hackAIR

OPEN PLATFORM

D1.4: 2nd Data Management Plan

WP1 – Project Management

Document Information

Grant Agreement Number	688363	Acronym	hackAIR	
Full Title	Collective awareness platform for outdoor air pollution			
Start Date	1st January 2016	Duration	36 months	
Project URL	www.hackAIR.eu			
Deliverable	D 1.4 – 2nd Data Management Plan			
Work Package	WP 1 – Project management			
Date of Delivery	Contractual	30 June 2017	Actual	30 June 2017
Nature	Report	Dissemination Level	Public	
Lead Beneficiary	DRAXIS Environmental S.A.			
Responsible Authors	Polimachi Simeonidou (DRAXIS), Ioulia Anastasiadou (DRAXIS), Christodoulos Keratidis (DRAXIS), Panagiota Syropoulou (DRAXIS)			
Contributions from	Philipp Schneider (NILU), Sonja Grossberndt (NILU), Laurence Claeys (VUB), Gavin McCrory (VUB), Wiebke Herding (ONSUB), Eleftherios Spyromitros-Xioufis (CERTH), Marina Riga (CERTH), Arne Fellermann (BUND)			

Document History

Version	Issue Date	Stage	Description	Contributor
1.0	25/5/2017	Draft	Request input from partners	Polimachi Simeonidou (DRAXIS)
2.0	5/6/2017	Draft	Acquisition of partners' input	Philipp Schneider (NILU), Sonja Grossberndt (NILU), Laurence Claeys (VUB), Gavin McCrory (VUB),Wiebke Herding (ONSUB), Eleftherios Spyromitros-Xioufis (CERTH), Marina Riga (CERTH)
3.0	20/5/2017	Draft	Integration of partners input	Ioulia Anastasiadou (DRAXIS), Christodoulos Keratidis (DRAXIS), Panagiota Syropoulou (DRAXIS)
4.0	28/6/20017	Draft	Internal review	Arne Fellermann (BUND)
5.0	30/6/2017	Final	Integration of internal review comments	Panagiota Syropoulou (DRAXIS)

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Table of Contents

1 Executive summary

The present document is a deliverable of the hackAIR project, funded by the European Commission’s Directorate – General for Research and Innovation (DG RTD), under its Horizon 2020 Innovation Action programme (H2020).

The deliverable presents the second version of the project Data Management Plan (DMP). This second version lists the various datasets that will be produced by the project, the main data sharing and the major management principles the project will implement around them. Thus, the deliverable includes all the significant changes such as changed in consortium policies and any external factors that may influence the data management within the project. It is submitted on Month 18 as a Mid –Term review of the hackAIR Data Management Plan. A third version of this document will be delivered by the end of the project in December 2018 (D1.5: 3rd Data Management Plan).

The deliverable is structured in the following chapters:

- Chapter 1 includes an introduction to the deliverable
- Chapter 2 includes a description of the methodology used
- Chapter 3 includes the description of the datasets along with the documented changes and additional information

2 Introduction

The Data Management Plan (DMP) is an essential document for the hackAIR project that addresses issues related to data management. By creating an earlier plan for managing data at the beginning of the project and updating it on a regular basis the consortium will save time and effort later on.

This deliverable D1.4: 2nd Data Management Plan aims to document all the updates of the hackAIR project data management life cycle for all datasets to be collected, processed or generated. A description of how the results will be shared, including access procedures and preservation according to the guidelines in Horizon 2020 projects. This is a living document and it evolves and gains more precision and substance during the lifespan of the project.

Although the DMP is being developed by DRAXIS, its implementation involves all project partners' contribution. The next version of the DMP, to be published in M36, will describe more in detail the practical data management procedures implemented by the hackAIR project.

3 Methodology

The DMP methodology approach that has been used for the compilation of the D1.4 has been based on the updated version of the "Guidelines on FAIR Data Management in Horizon 2020"[1] version 3.0 released on 26th of July 2016 by the European Commission Directorate – General for Research & Innovation.

The updated version of the hackAIR DMP addresses the following issues:

1 Data Summary

The Data Summary addresses the following issues:

- Outline the purpose of the collected/ generated data and its relation to the objectives of the hackAIR project
- Outline the types and formats of data already collected/ generated and/ or foreseen for generation at this stage of the project
- Outline the reusability of the existing data
- Outline the origin of the data
- Outline the expected size of the data
- Outline the data utility

2 FAIR data

According to the "Guidelines on FAIR Data Management in Horizon 2020" the DMP applies to the following types of data:

- Making data findable, including provisions for metadata
    - Outline the discoverability of data (metadata provision)
    - Outline the identifiability of data and refer to standard identification mechanism
    - Outline the naming conventions used
    - Outline the approach towards search keyword
    - Outline the approach for clear versioning
    - Specify standards for metadata creation (if any).
- Making data openly accessible
    - Specify which data will be made openly available and if some data is kept closed explain the reason why
    - Specify how the data will be made available
    - Specify what methods or software tools are needed to access the data, if a documentation is necessary about the software and if it is possible to include the relevant software (e.g. in open source code)
    - Specify where the data and associated metadata, documentation and code are deposited
    - Specify how access will be provided in case there are any restrictions.
- Making data interoperable
- Specify what data and metadata vocabularies, standards or methodologies will be followed in order to facilitate interoperability
- Address whether standard vocabulary will be used for all data types present in the dataset in order to allow inter-disciplinary interoperability

[1] European Commission, (26 July 2016*), Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0

- Increase data re-use
- Specify how the data will be licensed to permit the widest reuse possible
- Specify when the data will be made available for re-use
- Specify if the data produced and/ or used in the project is useable by third parties, especially, after the end of the project
- Provide a data quality assurance processes description
- Specify the length of time for which the data will remain re-usable

3 Allocation of resources

The objectives of this point address the following issue[2]:

- Estimate the costs for making the data FAIR and describe the method of covering these costs
- Identify responsibilities for data management in the project
- Describe costs and potential value of long term preservation

4 Data security

This point will address data recovery as well as secure storage and transfer of sensitive data.

5 Ethical aspects

This point will cover the context of the ethics review, the ethics section of the DoA and ethics deliverables including references and related technical aspects.

6 Other issues

Other issues will refer to other national/ funder/ sectoral/ departmental procedures for data management that are used.

The hackAIR project coordinator (DRAXIS) has provided on time all the work package leaders and rest of the partners with a template that includes all the above-mentioned issues along with instructions to fill the template.

[2] http://ec.europa.eu/agriculture/direct-support/iacs/index_en.htm

4 Datasets in hackAIR

4.1 Datasets in WP1 – Project management (DRAXIS)

For the purposes of WP1 the following dataset has been generated and will be updated if needed: Contact details of project partners and advisory board.

DMP component	Issues to be addressed
Data Summary	The database contains name, organisation and contact details for all project partners and advisory board members.
Making data findable, including provisions for metadata	Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP1_1_ Contact details of project partners and advisory board The data is stored in a simple table, with the following fields: - Name - Category (1: Partner, 2: Advisory Board Member) - Short description - Location - Link - Email - Comments Additional fields will be added as the project progresses.
Making data openly accessible	No
Making data interoperable	CSV
Increase data re-use	No
Allocation of resources	N/A
Data security	The data is collected for internal use in the project, and not intended for long-term preservation. The work package leader is keeping a quarterly backup on a separate disk.
Ethical aspects	The data is confidential for internal use within the consortium (as personal data is involved). The contact database is managed in a Google Spreadsheet, and linked from the project internal wiki.
Other issues	N/A

4.2 Datasets in WP2 - Analysis and requirements (VUB)

For the purposes of WP2 the following two datasets have been generated and will be updated if needed:

- User requirements (intake survey) data set
- User requirements (workshop) data

4.2.1 User requirements (intake survey) data set

DMP component	Issues to be addressed

Data Summary	The goal of the data collection is to be able to recruit the population that will co-create the hackAIR platform during the user requirement phase of the project. As hackAIR has specific user groups in mind for the usage of the application, it is important to attract a specific group of potential users for the co-creation session. The data are collected on an .xls file with information on demographics (gender, birth year, occupation), device ownership and internet access, personal innovativeness (scale) and air quality awareness (scale). The generated and collected data will not be re-used. The data will be used by the social science researchers of the project and pilot responsible (VUB, NILU, BUND) to recruit the citizens and to contextualize the co-creation sessions. This data was gathered by 20 citizens by filling a survey (paper questionnaire).
Making data findable, including provisions for metadata	In case of a report or paper submitted for publication with peer review, all research findings will be integrated into the report or paper. Datasets will never be added to the publication. Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP2_1_user requirements (intake survey) data set
Making data openly accessible	Because the dataset is very limited and only interesting in relation to the recruitment process of the users, the data will not be made openly available.
Making data interoperable	N/A
Increase data re-use	This data will not be licensed. The data is only used for one specific purpose on one specific time period and will not be updated, nor re-used.
Allocation of resources	The data is only stored on the computer of one of the researchers. Within the project no budget is foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Publications will be made during the duration of the project, or after the project has finished.
Data security	Personalized information is only gathered on the computer of one researcher. After anonymization the data is shared with other consortium partners.
Ethical aspects	N/A
Other issues	N/A

4.2.2 User requirements (workshop) data

DMP component	Issues to be addressed
Data Summary	The user requirement (workshop) data set contains all relevant data for designing, running and analyzing the co-creation workshop, to define the user requirements of the hackAIR platform. Co-creation sessions are organized with citizens from different cities to discuss the platform and application idea of the consortium and to generate ideas and reflections from potential users. The types of the data are qualitative insights into: • Experiences, practices and expectations with regards to measuring and retrieving air quality information • Expectations with regards to the hackAIR platform

	• Evaluation of the hackAIR platform • Contact information (name, email): Only known to the local organizers (BUND, NILU) for recruiting purposes. Data was collected during co-creation workshops of 24 to 32 data subjects (12 to 16 data subjects in M6, and 12 to 16 data subjects in M10). All data subjects are citizens of Berlin or Oslo. All participants were coded (by using pseudonyms) in the processing and reporting of the research results. This means that real names are not associated in any way with the information collected or with the research findings from this study. Aggregated and pseudonymized research findings will be discussed in **scientific research publications.** Only participants who signed **the informed consent statement** at the start of the workshop participated. By signing this form, they gave permission for the use and disclosure of pseudonymized information for scientific purposes of this study at any time in the future and for the audio-recording of the workshop only for post-processing purposes. The size of the following data were from 20 citizens (data subjects): • Text (open and closed questions) • Audio records. The workshops were **audio-recorded** for post-processing. This tape will be used by the involved researchers (BUND and NILU) only for the processing of the workshop findings. It will only serve research purposes and it will by no means be released to other persons. The data will be used by the social science researchers of the project and pilot responsible (VUB, NILU, BUND) to recruit the citizens and to contextualize the co-creation sessions.
Making data findable, including provisions for metadata	In case of a report or paper submitted for publication with peer review, all research findings will be integrated into the report or paper. Datasets will never be added to the publication. Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP2_2_ user requirements (workshop) data set
Making data openly accessible	Because the dataset is very limited and only interesting in relation to the recruitment process of the users, the data will not be made openly available.
Making data interoperable	N/A
Increase data re-use	N/A
Allocation of resources	The data is only stored on the computer of one of the researcher. Within the project no budget is foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Publications will be made during the duration of the project, or after the project has finished.
Data security	Personalized information is only gathered on the computer of the partners. After pseudonomization the data is shared with other consortium partners or reported of in publications.

Ethical aspects	N/A
Other issues	N/A

4.3 Datasets in WP3 - Collective sensing models and tools (CERTH)

For the purposes of WP3 the following five datasets will be generated:

- Geotagged Images Dataset
- Web cams Dataset
- Environmental measurements Dataset
- Twitter_v0.1
- Look-up Table

4.3.1 Geotagged Images Dataset

DMP component	Issues to be addressed
Data Summary	The dataset contains user generated images that are publicly available on Flickr and geo-tagged in Europe. The images are collected with the purpose of being analyzed by specialized computer software that detects images with a sky region appropriate for air quality estimation and extracts pixel color statistics (i.e. mean R/G, G/B ratios) from that region. The computed statistics are then given as input to the air quality estimation model developed within the project. The images are downloaded, downscaled to a maximum size of 500X500 pixels and stored until image analysis is performed (<1 hour). After this process, the images are permanently deleted from our servers. All image metadata are permanently stored in a database. Given the current data collection rate (~5,000 images per day), the dataset is expected to grow to more than 1.5 million items (~1Gb) in one year.
Making data findable, including provisions for metadata	Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP3_1_Geotagged_Images. The metadata that will be provided for each image are the following: • URL (of the image on Flickr) • R/G and G/B ratios of sky part of the image • Geo-coordinates • Timestamp The original images can be retrieved from the corresponding URLs. The metadata produced will be available through GitHub, figshare or Zenodo (the latter two providing DOIs) that will make them discoverable and identifiable. The dataset will also be discoverable by querying conventional search engines (e.g. Google) with the dataset name.
Making data openly accessible	All the metadata of the images described above (i.e. URLs, ratios, geo-coordinates, timestamps) will be made openly available through GitHub, figshare or Zenodo. The original images cannot be shared due to Flickr's privacy and copyright policies.

	Data access does not require any specialized software as images can be retrieved from the corresponding URLs using a web-browser and metadata will be provided in a csv formatted file that can be accessed with any text editor.
Making data interoperable	The metadata will be available in text-based machine-readable format (e.g. csv) that will allow easy parsing and information exchange.
Increase data re-use	The data will be made available at the end of the project and will be licensed with an open data license that allows re-distribution and re-use of the data on the conditions that the creator is appropriately credited and that any derivative work is made available under "the same, similar or a compatible license" (e.g. CC-BY-SA-4.0).
Allocation of resources	The cost of long term preservation is negligible as data will be hosted in free open repositories. Long term preservation will facilitate long term usability of the data for development of air quality estimation methods.
Data security	Recovery of the data is facilitated through the hosting services provided by GitHub, figshare or Zenodo. No sensitive data is included.
Ethical aspects	N/A
Other issues	N/A

4.3.2 Web cams Dataset

DMP component	Issues to be addressed
Data Summary	The dataset contains images extracted from static outdoor webcams located in Europe. The webcams are geo-tagged and, therefore, the images are also geo-tagged and time stamped. The images are collected with the purpose of being analyzed by specialized computer software that detects images with a sky region appropriate for air quality estimation and extracts pixel color statistics (i.e. mean R/G, G/B ratios) from that region. The computed statistics are then given as input to the air quality estimation model developed within the project. The images are downloaded, downscaled to a maximum size of 500X500 pixels and stored until image analysis is performed (<1 hour). After this process, the images are permanently deleted from our servers. All image metadata are permanently stored in a database. Given the current data collection rate (~10,000 images per day), the dataset is expected to grow to more than 3.5 million items (~2.5Gb) in one year.
Making data findable, including provisions for metadata	Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP3_2_Webcams. The metadata that will be provided for each image are the following: • URL (of the webcam) • R/G and G/B ratios of sky part of the image • Geo-coordinates • Timestamp The original images can be retrieved (only for webcams that provide historical data)

	from the corresponding URLs. The metadata produced will be available through GitHub, figshare or Zenodo (the latter two providing DOIs) that will make them discoverable and identifiable. The dataset will also be discoverable by querying conventional search engines (e.g. Google) with the dataset name.
Making data openly accessible	All the metadata of the images described above (i.e. URLs, ratios, geo-coordinates, time stamps) will be made openly available through GitHub, figshare or Zenodo. The original images cannot be shared due to copyright terms set by the webcam owners. Data access does not require any specialized software as images can be retrieved from the corresponding URLs (only for webcams that provided historical data) using a web-browser and metadata will be provided in a csv formatted file that can be accessed with any text editor.
Making data interoperable	The metadata will be available in text-based machine-readable format (e.g. csv) that will allow easy parsing and information exchange.
Increase data re-use	The data will be made available at the end of the project and will be licensed with an open data license that allows re-distribution and re-use of the data on the conditions that the creator is appropriately credited and that any derivative work is made available under “the same, similar or a compatible license”. (e.g. CC-BY-SA-4.0).
Allocation of resources	The cost of long term preservation is negligible as data will be hosted in free open repositories. Long term preservation will facilitate long term usability of the data for development of air quality estimation methods.
Data security	Recovery of the data is facilitated through the hosting services provided by GitHub, figshare or Zenodo. No sensitive data is included.
Ethical aspects	N/A
Other issues	N/A

4.3.3 Environmental measurements Dataset

DMP component	Issues to be addressed
Data Summary	The dataset contains environmental measurements (PM10/PM2.5) that were published in environmental data portals such as (EEA, openAQ and luftdaten.info). The dataset contains information such as the exact geolocation to which the measurements refer, the time stamp and the air pollutant. The data concerns measurements from stations across the whole Europe. The collected data will facilitate visualizations of current air quality status that will be developed with the project, acting as an additional information layer. Moreover, they can serve as ground truth annotations for the development of air quality estimation models using supervised learning techniques. The data is stored in a regularly updated MongoDB database whose current size is 500,000 records.
Making data findable,	Following the naming convention Data_<WPno>_<serial number of

including provisions for metadata	dataset>_<dataset title>_<version> the dataset is named: Data_WP3_3_Environmental. Each data record contains the following information: • Data source (e.g. openAQ) • Geo-coordinates of the station • Timestamp of measurement • Pollutant (pm10 or pm2.5) • Value of measurement The data will be available (in csv format) through GitHub, figshare or Zenodo (the latter two providing DOIs) that will make them discoverable and identifiable. The dataset will also be discoverable by querying conventional search engines (e.g. Google) with the dataset name.
Making data openly accessible	All the data described above will be made openly available through GitHub, figshare or Zenodo. Data access does not require any specialized software as they will be provided in a csv formatted file that can be accessed with any text editor.
Making data interoperable	The metadata will be available in text-based machine-readable format (e.g. csv) that will allow easy parsing and information exchange.
Increase data re-use	The data will be made available at the end of the project and will be licensed according to the licensing terms of the corresponding data sources.
Allocation of resources	The cost of long term preservation is negligible as data will be hosted in free open repositories. Long term preservation will facilitate long term usability of the data for development of air quality estimation methods.
Data security	Recovery of the data is facilitated through the hosting services provided by GitHub, figshare or Zenodo. No sensitive data is included.
Ethical aspects	N/A
Other issues	N/A

4.3.4 Twitter_v0.1

DMP component	Issues to be addressed
Data Summary	The dataset contains Tweets gathered by tracking about 100 English keywords using Twitter's streaming API. The data is collected to facilitate (along with data in Data_WP3_2_Environmental_v0.2) the development of Twitter-based air quality estimation models using supervised learning techniques. These models may serve as additional estimation sources in cases where the number of image-based estimations is insufficient. The content of the tweets as well as metadata such as the location and language of the Twitter account and the geolocation of the tweet (if available) are stored in a MongoDB database whose current size is 12.5 million records.
Making data findable,	Following the naming convention Data_<WPno>_<serial number of

including provisions for metadata	dataset>_<dataset title>_<version> the dataset is named: Data_WP3_4_Twitter. The dataset will contain the predictive features extracted from all the tweets that will be collected during the monitoring period. These features correspond to statistics of the tweets posted at a given timestamp-location combination such as: • Number of tweets • Number of air quality-related tweets • Number of tweets referring to high air pollution • Number of tweets referring to low air pollution • Number of tweets that contains specific keywords In addition, we will make available a subset of 50,000 tweet objects as an example of the type of tweets that we collect. Tweet objects contain various fields[3], including information such as: • Tweet text • Geo-coordinates of the tweet (if available) • Timestamp of the tweet • Account location • Account language The data will be available (in csv format) either through GitHub, figshare or Zenodo (the latter two providing DOIs) that will make them discoverable and identifiable. The dataset will also be discoverable by querying conventional search engines (e.g. Google) with the dataset name.
Making data openly accessible	All the data described above will be made openly available through GitHub, figshare or Zenodo. Data access does not require any specialized software as they will be provided in a csv formatted file that can be accessed with any text editor.
Making data interoperable	The metadata will be available in text-based machine-readable format (e.g. csv) that will allow easy parsing and information exchange.
Increase data re-use	A version of the data that complies with Twitter's terms of service will be made available at the end of the project. If possible, the data will be licensed with an open data license that allows re-distribution and re-use of the data on the conditions that the creator is appropriately credited and that any derivative work is made available under "the same, similar or a compatible license" (e.g. CC-BY-SA-4.0).
Allocation of resources	The cost of long term preservation is negligible as data will be hosted in free open repositories. Long term preservation will facilitate long term usability of the data for development of air quality estimation methods.
Data security	Recovery of the data is facilitated through the hosting services provided by GitHub, figshare or Zenodo. All data come from Twitter's public stream and are therefore not considered sensitive.

[3] https://dev.twitter.com/overview/api/tweets

Ethical aspects	N/A
Other issues	N/A

4.3.5 Look-up Table

DMP component	Issues to be addressed
Data Summary	Radiative transfer calculations have been implemented using the SBDART (Santa Barbara DISORT Atmospheric Radiative Transfer) radiative transfer model in order to create a Look-up Table (LUT) with Red (700nm) / Green (550nm) band ratios and Green (550nm) / Blue band (450nm) ratios. The LUT consists of one ASCII file per geographical grid cell of 5x5 degrees (a total of 2592 files for the whole globe / ~150MB). Each ASCII includes R/G and G/B ratios on an hourly basis, per specified AOD_{550} bins, sky viewing angles and directions relative to the sun. The input data used for the radiative transfer calculations leading to the LUT are from the well-established MACV1 aerosol climatology (AOD_{550}, single scattering albedo and asymmetry parameter) and the ERA-Interim reanalysis (total column ozone, water vapour, surface albedo). The LUT is used in order to calculate the AOD550 levels (particulate pollution) in the atmosphere from photos which are available in social media. After calculating the R/G and G/B ratios of the photos the LUT allows for the attribution of the ratios to AOD550 values. The radiative transfer calculations are driven by a script code which allows for the automatic generation of LUTs of various spatial and temporal resolutions depending on the needs of the project. The size of the data is a few Mb only.
Making data findable, including provisions for metadata	Readme files are generated with the parameters included in the LUT files and the method followed for the production of the LUT. The LUT dataset is for use only within the project. The LUT is a product of collaboration between DRAXIS and DUTH and should remain available only to project members as the LUT could potentially be used in the future for other scientific or commercial activities. Hence, No DOI required. No versioning needed, updates of the data for any reason, will overwrite the original version.
Making data openly accessible	The LUT dataset is for use only within the project. The LUT is a product of collaboration between DRAXIS and DUTH and should remain available only to project members as the LUT could potentially be used in the future for other scientific or commercial activities.
Making data interoperable	The LUT data could be easily used in the future by atmospheric aerosol retrieval algorithms that use photos, sky camera images, etc. As already mentioned above, the LUT is a product of collaboration between DRAXIS and DUTH and should remain available only to project members as the LUT could potentially be used in the future for other scientific or commercial activities.
Increase data re-use	The LUT data should not be useable by third parties even long time after the end of the project. The LUT should only be used in the future for other scientific or commercial activities from DRAXIS and DUTH.

Allocation of resources	The LUT data should not be useable by third parties even long time after the end of the project. The LUT should only be used in the future for other scientific or commercial activities from DRAXIS and DUTH. Data are managed by DUTH and DRAXIS. Both have copies of the data. As the data are a few Mb only, no particular costs are associated with long-term preservation. Potential value of preservation will depend on the project outcome and potential spin-off uses.
Data security	Multiple backups of the data are stored in removable hard disks in different locations.
Ethical aspects	N/A
Other issues	N/A

4.4 Datasets in WP4 - Data fusion model and reasoning services (NILU)

For the purposes of WP4 the following three datasets has been generated:

- CAMS regional modelling results
- Data fusion maps
- OntologicalData_v0.1

4.4.1 CAMS regional modelling results

DMP component	Issues to be addressed
Data Summary	The WP uses concentration fields provided by the Copernicus Atmosphere monitoring service which guides the interpolation of the hackAIR observations in the data fusion module. The data is required to provide a reasonable estimate of air quality with complete spatial coverage. The type of data is NetCDF. The existing CAMS data is re-used and its origination is from http://atmosphere.copernicus.eu/. The size of the data is Ca. 20-50 MB per day and it will be useful internally for processing in the data fusion module.
Making data findable, including provisions for metadata	The data will be discoverable on http://atmosphere.copernicus.eu/ but not on public hackAIR servers. The naming conventions used are the standard CAMS conventions.
Making data openly accessible	This data itself in its original form is only used internally for processing. Value-added derivatives of the dataset will be shown on the hackAIR website. Some figures in public hackAIR reports might contain subsets of the data.
Making data interoperable	The interoperability of the date will be achieved through the Standard CAMS metadata.
Increase data re-use	Data will only be used internally. The quality assurance process that will be used is the specific QA which is carried out at CAMS. The data will be re-useable during the lifetime of CAMS.
Allocation of resources	N/A

Data security	The CAMS data is publicly available and as such cannot be considered as sensitive. The dataset will be used only on the hackAIR server and as such falls under the same security standards as the rest of the hackAIR observations.
Ethical aspects	N/A
Other issues	N/A

4.4.2 Data fusion maps

DMP component	Issues to be addressed
Data Summary	Data fusion maps will be created as part of the hackAIR project by combining hackAIR observations with CAMS modelling information. These maps will be displayed to the user through a leaflet-based interface. The data adds value to the hackAIR observations by interpolating them in space. The types of data are: • GeoTIFF • GeoJSON • Shapefile The expected size of the data will be a few MB per day. This data is useful especially for hackAIR users who are interested in air quality in locations where no observations are available.
Making data findable, including provisions for metadata	The data will be discoverable through the main hackAIR web interface. The spatial metadata will be available within the provided GeoTIFF files: they contain georeferenced information of the data, such as projection, bounding box, etc.
Making data openly accessible	The data will be made openly available. It will be made available through the main hackAIR web interface and only a web browser is needed. The data and associated metadata, documentation and code are deposited on the main hackAIR server.
Making data interoperable	Metadata will not be exposed to the users.
Increase data re-use	The data will be licensed with the standard hackAIR license and a quality assurance process will be carried out once the system is operational.
Allocation of resources	N/A
Data security	N/A
Ethical aspects	N/A
Other issues	N/A

4.4.3 OntologicalData_v01.

DMP component	Issues to be addressed
Data Summary	The dataset contains heterogeneous information, such as *user profile details* (gender, health status, preferred activities, etc.), as well as *environmental data*

	measurements. The nature of the data is either textual or numerical. The purpose of the collection of the aforementioned data is for describing in an integrated and formal way all information that are considered as important for providing personalised decision support services automatically, within the framework of the hackAIR project. The representation and storage of such information in the hackAIR knowledge base, follows the principles of the hackAIR ontological-based framework, implemented within WP4 (T4.2). We separate abstract information (schema) from actual realisations (individuals) by following a multi-layered approach in ontology development process: the structure and concepts are represented in the TBox (terminological component) and the applied content is represented in the ABox (assertion component). The ontology is described in OWL[4] (Web Ontology Language) and can be delivered as separate files in any of the known ontology formats (.rdf, .owl, .ttl). The dataset targeted to be represented via ontology notions is provided by the *user profile module* and the *data fusion module* of the hackAIR framework, whenever a request for personalised recommendation needs to be served. The initial content is provided in JSON format[5], a language-independent open data format that uses human-readable text to express data objects consisting of attribute-value pairs. These data will serve as input for the rule-based reasoning module, for providing personalised recommendations to the users.
Making data findable, including provisions for metadata	Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version>, the dataset is named: Data_WP4_1_OntologicalData _v0.1. Generally, any information stored in ontologies can be referenced/tracked via the unique resource identifier (URI) and the name assigned to the concept (class), instance (individual) or relation (semantics) of interest. However, in order for this to be feasible, data should be publicly available on the web. Concerning the hackAIR ontological framework, it has been implemented in separate layers, enabling thus the selective availability of the content. The schema described in the ontology (abstract information) will be publicly available for further adoption by other interested parties, following the ontology-reuse principles. The ontology schema will be kept in project servers as well as in online ontology repositories (ontohub[6], LOV[7]) for easier track, search and discoverability. On the other hand, the populated instances (actual information) will not be permanently stored in the ontology nor will they be publicly available due to the personal content of the represented data (sensitive information about actual users of the system).

[4] OWL is an ontology language designed for representing rich and complex knowledge about things, semantics and relations between them. Available details at: https://www.w3.org/OWL/

[5] http://json.org/

[6] https://ontohub.org/

[7] http://lov.okfn.org/dataset/lov

Making data openly accessible	As already mentioned, abstract (TBox) and actual (ABox) data will be separately represented in the ontological framework of the project. The TBox describing the ontology schema will be kept in both owned servers and online ontology repositories for easier track, search and discoverability. Actual data will be instantly populated in in-memory ontology models for further post-processing by the reasoning module. No direct public access is feasible. Results of the recommendation process will be available to the users only through the hackAIR UI and only for requests produced through the hackAIR UI utilities. Developed web-services will enable the triggering of the recommendation module under specific data provision and request. The integration between the recommendation module and other involved hackAIR modules will be achieved in a low level communication; no expert's knowledge is needed for interacting with the recommendation module, at a higher level, only access to the hackAIR utilities.
Making data interoperable	The structure and relations of abstract data stored in the ontology is based on the specification of ontology requirements (what, why and how the ontology aims to represent its content). Interoperability of hackAIR ontological data and of third party ontological concepts can be feasible via a *direct mapping* between relevant notions of the domains of interest. Connection with existing or new ontologies can be made with the use of common OWL/RDF representations: (i) the property owl:sameAs may be used to connect concepts from different ontologies that could be considered as the same; (ii) the property rdfs:subClassOf/rdfs:subPropertyOf may be used in order to inherit the semantics of existing super-class/property.
Increase data re-use	The ontology schema (TBox) will be publicly available for further adoption, reuse or even extension of the represented content into third party ontologies. Its data will be licensed with an open data license that allows re-distribution and re-use of the data on the conditions that the creator is appropriately credited and that any derivative work is made available under "the same, similar or a compatible license" (e.g. CC-BY-SA-4.0). The data will be available after the 1st stable version of the hackAIR platform.
Allocation of resources	The cost of long term preservation is negligible as abstract data (i.e. ontology schema) will be hosted in owned servers as well as in free open ontology repositories (ontohub, LOV). Long term preservation will facilitate long term usability and potential extensibility of ontology notions of the project's ontology.
Data security	Sensitive information will not be stored in the ontology; such data will be provided directly by relevant hackAIR modules (i.e. user profile, fused data) and will be populated in an in-memory ontology model for performing instant calculations (interpretation of relations and inference of new knowledge) throughout the recommendation process. Thus, no security issues arise.
Ethical aspects	N/A
Other issues	N/A

4.5 Datasets in WP5 - Development of the hackAIR platform (DRAXIS)

For the purposes of WP5 the following dataset has been generated:

- Architecture and Integration Framework Definition Specification
- Mock ups

4.5.1 Architecture and Integration Framework Definition Specification

DMP component	Issues to be addressed
Data Summary	Functional and non-functional requirements, hardware requirements, component descriptions (inputs & outputs), component dependencies, API descriptions, information flow diagram, internal and external interfaces and testing procedures. All technical partners were asked to answer a set of questions, based on which further online and face to face discussions took place in order to form the final document. This will be the basis upon which the system will be built.
Making data findable, including provisions for metadata	There are no specific standards or metadata associated with these types of data. The data are available in D5.1: Architecture and Integration Framework Definition Specification. The dissemination level of D5.1 is public. It is available through the hackAIR wiki for the members of the consortium and when the project decides to publicize deliverables, it will be uploaded along with the other public deliverables to the project website or anywhere else the consortium decides.
Making data openly accessible	It will become both discoverable and accessible to the public once the consortium decides to do so.
Making data interoperable	N/A
Increase data re-use	Could be used as example for engineers who want to build similar systems.
Allocation of resources	N/A
Data security	All data will be saved in the DRAXIS premises and will be shared with all partners using the hackAIR wiki.
Ethical aspects	N/A
Other issues	N/A

4.5.2 Mock ups

DMP component	Issues to be addressed
Data Summary	Mock-ups have been created by UI/UX experts in order to help the development team design an intuitive and easy to use application. The type of the files produced is .psd. The expected size of each document is 1,5 KB.
Making data findable, including provisions for metadata	The mock-ups might become both discoverable and accessible to the public once it is delivered to the EU and the consortium decides to do so. In any case given that the mock-ups depict the look and feel of the final application, once the hackAIR product is out on the market, everyone who uses it will be able to view them.
Making data openly	The mock-ups will not be openly available, before the completion of the hackAIR

accessible	platform.
Making data interoperable	N/A
Increase data re-use	N/A
Allocation of resources	N/A
Data security	All data will be securely saved in the DRAXIS premises.
Ethical aspects	N/A
Other issues	N/A

4.6 Datasets in WP6 - Engagement strategies for user participation (VUB)

For the purposes of WP6 the following three datasets will be generated:

- Engagement survey data set
- Engagement qualitative insights data set
- Social Media Monitoring

4.6.1 Engagement survey data set

DMP component	Issues to be addressed
Data Summary	The engagement survey data set will contain relevant quantitative data for designing, running and analysing the engagement and behavioural change strategies. The project aims at engaging people to use the hackAIR application, to upload data that can be used for measuring air quality (citizen science) and to achieve behavioral change (changes in belief, changes in knowledge and behavior changes). To be able to do this in a good way we should first learn about (1) the characteristics of the hackAIR populations, (2) the initial belief, knowledge and behavior of the population, (3) the belief, knowledge and behavior of the population after using hackAIR, and (4) measure the effect of the used engagement strategies. The types and formats of the data are SPPS format .sav/ MS Office .xls format. The following **data** will be collected: • Demographic information: only used for the profiling of the research participants • Air quality awareness (use of standardized scales) • Motivations to start and to continue engaging into hackAIR and other citizen science initiatives (use of standardized scales) • Self-reported data on the factors that measure behavioral change (use of standardized scales) • Contact information: Only collected and used for recruiting purposes. Anonymous and aggregated research data will be **published** in internal project reports (accessible to all consortium partners via the hackAIR project Wiki) and in external scientific research publications from the project. No re-use of the raw anonymised data is foreseen, but is possible after request. The origin of the data is through online surveys with 100 up to 150 citizens (data subjects) that will fill in the survey. The data will be used by the social science researchers of the project and pilot

	responsible (VUB, NILU, BUND).
Making data findable, including provisions for metadata	In case of a report or paper submitted for publication with peer review, all research findings will be integrated into the report or paper. Datasets will never be added to the publication. Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the dataset is named: Data_WP6_1_survey data set.
Making data openly accessible	As the data will be limited and only about certain areas of Europe, we don't foresee to make the raw data openly available. The outcomes will be reported in (open access) journals. But we are open to share datasets upon request.
Making data interoperable	N/A
Increase data re-use	N/A
Allocation of resources	Within the project no budget is foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Publications will be made during the duration of the project, or after the project has finished. The data is only stored on the computer of one of the researcher.
Data security	Personalized information is only gathered on the computer of one researcher. After anonymization the data is shared with other consortium partners
Ethical aspects	N/A
Other issues	N/A

4.6.2 Engagement qualitative insights data set

DMP component	Issues to be addressed
Data Summary	The aim of this data collection is to gather qualitative insights data on the engagement of citizens and scientists for citizen science purposes in the field of AQ. All relevant qualitative data for designing, running and analysing the engagement and behavioural change strategies. Generating insights on engagement and behavioral change strategies in air quality citizen science projects to use lessons learned to create better strategies in hackAIR. The types of the data are qualitative insights into: • Experiences, practices and expectations with regards to measuring and retrieving air quality information • Expectations with regards to the hackAIR platform • Evaluation of the hackAIR platform • Contact information (name, email): Only known to the local organizers (BUND, NILU) No re-use of data is foreseen. The origin of the data is the following: **Data subjects of the expert interviews:** • 3 up to 7 experts in citizen science AQ projects. Data will not be anonymised. **Data subjects of the focus groups**:

	• Amount: 10 to 20 test subjects • Users who have successfully been engaged into using the hackAIR platform during the pilot tests (identified in WP7). All participants will be coded (by using pseudonyms) in the processing and reporting of the research results. This means that real names will not be associated in any way with the information collected or with the research findings from this study. Aggregated and pseudonymized research findings will be discussed in **scientific research publications.** Only participants who sign **the informed consent statement** at the start of the focus group will participate. By signing this form, they give permission for the use and disclosure of pseudonymized information for scientific purposes of this study at any time in the future and for the audio-recording of the workshop only for post-processing purposes. The following **data** will be collected: • Demographic information: only used for the profiling of the research participants • Insights into current air quality awareness and into motivations to start and to continue engaging into hackAIR and other citizen science initiatives • Overall project and expert information • Contact information: Only collected and used for recruiting purposes. **Data collection method**: focus group, expert interviews **Data subjects**: • Amount: 30 test subjects (approximately) • European citizens and domain experts **Data type**: • Text (written notes) • Audio-records: The focus groups might be **audio-recorded** for post-processing. This tape will be used by the involved researchers only for the processing of the workshop findings. It will only serve research purposes and it will by no means be released to other persons. Focus groups will be pseudonymized, expert interviews will not be anonymised as all experts consented to the use of their responses in the deliverable. The data will be used by the social science researchers of the project and partners responsible for the pilot (VUB, NILU, BUND) to create better strategies for engagement and behavioral change.
Making data findable, including provisions for metadata	Following the naming convention Data_<WPno>_<serial number of dataset>_<dataset title>_<version> the datasets are named: Data_WP6_2_qualitative insights data set experts Data_WP6_2_qualitative insights data set focus groups
Making data openly accessible	The raw data will not be made openly available, but all insights will be shared in publications. We didn't ask the experts if they were willing to make the raw data of their interviews public available.

Making data interoperable	N/A
Increase data re-use	N/A
Allocation of resources	Within the project no budget is foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Publications will be made during the duration of the project, or after the project has finished. The data is only stored on the computer of one of the researcher.
Data security	Personalized information is only gathered on the computer of the social science research partner. After pseudonymisation the data is shared with other consortium partners or reported of in publications.
Ethical aspects	N/A
Other issues	N/A

4.6.3 Social Media Monitoring

DMP component	Issues to be addressed
Data Summary	The dataset contains posts collected by the hackAIR's social media monitoring tool from Twitter, Facebook, Google+ and YouTube (using the public APIs provided by these platforms). Posts are organized in air quality-related collections (sets of keywords and/or accounts) that are regularly updated (every 15-30 mins) by pulling latest content from the respective social media platforms. In addition, the dataset contains refined lists of accounts and posts of interest for hackAIR that are identified using text mining methods. The dataset supports hackAIR's dissemination and engagement activities through the discovery of online hackAIR-related communities to target. In addition, several statistics are calculated on top of the collected posts, facilitating measurement of audience reach and impact from the hackAIR channels (e.g. Facebook page, Twitter account). Although the collected posts are publicly available on the respective social media platforms, they may potentially include personal information. Therefore, we do not intend to make the dataset findable, accessible, interoperable and re-usable (FAIR). For the duration of the project, the data is stored in secure servers maintained by CERTH and access to project partners involved in hackAIR's dissemination and engagement activities is provided through a protected web interface. After the end of the project all the collected data will be deleted.
Making data findable, including provisions for metadata	N/A
Making data openly accessible	N/A
Making data interoperable	N/A

Increase data re-use	N/A
Allocation of resources	N/A
Data security	N/A
Ethical aspects	N/A
Other issues	N/A

4.7 Datasets in WP7 - Pilot operation and evaluation (NILU)

For the purposes of WP7 the following five datasets will be generated:

- PM measurements by Arduino hackAIR sensors
- PM measurements by PSOC hackAIR sensors
- PM measurements by Commercial off the Shelf (COTS) sensors
- Content generated through the hackAIR platform
- Photos uploaded through the hackAIR mobile app

4.7.1 PM measurements by Arduino hackAIR sensors

DMP component	Issues to be addressed
Data Summary	PM measurements by Arduino hackAIR-sensors. The Arduino hackAIR node measures the PM (2.5 and 10) concentration using one of the supported sensors. The values are provided in μg/m^3. For each value, the following metadata are provided: Date, time, location (using user's profile geolocation details), sensor id, and user id. The data packet that is submitted from the hackAIR Arduino sensor fulfills both the Ethernet and wifi protocol and may be connected to any AP in wired or wireless mode using the most common security protocols like WPA. The data structure and protocol is available in the technical deliverable that describes the hackAIR Arduino node. The data (PM measurements and metadata) will be available to other Researchers through the Open Access to Research Data Pilot. The measurements and metadata will be made available for use by the hackAIR applications through the secure API that we will create.
Making data findable, including provisions for metadata	The data along with the metadata are stored securely in DRAXIS servers. Database and storage areas are set to take differential daily backups and a complete weekly one keeping up to the three last weeks. These measurements are discoverable through the hackAIR platform, only for users who are registered through the platform.
Making data openly accessible	The measurements and metadata are made available for use by the hackAIR applications through the hackAIR platform. The data (PM measurements and metadata) are available to other researchers through the Open Access to Research Data Pilot. The measurements and metadata are made available for use by the hackAIR applications through the secure API that is created.

Making data interoperable	N/A
Increase data re-use	The datasets are saved for a long time after gathering them for statistical reasons and will be available to other research groups upon request. Measurements can be shared with other researchers along with the geospatial information and time they were made. Metadata will not contain any information of the users who contributed it to hackAIR.
Allocation of resources	N/A
Data security	Data from the sensors are submitted directly to the hackAIR platform databases through Internet. The most common security protocols are adopted on the Arduino implementation regarding the wireless communication (like WPA etc.).
Ethical aspects	N/A
Other issues	N/A

4.7.2 PM measurements by PSOC hackAIR sensors

DMP component	Issues to be addressed
Data Summary	PM measurements by PSOC-hackAIR sensors. The PSOC Bluetooth Low Energy (BLE) hackAIR node measures the PM (2.5 and 10) concentration using one of the supported sensors. The values are provided in μg/m^3. For each value, the following metadata are provided: Date, time, location (using the GPS unit of the smartphone), sensor id, and user id. The data packet that is submitted from the hackAIR PSOC sensor fulfills the Bluetooth low energy (BLE) protocol and may be captured by any Bluetooth mobile device that runs the hackAIR mobile phone application. The data structure and protocol is available in the technical deliverable that describes the hackAIR PSOC node. The data (PM measurements and metadata) will be available to other researchers through the Open Access to Research Data Pilot. The measurements and metadata will be made available for use by the hackAIR applications through the secure API that we will create.
Making data findable, including provisions for metadata	The data along with the metadata are stored securely in DRAXIS servers. These measurements are discoverable through the hackAIR platform, only for users who are registered through the platform.
Making data openly accessible	The measurements and metadata are made available for use by the hackAIR applications through the secure API that will be created. The data (PM measurements and metadata) are available to other Researchers through the Open Access to Research Data Pilot. The measurements and metadata will be made available for use by the hackAIR applications through the secure API that we will create.
Making data interoperable	N/A
Increase data re-use	The datasets are saved for a long time after gathering them for statistical reasons and will be available to other research groups upon request. Measurements can be

	shared with other researchers along with the geospatial information and time they were made. Metadata will not contain the information of the users who contributed it to hackAIR.
Allocation of resources	N/A
Data security	Data from the sensors are submitted in a beacon mode. Thus no specific security measures were taken. When the data are received from the mobile device that is responsible to send the data to the hackAIR databases after adding all the required metadata the security aspects are included in the data transmission protocol as well as in the packet structure that is sent over the GSM/ Wifi network.
Ethical aspects	N/A
Other issues	N/A

4.7.3 PM measurements by Commercial off the Shelf (COTS) sensors

DMP component	Issues to be addressed
Data Summary	PM measurements by Commercial off the Shelf (COTS) sensors. The COTS estimates the PM concentration using consumable materials in order to build the sensing device. Sequentially a mobile device that is capable of capturing photos and it is equipped with a macro lens is used to analyse a photo of the sensor by adopting computer vision algorithms. The data are analysed and the results are provided in five levels (low, mid-low, middle, mid-high, high). For each of the levels, the following metadata will be provided: Date, time, location (using the GPS unit of the smartphone), sensor id, and user id. The data packet that is submitted from the hackAIR COTS sensor are the captured image and the estimation of the level of PM concentration.
Making data findable, including provisions for metadata	The data along with the metadata are stored securely in DRAXIS servers. These measurements are discoverable through the hackAIR platform, only for users who are registered through the platform.
Making data openly accessible	The photo and metadata are made available for use by the hackAIR applications through the secure API that will be created. The data (PM level and metadata) are available to other researchers through the Open Access to Research Data Pilot. The measurements and metadata will be made available for use by the hackAIR applications through the secure API that we will create.
Making data interoperable	N/A
Increase data re-use	The dataset are saved for a long time after gathering them for statistical reasons and will be available to other research groups upon request. Measurements can be shared with other researchers along with the geospatial information and time they were made. Metadata will not contain the information of who contributed it to hackAIR.
Allocation of resources	N/A
Data security	N/A

Ethical aspects	N/A
Other issues	N/A

4.7.4 Content generated through the hackAIR platform

DMP component	Issues to be addressed
Data Summary	Data like users' personal information, user perception of the atmosphere on a given day, posts and comments will be generated by the hackAIR users via the platform. The data described above will be saved in the hackAIR central database. Detailed log of user actions (login, logout, account creation, visits on specific parts of the app) will be kept in the form of a text file. This log will be useful for debugging purposes. Reports containing information on user devices (browsers and mobile phones) as well as number of mobile downloads (taken from play store for android downloads and app store for mac downloads) will be useful for marketing and exploitation purposes, as well as decisions regarding the supported browsers and operating systems. No existing data will be reused.
Making data findable, including provisions for metadata	As part of any stored data, meaningful metadata (time and date of posts and comments, owner, dates for the logs and the user perception statement) will be generated to assist the discoverability of the data and related information. Discoverability will be possible only for the administrator of the app for all the data and for some of the data (user name, anonymous contributed user perception of the atmosphere, posts and comments) only for registered users. The database will not be discoverable to other network machines operating on the same LAN, VLAN with the DB server or other networks. Therefore only users with access to the server (hackAIR technical team members) will be able to discover the database.
Making data openly accessible	Some of the data produced by the platform (posts, comments, users' display names and users' perception of the atmosphere) will be accessible through the hackAIR platform, only for users who are registered through the platform. On the technical level only authorized hackAIR technical team members will have access to the database.
Making data interoperable	N/A
Increase data re-use	Specific logging data will be used to identify the level of participation of a specific group of test users to measure behavioral change (WP6 T6.2).
Allocation of resources	N/A
Data security	Any personal data will be anonymized and encrypted. The following standards will be used: • RSA for generating public keys • AES for private data encryption SHA hashes for storing passwords All data are transferred via SSL connections to ensure secure exchange of

	information. Database will be set to take daily backups and a complete weekly one keeping up to the three last weeks.
Ethical aspects	N/A
Other issues	N/A

4.7.5 Photos uploaded through the hackAIR mobile app

DMP component	Issues to be addressed
Data Summary	hackAIR users will be asked to upload pictures depicting the sky. These pictures will be geotagged and time stamped and will be stored in the hackAIR DB or a hackAIR designated secure network storage area. They will be used to calculate the AOD for the specific location.
Making data findable, including provisions for metadata	As part of any stored data, meaningful metadata (image URL, timestamp and location) will be generated to help internal users locate the data and related information. Those data will be saved in the hackAIR database.
Making data openly accessible	The data (metadata and photos) will be available to other researchers through the Open Access to Research Data Pilot. The photos and metadata will be made available for use by the hackAIR applications through the secure API that we will create. For cases that there will need to be a mass use of photos (e.g. for conversion reasons) applications will be granted the right to read only from the storage area and bypass the API. These images might (a pending decision) be accessible through the hackAIR platform, only for users who are registered through the platform. Discoverability will be possible only for the administrator of the app for all the data. Some of the images might be posted to social media networks like Instagram only if the user authorizes that action.
Making data interoperable	Anyone who wants to have access to the photos and the metadata will be able to use them as a whole or a subset of it as it fits to his/her purpose.
Increase data re-use	The dataset will be saved for a long time after gathering them for statistical reasons and will be available to other research groups upon request. Images can be shared with other researchers along with the geospatial information and timestamp the images were taken at. Images will not contain the information of who contributed it to hackAIR.
Allocation of resources	N/A
Data security	The photos along with the metadata will be stored securely in the hackAIR database and/or storage system. The photos might be resized and changed in format in order to reduce storage but also be able to be reused by hackAIR if needed. Database and storage areas will be set to take daily backups and a complete weekly one keeping up to the three last weeks. The database will not be discoverable to other network machines operating on the same LAN, VLAN with the DB server or other networks. Therefore only users with access to the server (hackAIR technical team members) will be able to discover the

	database. The images folder will not be discoverable by systems or persons in the same or other servers in the same LAN/VLAN as the storage/database server. On the technical level only authorized hackAIR technical team members will have access to the database and database storage.
Ethical aspects	Photos will not be shared with other users unless the person who uploaded the photo wants to.
Other issues	N/A

4.8 Datasets in WP8 - Communication, Dissemination and Exploitation (ONSUB)

DMP component	Issues to be addressed
Data Summary	Work package 8 (Communication, Dissemination and Exploitation) does not generate research data. For the purposes of implementing and monitoring the communications strategy, the work package leader manages the following two datasets: 1. Contact database: The database contains name, organisation and contact details for all relevant contacts of the project. This includes members of the network of interest, related initiatives, participants of events, recipients of the newsletter etc. 2. Communications monitoring data: The database contains quarterly information on key communications indicators, including the number of stakeholders on contact list, visits to project website, newsletters, social media impressions, media impressions and events. Data sources include: automatic monitoring of Google Alerts, Twitter Analytics, Piwik and partner reports. Both datasets are confidential for internal use within the consortium (as personal data is involved). Data is managed in Google Spreadsheets and can be exported as an Excel or CSV file. Each is expected to contain between around 1.000 entries by the end of the project.
Making data findable, including provisions for metadata	The data is collected for internal use in the project, and not intended to be findable or accessible to parties external to the project.
Making data openly accessible	Contact data will not be made openly available in line with privacy protection legislation. A summary of the communications monitoring data is made available as part of the regular project reports.
Making data interoperable	While not intended for interoperability due to its confidential nature, the data is stored in simple tables that can be exported in CSV format.
Increase data re-use	The data is not intended to be shared or re-used.
Allocation of resources	As the work package only manages confidential data sets, no further costs are involved in making the data FAIR.
Data security	The data is collected for internal use in the project, and not intended for long-term preservation. The work package leader is keeping a quarterly backup on a separate

	disk.
Ethical aspects	Before receiving the regular newsletter of hackAIR, all contacts on the contact list have opted-in to receive the information.
Other issues	CSV

5 Conclusion

This second deliverable reflects the updated procedures to be implemented by the hackAIR project to efficiently manage the data it will produce. In particular, the 2nd DMP anticipates the data management strategy regarding the collection, management, sharing, archiving and preservation of data. The DMP is not a fixed document but it will be updated one more time during the project lifespan (M36).

Abbreviations

AES	Advanced Encryption Standard
AOP	Aerosol Optical Depth
API	Application Programming Interface
AQ	Air Quality
BLE	Bluetooth Low Energy
CAMS	Copernicus Atmosphere Monitoring System
COTS	Commercial off the Shelf
DISORT	Discrete Ordinates Radiative Transfer Program
DMP	Data Management Plan
DOIs	Digital Object Identifier System
EEA	European Environment Agency
GSM	Global System for Mobile Communications
LOV	Linked Open Vocabularies
LUT	Look-up Table
LAN	Local Area Network
MACV1	Max-Planck-Institute Aerosol Climatology version 1
N/A	Not Applicable
NetCDF	Network Common Data Form
OWL	Web Ontology Language
PM	Particulate Matter
PSOC	Programmable System-on-Chip
SHA	Secure Hash Algorithms
SSL	Secure Sockets Layer
QA	Quality Assurance
UI	User Interface
URL	Uniform Resource Locator
URI	Unique Resource Identifier
UX	User Experience design
VLAN	Virtual LAN
WPA	Wifi Protected Area
WP	Work Package