Attributing and Verifying European and National Greenhouse Gas and Aerosol Emissions and Reconciliation with Statistical Bottom-up Estimates

Deliverable 7.5
Avengers Data Management Plan v2

Authors: Alex Vermeulen, Marko Scholze & Jean-Nicolas Poussart (ULUND)

Date: 30/06/2024

Dissemination: Public

Work package: 7.4

Version: 2.0

Funded by the European Union

Attributing and Verifying European and National Greenhouse Gas and Aerosol Emissions and Reconciliation with Statistical Bottom-up Estimates (AVENGERS)

Call: HORIZON-CL5-2022-D1-02

Topic: HORIZON-CL5-2022-D1-02-01

Type of action: HORIZON Research and Innovation Actions

Granting authority: European Climate, Infrastructure and Environment Executive Agency

Project starting date: 01/01/2023

Project end date: 30/06/2026

Project duration: 42 months

Contact: Dr. Marko Scholze, Coordinator
Lund University, Sweden

Document history:

Version	Author(s)	Date	Changes
1.0	A.T. Vermeulen	07/07/2023	Final draft
2.0	A.T. Vermeulen	01/07/2024	Updates reflecting the project progress, clarifying data releases/open data

Internal review: VUA, EMPA

Table of Contents

PLAN OVERVIEW

A Data Management Plan created using DMPonline

Title: AVENGERS - ATTRIBUTING AND VERIFYING EUROPEAN AND NATIONAL GREENHOUSE GAS AND AEROSOL EMISSIONS AND RECONCILIATION WITH STATISTICAL BOTTOM UP ESTIMATES

Creator: Alex Vermeulen

Principal Investigators:
> **Marko Scholze (coordinator) : Lund University**
> **Sander Houweling (co-coordinator): Vrije Universiteit Amsterdam**
> Andre Butz : Heidelberg University
> Anna Maria Jönsson – Lund University
> Christian Mielke : UBA
> Dominik Brunner : EMPA
> Hugo Denier van der Gon : TNO
> Ines Marinosci : ISPRA
> Carolina Gallo : CMCC
> Thomas Kaminski : Inversion Lab

Funder: European Commission
Template: Horizon Europe Template
ORCID iD: 0000-0002-4007-6902

Project abstract:
Regional and national authorities and scientists are attempting to characterise greenhouse gas (GHG) and aerosol emissions that contribute to global climate change, using both bottom-up and top-down techniques, often with differing results. Bottom-up methods, for example, focus on measurements of a representative sample of equipment; this produces statistical estimates of atmospheric emissions. Top-down methods can be performed at a regional scale, based on measured concentrations in the air at specific sites or via satellite, for instance. The EU-funded AVENGERS project will advance the top-down approach for quantifying GHG and aerosol emissions, and will reconcile and integrate approaches into a joint bottom-up and top-down framework. The enhanced estimations will support policymaking by local/regional authorities, countries and international organisations.

ID: 125938
Start date: 01-01-2023
End date: 31-12-2026
Last modified: 01-07-2024
Grant number / URL: https://doi.org/10.3030/101081430

DATA SUMMARY

Will you re-use any existing data and what will you re-use it for?

We use datasets of surface fluxes of greenhouse gases and aerosols specified in time, space and by process category. Furthermore, we use atmospheric measurements from instruments at the surface, on mobile platforms and satellites.
The datasets will be used to perform atmospheric inversions, generating new datasets of surface fluxes.

We will re-use land/use cover data from Corine Land Cover dataset and other Copernicus land products and UNFCCC Inventory data of Italy, Sweden and Netherlands. The datasets will be used to improve the spatialization of national inventories bottom-up estimate of GHG emissions and removals (LULUCF), and N_2O emissions (agricultural soils) in WP 4.3.

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

We expect the data to be mostly in netcdf format, which will be true at least for those generate by the project. Some of the re-used data will be in GIS (ArcInfo) format and specific community standard CSV text files. The land/use cover data are at national level (regional level for Italy) in vector type (shapefile or geopackage format) or raster type (geotiff format), the inventory data are at regional and provincial level in tabular format or gridded (0.1°x 0.1°).

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

The project objectives require estimation of greenhouse gas and aerosol fluxes from atmospheric measurements in combination with atmospheric transport models as additional information to be used by environmental agencies to improve their national inventory reports. In task 4.3 the specific purpose is to produce spatially explicit estimates of GHG emissions for the AFOLU sector.

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

The following three main types of data will be used or produced. We provide a first estimate of data volume for each.

1. Observation data (re-used):
   a. In situ observations: 1-2 Gb
   b. satellite observations: xx
2. A priori emission fluxes (re-used and produced)
   c. Global data at 0.1°x0.1° and monthly resolution (2010-2022): ~500 Gb (50 Mb/field x 3 gases x 15 categories/gas x 156 months)
   d. European data at 5 km x 5 km and annual resolution (2018-2022): ~5 Gb (16 Mb/field x 3 gases x 15 categories/gas x 5 years)
      - National inventories for 5 countries at 1 km x 1 km resolution: ~2 Gb

3. A posteriori emission fluxes over Europe and model simulated observations (produced): ~1 Tb.
The exact number is difficult to estimate as it depends on grid output resolution, number of simulations, number of models, etc.

In addition to the above, each model group will produce multiple Tb of 3D intermediate model output fields, which however, will not be provided as output from the project.

All data generated within the project will become available as open data using a Creativ Commons Attribution 4.0 International (CCBY4) licence shortly after production and testing or only when needed with a moratorium set for release at the end of the project. A moratorium might be requested by the publisher of a journal to which the data is supplemented or by the data provider until the interpretation of the results based on the data has been published.

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

The origin of the data will be ascertained after the data collection process.

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

Other scientists and experts in greenhouse gas emissions at environmental agencies.

FAIR DATA

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Yes, all data to be retained will be provisioned with a persistent identifier, in the default case of using the ICOS repository this will at minimum be a Handle PID (ePIC), and in most cases also a Datacite DOI PID will be minted. In some cases data will be published using other public archives with similar functionality that mints PIDs or DOIs, like Pangea or Zenodo. In case of use of Pangea or Zenodo their data management practices apply, that we assume are already accepted as best practice.

2.1. Making data findable, including provisions for metadata: 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.

Yes, all PIDs will resolve to a landing page that contains human and machine readable rich data that allows the users to find and access the data using detailed (faceted) search criteria, selecting on keywords, variable names and other elements of the rich metadata provided at ingestion, that will follow where possible the INSPIRE themes.

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, keywords will be provided, users are encouraged to select keywords from the GCMD (Global Change Master Directory) standard. In the Carbon Portal data portal files can be selected based on a faceted search that includes the keywords, data type and other metadata elements.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

Yes, at the data portals used like Zenodo, Pangaea and ICOS Carbon Portal (CP) all metadata is accessible to both human and machine to machine communication, at CP using the following standards: iso19115, Schema.org, and by content negotiation as xml, json-ld, rdf/turtle, rdf/xml. All metadata is served from the landing pages and the open sparql endpoint.

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

The ICOS Carbon Portal is the trusted data repository for the large ICOS community and beyond, CP guaranties the availability of the metadata and data for the foreseeable future (2035 and beyond). All data is securely also copied using the EUDAT CDI B2SAFE service, that stores two copies of each digital artefact at two locations (one at CSC, Finland and one at

FZJ, Germany). Next to that ICOS Carbon Portal stores its own copy of the data, with a separated backup at two different buildings at Lund University, Sweden.

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

ICOS ERIC is in full control of the ICOS Carbon Portal as owner of the repository. ICOS is a long term research infrastructure, legally presented by ICOS ERIC, that is foreseen to exists permanently, for at least 20 years (starting 2015).

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, each data object is assigned a persistent identifier that contains (part of) its digital SHA256 checksum. Through the landing page that the PID resolves into, a direct link to the digital object is provided together with license and access metadata.

2.2.   Making data accessible - 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 raw and higher level data products provided will follow the ICOS data licence which is CC-BY 4.0. Other data licenses are possible, like CC0.
All data will become available as open data using a Create Commons licence International with Attribution (CCBY4), immediately after production and test. Only when needed a moratorium will be set for release at the end of the project. Such moratorium might be requested by the publisher of a journal to which the data is supplemented or by the data provider until the interpretation of the results based on the data has been published.
The temporarily restricted data will be open to project participants.

2.2.   Making data accessible - Data:
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.

The word used for embargo used in ICOS CP is moratorium. This is answered in the previous section.

**2.2.   Making data accessible - Data:
Will the data be accessible through a free and standardized access protocol?**

All data (and metadata) is available using standard https protocol using semantic web and linked open data technology.

2.2.   Making data accessible - Data:
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

There are no restrictions on use, users are requested to provide results using the data using the same data license and use the provided citation and acknowledge the data use and offer co-authorship if the data used are significant for the end result. Re-distribution is allowed as long as the license and conditions are passed on.

2.2.   Making data accessible - Data:

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

If user identification is required for data with restricted access, users will have to identify through the Carbon Portal login, using either Edugain, ORCID or local account login. All other data is available without user identification. We log data access purely based on IP and date/time.

2.2.   Making data accessible - Data:

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

Not required for the environmental data gathered in the framework of this project.

2.2.   Making data accessible - Metadata:

Will metadata be made openly available and licensed 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?

Yes, all ICOS CP metadata is licensed through CC0. Part of the metadata landing page information is an explicit reference to the data object for direct access.

2.2.   Making data accessible - Metadata:

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

ICOS is a long term research infrastructure, legally presented by ICOS ERIC, that is foreseen to exists permanently, for at least 25 years. ICOS Carbon Portal will preserve data and metadata for the lifetime of ICOS and beyond. After the termination of ICOS all data and metadata will be transferred to a static website and/or other trusted repositories during the decommissioning period of at least one year.

2.2.   Making data accessible - Metadata:

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)?

Data providers can add additional documents or links to software (e.g. in a Github repository) to any data object.

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?

Interoperability can only be defined in relation to an actual implementation and solution, and is thus by definition dependent on technology and standards. All ICOS data objects are stored in the B2SAFE repository. the associated metadata linked to the data is stored in a RDF database, based on an open ontology based on OWL, that is part of the RDF database. Read-only access to the metadata repository is given through an open SPARQL endpoint. All metadata is also exported to the B2FIND repository where it is also linked with the PIDs of the data objects in B2SAFE through CKAN. The B2FIND repository is again linked to the GEOPortal for global access to the metadata form other portals and portals of portals like GEOSS and the EOSC Marketplace.
The landing pages of the data object will allow for content negotiation to deliver the metadata in the format and vocabulary of the respective community standards. This translation using equivalences will be dynamic and online, will be anchored in the ontology and thus open and easy to maintain and update. Keywords follow in general the GCMD standard, data formats follow the community standards such as cf-netcdf, FLUXNET, WMO GAW etc.

2.3.   Making data interoperable:
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?

The ICOS ontology is modelled in owl and can be queried through the open sparql endpoint. The ontology is published at https://lov.linkeddata.es/dataset/lov/vocabs/cpmeta. All metadata is mapped to schema.org and iso19115.

2.3.   Making data interoperable:
Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1]*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/*)*

At the moment we only allow for association of data with other data without further qualification. The idea is that the cross-referencing can be performed better using the rich metadata attached to each data object, including the associated data. Associated data can be publications, protocols and also other documents

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.)?

Same as ICOS data providers, AVENGERS data providers are encouraged to provide links to supporting documentation with the data together with relevant descriptions (abstracts) and title of the dataset that also include references to methodology. The data will be foreseen with detailed provenance data on instrumentation and methods used and includes detailed information on location, variables, units, data providers and their roles, all modelled over time.

2.4.   Increase data re-use:
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?

Data will be provided in general according to the ICOS data Policy and using the Creative Commons Attribution 4.0 International (CC4BY) licence.

2.4.   Increase data re-use:
Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, all data and metadata will be stored permanently at the ICOS Carbon Portal for the foreseeable future (until 2035 and further).

2.4.   Increase data re-use:

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

Carbon Portal provides provenance of all data objects following the PROV-O ontology.

2.4.   Increase data re-use:

Describe all relevant data quality assurance processes.

Quality assurance is at the heart of ICOS and the reason of existence of the research infrastructure. The quality assurance procedures are described in the relevant papers and reports that are published by ICOS, the Thematic Centers and their contributors.
The ICOS data life cycle is documented in the Improved ICOS data Life Cycle document (https://doi.org/10.18160/D2JV-KB6B).

2.4.   Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

This is addressed in the following sub-sections

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.).

Software developed in the project will in principle be provided as open source through platforms such as Github using open licenses such as GPL. Exceptions are existing model codes that are background IP and modifications to that software. Workflows and protocols developed in the project will be published as project results through the deliverables and will be open without restriction.

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.

All scientific results will be shared using scientific best practices in open peer reviewed literature and the appropriate deliverables for public dissemination. All links to the re-used data and data produced will be referenced in the articles and reports, either directly or through data supplements through links with the PIDs, that will reference to the data and metadata as explained in this DMP, section 2.

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.) ?

No additional costs are associated for the project with the FAIR implementation at ICOS Carbon portal

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)

not applicable

Who will be responsible for data management in your project?

Data management is the responsibility of all actors in the project. (Meta)data curation for submission of data to ICOS CP will be the responsibility of the ICOS CP data stewards.

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)?

ICOS Carbon Portal has a long term perspective of more than 20 years. In case of winding down of ICOS or ICOS Carbon Portal, a contingency plan exists to provide a safe harbour to data and metadata. ICOS CP has an annual budget of about 2 M€ that allows for 15 fte, that allows for continuous development of the data services (5 fte), provision of Jupyter VREs (3 fte), data stewards (2 fte), data scientists (3 fte) and the costs for the CP cloud services and B2SAFE. Budget is decided by the ICOS General Assembly, consisting of representatives of the currently 16 European member states. Most of its core budget is provided by the Royal Kingdom of Sweden.
All data is already replicated at two instances using the EOSC service by EUDAT CDI B2SAFE at servers in both Finland and Germany.

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)?

Data at Carbon Portal is back upped daily through a server dislocated from the other main servers for data recovery. Besides that there is the B2SAFE replication discussed under the previous items. There is no sensitive data handled by CP. Data integrity is ensured by using the SHA256 checksums used at all transfers by sender and receiver and inclusion of this information in the object Handle PID.

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).

The ethical issues related to qualitative data collection include the handling of personal data (GDPR, anonymity, confidentiality, informed consent) are discussed in the ethics deliverables. They are related to collection of climate risk data in the task 2.3.
- The grant proposal and agreement include an ethical questionnaire
- ICOS-ERIC has ethical guidelines and a gender equality plan which can be referred to.
- ICOS-ERIC has an ethical advisory board that can provide guidance if needed.

In addition to these, each partner would ensure that the ethical guidelines are followed as pertaining to their country and/or institution context.

Due to the geographic dimensions of this project (including different continents and time zones), communication and qualitative data collection will partly rely on online platforms (as opposed to e.g. face to face project meetings). This presents an ethical issue in the form of data storage and access (e.g. in situations where the contents are recorded). As in all activities in the project, GDPR and inclusivity are applied in these situations also.

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

Yes. Any interview / survey / workshop participant will be provided, in a relevant language, with information of the purpose of data collection and the handling of personal data, and the study's aims and content, and asked for an informed consent. Participants will be informed of the voluntary basis of participation. Obtaining informed consent will take into account the specificities related to culturally perceived concept of personhood (individual consent versus collective consent). All personal information will be handled according to the GDPR regulation.

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)?

None