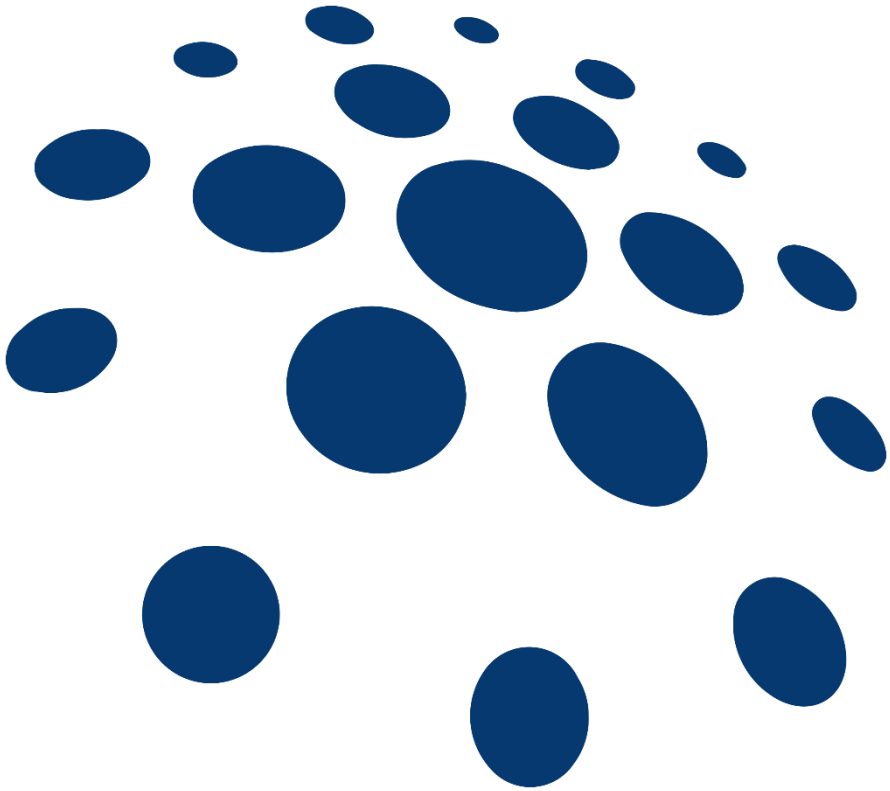
Horizon Europe

Project: 101138405 — Net4Cities

Deliverable 7.2
Data Management Plan



Net4Cities Consortium

Imprint

Suggested citation:
Erika von Schneidemesser, Seán Schmitz, Jörn Quedenau, Rune Ødegård (2024). D7.2 **Data Management Plan**. Horizon Europe Project Net4Cities.

Main findings and deliverables of the Net4Cities project will be available at www.net4cities.eu

This project has received funding from the European Union's Horizon Europe funding programme under the call HORIZON-CL5-2023-D5-01 – No. 101138405

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Document Control Information

Settings	Value
Document Title:	**Data Management Plan**
Project Title:	Net4Cities
Document Author:	Erika von Schneidemesser
Doc. Version:	1.0
Sensitivity:	Public
Date:	30/06/2024

Document Approver(s) and Reviewer(s):

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Action	Date
Kris Vanherle	Beneficiary	Review	21.06.24
Pablo Garcia	Associated Partner	Review	19.06.24
Erika von Schneidemesser	Coordinator	Review and Approve	21.06.24

Document history:

The Document Author is authorized to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner.
Changes to this document are summarized in the following table in reverse chronological order (latest version first).

Revision	Date	Created by	Short Description of Changes

Configuration Management: Document Location

The latest version of this controlled document is stored in the Sync&Share file-sharing platform at “Net4Cities/WP7/T7.5 OpenScience DataManagement/D7.2 data management plan”.

Table of Contents

Executive Summary

In Net4Cities, a large amount of data will be produced throughout the course of the project. These will largely be measured and modelled air, noise, and traffic data, some of which will be re-used from existing sources, and some which will be created by the actions of the project. To accommodate this data, the project beneficiary NILU will use its existing data storage infrastructure to store, process, and provide access to Net4Cities data. In its role as a leading environmental institute, NILU currently hosts and maintains multiple large, thematic databases for various international organisations, including the United Nations (UN), the World Meteorological Organization (WMO), and the European Space Agency (ESA). As such, their experience and resources will be leveraged to store all relevant data for Net4Cities.

NILU's infrastructure is supported by the Norwegian Agency for Shared Services in Education and Research (Sikt) and uses an advanced security system provided by a leader in firewall security. There are multiple firewalls in an active/passive configuration, including full redundance of the network architecture. Any data not hosted in NILU's servers is expected to not be project critical or sensitive to security issues.

A core component of the storage and usage of data will be the adherence to FAIR principles. Key amongst these in Net4Cities will be the public availability of data and the use of open-source software to ensure accessibility. Existing standards for the findability and accessibility of data and metadata, such as the ISO 19115 standards or the ACTRIS (Aerosol, Clouds and Trace Gases Research Infrastructure) vocabulary, as well as standard data formats, such as CSV, JSON, XML, or NetCDF, will be used. Data will be sent and received using the existing NILU REST-API. The metadata will be made available under a Creative Commons Zero (CC0) license, to prevent any restrictions on their use. Access instructions will be provided for both data and metadata, which will ensure seamless machine-to-machine interaction.

A key component of exploitation in the Net4Cities project is the long-term preservation of and access to all project data. To achieve this, open-source, scalable technologies for big data provided by the Apache Software Foundation will be used. This will provide a robust long-term infrastructure for the data.

Given the nature of Net4Cities and its role in establishing brand new datasets on previously under-evaluated pollutants, a key goal of data management in this project will be to maintain the datasets over the long-term. Using the NILU infrastructure for this purpose is ideal, as resources are already in place for ensuring longer-term storage. As the project develops and datasets are produced, this Data Management Plan (DMP) will be altered and revised accordingly.

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

Net4Cities will re-use existing data from urban air quality and noise monitoring infrastructures, as well as datasets (including emissions data and satellite data). The existing data will mostly be focused on regulated air pollutants, noise, and emissions in the 11 partner cities. Most of the air pollution data is reported to and can be openly accessed through Airbase, albeit with a delay of a couple years. We will access the data from the individual city data repositories in real-time.

Below are several links which are examples of the type of existing data that will be used in Net4Cities:

- Berlin: https://luftdaten.berlin.de/lqi; https://luftdaten.brandenburg.de/; Berlin emissions inventory (https://www.berlin.de/sen/uvk/umwelt/luft/schadstoffausstoss-emissionen/emissionskataster-2015/)
- Düsseldorf: https://www.lanuv.nrw.de/umwelt/luft/immissionen/aktuelle-luftqualitaet/
- Tbilisi: https://air.gov.ge/en/
- Telraam: www.telraam.net (traffic counting data)

What types and formats of data will the project generate or re-use?

Several data formats have already been identified, but these will likely change as the project develops. Much of the data will be timeseries data (air pollution, noise, traffic counts). The possible and likely data formats will be CSV, JSON, XML and/or NetCDF. Data will be sent and stored in NILU's sensor data platform, which support JSON (application/json MIME type) and XML (application/xml MIME type) format using the data platform's web API https://sensors.nilu.no/api/doc. Stored data can be download as JSON files from sensor platforms API or downloaded as csv file manually.

In addition to the timeseries data, emissions datasets will be used and activity data and/or emission factors generated and updated, as well as satellite data, and source apportionment modelling data. The data formats for these would be Geopackage or Geoparquet for GIS vector data, GeoTiff for rasters, and CSV or similar for emission factors. This data will be visualised in real-time maps. Finally, reports (.docx and .xlsx) and presentations (.pptx) will also be generated during the project; these will be stored as files in the Net4Cities Sync & Share folders (HIFIS - Helmholtz Federated IT Services Nextcloud Instance[1]).

[1] https://helmholtz.cloud/services/?serviceID=6dd798c4-72cc-4661-aae8-f47a1f3852ce

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The overall aim of Net4Cities is to facilitate the realization of the EU Green Deal's Zero Pollution Action Plan (ZPAP) by advancing air and noise pollution monitoring infrastructure and providing evidence-based support for implementing effective transport policies and thereby improving air quality and mitigating noise pollution. The data generated will provide information on new/emerging pollutants in urban areas related to transport sources, enable source apportionment analysis, allow for the assessment and improvement of transport-related emission inventory information, and inform real-time decision-support.

What is the expected size of the data that you intend to generate or re-use?

This is yet to be determined and will vary across the WPs, with WP2 and WP3 in particular, generating a substantial amount of data, as well as WP5 both using and generating modelled data. Somewhere in the realm of 10-100 TB is a rough estimate.

What is the origin/provenance of the data, either generated or re-used?

See the answers to the first question for the origin of the re-used data. For the generated data, this will be from in-situ measurements and from satellite data.

To whom might your data be useful ('data utility'), outside your project?

The generated data will also have utility for the public health and epidemiology communities, related to objective 2, supporting health assessments. Beyond that it will also have direct utility for municipalities, environmental agencies and similar government bodies to inform air quality and noise policy. It will also be relevant to the broader scientific community and other stakeholders in atmospheric science, urban planning, environmental science.

2. Fair Data

2.1 Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

To ensure effective data management within Net4Cities, and to guarantee long-term accessibility, discoverability, accurate citation and attribution, we will employ Digital Object Identifiers (DOIs) as persistent identifiers. This will also facilitate interoperability between different systems and platforms, including versioning, granularity, and tracking of data usage and impact.

NILU has the capability and resources to generate DOIs for datasets and will allocate unique identifiers to new datasets upon request. If difficulties arise in utilizing DOIs, Zenodo IDs (https://zenodo.org/) will be considered as an alternative option.

For open and shared source code, we will use the automatic Zenodo-GitHub integration supported by Zenodo to create DOIs for the Net4Cities code repositories and archives.

Data generated in the project will also be accessible by an open API, to be developed in the project.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Net4Cities will be dedicated, as far as possible within the project, to utilizing the existing standards. We will strive to be compliant with one of the two listed below:

- ACTRIS (Aerosol, Clouds and Trace Gases Research Infrastructure) Vocabulary[2] . ACTRIS metadata follows the guidelines set out in the ISO 19115 standard, which is dedicated to the documentation of geographic information and services. This standard ensures thorough documentation of various aspects such as identification, extent, quality, spatial and temporal attributes, and other relevant details of digital geographic datasets. Key components of the ISO 19115 standard include:

  ISO 19115-1:2014 - Core requirements for fundamental metadata.

  ISO 19115-2:2019 - Extensions specifically for imagery and gridded data.

  ISO 19139 - XML schema implementation, facilitating the encoding of metadata according to the ISO 19115 standards.

  These standards collectively ensure that metadata is interoperable, easily accessible, and properly documented across different platforms and systems, enhancing data management and usability.

- Air Quality EioNET Data Dictionary[3] . The EioNET Data Dictionary also use the ISO 19115.

A decision between the two options will be made later in the project, as will the level of compliance with these meta data vocabularies.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

There will be some search keywords implemented to a varying degree where metadata and references to relevant external vocabularies and code lists are available.

Will metadata be offered in such a way that it can be harvested and indexed?

To make the Net4Cities metadata vocabulary machine-actionable, it will be served through NILU's Data Platform's REST-API endpoint.

[2] https://vocabulary.actris.nilu.no/skosmos/actris_vocab/en/

[3] https://dd.eionet.europa.eu/vocabularies?expand=true&expanded=&folderId=1#folder-1

2.2 Making data accessible

Repository:

Will the data be deposited in a trusted repository?

The access protocol for the metadata will be clearly defined. Both the data format and access protocol will be provided as machine-readable metadata. Access to the monitoring data (noise, air quality and traffic) and model data (noise and air quality inventory data and model results), including metadata will be available directly through the sensor data REST-API endpoints and the Sensor Data web portal, which enables users to search, analyse, and download data produced within the Net4Cities project.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Appropriate arrangements will be made throughout the project to establish a repository where the data will be deposited. This includes agreements on data format, access protocols, and compliance with repository standards to ensure seamless integration and accessibility.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

See answer above.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All data that is being re-used is already openly available, including the air quality and noise monitoring data. In general, much of the emissions data for cities is also available, although in some cases for some cities this will not be open. Data regarding air quality and noise monitoring will also be made available, much of it in 'near real-time'. For data sharing we will require people to register for a login so that we know who has downloaded data and can provide appropriate use attribution requirements.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Embargo periods are optional and are specifically envisaged for early-career researchers in order to give them the appropriate amount of time to publish their results. Embargo periods should not extend 12 months after the data has been processed.

Will the data be accessible through a free and standardized access protocol?

Data collected or generated by the Net4Cities project will be accessible through a standardized access protocol. The open data will be freely available without requiring any authentication.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

If there are commercial interests in using specific datasets or parts of datasets, token authentication will be required for the Sensor REST-API endpoint, and user/password authentication will be necessary for the web portal to access the data. In the case of sensitive data being reused or generated within the project, the implementation of two-factor or multifactor authentication will be considered. This authentication solution will be applicable both during the project and after its completion.

How will the identity of the person accessing the data be ascertained?

To get access to restricted data and services, all users of NILU's Data Platform and Net4Cities' code repository/archive must first register to login.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

NA.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

To ensure the widest possible dissemination and unrestricted use of the metadata, a public domain dedication will be employed. Specifically, the Creative Commons Zero (CC0) license will be utilized, which effectively places the content in the public domain. This dedication enables users to freely use, modify, distribute, and build upon the content without any legal restrictions.

Access instructions for the data will be provided via the metadata, to enable seamless machine-to-machine interaction. This should not be the case, but if the size of the data is too large or the data is sensitive, direct access will be restricted and contact information on institutional and/or personal level will be provided.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Timeseries datasets generated from the Net4Cities monitoring infrastructure will be stored in NILU's Data Platform, which is a research infrastructure that is used for several research projects. The ambition for the infrastructure is to be operational for many years after the end of Net4Cities. In other words, as long as the infrastructure is in use, the storage and access of metadata and data will be available and findable.

If NILU cease its operation the plan is to transfer data and metadata to another data repository. For this we could use the GFZ Data Services. There data and metadata are stored separately which would also provide the opportunity to only use the metadata and link (on the specific landing pages) to external data sources.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

The aim in Net4Cities is that no specific or proprietary software will be required for data access and/or downloading/analysing data. Any software, tools, and code produced in Net4Cities will be available through open access or restricted repositories like GitHub/GitLab. An open-source license for the software will be strongly preferrable and encouraged. All developments related to EarthSense MappAir tools will not be open-source as these are a commercial product. This however, should not affect the data access for data that is used in MappAir, as this would be accessible via the data hub.

2.3 Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

As noted in section 2.1, Net4Cities will aim to use existing metadata standards and vocabularies, such as from ACTRIS or Air Quality EioNET Data Dictionary. This will ensure interoperability, exchange and re-use within and across disciplines. The above-mentioned vocabularies and standards are community-accepted best practice.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

NA

Will your data include qualified references[4] to other data (e.g. other data from your project, or datasets from previous research)?

Cross-references between different data within the project should be created wherever technically possible and useful for subsequent integrative analyses. The exchange formats are selected in such a way that this is fully supported.

2.4 Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

This documentation will be linked to the publication strategy. The documentation for validating data analysis and facilitating data re-use will either be included in a detailed methodology part of the publication or if not, as part of a corresponding data description document.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Net4Cities will publish data with the most open CC license possible, which will generally be CC BY 4.0 or CC0, unless this is not feasible for IP reasons.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes. See also answer to the previous question.

Will the provenance of the data be thoroughly documented using the appropriate standards?

See section describing data formats.

Describe all relevant data quality assurance processes.

Already integrated into the NILU Data Hub are automated data processing routines that include QC flags. In addition, the Helmholtz Centre for Environmental Research (UFZ) has an open-source quality control system (SaQC) that is available for use and provides extensive QaQc procedures in the Python programming language. These can be easily implemented into QaQc protocols in Net4Cities, where such protocols are not yet available.

---

[4] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

For outputs such as methodologies or code, the GitLab service provided by the Helmholtz Cloud could be used, which would enable access for all partners via the virtual organization. This should be publicly readable without additional costs associated.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

The project team includes partners with existing data infrastructure which will be used for the project. Cost related to acquisition of data & operating the infrastructure are budgeted in the grant.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

The costs associate to research data/output management were planned for in the project budget and will be covered under that. The necessary financial resources will be allocated across WPs and partners as appropriate.

Who will be responsible for data management in your project?

The Project Coordinator (GFZ-RIFS) and the host of the DataHub (NILU) will be responsible.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

We will aspire to follow the German Research Foundations' (DFG) good scientific practice guidelines for long-term preservation of data. This means that we are obliged to store data for at least 10 years and everything beyond the 10 years will be assessed as art of process to evaluate the need for longer-term archiving, which aims as safeguarding primary and potential secondary usage. Workflows are in place to address this.

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

All data stored in the sensor data platform generated by the Net4Cities and source code for the sensor data platform will be stored within NILU's infrastructure. NILU is a leading environmental institute and are hosting and maintaining several large thematic databases for on behalf of, among others, the UN, the World Meteorological Organization (WMO), several European research infrastructures and the European Space Agency (ESA). The data center is distributed across two connected buildings and daily off-site backups of servers, databases, and files are conducted to a third location. The internet infrastructure is supported by Sikt, the Norwegian Agency for Shared Services in Education and Research (https://sikt.no/en/home). NILU is using an advanced network security system from a leading firewall security provider. The internal network is protected by two firewalls in an active/passive configuration, with the network architecture being fully redundant from the firewalls to the servers. Virtualization is utilized to maintain system stability and redundancy.

Any data related to the project which is not hosted in NILU's system, but hosted by other project partners or external entities, is expected to not be project-critical nor sensitive to leaks or other security issues. This will be reviewed regularly (at a minimum for the regular reporting periods) and updated if this is identified to not be the case.

Will the data be safely stored in trusted repositories for long term preservation and curation?

The exploitation plan for the Net4Cities project will be designed to ensure the continued use and development of project results after its completion. To achieve long-term preservation the sensor data platform is built on open-source big data scalable technologies provided by Apache Software Foundation, which will ensure a robust and available infrastructure for the collected and generated data from the Net4Cities project and its future exploitation.

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

We have not yet identified any ethics or legal issues that will impact data sharing. We will regularly re-assess this over the course of the project.

Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

NA

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

NA

Conclusions

The Net4Cities Data Management Plan outlines key aspects of the processing, storage, and management of all data that will be generated or used during the course of the Horizon Europe project. As these data have not yet been collected or generated, the DMP will initially remain more general. This deliverable (D7.2) is the first of several that will address data management in the project. First, second, and final revisions (D7.6, D7.7, and D7.8, respectively) will be made to this plan as the project progresses to ensure all relevant aspects of data management are covered. All data in this project will adhere to FAIR principles. All data will be open-source and protocols open-access, except in a select few cases where intellectual property for commercial products is of relevance.

Acronyms

Table 1. Table of acronyms used in this template.

Acronym	Meaning
ACTRIS	Aerosol, Clouds and Trace Gases Research Infrastructure
CC0	Creative Commons 0
DFG	German Research Foundation
DMP	Data Management Plan
DoA	Description of the Action
DOI	Digital Object Identifier
EioNET	European Environment Information and Observation Network
ESA	European Space Agency
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reuse
ISO	The International Organization for Standardization
QaQc	Quality Assurance and Quality Control
UFZ	Helmholtz Centre for Environmental Research
UN	United Nations
WMO	World Meteorological Organization
WP	Work Package
ZPAP	Zero Pollution Action Plan