



DELIVERABLE

D1.4 Data Management Plan 2

Project Acronym:	COMPAIR	
Project title:	Community Observation Measurement & Participation in AIR Science	
Grant Agreement No.	101036563	
Website:	www.wecompair.eu	
Version:	1.0	
Date:	31 October 2022	
Responsible Partner:	DV	
Contributing Partners:	-	
Reviewers:	Internal reviewers: Alexopoulos Charalampos (UAEG) Jiri Bouchal (ISP) External reviewers: Andrew Stott	
Dissemination Level:	Public	X
	Confidential, only for members of the consortium (including the Commission Services)	

Revision History

Version	Date	Author	Organisation	Description
0.1	13/10/2022	Tom Callens	Digital Flanders	Updates to all chapters
0.2	19/10/2022	Charalampos Alexopoulos	UAEG	Review
0.3	19/10/2022	Jiri Bouchal	ISP	Review
0.4	21/10/2022	Andrew Stott	external	Review
1.0	31/10/2022	Tom Callens	Digital Flanders	Final version

Table of Contents

Executive Summary	4
1. Introduction	5
2. Data Summary	6
3. FAIR data	7
3.1. Making data findable, including provisions for metadata	7
3.2. Making data accessible	8
Data storage	8
Data sharing	9
Metadata sharing	10
3.3. Making data interoperable	11
3.4 Increase data re-use	12
4. Other research outputs	13
5. Allocation of resources	13
6. Data security	14
7. Ethics	14
7.1 Data protection	14
8. Conclusions and future work	15

List of Tables

Table 1: COMPAIR Data usage scenarios - data summary overview	6
Table 2: COMPAIR Data storage scenarios	8
Table 3: COMPAIR Data sharing scenarios	9
Table 4: COMPAIR Data interoperability scenarios	11
Table 5: COMPAIR Data re-use increase scenarios	12

Executive Summary

According to the COMPAIR DoA (2021), a Data Management Plan (DMP) is part of WP1 “Risk & Quality Management” Work Package. This is the second version after 12 months of ongoing project. The first version was available by 29 April 2022 and the final version will be delivered by the end of the project, by 31 October 2024.

The goal of this DMP is that, during the project, the COMPAIR partners take into account the many aspects of data management, metadata generation, data preservation, and analysis. The COMPAIR project follows the FAIR principles: data produced by the consortium will be made Findable, Accessible, Interoperable and Reusable to maximise the added-value of the project results.

In this DMP, we differentiate between input data and output data (e.g. created Citizen Science data). Input data can be derived from external sources (e.g. external APIs or web services) or can be uploaded on the consortium's data infrastructure. The output Citizen Science data (e.g. sensor data and dashboard data) will be downloadable and reusable. However, the consortium is still scrutinizing and elaborating the different output formats and linked privacy measures.

An essential element related to improving the reusability of Citizen Science data is the further standardisation on the data input and output side. Data interoperability and semantics could drastically reduce the need for data integration, e.g., using standardised sensor data formats like Sensor Observation Services (SOS) and SensorThings API.

Next steps/next iterations:

- Data security: in the third iteration of the DMP, the data security approach, outlined in this deliverable, will be revised.
- Data protection: in the third iteration of the DMP, the data protection aspects, outlined in this deliverable, will be revised.
- Data identification: detailed list of datasets that will be used and created within the project will be added to the last version of the DMP.

1. Introduction

The Horizon Europe Model Grant Agreement requires that a Data Management Plan (DMP) is established and regularly updated. According to the COMPAIR Description of Action (2021), a DMP is part of WP1 - the Project Risk & Quality Management Work Package. We use the recommended Horizon Europe DMP template (version 1.0 - 5 May 2021)¹.

In this DMP, we differentiate between input data and output data (e.g. calibrated Citizen Science sensor data). Input data can be derived from external sources (e.g. external APIs or web services) or can be uploaded on the consortium's data infrastructure. COMPAIR uses a combination of non-geospatial and geospatial datasets. Non-geospatial formats are open formats like XML², JSON(LD)³, CSV⁴ as non-geospatial data formats. The used geospatial formats combine data file formats and services like Shapefiles⁵, WMS⁶, WFS⁷, GeoTIFF⁸, CityGML⁹ and GeoJSON¹⁰. APIs are also used and are especially relevant for live data streams. Simulation models use a combination of model-specific formats, e.g. Cube Voyager¹¹ and more generic scripting formats like GDX¹² scripting.

The COMPAIR output data and Citizen Science data will be downloadable. However, the consortium is discussing the different output formats (D2.3 Sensor device functional and technical design report and privacy measures¹³).

Finally, there is the need for new standardisation initiatives. An essential element related to improving the reusability of Citizen Science data is the further standardisation of data. Data interoperability and semantics could drastically speed up data integration, e.g., using standardised sensor data formats like Sensor Observation Services (SOS)¹⁴ or SensorThings API¹⁵.

Any research data and papers will be published according to the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities¹⁶.

¹ https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template_he_en.docx

² <https://www.w3.org/XML/>

³ <https://www.w3.org/TR/json-ld11/#introduction>

⁴ <https://www.w3.org/TR/tabular-data-primer/>

⁵ <https://desktop.arcgis.com/en/arcmap/10.3/manage-data/shapefiles/what-is-a-shapefile.htm>

⁶ <https://www.ogc.org/standards/wms/introduction>

⁷ <https://www.ogc.org/standards/wfs>

⁸ <https://www.ogc.org/standards/geotiff>

⁹ <https://www.ogc.org/standards/citygml>

¹⁰ <https://www.ogc.org/standards/eo-geojson>

¹¹ <https://www.bentley.com/en/products/product-line/mobility-simulation-and-analytics/cube-voyager>

¹² https://www.gams.com/35/docs/UG_GDX.html

¹³ https://www.wecompair.eu/files/ugd/725ca8_56258bbb7a234badb35d1947230f8d41.pdf?lang=nl

¹⁴ <https://www.ogc.org/standards/sos>

¹⁵ <https://www.ogc.org/standards/sensorthings>

¹⁶ <http://openaccess.mpg.de/Berlin-Declaration>

2. Data Summary

To ensure both public and policy impact, COMPAIR enhances the value of data in 3 ways:

1. Improving publication availability by aligning datasets with existing international and EU data and metadata interoperability standards
2. Raising data quality (e.g. by utilising expert calibration algorithms for automated quality assessment and validation to enhance accuracy of IoT sensors)
3. Broaden flexibility of APIs for more tailored policy use (e.g. changing hourly results to a near real-time window for more operational decision making)

The following table summarises the typologies and contents of data collected and produced. For each distinct category, you can find the status by month 6.

Table 1: COMPAIR Data usage scenarios - data summary overview

Nature of datasets Data usage scenarios	Confidential	Anonymised and Public	Non anonymised and Public
Original data produced by the COMPAIR consortium	Internal meeting minutes Confidential deliverables Personal email communication Login data/user management (dashboards/gamification) Subscriptions data Consent forms about participating citizens	Traffic counts (from new Citizen Science sensor devices) Dynamic exposure data from wearable air quality sensors Integrated air quality data from stationary sensors (official measuring data combined with calibrated CS AQ data) Combined air and traffic data from policy dashboard (schoolstreets and playstreets) Carbon footprint dashboard data	COMPAIR Publication list COMPAIR Newsletters
Existing data already in possession of the COMPAIR consortium and/or partners and open data	Personal email communication Shared access to software repositories	Traffic data (from existing CS sensor devices) Transport network data Air Quality data from official measuring stations (Flanders) Air Quality data from official measuring stations (Sofia and Plovdiv)	N/A

		3D model of the buildings, Digital surface model, Digital terrain model, base map (Flanders)	
Existing data sourced/procured by the COMPAIR consortium or obtained from external suppliers	Licensed access and use during project execution	Free and open access and use during project execution - via obtaining data from open data portals and via application forms.	N/A

An implication of the above table is that **every partner is responsible for the behaviour of all team members**, which may also include subcontracted organisations (e.g. specialised press agencies) or volunteers.

Detailed tables of relevant existing datasets for the COMPAIR pilots are included in Annex 2 of the COMPAIR deliverable D2.1 Value Network Canvas (pp. 47-61)¹⁷.

In the final version of the DMP a detailed overview will be added with data sources created during the COMPAIR project.

3. FAIR data

The COMPAIR project follows the FAIR principles: data produced by the COMPAIR consortium will be made Findable, Accessible, Interoperable and Reusable to maximise the added-value of project results.

3.1. Making data findable, including provisions for metadata

The strategy to make the newly created sensor data, the calibrated sensor data and the enhanced API's findable is by **using the existing regional and national Open Data Portals** which are **harvested by the official portal for European data**¹⁸ The idea is to create metadata according to ISO 19115 or DCAT-standard (Data Catalog Vocabulary) following the regional or national setup, procedures and good practices. Keywords, common in the domain, will be used (e.g. the GEneral Multilingual Environmental Thesaurus - GEMET¹⁹) to increase findability of the data. (We will not use the metadata publishing solutions provided through the European INSPIRE initiative²⁰ because the COMPAIR-research data falls out of the scope of the INSPIRE directive²¹).

¹⁷ https://www.wecompair.eu/files/ugd/68109f_cb57a372cebd4ceb9020d55dfa765aca.pdf

¹⁸ <https://data.europa.eu/en>

¹⁹ <https://www.eionet.europa.eu/gemet/en/themes/>

²⁰ <https://inspire.ec.europa.eu/>

²¹ <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002>

Metadata of sensors will also be available through the SPARQL endpoints of the API.

As a result metadata will be created close to the data source, findable on different governmental levels and publicly available in all member state languages. In the last version of the DMP we aim to include the URL's of the metadata records on the European data portal.

The same strategy applies to the other research data (e.g. dashboard data) that will be created within the COMPAIR project.

While Telraam sensor-data is available via its own API (www.telraam-api.net) and well-documented including metadata, all sensor-data used in the COMPAIR-project is governed by the COMPAIR Data-Manager, where compliance to data standards is ensured.

3.2. Making data accessible

Data storage

The following table summarises the procedures for storing project-related data during the first 12 months of the COMPAIR project and provides an overview of the most frequently used repositories.

Table 2: COMPAIR Data storage scenarios

Nature of datasets Data storage scenarios	Confidential	Anonymised and Public	Non anonymised (temporary status)
Original data produced by the COMPAIR consortium	Common project repository	Repositories of the data provider or common project repository	Individual partner repositories Common project repository
Existing data already in possession of the COMPAIR consortium and/or partners and open data	Common project repository, or repositories of the data provider.	Repositories of the data provider or common project repository	N/A
Existing data sourced/procured by the COMPAIR consortium or obtained from external suppliers	Individual partner repositories Third party repositories of the data provider.	Repositories of the data provider or common project repository	N/A

Data storage and management of the COMPAIR datasources

The main principle is that data created by COMPAIR will be stored on existing data platforms, maintained by the partners, as much as possible. In case data needs to be stored commonly we will tackle this in the next version of the DMP.

COMPAIR will also use data sources stored on external data platforms as much as possible. These can be hosted in the cloud or on-site by the data providers.

It is not an objective of the COMPAIR project to make the input widely accessible. In the case of existing data that are open, accessibility is already offered by the provider. In the other case the COMPAIR partners will need to procure the data from external suppliers, and maybe pay fees, for the use within the project. It is beyond the scope of the COMPAIR project to open up commercial data. All data used will be listed in a next version.

The data created by COMPAIR will follow standardisation principles based on ISA² and W3C. The OGC/INSPIRE standards will be used for managing geospatial data. W3C Linked open data principles will be used for linking data semantically.

Data storage and management of the COMPAIR consortium project data

Google Drive™ is the selected tool as COMPAIR’s data and information repository. This includes the project deliverables (including relevant references utilised for their production or generated from them as project publications, e.g. journal articles, conference papers, e-books, manuals, guidelines, policy briefs etc.) and any other related information, including relevant datasets. It also implies that the privacy and security measures of Google Drive™ must be GDPR compliant. The verification of such circumstances is the responsibility of the coordinator. The procedures are described in the D1.2 Project Management Handbook.

Data sharing

All created data will be shared according to the directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (Open Data Directive)²².

The following table summarises the procedures for sharing data during the first 12 months of the COMPAIR project.

Table 3: COMPAIR Data sharing scenarios

Nature of datasets	Confidential	Anonymised and Public	Non anonymised (temporary status)
Data sharing scenarios			
Original data produced by the COMPAIR consortium	Analytics (login and subscription data)	Anonymisation Statistical evaluation Metadata generation Visualisation Analytics	Selection/ destruction

²² <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L1024>

		Publication as map services/APIs	
Existing data already in possession of the COMPAIR consortium and/or partners and open data	Visualisation Analytics	Anonymisation Statistical evaluation Metadata generation Visualisation Analytics Publication as map services/APIs	N/A
Existing data sourced/procured by the COMPAIR consortium or obtained from external suppliers	Visualisation Analytics	Anonymisation Statistical evaluation Metadata generation Visualisation Analytics Publication as map services	N/A

The necessary procedures (e.g. applying privacy by design principle) will be applied to ensure that the maximum of data can be shared (for as far as the procured access rights allow). Personal data will be destroyed after 3 months.

Example of existing Open dataset (with privacy by design) used in COMPAIR:

Data from the **Telraam sensor devices** used within the COMPAIR project will be aggregated with existing Telraam sensor data and will be made available publicly via the existing distribution channels: homepage interface (interactive map)²³ and API²⁴.

Telraam sensor devices count road users passing in front of them. The **images** of the road users are **processed by the onboard CPU** and **erased when processed**. Only count data (along with basic object properties such as size and duration of visibility) are uploaded to the servers. This raw data is classified into groups of road users (pedestrians, bikes, cars and lorries) and aggregated on an hourly basis. These hourly counts are publicly accessible through the homepage/map and API.

Metadata sharing

For the public partners the idea is to create metadata according to ISO 19115 or DCAT-standard following the regional or national setup, procedures and good practices. The method of sharing metadata from the private sector partners is part of further negotiations. National Open Data portals can be used for data from public bodies. A metadata portal like Zenodo²⁵ can be used to share metadata from private sector partners.

In principle data from anonymised and public data will be available and findable during the life cycle of the data. Metadata will remain available, even after the data is classified as obsolete.

²³ <https://telraam.net/>

²⁴ <https://telraam-api.net/>

²⁵ <https://zenodo.org/>

Personal data will be destroyed after 3 months, no metadata will be created.

3.3. Making data interoperable

The following table summarises the procedures for making data interoperable during the first 12 months of the COMPAIR project.

Table 4: COMPAIR Data interoperability scenarios

Nature of datasets	Confidential	Anonymised and Public	Non anonymised (temporary status)
Original data produced by the COMPAIR consortium	Statistical process (login and subscription data)	Publication as OGC service Export dashboard data as CSV	N/A
Existing data already in possession of the COMPAIR consortium and/or partners and open data		Publication as OGC service Export dashboard data as CSV	N/A
Existing data sourced/procured by the COMPAIR consortium or obtained from external suppliers		Publication as OGC service Export dashboard data as CSV	N/A

Table 4: COMPAIR Data interoperability scenarios

The anonymised and public data will be made interoperable by using standardised formats (see introduction) and (OGC compliant) API's.

Sensor data of traffic and air quality will be transformed via the COMPAIR Data Platform to an OGC Sensorthings API²⁶, ensuring and enhancing interoperability. The COMPAIR Data Platform and its components, like the Data Manager, are described in Deliverable D4.1 Solution Architecture Report, to be published by 31 October 2022.

Examples of enhancing interoperability:

As a deliverable of the COMPAIR project the **Telraam** API²⁷, which also includes sensor data from the European We Count-project and other Citizen Science initiatives, by

²⁶ <https://www.ogc.org/standards/sensorthings>

²⁷ <https://telraam-api.net/>

integrating the Telraam sensor data in the COMPAIR Data Manager, we make it more interoperable by applying an OGC API standard for sensor data.

As a deliverable of the COMPAIR project the **SODAQ AIR²⁸** database will be made accessible through an OGC API standard for sensor data, via the COMPAIR Data Manager.

The choice of which OGC sensor data standard (Sensor Observation Services or SensorThings API) to apply is part of ongoing technical analysis.

3.4 Increase data re-use

The following table summarises the procedures for making data interoperable during the first 12 months of the COMPAIR project.

Table 5: COMPAIR Data re-use increase scenarios

Nature of datasets	Confidential	Anonymised and Public	Non anonymised (temporary status)
Data re-use scenarios			
Original data produced by the COMPAIR consortium	Anonymisation	Technical documentation Open Data licence	Selection/ destruction
Existing data already in possession of the COMPAIR consortium and/or partners and open data	Anonymisation	Technical documentation Open Data licence	N/A
Existing data sourced/procured by the COMPAIR consortium or obtained from external suppliers	Anonymisation		N/A

Table 5: COMPAIR Data re-use scenarios

The anonymised and public data will be made re-usable by providing technical documentation and applying an Open Data licence (e.g. CC0, CC BY and CC-BY-SA) .

Example of increasing data re-use:

Data from the **Telraam sensor devices** used within the COMPAIR project will be aggregated with existing Telraam sensor data and will be made available publicly via the existing distribution channels: homepage interface (interactive map) and API.

²⁸ <https://sodaq.com/products/air/>

The API is **documented** extensively by means of the “**Postman API Platform**”²⁹. All API calls methods are described together with an introduction to basic definitions of the object involved and how traffic is modelled, requests examples and frequently asked questions³⁰.

All Telraam data are available under a CC-BY-NC licence³¹. That means that data can be used, adapted and published freely for non-commercial purposes. Telraam opts to limit to a non-commercial licence to allow their participating citizens to co-benefit if 3rd parties develop a commercial application or product using Telraam data. However, their default position is to share data with private individuals, companies and authorities freely for whatever purpose. They invite everyone to **explore commercial or non-commercial opportunities** with Telraam data; they only want their efforts and those of their participating citizens to be rewarded fairly if there's a commercial success.

4. Other research outputs

Any research data and papers will be published according to the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities³².

We consider the use of the open repository Zenodo³³ for publishing research papers and data sets.

5. Allocation of resources

The costs for implementing the FAIR principles in the project are fully integrated into the project budget. They are shared amongst the (pilot) partners responsible for the data and potentially involved in data capturing and modelling (SODA, TELR, VMM, IMEC, HHI, DAEM, SDA, EAP, INT3 and DV). The consortium partners have a budget to finance open access to research data and research results like articles published in scientific journals. Data management is the responsibility of every consortium partner. Still, there are specific tasks for each pilot lead and the consortium coordinator responsible for the data management plan and ethics. The final long-term preservation decision is on the project direction level (consortium coordinator, senior user representative, and senior supplier representative).

Financial means for applying security and re-use measures are foreseen in WP4 “Development & Technical Integration”, especially within T4.1 “Solution Architecture and Sprint Plan” and T4.3 “Digital City Twin, Citizen Science Data and Message Broker and AR app Integration”.

²⁹ <https://telraam-api.net/>

³⁰ <https://telraam.zendesk.com/hc/en-us/sections/4403919967249-Data>

³¹ <https://telraam.zendesk.com/hc/en-us/articles/360056754532-Telraam-Data-License>

³² <http://openaccess.mpg.de/Berlin-Declaration>

³³ <https://zenodo.org/>

The costs for storage and archiving are currently estimated to 2.000 Euro per month, but can evolve during the project. Personal data collected from users or involved citizens will be stored no longer than 3 months, according to the GDPR. Anonymised data will be stored 'indefinitely'.

6. Data security

Data security strategy is part of ongoing technical analysis. The IT technical data security aspects will be - partly - covered in deliverable D4.1 of month 12. Regarding the aspect of confidentiality or the degree to which a product or system ensures that data are accessible only to those authorised to have access, the general approach is to design tools that do not need a link between persons and data as much as possible. Take, for example, the dynamic air quality measurements. If a Citizen Scientist wants to see his or her measured dynamic exposure of air quality, no authentication will be required in the Dynamic Exposure Visualisation Dashboard. The user will just have to select the ID of the sensor he/she wants to see the readouts from. The ID should be marked on the sensor, for ease of use.

7. Ethics

The ethical and legal aspects will be scrutinised extensively during the project. Work package nine, "Ethics requirements", together with work packages one and seven, contains the legal and ethical requirements. The ethical aspects are about identifying legal and ethical considerations and providing an overview of ethical aspects. In WP 9, especially in D9.1 H - Requirement No. 1 the procedures and criteria that will be used to identify/recruit research participants are explained. The informed consent procedures that will be implemented are described. And also the measures that will be taken to protect and minimise the risk of stigmatisation of vulnerable individuals/groups that will be involved. Next to that, measures and procedures are described to ensure legal compliance. Each of the pilots has been scrutinised on the impact of ethics and legal issues. In WP 7, procedures and criteria will be described. Especially D7.2 Legal Requirements and Guide to Legal Compliance for Data-Driven Decision Making deals with ethical aspects applied to each of the pilots.

7.1 Data protection

The pilots should comply with the requirements of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (GDPR), and all partners have been instructed to keep this in mind. The Data Protection Officers of each partner are invited to check on the activities of the partner in the project. The Privacy and Ethics manager of the COMPAIR-project is Vasiliki Diamantopoulou (UAEG).

The only personal data that research participants would need to share in the context of these activities are: name, and surname, email address and, where applicable, organisation. Taken together, this data falls under the definition of personal data in the GDPR. No highly sensitive data will be collected in the context of such surveys.

How the collection and sharing of dynamic exposure data from wearable air quality sensors will be tackled in respect to GDPR is part of technical analysis. The general approach is to design tools that do not need a link between persons and data as much as possible, and so avoid the creation of personal data. Take, for example, the dynamic air quality measurements. During the measurements, the GPS tracking of the sensor will be on, but the sensor and its readouts will not be coupled to a specific Citizen Scientist. If a Citizen Scientist wants to see his or her measured dynamic exposure of air quality, no authentication will be required in the Dynamic Exposure Visualisation Dashboard. The user will just have to select the ID of the sensor he/she wants to see the readouts from. To avoid indirect identification of the participant, their track's beginning and end position will be cut off. The truncation method and parameters will be decided later in cooperation with the COMPAIR DPO.

Specifically with respect to the data from Telraam, there is a clear distinction between sensor-collected data and sensor-owner data. The former holds no privacy sensitive data (i.e. traffic volume per hour on a specific location), while the latter does hold privacy sensitive data (name, e-mail, address of sensor owners). The former data is the more bulky data set, requiring a strong data infrastructure, the latter holds a light database which requires stronger security. The Telraam data-infrastructure is fully compliant (see [Telraam privacy-policy](#)) and only sensor-data is shared with the Telraam Data Manager.

8. Conclusions and future work

This document is the second of a series of three planned deliverables concerning the COMPAIR Data Management Plan (DMP) in fulfilment of the requirements of the project's work plan. The main reason for planning three versions of the DMP (at months 6, 12 and 36) and particularly two of them during the first project year, is evidently related to the need to hold on until the development as well as piloting activities of COMPAIR gain further momentum, in order to:

- Secure the current, proposed structure of contents against any changes suggested by the gradual and incremental start up of the core project activities, and
- Colour the already existing contents with important add-ons based on the learning process that the COMPAIR partners will activate throughout the project's lifetime, considering also that a lot of project work will be oriented to operationalizing the connection between data handling (including data collection, integration, processing, analysing and visualizing).

This edition of the DMP has, fulfilled the immediate goals of such a stepwise approach to data management, by:

- Illustrating progress regarding strategy choices for security aspects;
- Illustrating progress regarding data protection measures;

The following steps in the COMPAIR data management process that will be reflected in the next version of the DMP are:

- **Data security:** in the third iteration of the DMP, the data security approach, outlined in this deliverable, will be revised.

- **Data protection:** in the third iteration of the DMP, the data protection aspects, outlined in this deliverable, will be revised.
- **Data identification**
A detailed list of datasets that are used within the project will be updated in the last version of the DMP.
A detailed list of datasets that are created by the project will be added to the last version of the DMP.