Funded by the European Union

Project Number: 687818
Project acronym: O4C
Project title: OPEN4CITIZENS - Empowering citizens to make meaningful use of open data
Contract type: H2020-ICT-2015 - RIA

Deliverable number:	**D4.5**
Deliverable title:	**Data management plan**
Work package:	WP4
Due date of deliverable:	M6 – June 2016
Actual submission date:	30/06/2016
Start date of project:	01/01/2016
Duration:	30 months
Reviewer(s):	Marc Aguilar (i2CAT), Francesco Molinari (POLIMI), Nicola Morelli (AAU), Grazia Concilio (POLIMI)
Author/editor:	Anne Sofie Juul Sørensen (Dataproces)
Contributing partners:	Dataproces

Dissemination Level of this Deliverable:	**PU**

Public	*PU*
Confidential, only for members of the consortium (including the Commission Services)	*C0*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687818. Further information is available at www.open4citizens.eu.

Document history

Version nr.	Date	Authors	Changed chapters
0.1	15/05/2016	Anne Sofie	Rough draft of contents
0.2	09/06/2016	Anne Sofie	1st version send to review
0.3	15/06/2016	Anne Sofie	Iteration based upon feedback
0.35	20/06/2016	Marc Aguilar (i2CAT), Francesco Molinari (POLIMI)	Internal Review of deliverable
0.4	28/06/2016	Anne Sofie	Final version
1.0	30/06/2016	Anne Bock, Nicola Morelli (AAU)	Final revision and submission to EC

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

List of Figures

List of Tables

Glossary

Acronym	Definition
[Mobile or Web] App	A self-contained program or piece of software, especially designed to be downloaded by a user on a mobile device or personal computer.
Challenge	A widespread call to action to participate in an open contest (like a Hackathon) for improving or renovating an existing situation.
Citizen	An inhabitant of a particular town or city.
Hackathon	A collaborative pressure-cooker event (see definition), most often involving software developers and domain experts, which is typically lasting several days and is aimed to the production of one or more apps.
Non-expert user	A person without professional or specialized knowledge in a particular subject (usually computer programming).
O4C approach	An approach based on the progressive interaction between three different activities: 1) Explore; 2) Learn/apply; 3) Consolidate.
Open Data	Data that can be freely used re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. Source: http://opendatahandbook.org/guide/en/what-is-open-data/
OpenDataLab	In Open4Citizens, a virtual/physical playground, being a point of reference for local citizens, interest groups, grassroots communities, service providers, IT experts, start-up businesses and students wanting to have access to Open Data to generate new services and applications
Public service	A service provided by a government body to people living within its jurisdiction, either directly (through the public sector) or through financing a third party (agency or subcontractor).

1. Executive Summary

This deliverable contains the first version of the Data Management Plan (M6). It describes the nature of the data that will be collected and used in the Open Data Lab Platform. The open data has been divided into two overarching data categories:

1) Open Data
    - **Category A**: *Open and ready to use*
    - **Category B**: *Open - but not yet in a ready to use format*
2) Closed Data
    - **Category C**: data that will not be used in this project

It also describes the overall architecture of the OpenDataLab Platform that will contain the data and touch base with the current development on how the data will be stored, made available, shared, reused and used during and after the project. It also outlines the ethics related to working with Open Data as well as presenting an outlook for the further work with the deliverable.

2. Introduction

The purpose of this deliverable is to 'provide an analysis of the main elements of the data management policy that will be used in the project with regards to all the datasets that will be generated by the project' as stated in the document 'Guidelines of data Management in Horizon 2020[1]. According to the DoA (p. 37), the aims of this deliverable (in M6) are to 'describe the nature of the data that will be collected and integrated and the way it will be stored, made available and used during and after the project.' According to the 'Guidelines of data Management in Horizon 2020[2]' p.5, the Data Management Plan should be written from a dataset-by-dataset point of view. However since we do not yet have the specific datasets this initial draft of the Data Management Plan will focus on describing the overarching data categories that will be encountered during the hackathon cycles as well as describing the overall architecture of the OpenDataLab Platform that will contain the data. As each pilot comes closer to defining its specific challenges during the pre-hack phase, we will gain a clearer picture on which datasets to collect and upload to the OpenDataLab Platform. This will contribute to shaping the next editions of this deliverable, which are contractually bound to appear by Month 15 and Month 30.

[1] Guidelines on Data Management in Horizon 2020, Version 2.0, October 2015
[2] Guidelines on Data Management in Horizon 2020, Version 2.0, October 2015

2.1. Overview on structure

This initial draft of the Data Management Plan focuses on describing the overarching data categories that will be encountered during the hackathon cycles followed by how each category will be managed.
Firstly the two data categories are presented in section 3.0. The subsections describe the origin of the data as well as the possible outcome of using the particular category (concept, mobile application, use case, etc.). Then the structure of the Open Data Lab Platform where the data will be managed is described in section 4.1. The subsections elaborate on access, data standards, data sharing, archiving and preservation, reuse of the data and ethics. Lastly the outlook for the development is presented.

3. Description of the Data Categories

A large quantity of data concerning e.g. urban environments, weather, census, etc. is being produced every day. Such data can be found, for example as published government data sets, in private companies who have recorded the behaviour of their customers or by scientific research, or generated by users, who often record their activities and in some cases share such data with friends or social networks. The term 'Open Data' refers to those data sets that are made publicly available. These data have a large potential to generate new applications and to enhance several aspects of human life, including transport, healthcare, climate and even human behaviour. The way to manage these data as well as realising the real potential of open data is still to be fully discovered.
The O4C project intends to contribute to create a demand for applications based on open data and to include citizens as a driver for innovation in the open data arena. However many citizens are not necessarily able to handle Open Data, or even to imagine what to do with them. To help close this gap between users and data, a clarification of the different types of data is needed. In general 'Open Data' means that the data is publically available and can be used, modified, and shared freely by anyone for any purpose[3]. With this in mind the following categories have been created:

3) Open Data
    - Category A: *Open and ready to use*
    - Category B: *Open - but not yet in a ready to use format*
4) Closed Data
    - Category D: data that will not be used in this project

[3] http://opendefinition.org/

The categories will be described in detail in the following sections. The colour: green, yellow and red corresponds to the level of employment of the data.

- Green: the data is ready to use.
- Yellow: the data needs to be converted to a usable format or there are restrictions on the use
- Red: the data is not available in the O4C project

3.1. Open Data - Category A

The general topic for this category is that the data can be classified as open. It means that it can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike[4].

Category	Public Data	User Generated Data	Datasets generated by Hackathon Participants
Category A: Open and ready to use	Publicly available datasets that are available in a structured format and ready to be used in the Platform E.g. CSV - Comma Separated Values	Basically any content that users of a service have created that is accessible through a graphical user interface Examples: - GPS data	Datasets that have been generated by the users for the O4C project Merged datasets also belongs in this category

Table 1: Category A - Open and ready to use data

3.1.1 Public data

The data is presented in a repository – an online platform – owned and maintained by e.g. a company, an organization or by local, municipal or national governments who have decided to make a selection of data available to the public. Examples on this type of data may include data related to demographics or urban planning, as well as to public transport or real-time road usage. Data from this category can be located through repositories like Open Data DK, Humanitarian Data Exchange, DR Archives, HIP.se – Health Innovation Platform, etc. and is typically presented in a structured CSV format. If the data has a restricted/limited use or isn't free of charge it belongs in Category B.

[4] http://opendatahandbook.org/guide/en/what-is-open-data/

3.1.2 User Generated Data

When a user creates content that will be visible to the general public they usually agree to this in the Terms of Use of the service. Examples for this category are posts that are publically available via a graphical user interface on e.g. Twitter. If the data has a restricted/limited use or isn't free of charge it belongs in Category B.

3.1.3 Data Generated by Hackathon Participants

This category refers to datasets that have been created during the hackathons. The datasets can e.g. come from merging 'Public' or 'User Generated Data' into new datasets. Since the data in this category is created from Open Data there are no restrictions related to use, modification or sharing of the data.

3.1.4 Possible Outcomes and Available Datasets

It will be possible to manipulate Data from this category in the Open Data Lab Platform. The outcome from this is mainly helping the hackathon participants gain understanding of Open Data as well as aiding the development of new services/improve existing ones during the hackathon cycle.
The data can be used as:

- Components in mobile or web applications
- Concepts - i.e. mockups of mobile or web application
- Data examples for the participants to gain a greater understanding of Open Data
- Some data may not be off limits because of privacy reasons, as explained in category D, but due to the fact that the owners of the datasets have not yet made them available. Thus existing Open Data can be used as components in use-cases - proof of concept - that can clear the way for opening data that is not currently available as open data.

3.2. Open Data - Category B

This category refers to data that needs to be converted or have certain restrictions on its use. It means that the data is publically available but it is not ready-to-use as the data from category A.

Category	Public Data	User Generated Data	Datasets generated by Hackathon Participants
Category B: Open - but not yet in an optimal format	Publically available unstructured data like: - PDF - Addresses - Menus - Library Catalogues Formats that need to be converted. As well as CSV files with restricted downloads (e.g. number of rows/file size)	Data generated by everyday users of services and technology. Examples: - Twitter	Hackathon participants creating datasets as PDF or XLS files or merging of datasets with restricted use

Table 2: The three sub-categories of Category B

3.2.1 Public Data

Public data refers to the often *unstructured* data that is publically available like online menus, libraries of information, catalogues, etc. Data from this category is typically owned by a for-profit company or an organization (for-profit or non-profit). This category also includes data owned by public institutions, which have not published the data in a format that makes them immediately useable (e.g. PDF)

The unstructured data requires more preparation than structured data, since it will have to be converted into a useable format and perhaps be extracted from the websites by using Dataproces' software robots.

If the use is restricted it means that there can be a limited number of downloads. The restriction can also refer to the ownership of the data, which means that the data is free to use but is owed by someone.

3.2.2 User Generated Data

User generated data is data that is quite literally generated directly by users of everyday services and technology. Examples from this category can be GPS data, Geo-tagged Twitter data. Also for this category the unstructured data requires more preparation than structured data, since it will have to be converted into a useable format and perhaps be extracted from the websites by using Dataproces' software robots. Restricted use can also apply which means that there can be a limited number of downloads or a payment is required to use the data. The restriction can also refer to the ownership of the data - that the data is free to use but is owed by someone.

3.2.3 Data Generated by Hackathon Participants

This category encompasses Hackathon participants creating datasets as PDF or XLS files or merging of the Open Data that have restricted use.

3.2.4 Possible Outcomes and Available Datasets

It is more time consuming to get hold of the data from Category B than Category A. It is possible to retrieve, collect or create datasets during the pre-hack, hack or post-hack phase but as a general rule the possible final outcome when using data from this category is *concepts*. If the data is converted it "moves" to category A. Data from this category can be used as concepts - i.e. mock-ups of mobile or web applications.

3.3. Closed data - Category C

Data that is not open refers to personal, medical and other sensitive data that is not open to the public and cannot be found by the general public online. It also encompasses data that is only meant as internal information to public administration. Data from this category will *not* be available in the O4C project. The same rule applies to data that are only available after purchase, such as business data, etc.

Category	Description
Category C: Closed data - data that will not be used in this project	Any data that contain personal information e.g. medical data or data that is only meant as internal information to public administration

Table 3: Category C

4. The OpenDataLab Platform

The OpenDataLab will be the place or playground where citizens can make their ideas more concrete. The Platform will help support this vision by leaning against the O4C approach: Explore, Learn/Apply, Consolidate and letting the participants build on different levels with Open Data. Starting out with building an understanding of Open Data, then moving on to building concepts and finally building applications and possibly a viable business (figure 1).

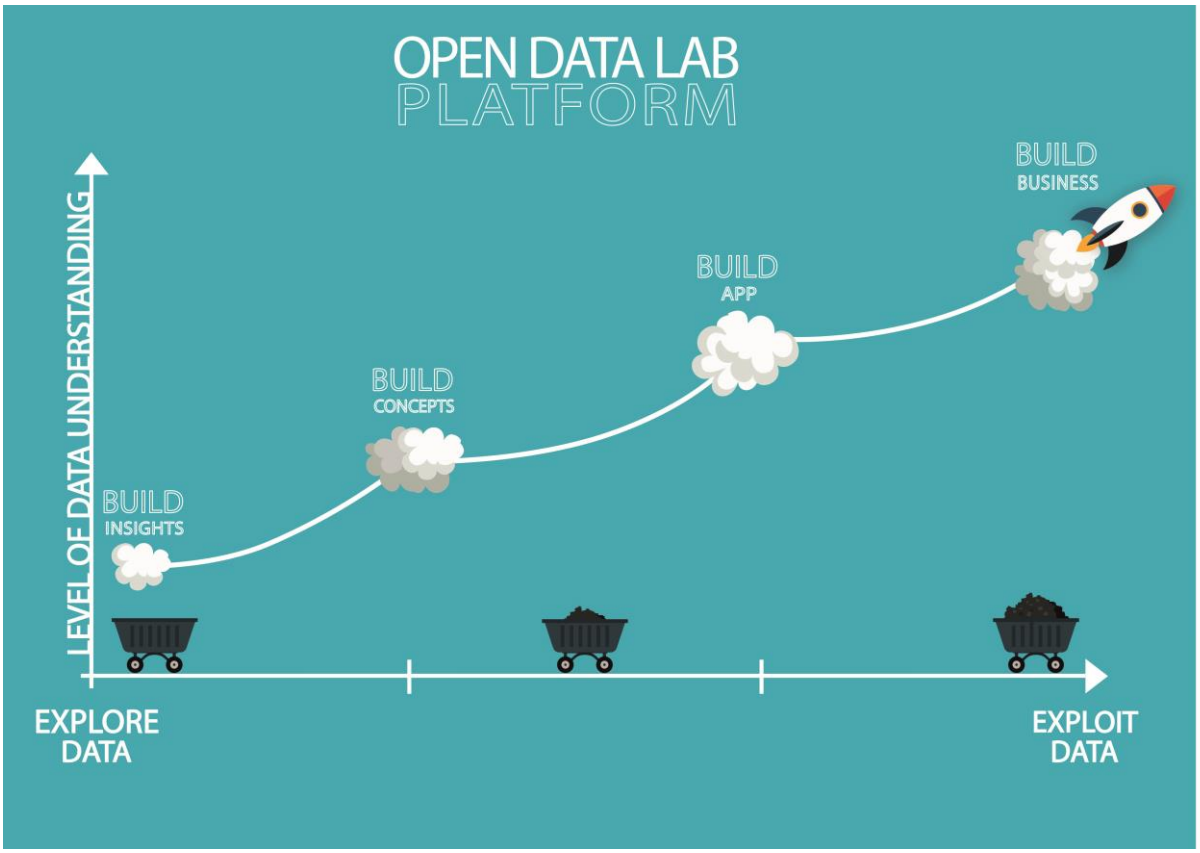

Figure 1: How the Open Data Lab Platform supports the O4C approach.

The figure (1) shows an ideal process. In real life the path will be more intricate and winding – however, one of the goals of the platform, as well as the O4C project, is "straightening" out the process – helping citizens getting data and getting there!
By helping citizens concretize their ideas, concepts and applications can be developed and where the data is not yet open a well-founded use-case can be created and used as an argument to ask the data owner to open the data. In order to let the participants to build with Open Data both datasets and selected tools will be available inside the Open Data Lab Platform.

The basic function of the Platform is to allow citizens to work with Open Data. In order to work with Open Data you must have access to datasets and you must have access to relevant tools.

Figure 2 shows a representation of the content of the Platform. It shows the relationship between the local hackathon teams/the hackathon participants (the box to the right – figure 2) and the Platform (the box to the right – figure 2).

- Upload of datasets: The local hackathon teams will at all times be able to upload datasets to the platform while the hackathon participants - for now - will not have this opportunity. This division of roles is a measure to ensure relevance and quality of the uploaded datasets. It also helps ensure that there is sufficient storage space so that the platform will run smoothly.
- Access to the uploaded datasets: The Open Data Lab Platform contains five data storages - one storage for each location/country - and a selection of tools to work with Open Data. The participants at the hackathon can then access the platform, fetch datasets and use the tools to work with the Open Data. For now, each local hackathon team will only be able to access its own data storage. This is again to ensure quality, structure and stability. At a later time a feature might be added to the platform that allows mixing datasets from the different local hackathon teams.
- The platform will be linked to a developer toolbox from IBM, called Bluemix. Arrangements are being made at the time of the editing of this deliverable, to make sure that the toolbox could be linked to the platform, in order to give access to a large range of tools to analyse, search, visualise and working with data.

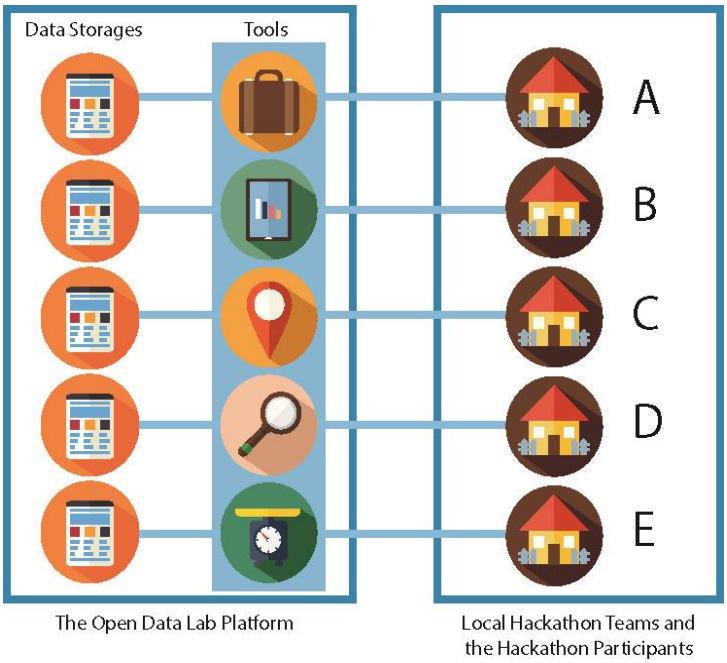

Figure 2 The OpenDataLab platform structure

4.1. Access to the Data inside the Open Data Lab Platform

All hackathon participants – and anyone else who accesses the Open Data Lab Website www.opendatalab.eu – can use the datasets and the tools inside.

4.2. Standards and metadata

Data standards for Category A (Table 1): CSV, XML, RDF, JSON, OIS, API queries, etc.
Data standards for Category B (Table 2): formats that require transformation to become machine readable such as PDF, JPG, TIFF, etc.

4.3. Data sharing

The datasets that are created during the hackathons will be shared through the Open Data Lab website www.opendatalab.eu. The essence of Open Data is that it is open – which means that anyone can use it without asking permission or informing the data owner about how or where it is being used and for what purpose. The datasets that belong to Category A are therefore embraced by unrestricted and unlimited use as well as unlimited sharing and manipulation. The datasets from Category B might present some restrictions in terms of usage. This has been elaborated upon in section 3.2.1.

4.4. Archiving and preservation (including storage and backup):

The plan is that all datasets that are uploaded to the Open Data Lab Platform will be stored on a server at Dataproces who will ensure preservation and backup throughout the project.

4.5. Tools

The tools inside the platform will allow the participant to perform different actions, including (among others):

- visualise data,
- build understanding through data examples and use cases
- construct concepts and mobile applications

The specific data visualisations tools will be described in detail in the next deliverable (M15). In order to build mobile applications the participants will additionally have access to Bluemix. This toolbox includes a large number of tools to handle, transform and analyse data, including tools to:

- transform data in HTML, PDF and Word format into Json and other appropriate formats
- create on-demand relational databases
- create apps for Web or mobile
- perform geospatial analysis

The work on the specific cases and the definition of detailed requirements will give the project team the possibility to better define and select the most useful tools.

4.6. Reuse of the data

The Open Data Lab Platform will form a virtual component of the five Open Data Labs. The platform is also meant to be active after the O4C project has been completed. This means that the Open Data that has been collected, generated and uploaded to the Platform during the project lifetime will be accessible both after each hackathon cycle and after the end of the funding period of the O4C project. This will be elaborated in the upcoming version of the Data Management Plan.

4.7. Ethics

When organising the events Open4Citizens collects information from public repositories, which contain Open Data. Open Data consist of information databases that are public domain, and therefore data that can be freely used and redistributed by anyone. Open4Citizens is thus not subject to any regulations regarding confidential or sensitive data storage, including principles stipulated in The Data Protection Directive 95/46/ECP and the General Data Protection Regulation (EU) 2016/679.

When providing Open Data for the events, Open4Citizens, its employees, and any contributing partners of Open4Citizens or its employees, shall not be liable for any harm arising from the use of the collected datasets shared through the Open Data Lab Platform, including but not limited to, how participating parties handle and develop the Open Data available on the Open Data Lab Platform.

In regards to the Open data from various data sources that is made available on the Open Data Lab Platform, Open4Citizens does not guarantee that this data has been published with the prior, necessary and informed approval that it requires. However the Open4Citizens team and the hackathons' team will verify the reliability of the source of the publication case by case.

5. Conclusions and Outlook

As mentioned in the preface this first version of the Data Management Plan presents the overarching categories and how data will enter into the Open Data Lab Platform. The next iteration of the Data Management Plan (M15) will provide a more detailed description of specific datasets while the final version (M30) will present the complete and detailed description of datasets and the final architecture of the platform. The next iteration will also elaborate on the concrete tools that will be available in the Open Data Lab Platform.