eco-bot
let's talk energy!

Deliverable 1.3
Data Management Plan

Risa, Stephanos Camarinopoulos
March, 2018

www.eco-bot.eu

D1.3: Data Management Plan

Summary

This document represents the Data Management Plan for the project Eco-Bot.
The aim of this deliverable is to describe the plan for managing the data generated and collected during the project. Spesifically, this deliverable provides the data management life cycle for all datasets to be collected, processed and generated by a research project.

DELIVERABLE NUMBER	WORK PACKAGE
D1.3	WP1
LEAD BENEFICIARY	**DELIVERABLE AUTHOR(S)**
RISA	Stephanos Camarinopoulos (RISA)
QUALITY ASSURANCE	
Reviewer 1: Ulrich Hussels	RISA
Reviewer 2: Stephanos Camarinopoulos	RISA
PLANNED DELIVERY DATE	**ACTUAL DELIVERY DATE**
30/3/2018	21/3/2018
DISSEMINATION LEVEL	□ PU = Public □ PP = Restricted to other programme participants x CO = Confidential, only for members of the consortium

I

Table of contents

List of Tables

CA: Consortium Agreement

CO: Confidential

DoW: Description of Work, referring to the Annex I of the Grant Agreement

EC: European Commission

DMP: Data Management Plan

GA: Grant Agreement

IPR: Intellectual Property Rights

OA: Open Access

PPR: Project Progress Reports

PSB: Project Steering Board

PU: Public

QA: Quality Assurance

SAB: Security Advisory Board

SOA: Service-oriented Architecture

STC: Scientific and Technical Committee

WP: Work Package

1. Introduction

1.1.   Purpose

This document is the Eco-Bot Data Management Plan (DMP). It covers:

- The handling of research data during and after the project.
- What data will be collected, processed or generated.
- What methodology and standards will be applied.
- Whether data will be shared and how.
- How data will be curated and preserved.

In principle, publicly funded research data are produced for the public interest and should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property and confidentiality. On this basis, the DMP intends to help researchers consider at an early stage, when research is being designed and planned, how data will be managed during the research process and shared afterwards with the wider research community.

The DMP specifies the availability of research data, describes measures to ensure that data are properly anonymized and protect the privacy of informants and respondents. The developed tools and models within the project shall not contain any case or consumers specific information.

With regard to access to research data, Eco-Bot will make the data and metadata available on the project internal website research data repository (https://redmine.risa.eu/projects/eco-bot). Project members will, in this repository, have access to both data and metadata. For the time being, research data is planned to be archived at a web server (soon to be defined) in future research projects and follow-up studies.

With regard to open access to scientific publications, Eco-Bot aims to publish in well-respected and highly rated peer-reviewed journals. This task will be undertaken mostly by the research partners, and the publications will cover several project fields of work. Particular effort will be made to secure Open Access (OA) to all interested persons, mainly through the project website but also through respective OA repositories such as OpenAIRE. Partners will be encouraged to forward press releases and announcements about project work to their contacts. This is expected to produce a regular flow of news articles in traditional and digital media, local and national. In addition, the Eco-Bot consortium will seek every opportunity, always in close collaboration with the EC R&D personnel, to disseminate the project vision and main results through various means offered by the EU, e.g. Horizon Magazine, research*EU results magazine, EuroNews TV etc. Partners will also investigate possibilities to participate in EU research conferences and public events, e.g. EU Energy Forum 2017, Open Door Days and H2020 Researchers Nights events during the project life. Eco-Bot presentations and demos will appear in relevant international conferences and other events.

The benefits of a well-designed DMP not only concern the way data are treated but also the successful outcome of the project itself. A properly planned DMP guides the researchers first to think what to do with the data and then how to collect, store and process them. Furthermore, a planning in data treatment is important for addressing timely security, privacy and ethical aspects. This way the research data are kept in track in cases of possible staff or other changes. The DMP can also increase preparedness for possible data requests. In short, planned activities, such as implementation of well-designed DMP, stand a better chance of meeting their goals than unplanned ones.

Once accepted by the Consortium, this Data Management Plan becomes an official project document, which should determine all beneficiaries' and consortium's actions.

1.2.   Scope of the Data Management Plan

This Data Management Plan is aimed to be initially used by the members of the consortium, including the Commission Services and provides guidance on data management to the project partners and participants. It is particularly relevant for partners responsible for data collection and pilots. The DMP will evolve throughout the project as new procedures are added or existing ones are updated. The deliverable's dissemination level is public.

The process of planning is also a process of communication, increasingly important in a multi-partner research. The characteristics of collaboration should be accordingly harmonized among project partners from different organizations or different countries. The DMP also provides an ideal opportunity to create best practice with regards to file formats, metadata standards, storage and risk management practices, leading to greater longevity and sustainability of data and higher quality standards.

Ultimately, the DMP should engage researchers in conversations with those providing the services (e.g. energy suppliers). In this context, the DMP becomes a document in accordance with relevant standards and community best practice.

This Data Management Plan is authorized by the Steering Committee. All subsequent changes / revisions should also be approved and authorized by the Steering Committee.

1.3.   Procedure Description

Data Management Planning is an integral part of management planning. The Data Management Plan regards all the data sets that will be collected, processed and/or generated within the project. The methodology the consortium follows to create the DMP is as follows:

1. Create a data management policy. To this end, we describe

    a) The elements that the EU proposes to address for each data set.
    b) The strategy that is used by the consortium to address each of the elements.

2. Analyze the completed by the project's partners DMP templates.

To ensure relevance of the Data Management Plan, the coordinator should ensure that the DMP is available to all concerned and that its requirements are met.

2. Guiding principles

The legal requirement for open research data in the Grant Agreement is not applicable. However, the legal requirements for publications should not represent a risk for compromising the privacy of informants participating in the Eco-Bot project, by following ethical requirements. This DMP assesses when and how data can be shared within a sound research ethical framework, where directly or indirectly identifiable information is not disclosed in the research process at any stage.

Furthermore, all the data produced or used in the project should follow the basic principles of FAIR data, as described on H2020 "Guidelines on FAIR Data Management in Horizon 2020":

- To be **Findable**, including provisions for metadata.
- To be **Accessible**.
- To be **Interoperable**.
- To be **Re-usable**.

2.1. Findable data

The data produced or used in the project shall be discoverable with metadata, identifiable and locatable by means of a standard identification mechanism, as it will be described in section 3.1 below.

The naming conventions will be described in the section 3.1 below, and shall be followed by every member of the consortium. The version numbers should be provided in every file and the data files shall be described by rich metadata that will make every data a searchable resource.

2.2. Accessible data

Data shall be easily retrievable by their identifier using a standardized communications protocol. The protocol should be open, free and universally implementable, allowing for an authentication and authorization procedure, where necessary.

The dataset will be as open as possible and as closed as necessary, according to the EC regulations. Therefore, every data shall be available to the consortium and some of them will be of public access.

The datasets will be available from https://redmine.risa.eu. From there the fully anonymized data are accessible to anyone within the project, free of charge. This repository allows for the deposition of all kinds of digital content: publications, data, software, multimedia etc. All published data will be provided with a Digital Object Identifier (DOI). This means that once published, datasets remain fixed over their lifetime, while the metadata can change.

2.3. Interoperable data

Data shall use a formal, accessible, shared and broadly applicable language for knowledge representation. Data and metadata should use vocabularies that follow FAIR principles and should include qualified references to other data/metadata.

The purpose of this is the facilitation of data exchange and re-use between researchers, institutions, organisations and countries. In case uncommon vocabularies or generated project specific ontologies are used/created, these vocabularies/ontologies shall be provided.

Eco-Bot builds of the principles of the European Interoperability Framework by adopting a Service Oriented Architecture and will make efforts to achieve

- Interoperability governance (creation of a service registry, e.g. the open source Apache jUDDI).
- Semantic interoperability, by defining vocabularies and schemata to describe data exchanges.
- Technical interoperability. Eco-Bot will make efforts to use formal technical specifications for data exchange, service interconnection and secure communication protocols (e.g. JSON, RESTful, HTTPS)

In general, the Eco-Bot aims at data interoperability, by limiting the complexity generated by multiple types of interfaces. Eco-Bot's web interface and API will use HTML and RESTful web services to provide interoperability. A standard, like OGC Publish/Subscribe Interface Standard will be followed for the coordination of communication between the central node and the other systems. JSON will be used for the description of datasets. The OASIS Universal Description, Discovery, and Integration (UDDI) specification for web services will

be used for the creation of a service registry. Using API endpoints makes the integration interoperable as long as other users or systems can use a REST API integration and have the available data to feed the different system mechanisms.

The application is designed with the concept of SOA as a Cloud Service, allowing for reusability, adaptability and maintainability. Thus, it combines web-based services and APIs, database systems and users. Software components are loosely coupled and they can be easily reused. A client-consumer can find a service in the registry and consume it. For instance, the energy disaggregation module consumes data deriving from smart meters in a format that will be clearly predefined and will be uniform for all providers (DEXMA, SEC, and EYPESA). As far as the chat-bot technology is concerned, Microsoft BOT Framework will be used, which offers predictable costs and scaling but efforts will also be made to be able to integrate with other open source NLP tools and BOT solutions.

With regard to communication, the system includes a transport layer (transfer of raw and disaggregated data), a session layer (session and connection coordination) and application layer (management of data, user authentication and service quality verification).

To test the interoperability, we plan a series of actions based on real case scenarios that include message structure definition and other test parameters such as time to respond, message exchange thresholds, connection drop actions, etc. A series of unit tests during the development will take place. Before adaptation and publishing services or data, a set of stress tests will take place to assess the capabilities and abilities of the system. After the successful pass of the test cases, the phase of validation will take place in order to validate the efficiency of the software, the service endpoints and the coordination with the other system components.

Nevertheless, the final architecture and the way the Eco-Bot platform will be connected with the various technological components of the project, as well as the legacy systems has not been finalized at the time of writing this deliverable. Different options will be investigated in order to ensure interoperability and at the same time to serve the objectives of the project.

2.4. Re-usable data

Data and metadata should be easily re-usable, licensed to permit the widest re-use possible. Therefore, metadata shall be richly described with a plurality of accurate and relevant attributes, as well as be released with a clear and accessible data usage license. Data should be associated with detailed provenance and meet domain-relevant community standards.

Eco-Bot collects data from the consumers. These data will be used to develop, test and benchmark the Eco-Bot algorithms and modules regarding the chat-bot and the back end algorithms (energy disaggregation). The type of the datasets includes energy consumption, CO_2 emissions and energy costs, while in cases of smart metering more information are included. The energy consumption data will be gathered through specific APIs and the energy will be aggregated in *csv files. Annotated and anonymised datasets are created that will be uploaded on Eco-Bot web portal to be accessed for free from other research communities and project for benchmarking. The only obligation for the third parties is to cite with an appropriate reference whenever these data are used. The datasets will be created at three stages through the project activities.

This document, being in line with the EU's guidelines regarding the DMP, should address the following elements for each data set collected, processed and/or generated in the project:

1. Data set reference and name.

2. Data set description.

3. Standards and metadata.

4. Data sharing.

5. Archiving and preservation.

To this end, the consortium develops a number of strategies that will be followed in order to address the above elements.

In this section, we provide a detailed description of these elements in order to ensure their understanding by the partners of the consortium. For each element, we also describe the strategy that will be used to address it.

3.1.   Data set Reference and name

In order to be able to distinguish and easily identify data sets, each data set is assigned with a unique name. This name can also be used as the identifier of the data sets.

3.2.   Data set description

Each data set that will be collected, processed or generated within the project will be accompanied by a brief description. The description will provide information regarding:

- The nature of the data set.
- The scale of the data set.
- To whom the data set could be useful.
- Whether the data set underpins a scientific publication.
- Information on the existence (or not) of similar data sets.
- Possibilities for integration with other data sets and reuse.
- In case the data set is collected, the origin of the data set will be also provided.

Table 1 Overview of datasets

#	Dataset Name	Brief Description	Format	Dissemination
1	Energy consumption data	Energy consumption of each informant	.json	public
2	Energy costs	The costs of energy in every household	.json	public
3	Smart metering data	Energy consumption of basic household devices	.json	public
4	User profile	Includes some user's information, in order to build his profile (e.g. age, marital status, sex)	(not yet defined)	confidential
5	Chat-bot profile	Includes the profile's personalized questions, the recommendations, the frequency of logging in etc)	(not yet defined)	confidential

3.3.   Standards and metadata

This field will describe suitable standards that will be used to describe the data as well as the metadata of the data sets. In Eco-Bot, we use data sets collected from customers (e.g. energy consumption, CO2 emissions and energy costs, in *csv files) as well as other datasets, including all the data sets published in other formats (e.g. excel, pdf, txt etc.). These data sets will also by accompanied, is some cases, by metadata.

Technical metadata may be specific to a content type, such as image, sound, video, text, spreadsheet or e-mail. Some content type-specific metadata is essential for rendering a digital object representation. For example, it is essential to know the attributes of a database. Some file formats enable the capture of technical, and other, metadata within their files, which has the advantage of keeping the files self-descriptive. However, by extracting and storing metadata explicitly we may also benefit. Separate metadata can:

- Be kept small and processed efficiently.
- Be distributed separately.
- Have different access rights and licensing arrangements than the content.

- Help to account for the whole life cycle of digital objects.
- Have its description standardized across file formats.
- Be managed and preserved by preservation systems.

3.4. Data sharing

The partner responsible for drafting the DMP along with all work package leaders, will define how data will be shared and more specifically the access procedures, the necessary software and other tools for enabling re-use, for all datasets that will be collected, generated, or processed in the project. In case the dataset cannot be shared, the reasons for this will be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

The data collected will be useful at two phases.

The first refers to the initial preparation phase in which we collect data to deploy the knowledgebase and the expert system. These data are some representative examples able to characterize all the variety of the multi-factorial models, the complexity of the content and the market segmentation characteristics that affect consumers' needs and preferences.

The second phase refers to the collection of data across the large-scale campaign launched by the project. In this phase, we evaluate the response of the ICT tools, the effectiveness of our system to change users' behaviour and the degree of penetration of our system to the final model.

3.4.1. Public Availability of Data

In view of the precautions for protection of personal data, it is explicitly confirmed that the data collected will be publicly available, after care is taken with regard to rules of confidentiality, anonymity and protection. Anonymised final data sets will be open access and procedures are set as to how data will be preserved and archived after closure of the programme in the Redmine repository. We are aware of post-publication risks to local researchers and end-users in our research sites and will mitigate all reasonable risk before publication. We will not use any images or names of any research subjects or local researchers without their prior consent and agreement.

3.4.2. Opt Out

It is important to note that even though comprehensive measures are taken to ensure the safety of participants, researchers and their environment, it is only after an Eco-Bot report or scholarly article is published, that the question of open access arises. Open access does not

entail an absolute obligation to publish all data: it is up to researchers and EU to decide whether data is suitable and ethical to be published or not.

3.5. Archiving and preservation

The partners will decide and describe the procedures that will be used in order to ensure long-term preservation of the data sets. This field will provide information regarding the duration of the data preservation, the approximate end volume, the associated costs and the plans of the consortium to cover the costs.

To protect personal data, RISA will assign to an expert-legal from the company to establish the legal framework that describes the rules and conditions that people can participate in the demonstration/ validation phase, in close collaboration with the responsible partners from the demo sites. The relevant partners (ESTB, SEC and DEXMA) will assign at least one person from their organisation to act as the Legal and Ethical Committee (please see Section 5).

Any personal datum will be processed in compliance with Directive 1995/46 on the processing of personal data, and the ePrivacy Directive 2002/58/EC.

The following guidelines will be used in order to ensure the security of the data:

- Keep anonymized data and personal data of respondents separate.
- Encrypt data if it is deemed necessary by the local researchers.
- Store data in at least two separate locations to avoid loss of data.
- Limit the use of USB flash drives.
- Save digital files in one of the preferred formats.
- Label files in a systematically structured way in order to ensure the coherence of the final dataset.

For some of the activities to be carried out by the project, it may be necessary to collect basic personal data (e.g. full name, contact details, background), even though the project will avoid collecting such data unless deemed necessary. Such data will be protected in compliance with the EU's Data Protection Directive 95/46/EC1 aiming at protecting personal data. National legislations applicable to the project will also be strictly followed.

All data collected by the project will be done after giving data subjects full details on the experiments to be conducted and only after obtaining signed informed consent form

4. Ethics

The Eco-Bot partners are to comply with the ethical principles as set out in Section 5.1 of the Grant Agreement.

All data will be stored, processed and analyzed at protected data centers following all data protection guidelines. Before the project start, the participants will provide the EU with templates of the informed consent forms and information sheet to be used during the project research. The data collected in the Eco-Bot project will consist of:

1) Energy usage data, which will be used only as described in the informed consent form.

2) The participants' names and email addresses indicated on the informed consent forms.

The hard copies of the informed consent forms (and the personal data they contain) will remain confidential. They will be gathered at restricted data centers. Only a very limited number of persons will have access to them (notably in case of participant's withdrawal). There will be no publication or disclosure of the said data.

4.1.   Data Handling

4.1.1. Data Collection

The data will be collected at the premises in close areas where only the volunteer adult participant will take place in the process. Data will be collected under the responsibility of respective task leader. In case that we need to need to store data using temporary devices, the collection will be performed under the responsibility of the coordinator and/or the partner responsible for conducting the experiment. In particular, a signed declaration form will be delivered named:

(1) Description of the pilot.
(2) Date(s) on which the pilot test carried out.
(3) Pilot place.
(4) Type of content captured.
(5) Storage device name.
(6) Brand name of the storage devices.
(7) Its serial number.

4.1.2. Data Storage

The content will be securely stored at the data center. Authorized access to the content is allowed only by authorized project members. In order for a partner to have access on the

content, an agreement is required. Remote networked access is allowed only under a VPN secure framework. However, the process will be conducted remotely.

The Informed Consent Form signed by the participant will normally allow snapshots of the content to be shared within the project dissemination (including outreach activities OR internet and social media).

Three years after the completion of the project, the data will be moved and securely kept as backup storage under the responsibility of the coordinator and/or RISA data protection officer.

4.1.3. Data Destruction

Upon any participant's request, the respective content will be erased from the system. The coordinator will be responsible to activate this deletion up to 5 working days.

4.2. Confidentiality

Eco-Bot project does not foresee processing of personal sensitive data (e.g. sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction).

The personal data of the consumers shall be strictly protected.

4.3. Intellectual Property Rights (IPR) Management

For every partner of Eco-Bot consortium, it is very important to have explicit rules on how to access Pre-Existing Know-How, knowledge and how to ensure the protection of intellectual property. Therefore, Eco-Bot partners are aware about the importance of the CA and IPR Issues in order to develop common and individual dissemination and exploitation strategies. As it shall be described in the Exploitation Agreement, the Eco-Bot partners will take into account the following agreements:

- Concerning exploitation of the project results, it is the understanding of the Consortium that knowledge and pre-existing know-how will be made available to the Consortium members in favourable conditions if they are necessary to perform the research and relative work in this project. The placement of Pre-Existing Know-How into the project will be detailed in the Appendix of the CA. Herein, every single partner is entitled to describe their own Pre-Existing Know-How.
- Foreground knowledge is owned by the contractor generating such information or result. Each contractor shall make available its foreground knowledge, on a royalty-free basis, to other contractors to the extent that such information is necessary for the production of their own foreground knowledge within the Eco-Bot project.
- Joint Foreground knowledge: If it is not possible to determine exactly the ownership of that foreground knowledge, i.e. several contractors participated in that specific

development ownership will be shared by the pro ratio effort invested by each contractor.

- Pre-existing know-how and foreground knowledge will be made available, on a royalty-free basis, to the other project partners for dissemination, research and academic purposes in respect to the intellectual property rights of the partner generating this knowledge.
- Pre-existing know-how and foreground knowledge will be made available to the other project partners for exploitation purposes at favourable conditions, with respect to the normal commercial conditions applied by the granting partner.

5. Conclusion

This deliverable consists the Eco-Bot project Data Management Plan and specifies the procedures to be applied by the partners and the project governing bodies, in order to guarantee the FAIR (Findable, Accessible, Interoperable, Re-usable) quality of data and the convenient transfer of information.

Specifically, the DMP described the data management life cycle for all datasets to be collected, processed and/or generated by a research project. It covered:

- The handling of research data during and after the project.
- The data which should be collected, processed or generated.
- The methodology and standards which will be applied.
- The access level of data and the ways to share them.
- How data will be curated and preserved.

Following the EU's guidelines regarding the DMP, this document shall be updated during the project lifetime (in the form of deliverables) in project-month 12, 18, 24 and 39.