Data Management Plan

D3.2

02/2017

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 689954.

Project Acronym and Name	iSCAPE - Improving the Smart Control of Air Pollution in Europe	
Grant Agreement Number	689954	
Document Type	Deliverable	
Document version & WP No.	V. 0.2	WP3
Document Title	Data Management Plan	
Main authors	Guillem Camprodon	
Partner in charge	IAAC	
Contributing partners	Francesco Pilla, UCD Andreas Skouloudis, JRC-EC	
Release date	10/02/2017	

Document Control Page	
Short Description	This report describes the management plan for the data collected

	during the project.		
Review status	**Action**	**Person**	**Date**
	Internal Review	Muhammad Adnan, UH and Luca Simeone, T6	15/02/2017
	QC&QA	Coordination Team	25/02/2017
Distribution	Public		

Revision history			
Version	**Date**	**Modified by**	**Comments**
V0.1	7/02/2017	Guillem Camprodon	Initial version
V0.2	20/02/2017	Guillem Camprodon	Updated version including input by internal reviewers
Final	25/02/2017	Francesco Pilla	QA&QC

Table of Contents

Table of Contents

List of Tables

LIST OF FIGURES

List of abbreviations

DMP Data Management Plan

GPS Global Position System

CC Creative Commons License

CSV Comma Separated File

JSON JavaScript Object Notation

REST Representational state transfer

DOI Digital Object Identifier

DWE Open Geospatial Consortium's Sensor Web Enablement

1 Executive Summary

This report describes the management plan for the data collected and managed on the project. It advocates for the effectiveness of openness and sharing, hence we strive to make data collected during the project as available as possible within the limit of personal privacy. The document will be updated regularly with the purpose of supporting the data management life cycle for all data that will be collected, processed or generated by the project.

2 Introduction

iSCAPE advocates for the effectiveness of openness and sharing, hence we strive to make data collected during the project as available as possible within the limit of personal privacy following the Fair Data Principles.

The project will collect the following types of data: quantitative data on the environment, collected using sensors in the different test sites, quantitative data related to day-to-day activity and travel scheduling decisions made by households and individuals, qualitative and quantitative data on levels of participation and user experience collected programmatically by the project websites or via interview and questionnaires.

These types of data have different characteristics: open sensor data access to the data is important since it contributes to the impact of the project. This data will be automatically collected and uploaded to the Internet and shared publicly and for free through an online visualization platform.

Quantitative data related to day-to-day activity and travel scheduling contains potentially more personal data that can be potentially linked to a person's location and lifestyle (e.g. when a person is at home). In order to make this data publicly available it will be aggregated since in order to guarantee the anonymization.

User experience data contains potentially more personal data, such as gender and age. Indicators as well can include user data since they aim at quantifying behavioral change and therefore refer to user behaviors, such as route to work. This data can be also aggregated since it is not so relevant on the level of a single individual. This fact allows to anonymize data for example by aggregating it in categories.

Databases will be deposited in recognized, international data repositories so that data will continue to be available if maintenance of the main web-platform is discontinued.

This Data Management Plan will be updated regularly thereafter with the purpose of supporting the data management life cycle for all data that will be collected, processed or generated by the project.

2.1 Fair Data Principles

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. (Wilkinson et al. 2016)

1. To be Findable:
    a. Metadata are assigned a globally unique and eternally persistent identifier.
    b. Data are described with rich metadata.
    c. Data (and metadata) are registered or indexed in a searchable resource.
    d. Metadata specify the data identifier.

2. To be Accessible:
    a. Data (and metadata) are retrievable by their identifier using a standardized communications protocol.
    b. The protocol is open, free, and universally implementable.
    c. The protocol allows for an authentication and authorization procedure, where necessary.
    d. Metadata are accessible, even when the data are no longer available.

3. To be Interoperable:
    a. Data (or metadata) use a formal, accessible, shared, and broadly applicable language for knowledge representation.
    b. Data (or metadata) use vocabularies that follow FAIR principles.
    c. Data (or metadata) include qualified references to other (meta)data.

4. To be Re-usable:

a. Data (or metadata) have a plurality of accurate and relevant attributes.
b. Data (or metadata) are released with a clear and accessible data usage license.
c. Data (or metadata) are associated with their provenance.
d. Data (or metadata) meet domain-relevant community standards.

3 Standards and Metadata

All the data within the project will be available using well known formats or documented accordingly. Data will be always including extensive metadata.

The data standard regarding sensor data will be the one provided by the Smart Citizen API. This is a JSON REST API web service fully documented on the platform webpage[1]. The service also supports batch data downloads as CSV. Metadata is accessible via the same interface. Data related to sensor location, sensor type or manufacturer is available within the same API. Connectors for other data standards can be built on top of the existing API. Specifically, the project will closely follow the development of Open Geospatial Consortium's Sensor Web Enablement (SWE) standards and in particular the Common Data Model.

The data standard related to other project activities might vary depending partners' internal tools and methodologies. However once certain datasets will be decided to be made public this will always follow a CSV tabular data standard. As part of the projects open data access and sustainability strategy once the project finishes sensor data will be programmatically downloaded as CSV files and metadata will be added following the same standard described above. To ensure the project data is made available using the same standards towards consistency and usability all the datasets should provide the following metadata. This will be then turned on the DataCite Metadata Schema[2].

- Digital Object Identifier (DOI)
- Publication date
- Title
- Authors
- Description
- Source
- Keywords

[1] Smart Citizen API Documentation http://developer.smartcitizen.me/

[2] DataCite Website http://schema.datacite.org/

All the data sets available within the project must provide a detailed level of metadata on their description. The following guidance provided by the NERC[3] will be followed:

Table 1 Guidance on metadata topics to address

Metadata topic to address	**Details**
Experimental Design / Sampling Regime	Metadata should be provided which details the experimental design and/or sampling regime, where applicable
Collection / Generation / Transformation Methods	Metadata should be provided covering methods used for collection of samples/observations. Alternatively, where data values are derived/generated/ transformed, then details of how this is achieved should be provided.
Fieldwork and / or Laboratory Instrumentation	Information should be supplied on instruments/machines used for collection/analysis of samples/observations where relevant.
Calibration Steps and Values	Details of the steps taken to calibrate any instruments/machines used, including use of any blanks, and the values used for calibration should be provided.
Nature and Units of Recorded Values	Metadata should be provided describing the nature of the recorded values contained and the units used sufficient to unambiguously define what has been measured and recorded in the dataset.
Analytical Methods	Full descriptions of any analytical methods used to generate the data values contained in the dataset

[3] Natural Environment Research Council http://www.nerc.ac.uk/

	should be included.
Quality Control	Any quality control measures undertaken to ensure the quality of the data values included in the dataset should be detailed.
Format of Stored Data	The format which was used to store the dataset during the lifetime of the project, and the format in which the dataset is made publicly available, if different, should be named in the contextual metadata.
Miscellaneous	Any additional information necessary to expand on that given in the discovery metadata record.

4 Datasets types

The following section describes the current datasets typologies forecast to be generated during the project. This section will be updated periodically as part of the DMP life cycle.

Table 2 Project dataset typologies for internal traceability

Trace code	**Label**

PD	Personal data
SV	Single value
TS	Time series
MA	Maps

Table 3 Project forecasted datasets at the deliverable release time

Data Set	**Description**	**Type**	**Source**	**Access**	**Partners Responsible**
DS_TS_001	iSCAPE SCK Low Cost sensors	Environmental	Sensors	Public	IAAC
DS_TS_002	iSCAPE SCK High End sensors	Environmental	Sensors	Public	IAAC
DS_PD_003	Users transport data collected over GPS	User activity	GPS	Public after aggregation for anonymization	UH
DS_PD_004	Users surveys	User opinions	Interviews	Public after aggregation for anonymization	FCC
DS_PD_005	Website visitors	User opinions	Counting	Public after aggregation for anonymization	T6

5 Sharing, accessibility and archiving

Public accessibility to data is a critical objective for the iSCAPE project, especially for sensor environmental data. This takes a dual approach focused on the two main project targets.

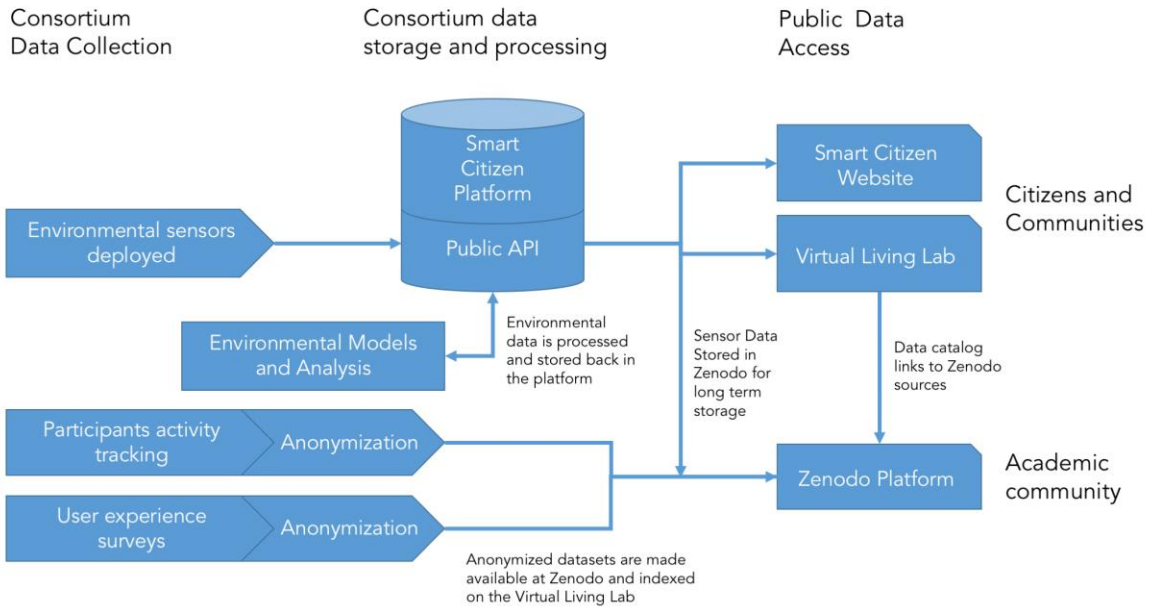

Figure 1 Project Data Lifecycle

First in order to reach out citizens and local communities, especially from the 7 pilot cities, the Virtual Living Lab will be developed on Task 8.1. This is especially important since having a sense of ownership over a technology intervention has been associated to sustained community engagement (Balestrini et al. 2014) The platform will provide access to the project environmental sensors in near real time and facilitate the exploration of data with other contextual data (maps, keywords) and processed reports.

Secondly in order to reach out the research and professional community data will be made available on the Zenodo[4] digital repository following the details specified on the **Standards and Metadata** section below. This includes the totality of the project data that will be made finally public including the collected environmental data on the Smart Citizen that will be dumped there periodically. This protocol aims at guaranteeing consistency on the use, reference and preservation of the sources, especially when the project finishes. Since this method does not provide real time neither interactive access to data, Environmental Sensor data will be also available using the Smart Citizen API described on Section 3 above. This JSON REST API allows to programmatically access sensors data in near real-time and to request batch data in CSV format.

As a default standard, the project will use Creative Commons Licenses for public data sharing. In particular Attribution 4.0 International[5] in order to support third party researchers the right to redistribute the material in any medium or format and build upon the material for any purpose, even commercially. However appropriate credit must be given to the original source, provide a link to the license, and indicate if changes were made.

As described above data archiving and sustainability will be guaranteed by the Zenodo digital repository. As a European Commission supported initiative and technically supported by CERN we trust this as the best way to ensure access to the generated data remains long after the project ends. Furthermore, environmental sensor data will be preserved on the Smart Citizen Platform as long as the project is alive. This guarantees any applications developed on top will remain accessible. In terms of data archiving and safety the Smart Citizen Platform relies on a distributed cluster of Cassandra databases split across different datacenters in Europe.

[4] Zenodo was launched at the CERN Data Centre in May 2013 with a grant from the European Commission with a special commitment to sharing, citing and preserving data and code. As a digital repository, Zenodo registers DOIs for all submissions through DataCite. The platform is based on the Invenio open-source software, Zenodo profits from and contributes to the foundation of code used to provide Open Data services to CERN and other initiatives around the world.

[5] A Creative Commons (CC) license Attribution 4.0 International license file https://creativecommons.org/licenses/by/4.0/

Table 4 Project forecasted datasets (Table 3) public access and archiving methods

Data Set	Consortium data storage and processing	Data Processing for Public Release	Public Data Access	Public Data Archiving
DS_TS_001	Smart Citizen Platform	-	Smart Citizen Website and Virtual Living Lab	Smart Citizen and Zenodo Platform
DS_TS_002	Smart Citizen Platform	-	Smart Citizen Website and Virtual Living Lab	Smart Citizen and Zenodo Platform
DS_PD_003	Partners Internal Facilities	Aggregation for anonymization	Zenodo Platform	Zenodo Platform
DS_PD_004	Partners Internal Facilities	Aggregation for anonymization	Zenodo Platform	Zenodo Platform
DS_PD_005	Partners Internal Facilities	Aggregation for anonymization	-	Zenodo Platform

6 Data Privacy and Security

6.1 Personal Data

All recordings, transcripts and notes from workshops and interviews will be anonymized by removing the names of participants. By the end of the fieldwork, material will be stored on the premises and the machines of the respective iSCAPE consortium partner organizations premise and machines. Data likely to contain personal information as those collected in WP4 will be aggregated deleting sensible information as home address and mobility patterns before is publicly released towards guaranteeing the anonymization of information.

Project participants through the different pilots and activities will provide some personal information to allow their identification, and the attribution of the sensing devices. This

process will be carried directly by the Smart Citizen platform which already has its own Data Management Plan being IAAC the data controller.

Dissemination activities might rely on several third-party platforms mostly operated by US companies (e.g., Google Analytics, MailChimp, Facebook). In order to comply with the EU data protection standards, the companies have to certify its compliance with the EU-U.S. Privacy Shield Framework[6]. Before any of the following services is used within the project, compliance will be verified at the Privacy Shield list[7].

6.2 Security

Sensitive data will be stored on premises by partners as described above except for the Smart Citizen platform. The platform operates on the EU on ISO 27001 certified data centers following data management best practices as encryption, pseudonymization by default and backups and failover mechanisms.

[6] Commission Implementing Decision (EU) 2016/1250 of 12 July 2016 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the EU-U.S. Privacy Shield
http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.207.01.0001.01.ENG
[7] List of companies in compliance with the EU-U.S. Privacy Shield Framework
https://www.privacyshield.gov/list

7 Plan Maintenance

The Data Management Plan will be updated regularly by the Data Manager (IAAC) and reviewed by WP-leads with the purpose of supporting the data management life cycle for all data that will be collected, processed or generated by the project

Table 5 Planned review dates

Review Dates	Project Month	DMP Version
September 2017	M13	v2.0
September 2018	M24	v2.0
August 2019	M36	v4.0

8. References

Balestrini, M. et al., 2014. Understanding sustained community engagement. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. Available at: http://dx.doi.org/10.1145/2556288.2557323.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, p.160018.