FWO DMP Template

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

1. General Information	
Name applicant	**Emilio José Palacios-García**
FWO Project Number & Title	210280/12ZZV22N – Providing energy metering infrastructures with secure extended services (PREMISES)
Affiliation	☒ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other:
2. Data description	
Will you generate/collect new data and/or make use of existing data?	☒ Generate new data ☒ Reuse existing data

Describe the origin, type and format of the data (per dataset) and its (estimated) volume

*If you **reuse** existing data, specify the **source** of these data.*
*Distinguish data **types** (the kind of content) from data **formats** (the technical format).*

Table 1. Data that will be generated

Type of data	**Format**	**How would it be created?**	**Volume[1]**
Code	py, ipynb	Development of project methodology. Open-source visual studio code will be the editor of choice	~10 KB per file. The number of necessary classes, modules and scripts is hard to estimate at this stage
Machine learning models	pkl	Tunned machine learning models developed in Python will be serialised as pickle objects.	100 - 500 kB per file. The number of models is hard to estimate at this stage.
Model training and other statistical results	csv	Output of models training process. It will depend on the chosen evaluation metrics, features and model complexity.	< 100 kB per file is expected.
Laboratory test profiles	csv	Power measurements of laboratory appliances activation simulation	< 10 MB per file is expected in each experiment
(UML) diagrams	drawio, mmd, dot	Class, sequence, state, flow or communication diagrams based on the system architecture and functionalities.	< 100 kB per file. Number of files is hard to estimate at this stage.
Graphic results	pdf, png, tif	Any result generated from training, evaluating or profiling the models and system architecture	< 5 MB for pdf, png. < 50 MB for large high-resolution TIFFs. Number of figures is hard to estimate at this stage.

Deliverables	tex, pdf	Reports of the project tasks and work packages	~ 1 MB per group of files. 14 deliverables in total

Table 2. Data that will be reused

Nr	Name	Format	Description	Volume[1]	License
1	UK-DALE Domestic Appliance-Level Electricity	csv (dat), yaml (metadata)	5 households, for 4.3 years. Disaggregated power every 6 s. Aggregated voltage and current at 16 kHz	14.9 GB	CC BY 4.0
2	RAE - The Rainforest Automation Energy	csv	2 households, for 2 months. Sampled at 1 Hz.	15.1 GB	CC BY 4.0
3	REFIT - Personalised Retrofit Decision Support Tools for UK Home Using Smart Home Technology project	csv	20 households from 2013 to 2015. Sampled every 8 seconds.	12.1 GB	CC BY 4.0
4	ECO - Electricity Consumption & Occupancy	csv, matlab	6 households for 8 months. Sampled at 1 Hz.	3.8 GB	CC BY 4.0
5	AMPds2 - The Almanac of Minutely Power dataset	csv	1 households for 2 years. Sampled every minute.	2.1 GB	CC BY 4.0

6	DRED - Dutch Residential Energy Dataset	csv	1 household for 6 months. Sampled at 1 Hz.	2.8 GB	Not mentioned
7	GreenD - Energy Metering dataset	csv	7 households for 1 year. Sampled at 1 Hz	9.8 GB	Not mentioned
8	ACS-F2 - Database of appliance consumption signatures	xml, mat	Diverse appliances profiles sampled every 10 seconds	232 MB	Non-commercial
9	Tracebase	csv	Diverse appliances profiles sampled every second.	2.9 GB	Open Database
10	IHEPCDS - Individual Household Electric Power Consumption dataset	txt	1 household for 47 months. Sampled every minute.	130 MB	CC BY 4.0
11	REDD - Reference Energy Disaggregation dataset	csv (dat)	6 hourshold for several weeks. Sampled at both low frequency (1 second) and high frequency (15 kHz)	2.8 GB	Not mentioned. Permission granted by authors
12	UKDA - One-Minute Resolution Domestic Electricity Use Data	csv	22 households for 2 years. Sampled every minute.	62.7 GB	Commercial use must be granted by data owner. Data obtained

					from the UK data service
13	Smappee Data[2]	csv	~150 households for 1 year. Sampled every 5 minutes with disaggregated appliances	~1 MB per household	Collaboration and privacy agreement must be drafted if this dataset is eventually used.

[1]Volumes are estimated based on uncompressed folders.

[2]Smappee dataset will only be requested if sources 1-12 are not sufficient for reaching relevant results.

3. Ethical and legal issues

Will you use personal data? If so, shortly describe the kind of personal data you will use AND add the reference to your file in your host institution's privacy register. *In case your host institution does not (yet) have a privacy register, a reference is not yet required of course; please add the reference once the privacy register is in place in your host institution.*	☒ Yes ☐ No If yes: - Privacy Registry Reference: G-2020-2389 (Approval pending) - Short description of the kind of personal data that will be used: The study will use datasets of energy demand (secondary data processing) in individual households with disaggregated appliances consumption. These data will always be provided in a pseudo-anonymised way to us as power timeseries and under no circumstances will it contain sensitive data such as racial or ethnic origin, political opinions, religious beliefs, sexual orientation, and health. Nevertheless, since power consumption profiles can reveal user behaviour, the datasets still fall into the category of general personal data of the GDPR. Thus, an application to KU Leuven Privacy & Ethical Review (PRET) has been submitted. Approval is still pending.
Are there any ethical issues concerning the creation and/or use of the data (e.g.	☒ Yes ☐ No

experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).	If yes: - Reference to ethical committee approval: G-2020-2389 (Approval pending): As indicated above, the main ethical concern of the project is that energy consumption measurements carry sensitive information from which consumer behaviour patterns might be detected. As such they should be consider as personal data and be treated adequately during the research project. In this regard, the project will also be presented to relevant ethical committee, in this case the Social and Societal Ethics Committee (SMEC) at KU Leuven for approval.
Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?	☒ Yes ☐ No If yes, please comment: The public data sources indicated in the data descriptions are provided by third parties. Therefore, neither the fellow nor KU Leuven are the owners of the data. Thus, no IP can be claimed on those datasets. Nevertheless, for some validations, synthetic laboratory test data will be generated. These datasets could be a helpful tool for other researchers to validate their own algorithms. Given the interest that the academic community can have on them, we will share these data under a CC-BY license.
Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?	☒ Yes ☐ No If yes, please comment: In section 2, Table 2 the following restrictions apply: - **Data sources 1-7, 9-11:** the corresponding licenses allow us to reuse, share, and publish the data, as far as we acknowledge the authorship. - **Data sources 8 and 12:** commercial use of the data is restricted. However, at this moment this is not foreseen within the project. - **Data source 13:** as indicated before, this source will only be pursued if the other datasets are not sufficient for our methodology. In this case, a data transfer agreement (DTA) will be set up with Smappee, where the different dissemination restrictions will be specified by the parties.

4. Documentation and metadata

What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?	- **Code:** All modules, classes and functions will be documented using *docstrings*, i.e., a structured comment, which is included in the same source file so code and documentation are a single unit. As most of the coding will be in python, we will follow the Google style docstrings guidelines. Moreover, the tool Sphinx will be used to convert the documentation into a separate PDF, which can be distributed to future end-users of the libraries. - **Machine learning models**: Each model or group of models developed with the same methodology will be accompanied of a Jupyter notebook that explains and illustrates the usage, training/test process, evaluation and hyperparameters selection. Additionally, the list of all notebooks will be document with a README file that indicates the content of each of them, related datasets, and other relevant information such as methodological procedures, derived publications, etc. - **Model training and other statistical results:** The evaluation and summary statistics will use a naming convection including at least the following fields: YYYYMMDD_MODEL_OUTPUT.csv. This will enable traceability and quick search over the results. In this naming model, YYYY indicates the year, MM the number of the month, and DD the day of the month. The hour and minute should also be included if multiple reports are generated for the same model, during the same day. Subsequently, the MODEL and OUTPUT sections should be used to provide an intuitive description of the data. As an example, we can have: *20220402_SVM_traningPerformance.csv*, which include the training performance metrics, for a support vector machine model obtained the 02/04/2022. In some cases, the model and/or output fields can be used in the folder structure naming rather than in the csv files names for simplicity. In addition to this naming convection, a README file will always accompany a group of files to indicate elaborate on the content of each csv, a description of the columns and rows, and any derivate material such as diagrams, graphics or deliverables. - **Laboratory test profiles:** Datasets generated in the laboratory will follow the same naming convection as above, where the MODEL and OUTPUT fields can be substituted by the name of the experiment and the content of the file, e.g., *20220201_syntheticProfiles_smartMeter.csv*. As before, README files will document the content, experiment settings, columns and rows fields and derived material. - **UML diagrams:** UML encompasses any diagram which describe our code, classes, system architecture, relationships, communications, etc. We will follow the guidelines of the Object Management Group to generate such diagrams, always employing open-source tools and formats. The software Diagrams.net (previously known as draw.io) will be used for general diagrams generation.

	Moreover, some graphs will be embedded into the README files using *Markdown* compatible format such as mermaid or Graphviz. - **Graphic results:** all the csv files, i.e., consumption datasets, as well as laboratory test and statistical results can be presented in a graphical manner for better comprehension. Graphic results will be generated using python compatible tools such as matplotlib or seaborn and stored alongside the raw csv data. As indicated above, the README file will describe the relationships between datasets and graphic results. - **Deliverables:** the project reports will be written using LaTeX, and changes tracked using git version control. The source codes and generated pdf will be stored on KU Leuven GitLab (one project per document). The reports should be self-contained, however, a README file will be placed on each GitLab repository with an executive summary of the document and related outputs such as articles, conference proceedings or published datasets. The same approach will be followed for any written documentation derived from the project such as articles or manuals.
Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.	☒ Yes ☐ No If yes, please specify: Data files, i.e., all csv files generated during the project will be accompanied of a metadata file. Although there are various databases and repositories for storing open energy data such as OEP (openenergy-platform.org) or Open Energy Data Initiative (OEDI) (openei.org), there is no standard metadata representation available. Therefore, we will use the Dublin Core Metadata Initiative (DCMI) as the basis for our metadata documentation as recommended by some European initiatives (BMWK - Open Data for Electricity Modeling (bmwi.de)). Additional elements will be added to the basic 15 core elements if required by our project.

5. Data storage & backup during the FWO project	
Where will the data be stored?	- **Reused public datasets:** Since these datasets are publicly available, we could simply download a copy to our local device and there no need for extra storage. However, some of the data are provided in personal webpages that might disappear or be taken down during the project. Therefore, a copy will also be kept on OneDrive as a backup.

	- **Datasets provided by Smappee:** If an agreement is set up with Smappee, the data will be transferred in a pseudo-anonymised way over a secure channel and stored in the ESAT network drives. This storage is secured and encrypted in accordance with KU Leuven policies. Access will be restricted to only the fellow and the supervisor. Furthermore, the data processing will always be carried out in a laptop provided by KU Leuven with an encrypted hard drive and password protected access. - **Data generated within the project:** As a general rule, the results of the project will be stored by default on OneDrive to boost the collaboration and review process between fellow, supervisor and collaborators. In case of privacy-sensitive results, however, the ESAT network drives will be used. - **Deliverables and code:** For version control tracking, all deliverables and code will be hosted on the ESAT GitLab services. The same procedure will be followed for any other publications or reports derived from the project.
How will the data be backed up?	- **Local devices**: no backup procedure in place. The fellow is responsible for backing up important information following the storage policies described above. - **OneDrive**: automated daily backup managed by Microsoft services. Delete items can be recovered up to 93 days later from the recycle bin. - **ESAT network drives**: automated daily backup and snapshots managed by ESAT IT Team. Data recovery can be requested the local IT helpdesk. - **GitLab**: backup procedures are in place and are managed by ESAT IT Local Team.
Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.	☒ Yes ☐ No If no, please specify: Based on the estimated volumes in section 2, the current quota is sufficient for the project: - 30 GB on ESAT network drive - 2 TB on OneDrive
What are the expected costs for data storage and backup during the project? How will these costs be covered? *Although FWO has no earmarked budget at its disposal to support correct research data*	- **OneDrive**: This storage service is part of the Microsoft 365 Education license which is financed centrally by KU Leuven for active KU Leuven staff. - **ESAT Network drives**: The cost of this service is directly paid by the research group (ELECTA-ESAT) based on the amount of data and number of clients. Therefore, it would only be used for privacy-sensitive data which need to adhere to strict security procedures.

*management, FWO allows for part of **the allocated project budget** to be used to cover the cost incurred.*	
Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?	All data will be store on password protected and encrypted devices and/or services, which follow KU Leuven IT security policies. Only the fellow and the supervisor will have access to these data. Privacy-sensitive data used for secondary processing will always be provided in a pseudo-anonymised manner, over a secure channel using Belnet sender, and stored in the password protected and encrypted ESAT network drives. Pseudo-anonymisation at source will make almost impossible to the KU Leuven researchers to identify any natural person since the tokenisation procedure will not be known. Nevertheless, we will follow the same security standard as if we were working with fully identifiable data.

6. Data preservation after the end of the FWO project FWO expects that data generated during the project are retained for a period of minimally 5 years after the end of the project, in as far as legal and contractual agreements allow.	
Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).	- **Public datasets** will not be retained after the finalisation of the project since they can be easily obtained from their origins at any stage. Nevertheless, if any time-consuming data processing was applied to these data the researcher will retain and securely store the resulting dataset for at least 10 years, following KU Leuven RDM policy. Furthermore, the metadata provided with the different models will clearly identify which public datasets must be employed with which model so the methodologies can be replicated. - **Smappee datasets** as private-sensitive data will be stored for 10 years after the finalisation of the project unless the data transfer agreement with Smappee concludes otherwise. - **Derived results** such as models, code, statistical analysis, or experiment datasets will also be retained for at least 10 years as they are the basis for the publications and are likely to be reused by other researchers both at the research group as well as within general academic community (if made publicly available).
Where will these data be archived (= stored for the long term)?	The data will be stored at ESAT network drives for at least 10 years as indicated above. This storage unit has automatic backup procedures as well as encryption and security measurements that guarantee the security and long-term availability of the datasets. It is important to highlight here that a separated folder with strict access rights will be kept for the privacy-sensitive datasets.

	The researcher and supervisor will also evaluate if some of the datasets or codebase generated during the project should be made available to other researchers by means of KU Leuven Research Data Repository and KU Leuven GitLab in the case of code. For article and conference proceedings, the accepted version will be deposited on Lirias to comply with FWO open-access policy.
What are the expected costs for data preservation during these 5 years? How will the costs be covered? *Although FWO has no earmarked budget at its disposal to support correct research data management, FWO allows for part of* ***the allocated project budget*** *to be used to cover the cost incurred.*	Following the estimated figures indicated in Section 2, the total storage requirements at the end of the project should not exceed 30 GB. The department and unit of the fellow (ESAT-ELECTA) have storage systems in place that could easily accommodate this volume of data. However, if extra storage is needed, it could be obtained at KU Leuven ICTS with an estimated cost of 54€/year for a 100GB server back-end storage Type 1 (archival purposes).

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?	☒ Yes ☐ No If yes, please specify: - **Publicly available datasets** can be shared without any restrictions other than the limitations in commercial use specified in some of them. (See Section 2, Table 2) - **Privacy-sensitive datasets** obtained from Smappee cannot be shared between KU Leuven and other third parties unless a specific data transfer agreement is set up between them and Smappee. - **Project results** can be shared with other researchers as long as no tech transfer and valorisation potential are identified.
Which data will be made available after the end of the project?	- **Machine learning models** if no IPR potential is identified. - **The codebase** providing no potential IPR is identified within the project. - **Experimental test data** with appliances profiles, aggregated and disaggregated measurements. - **Statistical results** derived model training and other analysis can be also shared. Nevertheless, the potential of these datasets to be reused is very low as they can only be understood in the context of a particular analysis.

Where/how will the data be made available for reuse?	☐ In an Open Access repository ☐ In a restricted access repository ☐ Upon request by mail ☒ Other (specify): - **Machine learning models** can be shared as pickle files (pkl) accompanied by a Jupyter notebook (ipynb) which exemplifies it usage. The content will be hosted on KU Leuven GitLab as public repositories. - **The codebase** will also be hosted on KU Leuven GitLab. However, it will be initially configured as a private repository. If the results are relevant and no potential IPR is identified within the project, the repositories will be made public. - **Experimental test data** which might results interesting for other researchers will be made available on KU Leuven Research Data Repository (RDR) following the FAIR principles. - **Statistical results**, if deemed relevant for other researchers, will be shared alongside the publications using the data services of publishers such as Elsevier or IEEE. In case the publisher does not have a data repository, KU Leuven RDR will be used.
When will the data be made available?	- **Models and codebase** will be made available only at the end of the project and providing no IPR potential is identified. - **Experimental test data and statistical results** will be shared upon publication of the results.
Who will be able to access the data and under what conditions?	Datasets created by the authors will be shared following a Creative Commons Attribution (CC-BY) license as a general rule.
What are the expected costs for data sharing? How will these costs be covered? *Although FWO has no earmarked budget at its disposal to support correct research data management, FWO allows for part of* ***the allocated project budget*** *to be used to cover the cost incurred.*	At the current stage no additional costs associated to data sharing have been identified. KU Leuven RDR allows researchers to store up to 50 GB of data for free, a figure that is within the estimated total storage requirements of the entire project.

8. Responsibilities

Who will be responsible for the data documentation & metadata?	The fellow will be responsible for the documentation of the dataset and the generation of the corresponding metadata.
Who will be responsible for data storage & back up during the project?	The fellow, together with the supervisor will ensure that the storing policies described in section 5 are closely followed. Regarding the backup procedures, they are automated and managed by the different services. The fellow and supervisor will be only responsible for backing up information from their local devices to the cloud services (OneDrive and ESAT network drives)
Who will be responsible for ensuring data preservation and sharing?	During the project, the fellow will ensure preservation and will work on the documentation of the data to guarantee later reuse. Once the project is finalised, the supervisor will hold the responsibility for data preservation, while reuse should have been procured at the previous stage with adequate documentation by the fellow.
Who bears the end responsibility for updating & implementing this DMP? *Default response: The PI bears the overall responsibility for updating & implementing this DMP*	The fellow under the guidelines of his supervisor bears the end responsibility of updating & implementing this data management plan.