hackAIR

OPEN PLATFORM

D1.5: 3rd 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 1.5 – 3rd Data Management Plan			
Work Package	WP 1 – Project management			
Date of Delivery	Contractual	31 December 2018	Actual	27 December 2018
Nature	Report	Dissemination Level	Public	
Lead Beneficiary	DRAXIS Environmental S.A.			
Responsible Authors	Christodoulos Keratidis, Panagiota Syropoulou, Maria Akritidou			
Contributions from	Philipp Schneider (NILU), Sonja Grossberndt (NILU), Laurence Claeys (VUB), Gavin McCrory (VUB), Wiebke Herding (ONSUB), Natasa Moumtzidou (CERTH), Symeon Papadopoulos (CERTH), Stefanos Vrochidis (CERTH), Marina Riga (CERTH), Polychronis Charitidis (CERTH), Markos Zampoglou (CERTH), Emmanouil Krasanakis (CERTH)			

Document History

Version	Issue Date	Stage	Description	Contributor
1.0	15/10/2018	Draft	Request input from partners	Christodoulos Keratidis (DRAXIS)
2.0	12/11/2018	Draft	Acquisition of partners' input	Philipp Schneider (NILU), Sonja Grossberndt (NILU), Laurence Claeys (VUB), Gavin McCrory (VUB), Wiebke Herding (ONSUB), Natasa Moumtzidou (CERTH), Symeon Papadopoulos (CERTH), Stefanos Vrochidis (CERTH), Marina Riga (CERTH), Polychronis Charitidis (CERTH), Markos Zampoglou (CERTH), Emmanouil Krasanakis (CERTH)
3.0	17/12/2018	Draft	Integration of partners input	Panagiota Syropoulou, Maria Akritidou (DRAXIS)
4.0	20/12/2018	Draft	Internal review	Arne Fellermann (BUND), Symeon Papadopoulos (CERTH),
5.0	27/12/2018	Final	Integration of internal review comments & submission	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

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 final version of the project Data Management Plan (DMP). This final version lists the various datasets that have been produced by the project, the main data sharing and the major management principles that have been followed. Thus, the deliverable includes all the significant changes such as changed in consortium policies and any external factors that might have influenced the data management within the project.

The deliverable is structured in the following chapters:

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

1 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 D1.5: 3rd Data Management Plan aims to document all the updates of the hackAIR project data management life cycle for all datasets to have been collected, processed and/ or generated. A description of how the results will be shared, including access procedures and preservation according to the guidelines in Horizon 2020 projects and General Data Protection Regulation (GDPR). This was a living document and it was evolved and gained 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. Since, this is the final version of the project Data Management Plan, all the Work Packages are included despite the fact that some of them might have not occurred any changes.

2 Datasets in hackAIR

2.1 Datasets in WP1 – Project management (DRAXIS)

DMP component	Issues to be addressed
Data Summary	The database contains name, organisation and contact details for all project partners and advisory board members. 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 Furthermore, interviews have been contacted with the Advisory Board members and webinars have been held in order to inform them about the project status and progress. Most interviews and webinars have been conducted remotely either using Skype or WebEx. The expected size of the data is not applicable, as the size is not a meaningful measure.
Making data findable, including provisions for metadata	The data collected for the project partners and the Advisory Board members are stored on DRAXIS server and are not directly accessible from outside. The naming convention used is: Data_WP1_1_ Contact details of project partners and advisory board

	The data with regards to the interviews, webinars and consortium meetings are also stored on DRAXIS server. Moreover, these data cannot be made available to third parties. However, the interviews are available in D1.6 1st Report of Advisory Board meetings and D1.7 2nd Report of Advisory Board meetings. The dissemination level of these deliverables is public and they are available in the project's website and Wiki and in Zenodo[1] through the Digital Object Identifier (DOI): • D1.6 1st Report of Advisory Board meetings: DOI: 10.5281/zenodo.2273304 • D1.7 2nd Report of Advisory Board meetings: DOI: Not yet created The naming convention used is: Data_WP1_2_Advisory Board. Regarding the input for the DMP, the data are also stored on DRAXIS server and are not directly accessible from outside. These data are presented in the respective deliverables, which are publicly available either through the project website and Wiki or through Zenodo with the following DOIs: • D1.3 1st Data Management Plan: DOI: https://doi.org/10.5281/zenodo.2250192 • D1.4 2nd Data Management Plan: DOI: https://doi.org/10.5281/zenodo.2251825 The naming convention used is: Data_WP1_3_Data Management Plan. As part of any stored data, metadata were generated, which include sufficient information with appropriate keywords to help external and internal users to locate data and related information.
Making data openly accessible	The datasets are not publicly available. All the data are made publicly available as part of the aforementioned deliverables and through the project's website, Wiki and Zenodo.
Making data interoperable	N/A
Increase data re-use	Data are publicly available as part of the aforementioned deliverables and can be accessed and re-used by third parties indefinitely without a license.
Allocation of resources	No additional costs are foreseen for making this dataset FAIR.
Data security	The data have been collected for internal use in the project, and not intended for long-term preservation. No personal information will be kept after the end of the project. Moreover, DRAXIS pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/ 679.
Ethical aspects	N/A
Other issues	N/A

[1] http://zenodo.org/

2.2 Datasets in WP2 - Analysis and requirements (VUB)

2.2.1 User requirements (intake survey) data set

DMP component	Issues to be addressed
Data Summary	The goal of the data collection was to recruit the population that co-created 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 was important to attract a specific group of potential users for the co-creation session. The data have been 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 have been 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	The data are stored on the VUB servers and labelled with the work package and type of data. As the dataset contains confidential and sensitive information, the raw data will not be made available from outside but anonymised data can be made available upon request and after an evaluation of the request (i.e. purpose, goals, etc.). 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. The naming convention used is: Data_WP2_1_user requirements (intake survey) data set. As part of any stored data, metadata were generated, which include sufficient information: • to link it to the research publications/ outputs, • to identify the funder and discipline of the research, and • with appropriate keywords to help external and internal users to locate data.
Making data openly accessible	The data will be kept closed until the end of the project due to data contain sensitive personal information and therefore, it cannot legally be made public.
Making data interoperable	N/A
Increase data re-use	The data will not be licensed and they have been used for one specific purpose on one specific time period and have not been updated nor re-used.
Allocation of resources	Within the project no budget has been foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Papers are likely to be published after the completion of the project.
Data security	The data are stored on VUB servers and are not directly accessible from outside. After anonymization, the data is shared with the rest of the consortium partners. Furthermore, VUB pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable

	national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.2.2 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 were 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 were 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 were 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 were 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 has been used by the involved researchers (BUND and NILU) only for the processing of the workshop findings. It only served research purposes and it have by no means been released to other persons. The data have been 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. The naming convention used is: 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 are not made openly available.
Making data interoperable	N/A
Increase data re-use	N/A
Allocation of resources	Within the project no budget has been foreseen to pay for open access publishing, but still we will look to publish results in (free) open access journals. Papers are likely to be published after the completion of the project.
Data security	The data are stored on VUB servers and are not directly accessible from outside. After anonymization, the data is shared with the rest of the consortium partners. Furthermore, VUB pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

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

2.3.1 Geotagged Images Dataset

DMP component	Issues to be addressed
Data Summary	The dataset contains geotagged Images in Europe retrieved from two sources; Flickr and webcams. Regarding Flickr images they are user-generated images that are publicly available on Flickr and geo-tagged in Europe, while webcam images are extracted from static outdoor webcams located in Europe. It should be noted that webcams are geo-tagged and, consequently the images retrieved are also geo-tagged and time-stamped.

	The images were collected and 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 were then given as input to the air quality estimation model, developed within the project, which produces the Aerosol Optical Depth (AOD). The images were downloaded, downscaled to a maximum size of 500X500 pixels and stored until image analysis was performed (<1 hour). After this process, the images were permanently deleted from CERTH servers. All image metadata were permanently stored in a database. Under normal circumstances, the data collection rate for Flickr was around 5,000 images per day which translates to more than 1.5 million items (~1Gb) in one year. Similarly, for the webcams, the data collection rate was around 10,000 images per day which results in 3.5 million items (~2.5Gb) in one year. However, the number of images with usable sky was significantly less. Specifically, based on experiments realized it seemed that the percentage of usable images was around 11% of the total images retrieved from Flickr. In of webcams given that they were selected manually the percentage was significantly higher and drops only in days with overcast. The dataset was created by images retrieved from Flickr and webcams starting from June 2016. The number of records from both sources was 262,255.
Making data findable, including provisions for metadata	The naming convention used is: Data_WP3_1_Geotagged_Images_1.0. The metadata that were generated for each image are the following: • source • URL (of the image on Flickr or the webcam) • R/G and G/B ratios of sky part of the image • Geo-coordinates • Timestamp • AOD For the images retrieved from Flickr, the original images can be retrieved from the corresponding URLs. However, for the webcam images the original images can only be retrieved for the webcams for the ones that provide historical data. Thus, in case such historical information the webcams URLs links to the webcam itself and not the image taken at the specific date-time. The data is available on GitHub that make them discoverable and identifiable in the following link: https://github.com/MKLab-ITI/hackair-data-retrieval/tree/master/data/Geotagged_Images_dataset and in Zenodo (https://doi.org/10.5281/zenodo.2222342). The dataset is also 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.

	However, the original images cannot be shared due to Flickr's privacy and copyright policies and due to the copyright terms set by the webcam owners. In order to access the images no specialized software is required as images can be retrieved from the corresponding URLs using a web-browser.
Making data interoperable	The dataset is available in text-based machine-readable format (csv) that allow easy parsing and information exchange.
Increase data re-use	The data and associated code are licensed with an open data license (Apache License v2.0) that allows re-use of the data and code.
Allocation of resources	The data and associated code are licensed with an open data license (Apache License v2.0) that allows re-use of the data and code.
Data security	The cost of long-term preservation is negligible as data is hosted on GitHub. Long-term preservation will facilitate long-term usability of the data for development of air quality estimation methods. Furthermore, CERTH pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.3.2 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 concern measurements from stations across the whole Europe. The collected data facilitate visualizations of current air quality status, acting as an additional information layer. Moreover, they 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. The collection period extends from 1 October 2018 till 1 November 2018 and the number of records is 6,223,421
Making data findable, including provisions for metadata	The naming convention used is: Data_WP3_3_Environmental_1.0. 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 is available on GitHub that make them discoverable and identifiable in the following link: https://github.com/MKLab-ITI/hackair-data-retrieval/tree/master/data/Environmental_dataset The dataset is 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. data source, timestamp, pollutant, value, geo-coordinates, timestamps) is made openly available through GitHub. Data access does not require any specialized software as they are provided in csv formatted files that can be accessed with any text editor.
Making data interoperable	The dataset is available in text-based machine-readable format (csv) that allow easy parsing and information exchange.
Increase data re-use	The data and associated code are licensed with an open data license (Apache License v2.0) that allows re-use of the data and code.
Allocation of resources	The cost of long-term preservation is negligible as data is hosted on GitHub. 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. Furthermore, CERTH pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.3.3 Twitter_AQ

DMP component	Issues to be addressed
Data Summary	The dataset is collected to facilitate the development of Twitter-based air quality estimation models using supervised learning techniques applied in the following paper: Charitidis, P., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Twitter-based Sensing of City-level Air Quality. In Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2018 IEEE 13th. IEEE. The dataset is created by processing tweets that were retrieved from Twitter Streaming API by tracking 120 air quality related terms and ground truth PM2.5 measurements from OpenAQ API. The collection period extends from 8 February 2017 to 19 January 2018.

	A location estimation method[2] is applied on tweets to infer the location of each tweet. The dataset contains tweets from five (5) English (London, Leeds, Liverpool, Birmingham and Manchester) and five (5) American (New York, Boston, Pittsburgh, Philadelphia and Baltimore) cities.
Making data findable, including provisions for metadata	The naming convention used is: Data_WP3_4_Twitter_AQ. The dataset contains predictive features extracted from tweets in a n-hour temporal bin during the monitoring period where $n = [6, 12, 24]$. 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 In addition, features include a descriptive representation of the tweets assigned to a city ***c*** during a temporal bin by adopting a Bag of Words scheme. First, all tweets are preprocessed by applying tokenization, lowercasing and stop-word removal. Then, we create a vocabulary $\mathrm{W} = \{w_1, \dots, w_n\}$ for each country (UK, US) that consists of the n=10,000 most frequently occurring words in a random 1 million sample of the collected tweets. Using this vocabulary, a BoW vector $\mathrm{x} = [x_1, \dots, x_n]$ is generated to represent all tweets in(c, t), where x_i denotes the number of tweets containing w_i divided by the total number of tweets in(c, t). The published dataset consists of: a) the vocabularies of each country in csv file format with the following fields: • Term: Words in the vocabulary • Index: a corresponding index and b) Twitter information for each city in csv files (<City_name>_<window>.csv) with the following fields:

[2] G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, "Geotagging text content with language models and feature mining," Proceedings of the IEEE, vol. 105, no. 10, pp. 1971–1986, Oct 2017.

	• min_timestamp: the starting <window>-hour window timestamp • max_timestamp: the ending <window>-hour window timestamp • tweet_ids: space separated tweet ids in the corresponding window • bow_10k_unigrams: space separated $index$->$x(index)$ • #aqs: Number of air quality-related tweets • #high: Number of tweets referring to high air pollution • #tw: Number of tweets • nearby_ground_truth_pm25: inverse distance weighted mean of PM2.5 pollution values from nearby cities in $\mu g/m^3$ • pm25: PM2.5 pollution values in $\mu g/m^3$ The data are available on GitHub that make them discoverable and identifiable in the following link: https://github.com/MKLab-ITI/twitter-aq/tree/master/datasets The dataset is also discoverable by querying conventional search engines (e.g. Google) with the dataset name ("Twitter-AQ" in quotation marks).
Making data openly accessible	All the data described above is publicly available on GitHub. Data access does not require any specialized software as they are provided in csv formatted files that can be accessed with any text editor.
Making data interoperable	The dataset is available in text-based machine-readable format (csv) that allow easy parsing and information exchange.
Increase data re-use	The data and associated code is licensed with an open data license (Apache License v2.0) that allows re-use of the data and code.
Allocation of resources	The cost of long-term preservation is negligible as data is hosted on GitHub. 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. Furthermore, CERTH pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.3.4 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. If there is need for updates, the old data are overwritten and all actions

	are audited in detail and a log is kept, containing the changed text for security reasons.
Ethical aspects	N/A
Other issues	N/A

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

2.4.1 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 is useful internally for processing in the data fusion module.
Making data findable, including provisions for metadata	The data is discoverable on http://atmosphere.copernicus.eu/ but not on public hackAIR servers. The naming conventions used is: Data_WP4_1_CAMS conventions. The data are available to the public through the following deliverable of WP4. The dissemination level of these deliverable is public and it is available in the project's website and Wiki and in Zenodo through the DOI: • D4.2 Semantic integration and reasoning of environmental data: DOI: https://doi.org/10.5281/zenodo.2272957
Making data openly accessible	The data in their original form are only used internally for processing. Value-added derivatives of the dataset are presented on the hackAIR website. Some figures in public hackAIR deliverables might contain subsets of the data.
Making data interoperable	The interoperability of the data is achieved through the Standard CAMS metadata.
Increase data re-use	Data are only used internally. The quality assurance process that have been used is the specific QA which is carried out at CAMS. The data are 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 has been used only on the hackAIR server and as such falls under the same security standards as the rest of the hackAIR observations. Furthermore, NILU

	pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.4.2 Data fusion maps

DMP component	Issues to be addressed
Data Summary	Data fusion maps have been created as part of the hackAIR project by combining hackAIR observations with CAMS modelling information. These maps have been 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 is a few MB per day. These data are 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 are discoverable through the main hackAIR web interface. The spatial metadata are available within the provided GeoTIFF files: they contain georeferenced information of the data, such as projection, bounding box, etc. The naming convention used is: Data_WP4_2_Fusion maps
Making data openly accessible	The data are openly available to the public either through the WP4 deliverables or through the main projects' web interface. A web browser is only needed to view the data. The data and associated metadata, documentation and code are deposited on the main hackAIR server.
Making data interoperable	Metadata are not exposed to the users.
Increase data re-use	The data are licensed with the standard hackAIR license and a quality assurance process has been carried out.
Allocation of resources	N/A
Data security	N/A
Ethical aspects	N/A

Other issues	N/A

2.4.3 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). Abstract information (schema) is separated 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 [3] (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[4], a language-independent open data format that uses human-readable text to express data objects consisting of attribute-value pairs. These data have been served as input for the rule-based reasoning module, for providing personalised recommendations to the users.
Making data findable, including provisions for metadata	The naming convention used is: Data_WP4_3_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) is publicly available for further adoption by other interested parties, following the ontology-reuse

[3] 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/

[4] http://json.org/

	principles. The ontology schema is kept in project servers as well as in online ontology repositories (ontohub[5], LOV[6]) for easier track, search and discoverability. On the other hand, the populated instances (actual information) are not 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).
Making data openly accessible	As already mentioned, abstract (TBox) and actual (ABox) data are separately represented in the ontological framework of the project. The TBox describing the ontology schema is kept in both owned servers and online ontology repositories for easier track, search and discoverability. Actual data are 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 are available to the users only through the hackAIR UI and only for requests produced through the hackAIR UI utilities. Developed web-services enable the triggering of the recommendation module under specific data provision and request. The integration between the recommendation module and other involved hackAIR modules has been 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) is publicly available for further adoption, reuse or even extension of the represented content into third-party ontologies. Its data have been 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" (CC-BY-SA-4.0).
Allocation of resources	The cost of long-term preservation is negligible as abstract data (i.e. ontology schema) are hosted in owned servers as well as in free open ontology repositories

[5] https://ontohub.org/

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

	(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 has not been stored in the ontology; such data are provided directly by relevant hackAIR modules (i.e. user profile, fused data) and are 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

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

2.5.1 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 face to face discussions took place in order to form the final document. This was the basis upon which the system was built.
Making data findable, including provisions for metadata	There are no specific standards or metadata associated with these types of data. The naming convention used is: Data_WP5_1_System Architecture and Integration Framework 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 website and Wiki and through Zenodo with the following DOI: https://doi.org/10.5281/zenodo.2250338
Making data openly accessible	All data are made publicly available as part of the D5.1 Architecture and Integration Framework Definition Specification.
Making data interoperable	N/A
Increase data re-use	Data are publicly available as part of the D5.1 Architecture and Integration Framework Definition Specification and can be accessed and re-used by third parties indefinitely without a license.
Allocation of resources	No additional costs are foreseen for making this dataset FAIR.
Data security	The data have been collected for internal use in the project, and not intended for long-term preservation. Furthermore, DRAXIS fully complies with the applicable national, European and international framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A

Other issues	N/A

2.5.2 hackAIR platform

DMP component	Issues to be addressed
Data Summary	Various data like users' email, user perception of the air quality on a given day, health sensitivities are entered in the hackAIR platform. These data described above are saved in the hackAIR central database. All user actions (login, logout, account creation, visits on specific parts of the app) are logged and kept in the form of a text file. This log is useful for debugging purposes. Reports containing information on user devices (which browsers and mobile phones) as well as number of mobile downloads (taken from play store for android downloads and app store for mac downloads) are useful for marketing and exploitation purposes, as well as decisions regarding the supported browsers and operating systems. Furthermore, users are able to see results from various sensors as well as results based on the photos uploaded from the various users of the application.
Making data findable, including provisions for metadata	The data are not directly accessible from outside. These data cannot be made available to third parties. However, the handling of these data is described in the deliverables D5.2 1st version of integrated and tested hackAIR open platform and D5.3 Final version of integrated and tested hackAIR open platform. The dissemination level of these deliverables is public and they are available in the project's website and Wiki and in Zenodo through the Digital Object Identifier (DOI): • D5.2 1st version of integrated and tested hackAIR open platform: DOI: https://doi.org/10.5281/zenodo.2273754 • D5.3 Final version of integrated and tested hackAIR open platform: DOI: https://doi.org/10.5281/zenodo.2276621 The naming convention used is: Data_WP5_2_hackAIR platform. Every action on the platform produces meaningful metadata such as time and date of measurement creation or measurement amendments as well as timezone capture of the uploaded photos. The database is not 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) are able to discover the database.
Making data openly accessible	Only registered users have access to the data. The data produced by the platform are sensitive private data and cannot be shared with others without the user's permission. No open data will be created as part of hackAIR. The database is accessible only by the authorized technical team.
Making data interoperable	N/A
Increase data re-use	Moreover, the language of the content and data are in the following languages (English, German and Norwegian).

	The raw data are not publicly available. However, the hackAIR platform is an open source platform and it is offered under the AGPL v3 open source license and it is accessible through Zenodo through the DOI: https://doi.org/10.5281/zenodo.1442608
Allocation of resources	Resources have been allocated according to the project plan and WP3 allocated resources. No additional costs are foreseen for making this dataset FAIR.
Data security	All platform generated data have been saved on the hackAIR database server. Encryption has been used to protect sensitive user data like emails and passwords. All data are transferred via SSL connections to ensure secure exchange of information. If there is need for updates, the old data are overwritten and all actions are audited in detail and a log is kept, containing the changed text for security reasons. The system is weekly backed up and the back-ups are kept for 3 days. All backups are hosted on a remote server to avoid disaster scenarios. All servers are hosted behind firewalls inspecting all incoming requests against known vulnerabilities such as SQL injection, cookie tampering and cross-site scripting. Finally, IP restriction enforces the secure storage of data. DRAXIS pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and international framework, and the European Union's General Data Protection Regulation 2016/679. Moreover, "Personal Data Protection Policy " and "Terms and Conditions" have been included in the hackAIR platform, in order to inform the users of how hackAIR collects, processes, discloses and protects the incoming information. The hackAIR platform will not keep personal data and other information after the end of the action that took place on 31-12-2018.
Ethical aspects	All users' generated data will be protected and will not be shared without the user's consent.
Other issues	N/A

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

2.6.1 Engagement survey data set

DMP component	Issues to be addressed
Data Summary	The engagement survey data set contains 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 behavioural change (changes in belief, changes in knowledge and behaviour 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 behaviour of the population,

	(3) the belief, knowledge and behaviour 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 was collected during May and July 2017: • 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 behavioural change (use of standardized scales) The aggregated research data was **published** in internal project reports (accessible to all consortium partners via the hackAIR project Wiki) and will be used in external scientific research publications from the project. The origin of the data is through online surveys with 372 citizens (data subjects) that filled 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. The naming convention used is: Data_WP6_1_survey data set. The data are also available to the public through the deliverables D6.1 Engagement strategy for hackAIR community involvement and D6.2 Behavioural Change techniques for hackAIR community. The dissemination level of these deliverables is public and they are available in the project's website and Wiki and in Zenodo through the DOIs: • D6.1 Engagement strategy for hackAIR community involvement: DOI: https://doi.org/10.5281/zenodo.2275221 • D6.2 Behavioural Change techniques for hackAIR community: DOI: https://doi.org/10.5281/zenodo.2275355
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	The data are stored on VUB servers and are not directly accessible from outside. After anonymization, the data is shared with the rest of the consortium partners. Furthermore, VUB pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.6.2 Engagement qualitative insights data set (expert group)

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 by interviewing experts in the field of citizen science. These qualitative data are used for designing, running and analysing the engagement and behavioural change strategies. The types of the data are qualitative insights into: • Experiences, practices and expectations with regards to measuring and retrieving air quality information • Engagement of citizens in citizen science project: best practices, strategies and tactics. • Contact information (name, email) No re-use of data is foreseen. The origin of the data is the following: **Data subjects of the expert interviews:** • 6 experts in citizen science AQ projects. Data will not be anonymised and will be used as such in publications. The citizen science experts were able to review the processed transcript of the interview, and also the scientific publication that was made out of this dataset. **Data collection method**: expert interviews **Data type**: • Text (written notes) • Audio-records: The expert interviews were **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. The data will be used by the social science researchers of the project to create better strategies for engagement and behavioral change.
Making data findable, including provisions for metadata	The naming convention used is: Data_WP6_2_qualitative insights data set experts

	The data are also available to the public through the deliverables D6.1 Engagement strategy for hackAIR community involvement and D6.2 Behavioural Change techniques for hackAIR community. The dissemination level of these deliverables is public and they are available in the project's website and Wiki and in Zenodo through the DOIs: • D6.1 Engagement strategy for hackAIR community involvement: DOI: https://doi.org/10.5281/zenodo.2275221 • D6.2 Behavioural Change techniques for hackAIR community: DOI: https://doi.org/10.5281/zenodo.2275355
Making data openly accessible	The raw data will not be made openly available, but all insights will be shared in publications.
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.
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

2.6.3 Quantitative dataset about user experience and acceptance of the hackAIR solution

DMP component	Issues to be addressed
Data Summary	The data set contains questions about the user experience, acceptance and usability of the hackAIR solution (web app, mobile application, and sensors). The survey invites hackAIR users to share their opinions and feedback about the developed solution through open and closed questions (e.g. standardized scales about perceived ease of use, user performance, perceived usefulness, attitude, intention, etc.). The results of this dataset are analysed using descriptive statistics, and the summary is provided to the consortium partners. The feedback is collected through an online survey link, and through a paper format that is filled in on the spot by workshop participants. The following **data** was collected during February and December 2018: • Demographic information: only used for the profiling of the research participants (country, gender age). • Frequency of usage of the hackAIR solution. • User performance questions with the hackAIR solution (type of tasks completed).

	- User satisfaction questions about the hackAIR solution. - E-mail address of interested participants for further contact: e-mail was entered by participants if they agreed to participate in the behaviour change study of hackAIR. Privacy statement was provided in the online survey.
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. The naming convention used is: Data_WP6_3_Survey_dataset The aggregated research data were **published** in internal project reports (accessible to all consortium partners via the hackAIR project Wiki and Zenodo) and will be used in external scientific research publications from the project.
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. The collected email addresses are not shared with third parties.
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.
Data security	The data are stored on VUB servers and are not directly accessible from outside. The personal identifiable information was used to contact interested participants in the behaviour change studies by BUND and NILU. These data will be deleted after the project lifetime. Furthermore, special attention was paid to security, the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.6.4 Behavioural analysis: experiment in Brussels, Oslo and Berlin

DMP component	Issues to be addressed
Data Summary	The data set contains data about the behaviour profiles of citizens in Brussels, Oslo and Berlin. The behavioural profiles are constructed through the following data, collected during June – December 2018:

	• Motivation to participate in the study, and to inform oneself about air pollution • Beliefs regarding air quality (self-efficacy, problem awareness, citizen voice) • Knowledge regarding air quality (perceived and practical knowledge) • Behaviours regarding air quality (protective and pro-active behaviours), and self-perceived change • Demographic characteristics: age, gender, level of education, place of living • Contact details: email address. Participants in Brussels were invited to also participate in an interview, therefore, their email addresses were collected. For the interviews in Oslo and Berlin, the pilot coordinators contacted and invited interviewees for VUB. The purpose of this study was to investigate whether changes in the behaviour profiles occurred of citizen in Brussels, Oslo and Berlin after participation in a sensor-assembling workshop, or after using the hackAIR solution. The data was collected via surveys and interviews in Brussels (50 participants), and solely through interviews in Berlin and Oslo (10 participants in each pilot city).
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. The naming conventions used are: Data_WP6_4_survey data set Brussels. Data_WP6_5_qualitative data set interviews Brussels. Data_WP6_6_ qualitative data set interviews Berlin. Data_WP6_7_ qualitative data set interviews Oslo. The aggregated research data were **published** in internal project reports (accessible to all consortium partners via the hackAIR project Wiki) and will be used in external scientific research publications from the project.
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. The collected email addresses in Brussels are not shared.
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.
Data security	The data are stored on VUB servers and are not directly accessible from outside. The personal identifiable information was used to contact interested participants in the behaviour change studies by VUB for Brussels, and by NILU and BUND for participants in Oslo and Berlin. These data will be deleted after the project lifetime. These data will be deleted after the project lifetime. Furthermore, special attention was paid to

	security, the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.6.5 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 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 and social network analysis methods. Finally, the dataset contains the network of followers of the hackAIR Twitter account and their followers. 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. Six months after the end of the project all 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

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

2.7.1 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 PM2.5 and PM10 concentration using one of the supported sensors. The values are provided in μg/m3. 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 are available in the technical deliverable that describe the hackAIR Arduino node. The data (PM measurements and metadata) are available to other Researchers through the Open Access to Research Data Pilot.
Making data findable, including provisions for metadata	The measurements and metadata are available for use by the hackAIR applications through the secure API that we have created. The naming convention used is: Data_WP7_1_ Arduino_PM_measurements The measurements and metadata are made available for use by the hackAIR applications through the hackAIR platform.
Making data openly accessible	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 and are also available in Zenodo: https://doi.org/10.5281/zenodo.2222342
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 do not contain any information of the users who contributed it to hackAIR.
Allocation of resources	N/A
Data security	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.

	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 (such as WPA).
Ethical aspects	N/A
Other issues	N/A

2.7.2 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 PM$_{10}$ concentration using one of the supported sensors. The values are provided in µg/m^3. 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.
Making data findable, including provisions for metadata	The data along with the metadata are stored securely in DRAXIS servers. The data structure and protocol are available in the technical deliverable that describes the hackAIR PSOC node. The data (PM measurements and metadata) are available to other researchers through the Open Access to Research Data Pilot. The measurements and metadata are available for use by the hackAIR applications through the secure API that was created and also in Zenodo: https://doi.org/10.5281/zenodo.2222342. For each value, the following metadata are provided: Date, time, location (using the GPS unit of the smartphone), sensor id, and user id. The PM$_{2.5}$ and PM$_{10}$ concentration measurements are discoverable through the hackAIR platform, only for users who are registered through the platform. The naming convention used is: Data_WP7_2_PSOC_PM_measurements
Making data openly accessible	The measurements and metadata are made available for use by the hackAIR applications through the secure API that was 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 was 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 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 are 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

2.7.3 PM measurements by Commercial off the Shelf sensors – hackAIR cardboard

DMP component	Issues to be addressed
Data Summary	PM measurements by Commercial off the Shelf sensors – hackAIR cardboard. This 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). 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 and for each of the levels, the following metadata are provided: Date, time, location (using the GPS unit of the smartphone), sensor id, and user id. The naming convention used is: Data_WP7_3_Cardboard_PM_measurements The photo and metadata are made available for use by the hackAIR applications through the secure API that was created.
Making data openly accessible	The measurements and metadata are made available for use by the hackAIR applications through the secure API that was 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 was created.
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

2.7.4 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 are generated by the hackAIR users via the platform. The data described above are saved in the hackAIR central database. Detailed log of user actions (login, logout, account creation, visits on specific parts of the app) are kept in the form of a text file. This log is 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) are 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) are generated to assist the discoverability of the data and related information. Discoverability is 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 is not 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) are able to discover the database. The naming convention used is: Data_WP7_4_Generated Content
Making data openly accessible	Some of the data produced by the platform (posts, comments, users' display names and users' perception of the atmosphere) are accessible through the hackAIR platform, only for users who are registered through the platform. On the technical level only authorized hackAIR technical team members have access to the database.

Making data interoperable	N/A
Increase data re-use	Specific logging data are 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 are anonymized and encrypted. The following standards are 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 is set to take daily backups and a complete weekly one keeping up to the three last weeks. Furthermore, special attention was paid to security, the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	N/A
Other issues	N/A

2.7.5 Photos uploaded through the hackAIR mobile app

DMP component	Issues to be addressed
Data Summary	hackAIR users upload pictures depicting the sky. These pictures are geotagged and time-stamped and are stored in the hackAIR DB or a hackAIR designated secure network storage area. They are 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) are generated to help internal users locate the data and related information. Those data are saved in the hackAIR database.' The naming convention used is: Data_WP7_5_Uploaded_photos The data (metadata and photos) are available to other researchers through the Open Access to Research Data Pilot. The photos and metadata are available for use by the hackAIR applications through the secure API that we created. The estimations from the photos and the relevant metadata are available in Zenodo: https://doi.org/10.5281/zenodo.2222342. 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.
Making data openly accessible	These images might (a pending decision) be accessible through the hackAIR platform, only for users who are registered through the platform.

	Discoverability is 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 is 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 are 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 are set to take daily backups and a complete weekly one keeping up to the three last weeks. The database is not 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) are able to discover the database. The images folder is not 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 have access to the database and database storage. Furthermore, special attention was paid to security, the privacy and confidentiality of the users' personal data by fully complying with the applicable national, European and International framework, and the European Union's General Data Protection Regulation 2016/679.
Ethical aspects	Photos are not shared with other users unless the person who uploaded the photo wants to.
Other issues	N/A

2.8 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 datasets are for internal use in the project, and not intended to be findable or accessible to parties external to the project. The data are available through the public deliverables and are accessible through Zenodo: • D8.1 Communication and Dissemination Strategy: DOI: https://doi.org/10.5281/zenodo.2249861 • D8.3 Dissemination pack: DOI: https://doi.org/10.5281/zenodo.2251583 • D8.4 Network of Interest established (1st meeting report): DOI: https://doi.org/10.5281/zenodo.2251736 • D8.5 Report on co-creation of services: DOI: https://doi.org/10.5281/zenodo.2273658 • D8.6 hackAIR workshop toolkit: DOI: https://doi.org/10.5281/zenodo.2275974 • D8.7 Network of Interest workshop report: DOI: https://doi.org/10.5281/zenodo.2276309 • D8.9 Network of Interest closing meeting report: DOI: Not yet created The naming convention used is: Data_WP8_1_Communication and dissemination data As part of any stored data, metadata were generated, which include sufficient information with appropriate keywords to help external and internal users to locate data and related information.
Making data openly accessible	Contact data are not openly available in line with privacy protection legislation. A summary of the communications monitoring data are 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

	and fully complies with the applicable national, European and international framework, and the European Union's General Data Protection Regulation 2016/679.
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

3 Conclusion

This final version of the hackAIR data management plan reflects the updated procedures that were implemented by the hackAIR project to efficiently manage the data it produced. In particular, the final DMP anticipates the data management strategy regarding the collection, management, sharing, archiving and preservation of data.

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