension of the DataverseNO standard metadata schema, has been defined to describe, discover, and track the existing data collected by the COLLECTiEF project and the data that will be generated by it in the coming years. A search metadata schema based on the following widely used search metadata standards in human- and machine-readable formats will be used: Dublin Core, Documentation Data Initiative (DDI),

[6] Wilkinson et al, M., *The FAIR Guiding Principles for scientific data management and stewardship*. Available at: https://www.nature.com/articles/sdata201618.pdf [Accessed 22 November 2021].

[7] Conzett, P., DataverseNO: *A National, Generic Repository and its Contribution to the Increased FAIRness of Data from the Long Tail of Research*. Available at: https://munin.uit.no/bitstream/handle/10037/18564/article.pdf?sequence=3&isAllowed=y [Accessed 22 November 2021].

DataCite, and Schema.org. To make documents findable within the repositories, metadata will be included with the document, metadata documentation will be provided in two ways, in the metadata fields and in a separate ReadMe file that must be uploaded along with the data files[8].

According to the **third principle (F3)**, to make data 'findable', **all metadata must clearly and explicitly include the identifier of the data it describes**. To meet this requirement the PID (DOI or Handle) of the dataset will be part of the metadata record presented on the dataset landing page; the PID of the file will be part of the metadata record presented on the file landing page; and finally, both the PIDs of the dataset and the file are included in the exported metadata files. All three implementations are also adopted in DataverseNO[9].

According to the **fourth principle (F4)**, to make data 'findable', **retrievable metadata is recorded or indexed in a searchable resource**. To meet this requirement dataset published with DOIs in a Dataverse-based repository are collected and indexed by DataCite Search. Through DataCite this metadata is made available to several other search services, including BASE (Bielefeld Aca-demic Search Engine) and the search system used by libraries at Norwegian universities and university colleges. Schema.org metadata is encoded in Dataverse dataset landing pages and indexed from there by Google Dataset Search[10].

4.1.1 Approach towards search keyword

Within the DataverseNO repository, documents can be searched in two distinct ways, basic and advanced. In the basic search mode, queries or exact phrases can be used (text character between quotes). In the advanced search mode, search terms can be entered for collections, dataset metadata (citation and domain-specific), and file-level metadata. It is also possible to search for tabular data files, with variable names and labels. For this, it will be important to use keywords relevant to the published content (e.g., energy consumption, POE, ventilation, relative humidity) and any other more specific keywords relevant to the content of the publication), as well as to implement an appropriate and content-relevant file naming strategy.

4.1.2 Approach for clear versioning

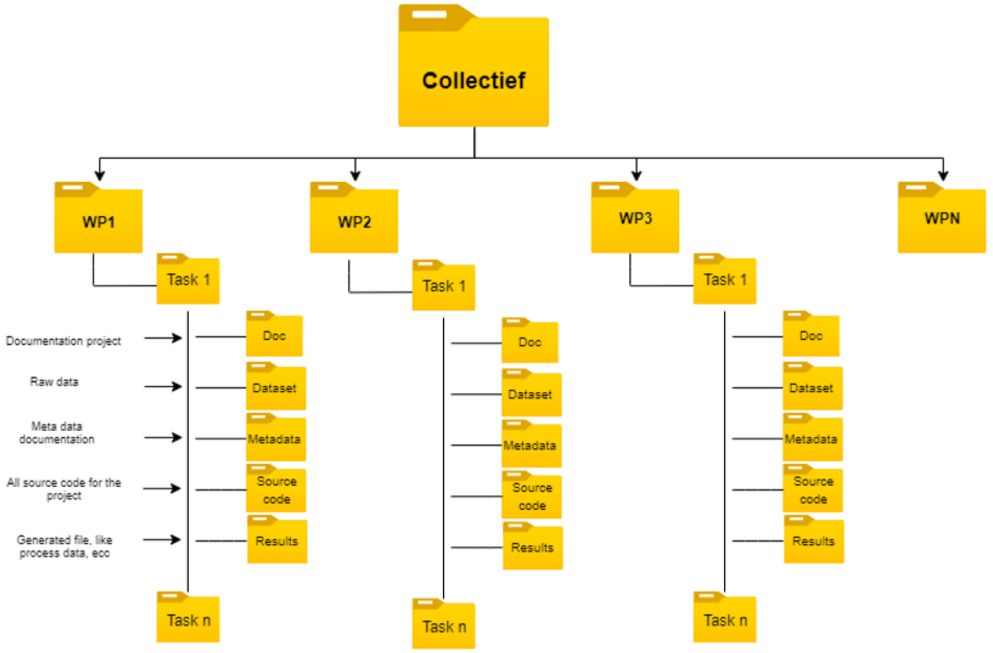
Individual file names will contain version numbers that will be incremented with each revision. Version control mechanisms are implemented within the DataverseNO repository to keep track of any changes to metadata or files (e.g., uploading a new file, changing file metadata, adding or changing metadata) once the dataset is published. This means that a new version of this dataset will be created once the published dataset has been modified.

4.1.3 Strategy for naming files and folders (and document it in metadata)

The folders in the data repository will be and structured using a structure that consisting of the project name as root, the next level is followed by the WP name, within it as a sub-level are the tasks associated with the specific WP. While the folders related to source codes, datasets, research results and the various publications generated are nested as a sub-level of the Task folder.

[8] Ibidem
[9] Ibidem
[10] Ibidem

Figure 5 – File folder organization



- **Doc's folder:** for text documents and additional documentation that will be associated with the project.
- **Dataset's folder:** for the collection of raw data (which should only be ready, do not edit the raw data directly).
- **Metadata folder:** for the collection of metadata documentation
- **Source code folder:** for collecting the sources of the project's scripts and programs.
- **Results folder:** for the collection of files generated during cleaning and analysis (such as processed data or visualizations)

File name organization: for the names of metadata, datasets, source code, templates for presenting results, a naming convention will be defined consisting of the following sections, explained below:

1) A prefix, indicating the type of publication, whether it is metadata, dataset, source code, template for presentation of results
2) A root consisting of:
   - The location where it was produced (Norway "NO", Italy "IT", Cyprus "CY")
   - The acronym/short name of the organisation that produced the data
   - The short and meaningful name of the dataset/template (e.g., Energy consumption "EC", Indoor physical environment "IPE", etc.)
   - Type of measurements (EC_ thermal, VS_ air_flow, etc.).
3) A suffix indicating the date of the last upload in the format YYYYMMDD and the corresponding version in the Repository and the file extension

Each of these elements is separated by an underscore: _

Example of name conventions:

[Type]_ [Location]_ [partner name]_[dataset name]_[measurements]_ [date]_[version]_[.ext]

[Dataset]_[NO1]_[CETMA]_[EC]_[thermal]_[2021_09_20]_[v01].[csv]

4.2 Making data openly accessible

To ensure that data is accessible (A) the following issues must be specified:

According to **the first principle (A1) and their sub-priciples (A1.1, A1.2)** to make data 'accessible', metadata must be retrievable from its identifier using a standardised communication protocol, i.e., a system of rules allowing the transmission of information between communication systems. To meet this requirement, the data and metadata stored in the DataverseNO repository are accessible through a number of protocols, including Hypertext Transfer Protocol (HTTP), rsync over Secure Shell (SSH), and Representational state transfer (REST) via Application programming interface (API), which provides access through, for example, cURL.

HTTP is the default protocol used when users access metadata in DataverseNO, while access via the other protocols must be clarified in advance with repository management. Accessible data uses a protocol that enables an authentication and authorisation process. DataverseNO provides machine-accessible public interfaces for searching data, accessing metadata and downloading data files, using a token to grant access when data files are restricted. For API access to data, the repository supports both session-based and API key-based authentication. Access to files in DataverseNO can be restricted to an embargo period[11].

According to **the second principle (A2),** to make data 'accessible', **metadata is accessible even when the data for some reason is no longer available.** To meet this requirement, by default datasets - including the files they contain - cannot be deleted in DataverseNO. However, there may be situations where data has been published that should not be openly available. In these cases, files in a dataset can be deaccredited, which means that access to these files is removed. Deaccessioning does not affect the dataset's citation metadata; therefore, the data is still findable and citable. After deaccessioning, the metadata includes information about why the data is no longer available[12].

4.2.1 Specify what methods or software tools are needed to access the data

Published datasets are discoverable and openly available to anyone with Internet access. Items are archived in formats that can be opened and read using freely available software and highly open data formats.

4.2.2 Time of conservation of data (public and not)

Public data will be openly accessible for at least 10 years after DOI assigned following the guarantee provided by DataverseNO.

4.2.3 How access will be provided in case there are any restrictions on use of data

In case there are restrictions on use of data**,** roles can be assigned to users that define the specific actions users can perform on data sets and/or files. Restrictions may differ depending on the users and the predefined rules, but generally relate to secure (locked) data storage and password-protected computers and prohibit the storage of data on the hard disks of computers that can be accessed

[11] Ivi
[12] Ibidem

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

through a network connection. These agreements may also limit the types of analysis that are done by the user.

4.2.4 Data will be made openly available

Details of which data will be made openly available, and the reasons for any data being kept closed, will be provided in future updates of the data management plan.

4.3 Making data interoperable

To ensure that data is interoperable (I) the following issues must be specified:

According to **the first principle (I1),** to make data 'interoperable', **metadata must use a formal, accessible, shared, and widely applicable language for knowledge representation.** To meet this requirement, the metadata of each record stored in the repository is represented internally in the JSON-LD for Schema.org or DDI (XML) schema as a rich schema to support extended variable metadata. At a general level, this is implemented in Dataverse with Linked Data support through JSON-LD for Schema.org, which means that the general metadata of a dataset and its files are represented in a format that allows the information to be searched along with other Linked Data supported data. Although available in DataverseNO, currently, interoperability can work well for general attributes that are provided through automatically encoded metadata in files (e.g., file type information), but not for more content- and domain-specific attributes[13].

According to **the second principle (I2)**, to make data 'interoperable', **the use of metadata vocabularies that follow the FAIR principles is necessary**. To meet this requirement, FAIR-controlled vocabularies and data models can be implemented manually in the selected repository, e.g., they can be implemented as keywords in the general metadata section. FAIR-controlled vocabularies can also be implemented in the DataverseNO repository through custom metadata schemas or be specified as precompiled or suggested values in metadata templates (which in Dataverse are called dataset templates). The repository identified in this proposal has made use of the latter approach, particularly in cases where a single project produces several datasets with related content, so that vocabulary values are largely common for all datasets (e.g., datasets covering time series). By default, however, Dataverse does not yet support controlled vocabularies and complex ontologies[14].

According to **the third principle (I3)**, to make data 'interoperable'**, metadata must include qualified references to other (meta)data**. To fulfil these requirements, the DDI schema supports references to other data. Such references can be added in two fields of the general metadata schema in Dataverse, one for related data and another for related materials (other research objects). Currently, information can only be entered as free text in these two fields, and the information is not exported to DataCite[15].

[13] Ivi
[14] Ibidem
[15] Ibidem

4.4 Increase data re-use (through clarifying licenses)

All research results are shared as openly as possible to maximize the use and reuse of research results. Reusable metadata is defined in the FAIR Data Principles as metadata that is richly described with a plurality of accurate attributes and relevant attributes.

According to the principle of re-usability (R1) and its sub-principles (R1.1, R.1.2, R.1.3), to make data 'reusable', **metadata must be released with a clear and accessible licence to use the data** (sub-principle R1.1). To meet this first requirement, DataverseNO offers the possibility to define a license when depositing the dataset and allows to display the license information in the descriptive metadata. The use of Creative Commons licenses (CC0 or CC BY) is recommended as they are considered the best to facilitate the reuse of datasets. The DataverseNO repository uses CC0 as the default license. However, it is possible to define a license other than CC0 by defining terms of use when depositing the data. Currently, licenses other than CC0 1.0 Universal (CC0 1.0) Public Domain Dedication are not predefined, and by default are not machine-readable.

According to the data reusability **sub-principle (R1.2)**, **metadata must be associated with a detailed provenance** (origin of the data, how it was obtained, processed and by whom). To fulfil this requirement, the repository has support for rich metadata that includes information about data authors and other contributors, data providers, data distributors, as well as related data (e.g., used as input data).

According to the sub-principle of data reusability (R1.3), **reusable metadata must meet community standards relevant to the domain**. To meet this requirement, different metadata standards are provided in Dataverse within the different domains of science. In the identified repository a few are supported (social sciences, life sciences, astronomy and physics). In addition to the metadata schemes provided by default in Dataverse, domain-relevant community standards can be implemented by creating custom metadata blocks. However, this option is not currently used in DataverseNO, as it is preferable to follow standardised domain approaches whenever possible.

4.4.1 Data embargo period

All data sharing and publication will respect international, European and national privacy laws, as well as the commercial interests and intellectual property rights of the project partners, which may lead to withdrawal from publication or embargo periods on some data produced by the project. Such decisions will be explained better in the next update of Data Management Plan.

4.4.2 How long data will be available for reuse once they are shared and quality assurance processes in place.

Similarly, the detail about how long data will be available for reuse once they are shared and the description of any quality data assurance processes, will be provided in future updates of the data management plan.

5. Allocation of resources

5.1 Resources and costs

Cost for data storage and backup: during the project, all data will be stored on secure, dedicated institutional servers provided by NTNU (HUNT Cloud) where they can only be accessed by project partners (budget is allocated for storage). GDPR rules will be respected and regular back-ups will also be made to ensure that data is not lost in case of technical problems with the server, accidental deletions and/or overwrites. Storage and backup will make use of servers provided by the HUNT Cloud - NTNU institution, through which data will be automatically replicated three times on three physical storage machines to protect against equipment failure. The cost of creating and maintaining the repository is included in standard indirect costs or overheads. Service costs are available at this link[16]. They vary according to a number of criteria (space required, storage time, desired security levels, transfer mechanisms, etc.) and according to 3 different service subscription profiles (white, orange, blue). These costs will be better defined in future updates of the data management plan.

Cost for data sharing: the costs of producing scientific publications, hosting a project website and depositing open access data are included in the COLLECTiEF budget as eligible costs. Partners have allocated a budget to cover publication costs for either 'gold' or 'green' access publications. Preference will be given to 'gold' open access schemes (i.e., free publication for readers). DataverseNO will be used as the FAIR repository of data for the project, GitLab will be used as the repository of confidential Data, only accessible to project partners (source code and documentation), and the website for publishing dissemination activities. No additional costs for data sharing are foreseen beyond those already indicated in the budget as eligible costs.

Cost for data transfer, access and security: security can be organised by the institutional IT services provided by HUNT Cloud. The service provides continuous monitoring of an extensive list of system activity, including access attempts and network activity. Individual logs are carefully maintained in accordance with their purpose of collection to enable retrospective audit capabilities. Consideration should be given to purchasing extra functionality to encrypt/anonymise confidential data prior to transfer (e.g., tape archiving), and/or to archive data with long-term retention requirements as required by regional ethics committees. Data anonymisation will be done on data primarily for structured data and not for unstructured data. The principles of avoiding or minimising the transmission of personal data will be followed in line with applicable legislation. Only data necessary for the research objectives of the project will be collected. No other sensitive information will be collected or stored. Therefore, a low cost is expected as many identifiers are excluded a priori from the data files. As far as transfer processes between "internal" user partners in the database are concerned, they can transfer data between HUNT Cloud and their local computers. This type of transfer is protected in encrypted communication channels (VPN). Simple and secure solutions are also provided for transfers to and from project collaborators outside the HUNT Cloud. The choice and cost of the solution for data transfer, access and security depend on the type of data, the expected duration of the transfer and the technical capabilities of the external party. These costs will be better defined in future updates of the data management plan.

The resources, costs, potential value, associated with long-term preservation strategies, as well as how the data will be preserved beyond the project and for how long, will be the subject of discussion in the forthcoming General Assemblies of the Consortium. The objective of COLLECTiEF is to align with the long-term preservation of the data described in the present management plan.

[16] NTNU, HUNT *Cloud Price List*. Available at: https://assets.hdc.ntnu.no/assets/prices/hunt-cloud-price-list-3-1.pdf [Accessed 22 November 2021].

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

6. Data access and security

6.1 Storage, Backup and security methods and procedure

A fundamental task is to manage the data in a secure way. COLLECTiEF will promote the use of Engineering best practices and state-of-the-art data security measures. The goal of these measures will be to ensure that data remains consistent over the lifetime of the project and there exist alternatives to the main files in case they disappear or get corrupted.

All ICT systems will be designed to safeguard collected data against unauthorized use and to comply with all national and EU regulations. An encryption component will add an extra layer of security to the data files and information.

All the generated data will be managed, processed, and stored in a secure environment (lockable computer systems with passwords, firewall system in place, power surge protection, virus/malicious intruder protection) and by controlling access to digital files with password protection.

In more details, the following mechanisms will be implemented:

- **Access Control**: the access to the system will be allowed only after controlling the level of access that each user has depending on their role. There will be appropriate mechanisms to define and enforce such access control (e.g., firewalls, file systems permissions, secure log-in) including physical control.

- **Data Confidentiality**: within the scope of the project, the protection of information from unauthorized access and disclosure must be preserved by restricting per-user access and encrypting the information during transmission and during storage. After the defined retention period expires, ensure information erasure/destruction.

- **Data preservation**: data backup and maintaining techniques will be used to assure long-term value and integrity of data.

- **Data backups** are expected to occur once a week, though this is subject to change based on the amount of data that will ultimately need to be backed up.

- **Data anonymization techniques** (such as data masking, pseudonymization or data swapping) will be used as well during the project lifetime.

7. Ethical aspects

7.1 Ethical aspects considered

Specifically, the ethical issues identified and highlighted within the COLLECTiEF project focus on two topics and are as follows:

- the collection, collation, and processing of data from demonstration buildings (to assess performance)
- the personal data of participants in dissemination events and workshops

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

In COLLECTiEF, the data to be collected focus on energy performance and indoor environmental quality and, as such, do not fall into the category of sensitive data. As for the data generated by sensors and/or home building equipment, although these are not sensitive measurements, with appropriate analysis (e.g., monitoring changes in C02, relative humidity or light), one could track the presence/absence of people in the monitored environments. Again, from energy consumption data with a high temporal resolution, personal daily routines could be identified, when collected at the level of a single household. Even more significant is the data generated by the POE questionnaire. The data that are requested are some related to the identification of some characteristic parameters of the respondent and information related to his/her thermal (but also visual) perception or reaction. Then there is a part of sensitive and personal information and other data related to biometric measures (weight, height, BMI).

In this sense, security and privacy issues are raised from two points of view:

- Data storage and processing (minimisation, anonymisation, encryption, aggregation).
- Related to the misuse of data or data theft (unauthorised access and/or databreach).

The consortium is aware of the potential risk of misuse of the technology and will therefore mitigate these risks from the start of the project by applying appropriate measures for data protection and security during the project and beyond. This responsibility (see section 9 - Responsibilities for more details) falls on the Collective's "controller" and related processors, who must:

- Obtain formal consent for the storage, sharing and re-use of data.
- Protect the identity of participants through minimisation, anonymisation, encryption and aggregation techniques.
- Ensure that data is stored, transferred and managed securely.

Informed consent: people's participation in the questionnaire will be on a voluntary basis, and a form will be prepared to obtain their informed consent prior to their involvement in the COLLECTiEF project. The informed consent form and its associated information sheets will be drafted using simple language and fully understandable, so that the potential participant has fully understood the information. The informed consent form will describe the objectives, methods and implications of the research. But also, the nature of the participation, the amount and nature of the data stored, the benefits, the risks that may arise, and finally the type of disclosure of the results that will be made.

Within the consent form, potential participants will explicitly state that their participation in the questionnaire is voluntary, and clauses will be included to guarantee anyone the right to withdraw/refuse their participation, at any time, and without any kind of consequences.
Finally, it will be made explicit what procedures will be put in place in case of risk, such as unexpected or accidental discovery of sensitive information by unauthorised parties**.** The informed consent procedures that will be implemented for the participation of humans will be submitted as part of deliverable D5.1: Performance Measurement & Verification Protocol – Concepts and methods for performance evaluation of COLLECTiEF solutions (due date: M12).

Protecting the identity of participants: the principles of avoiding or minimising the transmission of personal data will be followed in line with applicable legislation. Only data necessary for the research objectives of the project will be collected. No additional sensitive information will be collected or stored. In addition to this option, data collected in the project will be anonymised and aggregated as close to the most granular level as possible. Aggregation of data, e.g., at a less granular temporal resolution (once a day) or at a higher geographical resolution (e.g., energy consumption at cluster/district level) mitigates the risk of individual person/building identifiability. The same applies to data anonymisation, which will be ensured by separating identifiable data from anonymised data.

Data transmission, storage and retention: the partnership, within the scope of its competencies and the IT infrastructure at its disposal, will ensure the secure storage, delivery and access of personal information, as well as the management of users' rights. In this way, it can be ensured that the content accessed, delivered, stored and transmitted will be managed by authorised persons with clearly defined rights/obligations. In this sense, state-of-the-art firewalls, network security, encryption and authentication will be used to protect the collected data (specific details will be developed later in the project, within WP3 and 4).

Collected data will be stored on a secure server, accessible only to the COLLECTiEF partnership network. Network traffic intrusion detection systems will monitor anomalies and, if necessary, activate restriction mechanisms. This will be combined with a controlled access mechanism and, with respect to data transmission, efficient encryption and coding mechanisms. Data security will cover storage, encryption and transmission procedures for personal data in line with national and European data protection legislation.

Anonymized and identifiable data will be stored separately, and only authorized persons for the project will have access to the stored data. Anonymized data will be available to all project partners, while sensitive data will be available to partners directly involved as "processors" for the purposes of this project. In case some sensitive data are requested for research purposes by partners/researchers, part of the partnership, but not directly involved in the processing of personal data, access and distribution of the same will be granted only after the explicit permission of the "controller".

Researchers will have access rights to read and add datasets to the database. No editing rights will be granted to them, so as to minimize the risk of unauthorized alteration or dissemination of sensitive data. As an additional assurance mechanism, researchers handling personal and sensitive data will be asked to sign a statement committing them to ensure that project data are not provided to persons outside the project consortium.

8. Other

National strategy on access to and sharing of research data[17] and NTNU’s policy for open research data 2018-2025[18] are well in line with open access and data management guidelines in Horizon 2020[19].

[17] Government.no., *National strategy on access to and sharing of research data*. Available at: https://www.regjeringen.no/en/dokumenter/national-strategy-on-access-to-and-sharing-of-research-data/id2582412/sec1 [Accessed 22 November 2021].

[18] NTNU., *NTNU’s policy for open research data 2018-2025*. Available at: https://innsida.ntnu.no/documents/portlet_file_entry/10157/NTNU+Open+Data_Policy.pdf/42f1ed94-4d4f-4d6b-a033-dd42a02ccefc?status=0 [Accessed 22 November 2021].

[19] European Commission., Data management - H2020 Online Manual. Available at: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [Accessed 22 November 2021].

9. Responsibilities

9.1 roles and responsibilities in the project

Within this section of the DMP, data management responsibilities are assigned. In general, the project partners assume responsibility for collecting, managing, and sharing research data in the project in accordance with the data use requirements defined by the consortium agreement. The project coordinator (NTNU) will act as the data manager for the project administrative data. The CETMA partner will act as the overall data manager and monitor compliance with the data management plan.

Data controller: the project consortium has identified and appointed a data controller. CETMA as data controller, will be responsible for complying with obligations under both national and international data protection legislation.

Data processor: according to the General Data Protection Regulation 2016/679[20] data controllers and processors are fully responsible for processing operations, which means that each beneficiary is ultimately responsible for their own data collection and processing. Project partners are required to follow internal national data protection regulations and the European GDPR.

In line with the principles of the GDPR, **project partners are responsible for the data they produce (Data processors)**, which means **they must coordinate with the data controller** to ensure that procedures and protocols are adhered to with internal processes and national regulations (including how to obtain consent and procedures that must be put in place in the event of privacy breaches).

The following responsibilities apply to datasets on an individual basis:

Acquisition/capture of (raw) data responsibility

- **For pre-existing datasets owned by any of the COLLECTiEF partners or other parties,** the partner who owns that dataset and makes it available to the project (unless otherwise stipulated in section 25 GA), will be the data manager and will coordinate with the with the data controller to follow established means and purposes for processing.
- **For the creation or acquisition of new datasets,** the partner(s) creating or acquiring the dataset will be the data controller(s) and will coordinate with the with the data controller to follow established means and purposes for processing. This applies to all types of data research.

Metadata production responsibility: each partner is responsible for his own metadata, which will be agreed upon with the data controller according to the predefined and available standards.

Data sharing responsibility: the data processor will coordinate with the data controller to agree and establish the details of how the data will be shared, including access procedures, embargo periods (if any), guidelines for technical mechanisms for dissemination, and necessary software and other tools

[20] European Parliament and Council. *Regulation (EU) 2016/679*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [Accessed 22 November 2021].

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement No 101033683

to enable reuse, while also determining whether access will be broadly open or limited to specific groups. Similarly, it will identify the repository in which the data will be archived specifically indicating the type of repository.

Data archiving and retention (including storage and backup) Responsibility: depending on each data set, the data archiving and retention procedures that will be put in place for long-term data retention will be the responsibility of the data controller, after agreement with the partners involved. This includes stating how long the data is to be retained, what its approximate final volume is, what the associated costs are, and how these are expected to be covered.

Quality assurance of data responsibility: data and information that comply with data quality standards are critical to the success of the COLLECTiEF project. As data processors, project partners are responsible for the data they produce, and therefore also preside over specific procedures to ensure that data and information meet quality standards.

Data security responsibility: partners, as data processors, will pay special attention to routines to ensure the confidentiality of data storage and processing. In coordination with the Data Controller, they undertake to implement all appropriate technical and organisational measures necessary to protect potential personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorised disclosure or access, and against all other unlawful forms of processing, taking into account the particular nature of the processing operations carried out.

10. References

Conzett, P., DataverseNO: *A National, Generic Repository and its Contribution to the Increased FAIRness of Data from the Long Tail of Research*. Available at: https://munin.uit.no/bitstream/handle/10037/18564/article.pdf?sequence=3&isAllowed=y [Accessed 22 November 2021].

European Commission., Data management - H2020 Online Manual. Available at: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [Accessed 22 November 2021].

EUROPEAN PARLIAMENT AND COUNCIL. *Directive 95/46/EC*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=EN [Accessed 22 November 2021].

EUROPEAN PARLIAMENT AND COUNCIL. *REGULATION (EU) 2016/679*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [Accessed 22 November 2021].

Government.no., *National strategy on access to and sharing of research data*. Available at: https://www.regjeringen.no/en/dokumenter/national-strategy-on-access-to-and-sharing-of-research-data/id2582412/sec1 [Accessed 22 November 2021].

NTNU, HUNT Cloud Price List. Available at: https://assets.hdc.ntnu.no/assets/prices/hunt-cloud-price-list-3-1.pdf [Accessed 22 November 2021].

NTNU, HUNT Cloud service: *General information on cloud services*. Available at: https://www.ntnu.edu/mh/huntcloud [Accessed 22 November 2021].

NTNU., *NTNU's policy for open research data 2018-2025*. Available at: https://innsida.ntnu.no/documents/portlet_file_entry/10157/NTNU+Open+Data_Policy.pdf/42f1ed94-4d4f-4d6b-a033-dd42a02ccefc?status=0 [Accessed 22 November 2021].

OpenAIRE, Costing Tools for Data Management Plan, Available at: openaire.eu/estimating-costs-rdm-tool [Accessed 22 November 2021].

Wilkinson et al, M., The FAIR Guiding Principles for scientific data management and stewardship. Available at: https://www.nature.com/articles/sdata201618.pdf [Accessed 22 November 2021].