Ref. Ares(2023)1659804 - 07/03/2023

H2020 – LC – SC3 – EE – 2019 – GA 894240

Operating System for Smart Services in Buildings



D3.4 Rules and Methods for Data Anonymisation and GDPR Compliance

WP3 Common Ontology and Semantics

	Name	Date
Prepared by	Christian Thomsen (AAU)	15.02.2023
Peer reviewed by	Yves Stauffer (CSEM), Dominique Gabioud (HES-SO)	20.02.2023
Reviewed and approved by	Dominique Gabioud (HES-SO)	02.03.2023



This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 894240.

Distribution list

External		Internal	
European Commission	x	Consortium partners	x

Change log

Version	Date	Remark / changes
1.1	15.02.2023	First version by AAU
2.1	27.02.2023	Second version by AAU
3.1	02.03.2023	Final version by HES-SO

To be cited as

"D3.4 Rules and Methods for Data Anonymisation and GDPR Compliance" of the HORIZON 2020 project domOS, EC Grant Agreement No 894240.

Disclaimer

This document's contents are not intended to replace the consultation of any applicable legal sources or the necessary advice of a legal expert, where appropriate. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user, therefore, uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view.

Table of contents

Executive Summary **4**

1. Introduction **4**

2. Overview of Available Data Sets from Demonstrators **5**

2.1. Paris Demonstrator 5

2.2. Neuchâtel Demonstrator 5

2.3. Sion Demonstrator 5

2.4. Aalborg and Skive Demonstrators 5

2.5. Summary of Available Data Sets 6

3. Preparation Process **6**

3.1. Provide Building Descriptions 6

3.2. Pseudonymise Data 7

3.3. Ad-Hoc Cleanse the Data 7

3.4. Provide Metadata 8

3.5. Define Terms for Use 9

3.6. Define Process for Requesting Access to Data 9

4. Publication on an Open Data Repository **9**

5. FAIRness of Data **10**

6. Conclusion **10**

Appendix A: Template for Declaration of Disclosure of Personal Data **11**

Terms, definitions, and abbreviated terms

GDPR	General Data Protection Regulation
FAIR Principles	Findability, Accessibility, Interoperability, and Reusability Principles
UUID	Universally Unique IDentifier

Executive Summary

This document presents the data sets collected at the five demonstrators in the domOS project. The consortium wishes to follow the FAIR data principles and share data. The collected data, however, contains personal data and cannot be shared as open data. It is not feasible to anonymise any of the data sets, but three of them can be pseudonymised and shared for research purposes. The document also presents the steps needed to do that in way where the FAIR principles are met.

1. Introduction

In the domOS project, smart building technologies are demonstrated in five sites, namely: Paris (F), Neuchâtel (CH), Sion (CH), Aalborg (DK), and Skive (DK). In the demonstrators, smart services are designed, deployed, and evaluated. As part of this, data is collected from sensors and meters on the buildings. In addition, metadata describing the sensors, buildings, etc. is created for each demonstrator.

The data sets are of vital importance to the demonstrators to operate and evaluate their smart services. The data sets can, however, also be useful to the research community, e.g., to evaluate or simulate future smart services. According to the GDPR, "*any information relating to an identified or identifiable natural person*" is personal data. Since the measurements in the data sets are from specific households, the data sets thus contain personal data. As presented in "D1.2 Data Management Plan", the domOS project follows the principle "*as open as possible, as closed as necessary*". The raw data sets cannot be publicly shared when they contain personal data. In some circumstances it is, however, possible to share (derivatives of) the data sets. In particular, anonymisation and/or pseudonymisation should be applied. "D1.2 Data Management Plan" also foresees that anonymisation of the data sets will not be possible as it would require a stop of the operation of the smart services using historical data and erasing anything that can allow the identification of individual participants. This would also include links to the energy topology and addresses.

It should be noted that the data collection is done by the Demonstration Leader at each demonstrator. It is also the Demonstration Leader who is responsible for consent forms from the data subjects living in the buildings. The Demonstration Leader thus controls and possesses the data sets for a given demonstrator and ultimately decides whether the data sets can be shared.

The current document presents which data sets are available from the demonstrators and which of them can be shared. In addition, the needed steps to prepare the data sets for sharing are presented. These steps will be taken in the preparation of sharable data sets for the forthcoming "D3.6 Reference Data Sets".

The rest of this document is structured as follows. Section 2 presents the available data sets from the demonstrators and explains what can be shared. Section 3 presents the process to be followed to prepare for sharing of a given data set. Section 4 describes how and where metadata about the collected data sets should be publicly published. Section 5 summarises how the FAIR principles are followed by the approach outlined in the document. Finally, Section 6 concludes.

2. Overview of Available Data Sets from Demonstrators

2.1. Paris Demonstrator

The Paris Demonstrator has data about 10 buildings and data from responses to a questionnaire about physical properties of the buildings. According to the Demonstration Leader, only the latter questionnaire responses can be made available as open data. It is, however, deemed by the domOS project that this data is not useful for others without the possibility of combining it with actual measurements of energy consumption and indoor climate, i.e., sensor data. Thus, it has been decided not to provide open access to data from the Paris demonstrator.

2.2. Neuchâtel Demonstrator

The Neuchâtel Demonstrator has data from a single house with multiple apartments. The data includes measured values from a heat generator as well as values from smart thermostatic valves in the apartments. The Demonstrator Leader has concluded that complete anonymization of the Neuchâtel data is not possible and given the consent forms and information sheets agreed and signed by the relevant data subjects, it is then not possible to share the data with external parties. There will thus be no open data from the Neuchâtel Demonstrator.

2.3. Sion Demonstrator

The Sion Demonstrator has data from around 100 private residential houses. The data contains time series about electricity consumption from smart meters as well as sub-metering data for energy processes. In addition, metadata about the houses, including, e.g., size, type, and number of inhabitants, is available. As the data is related to identified or identifiable households, it is considered personal data.

As described earlier in the document “D1.2 Data Management Plan”, anonymising such data would have negative impacts on its usefulness and on the service operator. The data from the Sion Demonstration will thus follow the principle described in “D1.2 Data Management Plan” and be pseudonymised. As the data contains personal data and becomes pseudonymised rather than anonymised, the resulting data set is still considered to contain personal data. Following the principles of “D1.2 Data Management Plan”, the data cannot be made available on open data repositories. Instead, it will be stored locally at the Demonstrator Leader (i.e., OIKEN) from where external parties can request access for research purposes. Such requests will be evaluated carefully by the Demonstrator Leader, and if granted, a formal agreement will specify the terms for access and use of the data. Consent forms allowing this have been signed by the relevant data subjects.

OIKEN has started the collection of the data and it is expected that the data can be shared with domOS partners from end of 1st quarter in 2023. When the steps in Section 3 have been completed, the data can also be made available to other members of the research community.

2.4. Aalborg and Skive Demonstrators

NEOGRID is now Demonstrator Leader for both the Aalborg Demonstrator and the Skive Demonstrator. The two demonstrators are equipped with the same sensors, and thus generate similar data sets. They

also use the same consent form. The two demonstrators are therefore described together in this section. The Aalborg Demonstrator consists of 15 buildings of which three are single family houses and the remaining buildings have 152 apartments in total. The Skive Demonstrator consists of 12 single family houses. For both demonstrators, the available data in addition to building metadata (location, building type, …) includes measurements from indoor climate sensors and energy meters. As is the case for the Sion Demonstrator, the data is then related to identifiable or identified households and it is considered personal data.

In addition, the data from the Aalborg and Skive Demonstrators cannot be anonymised without negative impacts on its usefulness and the service operator which needs to keep the operational data (see "D1.2 Data Management Plan"). Instead, the data will be pseudonymised such that actual addresses are replaced with an integer ID not carrying any special meaning. The data sets will, however, still contain personal data and can thus not be published in an Open Data Repository. Instead, the Demonstrator Leader (i.e., NEOGRID) will store the data and external parties can request access for research purposes from NEOGRID. Consent forms allowing this have been signed by the relevant data subjects.

NEOGRID already collects data from the two demonstrators, and it can be shared with domOS partners immediately. When the steps in Section 3 have been completed, the data can also be made available to other members of the research community.

2.5. Summary of Available Data Sets

Three of the demonstrators (i.e., Sion, Skive, and Aalborg) have data that can be shared with the research community. The data sets, however, contain personal data and cannot anonymised. All of them will instead be pseudonymised. In the remainder of this document, we focus on the steps needed to make these three data sets sharable.

3. Preparation Process

In this section, we describe in more details the process to be carried out for the data sets that demonstrators make available to the research community.

3.1. Provide Building Descriptions

For all three demonstrators (i.e., Sion, Aalborg, and Skive), the data sets contain sensor measurements from several buildings. It is paramount for practical use to have information about the buildings. To ensure this, Building Descriptions (BDs) for the considered buildings must be provided by the demonstrators. As described in "D3.5 domOS Common Ontology, Final Version" *A Building Description (BD) is a machine-readable description that provides an abstract description of a building and its properties as needed by the smart services (or applications) controlling (the energy of) the building. It consists of a description of building level - specific metadata (e.g., total space, thermal capacity, envelope type), a description of the structure of the building (e.g., storeys, apartments, rooms), the energy/electricity/heat flow between spaces and devices in the building, the measurements and control points in that building and links to the things and services installed in the building. Specifically, The Building Description is a JSON-LD file based on the domOS Common Ontology.*

Before sharing data sets, Demonstration Leaders must thus ensure that valid Building Descriptions exist. It is a currently on-going task to create Building Descriptions for both the Sion and Aalborg/Skive Demonstrators and thus this requirement is expected to become fulfilled in the near future.

3.2. Pseudonymise Data

As described in Section 2, the data sets identified by the demonstrators for possible sharing cannot be anonymised and will instead be pseudonymised. This means that personally identifiable information (in particular, addresses) become replaced by artificial identifiers, so called “pseudonyms”. Some attributes with personally identifiable information may instead be completely removed, but only in a way where it still is possible to distinguish different buildings from each other. An address should thus typically not be removed, but rather replaced by a pseudonym. Using pseudonyms and possibly removing some attributes makes it more difficult – but not impossible – for others to see exactly where the data originates from. The mapping from actual addresses to pseudonyms is, however, known by the Demonstration Leader and may even be guessed by others and it is important to stress that the data thus not becomes anonymised by this.

The general principle is thus that before sharing a data set, the Demonstration Leader carefully ensures that attributes with personally identifiable information are replaced by pseudonyms.

For the Aalborg/Skive data sets, only the addresses must be replaced. They are currently replaced by an integer *locationID* which does not carry any meaning and only serves to make it possible to differentiate between different locations/addresses. The same *locationID* is used for the same building in both the sensor data and the Building Descriptions.

For the Sion data set, it is currently also expected that only addresses must be replaced by pseudonyms. It is expected that the addresses will be replaced both in the sensor data and the Building Descriptions by the pseudorandom UUIDs already used internally as primary keys for buildings.

3.3. Ad-Hoc Cleanse the Data

Some ad-hoc cleansing of the data sets may or may not be needed. For example, unnecessary attributes can be removed, duplicates can be removed, and illegal/impossible values can be identified and modified during cleansing.

Data from the Aalborg/Skive Demonstrators is already shared within the project and is not likely to need much cleansing before it can be shared with others. The Demonstrator Leader for both the Aalborg and Skive Demonstrators, i.e., NEOGRID, currently collects the data in plain JSON format. The data can be shared as-is in this format. Alternatively, it is possible to convert the data into RDF triples and serialize them in JSON-LD or Turtle format. By doing this, it is possible to use the domOS Common Ontology to explicitly express what the data represents. (Note that the Building Description exists and provides semantic information about what the sensor measurements represent no matter if the sensor data is represented in RDF or not).

Data from the Sion Demonstrator is expected to be ready for internal sharing in the project by end of 1st Quarter 2023. By then it will be evaluated whether the data needs to be cleansed and/or converted into another format.

3.4. Provide Metadata

The data sets should be described such that other parties can understand what the data represents. Note that this is different from the *semantic metadata* contained in the Building Description. There should thus, in addition the Building Description, be a *textual description* of the background:

- Where the data was collected
  - This should explain background information about the data, e.g., that the data was collected in a given number of apartments heated with district heating.
  - This should not necessarily include the exact address, but could instead, e.g., provide the area or city.
- When the data was collected
  - This should explain the time range covered by the data, e.g., from 2023-01-01, and how often data was sampled, e.g., every 100 ms.
- How the data was collected
  - This should explain how the data set's values were obtained, e.g., from a sensor of a certain type mounted in the different rooms.
- Any other information needed to understand the data
  - This includes a description of the filenames in the data set.

In addition to the background description, there should also be a description of *each attribute* including:

- The name of the attribute
- The datatype of the attribute
- The meaning of the value of the attribute
- The unit of the attribute
- The possible range of values for the attribute

Finally, the format(s) the data set is made available in should be described. This should be an open format, for example JSON, CSV, or RDF.

The metadata descriptions presented here (i.e., the metadata describing the background and each attribute, but not the Building Description) should be gathered in a single document for each demonstrator. The Demonstration Leader is responsible for doing that. The file should be provided in a format which can be read without requiring access to paid software. This can, for example, be plain text or XHTML, or even Word format which can be read with free alternatives such as LibreOffice and Google Docs.

3.5. Define Terms for Use

For open data, several licenses, such as the Creative Common licenses, exist and can be easily applied to define terms of use. As the data from the Aalborg, Skive, and Sion Demonstrators cannot be shared as open data, these licenses, however, cannot be applied by domOS.

Before data can be shared with an external party, the Demonstration Leader must instead define how and for what the data can be used. As the data is not anonymised, it cannot become open data and the terms should disallow further sharing of the data. Further, the terms should (at least) specify exactly who will get access for what purpose for how long, what will happen to the data afterwards, and which security measures should be taken by the receiving party. Since the data sets contain personal data, a declaration of disclosure of personal data should be filled out before a data set is shared with an external party. A template for such a declaration is shown in Appendix A for inspiration. The demonstrators may have to use another declaration of disclosure of personal data to accommodate their needs. According to "D1.2 Data Management Plan", the Demonstration Leader is also Data Controller and further each Demonstrator has a Data Protection Officer (DPO). It is their responsibility to ensure that all legal requirements (including GDPR-related ones) are met and in case of any doubt, they should seek legal assistance.

3.6. Define Process for Requesting Access to Data

The Demonstration Leaders are responsible for the demonstrators' data. The Demonstration Leader should thus define procedures for how to evaluate requests for access to data.

4. Publication on an Open Data Repository

As they contain personal data, the data sets from the demonstrators should *not* be uploaded to external repositories. To follow the FAIR principles, the metadata descriptions (described in Section 3.4) should, however, be uploaded to an open repository. This makes it possible for interested parties to become aware that the data sets exist and find out how to contact the owner. It is also possible to cite the data set. Many repositories exist and, for example, from https://www.re3data.org/ it is possible to search for repositories. In the domOS project, it has been decided to publish the metadata descriptions on Zenodo (https://zenodo.org).

The Demonstration Leader should upload the metadata description as a *Dataset* on https://zenodo.org/deposit/new and fill out the fields under *Basic information*. The *Access right* under *License* should be set to *Restricted Access* and the conditions for access should be given as well as an email address to the responsible for evaluating requests for data access. Under *Funding*, domOS should be added.

Any other relevant fields should also be provided before the metadata is published. When the metadata about the data set is published, a Digital Object Identifier (DOI) is assigned such that it is possible to uniquely identify and cite the data set even though the data set itself is not openly accessible.

5. FAIRness of Data

By following the steps in Section 3 and Section 4, the data sets fulfil the FAIR principles:

F) The data sets **Findable** as they have a DOI from Zenodo and machine-readable metadata is available.

A) The data sets are **Accessible** as it is described how one can get access and there is access to metadata

I) The data sets are **Interoperable** as descriptions of the data are provided such that it is possible to understand what the data represents and can be used for, and the data is stored in an open format.

R) The data sets are **Reusable** as metadata descriptions are available and procedures for getting access to the data are established.

6. Conclusion

The five demonstrators in the domOS project all collect data sets. All the data sets contain personal data. For the Paris and Neuchâtel demonstrators, it was concluded that data from sensors and meters cannot be anonymised and that it is not feasible to share the data in pseudonymised form. For the Sion, Aalborg, and Skive demonstrators, the data cannot be anonymised as this would negatively affect the service operators and hinder the usefulness of the data. Instead, the data sets from these three demonstrators can be shared in pseudonymised form. This means that the data sets will not be published on open data repositories. To follow the FAIR principles, metadata including descriptions of the data sets will, however, be published on the open data repository Zenodo. It is then possible for researchers to find information about the data sets and request access to them directly from the Demonstration Leaders.

Appendix A: Template for Declaration of Disclosure of Personal Data

Details about disclosing party	
Name of data owner **Title**	
Name of company	
Title of project, study, survey, database or register.	
Purpose	State the purpose of the project to which you want to disclose data.
Include reference to a cooperation agreement, grant agreement or the like, if relevant.	
What type of personal data do you want to disclose?	
Where and how have the personal data been collected or received?	For example: Aggregated data from Statistics Denmark's Division of Research Services, personal data collected through interviews, questionnaires, tests etc.

Has a written consent been obtained?	☐ No ☐ Yes *Please attach a scanned copy of the declaration of consent.* Consent was obtained on [enter date]: _____
What form are the personal data in?	Examples: Video recordings, images, statistics, completed questionnaires etc.
What categories of data subjects do the personal data relate to?	For example: 'Women aged 35-50 years, living in northern Jutland who receive or have received pharmacological treatment against increased metabolism.'
How many data subjects do the personal data relate to?	
Have the data subjects been informed that the data will be disclosed to the Receiving Data Controller?	☐ Yes ☐ No, the data subject has been informed that only XXX will process the data or that the data will not be disclosed to other parties.
What is your reason for disclosing personal data?	The reason may, for example, be that the disclosure is necessary for a partner's research project.

Confirm that only necessary data are disclosed. You only have statutory authority to disclose necessary data.	☐ I confirm that only data *necessary* for the Receiving Data Controller will be disclosed.
Will the personal data be pseudonymised before being disclosed? *Pseudonymisation may be omitted if this is strictly necessary for the Receiving Data Controller's scientific or statistical investigations. You must give reasons for this here.* *The personal data must, in so far as is as possible, be disclosed in a form in which they can no longer be attributed to the data subject by the Receiving Data Controller without the use of additional information. For example, by pseudonymisation.*	☐ No. The personal data are not pseudonymised. *Explain why the personal data are not pseudonymised or anonymised prior to disclosure:* ☐ Yes. The personal data are pseudonymised. *Describe how the data have been/will be pseudonymised:* ☐ Yes. The personal data are pseudonymised, but the additional information is forwarded so that the Receiving Data Controller can identify the data subjects. *Give reasons for the forwarding of the information:*

- **Signatures**

Date:

Name:

_____

Signature

Date:

Name:

_____

Signature

This form must be filled in by the receiving research institution.

Information about the Receiving Data Controller	
Details about the research institution	• Name • Address • Postcode and town/city • Registration number • Telephone number • Email
Details about authorised person(s)	Authorisation may be through any formal authorisation system and procedure established by the Receiving Data Controller. The authorised person may, for example, be a project manager: • Name • Address • Postcode and town/city • Registration number • Telephone number • Email
Title of project, study, survey, database or register.	

What are the personal data to be used for?	☐ Research project, study, or survey ☐ Database or register ☐ Statistics ☐ Other:
Purpose	State the purpose for which the disclosed data will be used.
Why are the data necessary for the project?	Describe in detail why the personal data are necessary for the project/study. Enclosing a project description is unfortunately not sufficient.
When will the data processing be concluded?	Enter date.
Declaration	As the Receiving Data Controller, I hereby declare, by my signature on this form, that I accept (tick): ☐ That the data will be used solely for statistical or scientific purposes. ☐ That the data are necessary for the research in question, that no more data will be processed than what is necessary for the performance of the research and that data that are not necessary will be erased or returned as quickly as possible.

	☐ That the results will be disseminated in such a way that it is not possible to identify individual data subjects. ☐ That the data will only be processed by persons authorised by the Receiving Data Controller to have access to the data. In addition, these persons have been provided with the necessary general and specific instructions on how to handle the personal data received, including the instruction that the data may be processed solely for the purpose of scientific or statistical studies. ☐ That the personal data will be erased, anonymised, or destroyed on conclusion of the study, so that it will not subsequently be possible for anyone to identify individual data subjects who participate in the study. ☐That, in accordance with Article 32 of the General Data Protection Regulation, taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, appropriate technical and organisational measures will be implemented to ensure a level of security appropriate to those risks.

Signature

Name and title of the authorised signatory from the Receiving Data Controller:

_____

Date:

_____

Signature