First Update of the DMP

Work Package:	8
Deliverable:	8.9
Due Date:	31.10.2024 (M24)
Submission Date:	31.10.2024 (M24)
Dissemination Level:	Public
Type:	DMP
Responsible:	NORCE
Author(s):	Ozan Mert Göktürk (NORCE), Ivan Puga Gonzalez (NORCE), Jesus Fernandez (CSIC)

Funded by the European Union

Contents

1 Summary for Publication 1

1.1 FAIR data 2
1.2 Making data accessible 3
1.3 Making data interoperable 5
1.4 Increase data re-use 6

2 Other research outputs 7

3 Allocation of resources 7

4 Data security 8

5 Ethics 8

6 Other issues 8

Abbreviations Used

xxx	

1 Summary for Publication

This is the first update of the Data Management Plan for the project Impetus4Change (I4C).

In this update, we highlight activities related to the internal sharing, interoperability and reusability of I4C data and tools. These include:

* Establishment of a data storage and manipulation area, by Norwegian Research Centre AS (NORCE), on the facilities provided by the Norwegian Research Infrastructure Services (NRIS / Sigma2).

* Deployment of a local, cloud-based data lab (I4C-Hub), by CSIC (Spanish National Research Council), where the production of FAIR data has started to be set up in the context of Empirical Statistical Downscaling (ESD) data.

* Establishment of I4C Github pages.

Details can be found in the "Repository" section.

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.
Some existing data will be re-used (as input to the project tools) to generate new data.

What types and formats of data will the project generate or re-use?
Data that will be re-used are global climate model output for historical and future periods as well convection-permitting regional climate model outputs from the CORDEX FPS-convection and EUCP project as well as global reanalysis data and observational data from weather stations. These will be used as input to various climate models and tools within the project, which will generate regional and local scale climate information. The data format will predominantly be NetCDF, which is one of the standards in the broad field of climate modeling. Conversion to other, less technical data formats such as CSV text files or spreadsheets will be considered and implemented during the project, depending on user and stakeholder needs. For example, WP4 which focuses on hazards and extremes, will generate datasets on climate risk indices which will be stored in NetCDF format, but these can be modified to other formats for use in the demonstrator cities and beyond.

Social scientists in the project will generate two types of data, one related to in-depth interviews and one related to Agent-Based Modelling (ABM). Data from the interviews will include transcripts of the interviews, this data will be anonymized. Transcripts of the interviews will be in text format. Data from the ABM will include measurements of the variables of interest under the different modelling conditions. The format output will be csv. In addition, a new dataset will be generated from WP1 T1.1, the knowledge network literature review. This dataset will contain information regarding the

knowledge networks identified during this task, and of the members of each of the knowledge networks, it will also be in CSV format.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
In our project, data re-use is a requirement to generate new data, as explained above. The purpose of generating new data is to obtain local scale climate information for selected European cities (and the Caribbean), in order to be able to assess the impact of weather and climate extremes under future climate change. These are direct objectives of the project.

The in-depth interviews information will be used to generate a knowledge network and make an in-depth analysis of its architecture. The data from the ABM model will help understand what are the best configurations of the knowledge network so that it can create and transfer knowledge among its nodes optimally. The data generated from the model will represent network architecture measurements, nodes' characteristics measurements, and network optimization measurements. The dataset on knowledge networks is generated to support the literature review in Task 1.1 by systematically documenting the members of identified knowledge networks in climate change adaptation. This dataset will also be valuable for other researchers and scientists interested in climate change adaptation and knowledge networks, as it offers a reusable resource for mapping existing knowledge networks across Europe.

What is the expected size of the data that you intend to generate or re-use?
This will be on the order of 100-300TB (subject to updates).

What is the origin/provenance of the data, either generated or re-used?
The origin/provenance of our data will be global and regional climate models, hydrological models, reanalysis datasets, data collected from weather observation stations and other climate tools such as climate model emulators. Some data will also be generated through in-depth interviews, agent-based modelling (ABM), and literature review.

To whom might your data be useful ('data utility'), outside your project?
Our data will be useful to any end user of climate information, from climate scientists to stakeholders, decision makers and public bodies. Further, social scientists interested in the architecture and functioning of knowledge networks in climate change adaptation. It may also be useful to social scientist modelers interested in network analysis and efficient transmission of information.

1.1 FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?
Yes, our data will have PIDs.

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.
The format of our data will predominantly be NetCDF, which is one of the standard formats of the broad climate modelling field. NetCDF is a 'self-describing' data format, which means that the metadata are included in the file itself and can be accessed using computer codes.

Where appropriate, open metadata standards and naming rules (CF-conventions, CORDEX guidelines, CMIP/CMOR attributes and file naming conventions, CMIP controlled vocabularies) will be adopted. Additional metadata standards have been proposed within the project for the new tools developed (emulators), following other more general standards (CF and CORDEX).

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

Will metadata be offered in such a way that it can be harvested and indexed?
For climate modelling search keywords are directly accessible from file names (facets following the thematic conventions – e.g., those for CMIP or CORDEX), and these can be used to feed catalogue indexes.

1.2 Making data accessible

1.2.1 Repository:

Will the data be deposited in a trusted repository?
Yes. Some of the generated model output will be published on the Earth System Grid Federation (ESGF). It will be accessible through both the ESGF and also the European Open Science Cloud (EOSC). Any other research output classifiable as data will be published on EOSC or partner archives and made publicly accessible through e.g., ftp.

NORCE has established a data storage and manipulation area on the facilities provided by the Norwegian Research Infrastructure Services (NRIS / Sigma2). This area is accessible by all project members and work packages through Secure Shell (SSH). It is intended for sharing project data sets (both re-used and generated) internally. As many I4C data sets are created in close collaboration of its partners, and because part of these data (namely, climate model output) will be the input for other scientific work at I4C, the establishment of the described storage space was crucial for the project to progress smoothly. Further, the storage area will serve as the intermediate platform from which the finalized I4C data sets will be uploaded to the publicly accessible data platforms such as the ESGF.

CSIC has deployed a local, cloud-based data lab (I4C-Hub) where the production of FAIR data has started to be set up in the context of ESD data. For the moment, we focus on the Interoperability and Reusability of the data by adopting community

(CORDEX) standards on NetCDF files and by providing notebooks with examples of exploiting these data. This initial seed of the I4C-Hub is in testing phase for internal use, and will be extended to other project results and migrated to dedicated resources benefiting from the EOSC services during the next reporting period.

The resources currently available are:
Data:
-Decadal predictions (CMIP6-DCPP) for the EC-Earth3 model hindcast
-The CERRA reanalysis as the ESD observational reference
-Seasonal Prediction System 5 Forecasts
-Sea level pressure of the ERA5 reanalysis
Software libraries:
-Python libraries (xarray, pandas, geopandas, …)
-R frameworks: CSDownscale and climate4R
-Command-line tools (CDO, NCO, …)
-Scripts and notebooks to apply ESD techniques to seasonal and decadal forecasts.

These are also available at the I4C Github pages (https://github.com/impetus4change/).

Have you explored appropriate arrangements with the identified repository where your data will be deposited?
Yes. The partners of our project have prior experience in publishing their research data on the aforementioned repositories.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?
The repositories we will assign digital object identifiers to data.

1.2.2 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 will normally be openly available. In the event of any deviation from this, explanations will be provided accordingly. However, due to privacy concerns data from the in-depth interviews will be anonymized.

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.
No embargos will be applied.

Will the data be accessible through a free and standardized access protocol?
Yes, we will create an open data space on EOSC. Also ZENODO, ESGF and other free access spaces will be used appropriately. The DMP task team will also explore other

FAIR solutions to ensure accessibility of I4C data. This will be better defined in the next DMP update.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? How will the identity of the person accessing the data be ascertained?
There will be no restrictions on use.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?
No.

1.2.3 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.
Yes.

Will metadata contain information to enable the user to access the data?
In the case of climate data, the metadata are not hosted externally but are attached to the data themselves. Therefore, this question is not really applicable to I4C. However, the host repositories will include information to enable user access.

How long will the data remain available and findable?
Long term preservation is a challenge/problem (storage costs, both financial and environmental are high and not covered by the project). 5-10 years is the industry standard for such archiving, and we will aim to meet this.

Will metadata be guaranteed to remain available after data is no longer available?
As our data will be self-describing (i.e. containing its own metadata), the metadata will stay available as long as the data are available. For data generated from interviews and in other format a solution is under discussion

Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code).
Software needed to access or read NetCDF files are widely and openly available. Relevant software will be provided for any other data or file types requiring such software for access. A software environment to access and read these data will be provided via the EOSC data space. Reproducible environment files will be provided for users to replicate locally the software environment to access and read the data.

1.3 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?

The NetCDF data format is one of the standard formats in climate science, therefore it can be accessed and read using many different programming languages and software. These include widely used languages and software such as Python or ArcGIS, therefore the interoperability of our data across many different disciplines is readily ensured. We will also employ standard vocabularies such as CF-conventions and CMOR in order to maintain consistency and exchangeability across other disciplines.

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?
It is unlikely that we will generate project-specific ontologies or vocabularies, but we will use existing ones (e.g., CF-conventions, CMOR). If we do, these will be mapped to common ontologies, and be openly published to allow reusing.

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?
Yes. For example, the metadata tied to the NetCDF files of regional climate model output will contain references to global climate model output used as input for downscaling. These follow standards established by previous and ongoing research (e.g., CORDEX).

1.4 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.)?
All documentation needed to facilitate data re-use will be published through the European Open Science Cloud (EOSC) and the project's GitHub organization.

Will your data be made freely available in the public domain to permit the widest re-use possible?
Yes, all data will be freely available, requiring only attribution for re-use (CC BY license).

Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?
Yes, where appropriate we will use Creative Commons license attribution.

Will the data produced in the project be usable by third parties, in particular after the end of the project?
Our data will be usable by all third parties once they are published. See answer to previous question regarding the challenges of long-term archiving. We aim for 5-10 years accessibility beyond the end of the project.

Will the provenance of the data be thoroughly documented using the appropriate standards? Describe all relevant data quality assurance processes.
Climate model output to be generated in the project will have the appropriate data and metadata standards. The quality of these will be assured through a quality control procedure, called PREPARE, which is routinely used for the output of global climate model simulations and others for regional models (https://github.com/IS-ENES-Data/QA-DKRZ).

2 Other research outputs

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.

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.). 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.

Any research output other than data will be managed and shared through appropriate channels, such as Zenodo, GitHub or EOSC. Details of this will be provided as updates in this data management plan, as the project generates said output in due course.

3 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.) ? 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).
It is a challenge to estimate these costs. Many services are provided by national bodies and data centers. Often these provide archiving services on a rolling basis free of charge, but that can always change. During the project period, costs have been set aside for e.g., establishing an EOSC data space.

Who will be responsible for data management in your project?
We have a data management task team, led by Ozan Mert Göktürk (NORCE).

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)?
Data generated in the project will be available for 10 years, which is the industry standard. As with the first question in this section it is not possible to estimate these costs.

4 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)? Will the data be safely stored in trusted repositories for long term preservation and curation?
The repositories we will be using have their own measures for long-term preservation ensuring data security through e.g., HTTPS, checksum, replication, etc. Long term storage/archiving will be through resources such as Norway's Research Data Archive (https://www.sigma2.no/research-data-archive).

5 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).
There are no ethical issues that can have an impact on the sharing of climate data. However, surveys and interviews will be done in accordance to the ethics regulations of NORCE and the EU and subjected to the approval of the ethics committee.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?
Yes. Participants will be given the option of being part or not of the survey/interview exercise.

6 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)?
No.

IMPETUS4CHANGE (I4C)

IMPROVING NEAR-TERM CLIMATE PREDICTIONS FOR SOCIETAL TRANSFORMATION

Grant agreement ID: 101081555

Call: HORIZON-CL5-2022-D1-02

Type of Action: HORIZON-RIA

Start date: 1 November 2022

Duration: 48 months

Website	impetus4change.eu
Twitter	@I4C_eu
LinkedIn	Impetus4Change
Zenodo repository for I4C open access documents	Impetus4Change Community