S2S4E

Climate Services for Clean Energy

Research and Innovation action

H2020-SC5-2017

D5.2 Data Management Plan

Deliverable D5.2

Version 1.2

Authors: BSC-CNS, CAPGEMINI

Document Information

Grant Agreement	776787
Project Title	Sub-seasonal to Seasonal climate forecasting for Energy
Project Acronym	S2S4E
Project Start Date	01/12/2017
Related work package	WP 5- Operational Climate Service: Decision Support Tool (DST) implementation
Related task(s)	Task 5.1 Data Management Plan - definition of data protocols and formats
Lead Organisation	BSC
Submission date	31/05/2018
Dissemination Level	PU

History

Date	Submitted by	Reviewed by	Version (Notes)
31/05/2018	BSC	Capgemini, BSC	1.2
30/04/2018		Capgemini, BSC	1.1[1]
22/03/2018		BSC, Capgemini, TCDF	1.0[2]

[1] https://tinyurl.com/ycom83n7

[2] https://tinyurl.com/yaqvldp5

Table of content

List of tables

Summary

This document is the first version of the Data management plan. This is a live document that will be revised twice during the project (D5.3 due at M18 and D5.4 due at M36). The purpose of this document is to present the management of data in the scope of the S2S4E project.

Keywords

Data, privacy, preservation, open access, storage

About S2S4E

The project seeks to improve renewable energy variability management by developing a tool that for the first time integrates sub-seasonal to seasonal climate predictions with renewable energy production and electricity demand.

Our long-term goal is to make the European energy sector more resilient to climate variability and extreme events.

Large-scale deployment of renewable energy is key to comply with the emissions reductions agreed upon in the Paris Agreement. However, despite being cost competitive in many settings, renewable energy diffusion remains limited largely due to seasonal variability. Knowledge of power output and demand forecasting beyond a few days remains poor, creating a major barrier to renewable energy integration in electricity networks.

To help solve this problem, S2S4E is developing an innovative service to improve renewable energy variability management. The outcome will be new research methods exploring the frontiers of weather conditions for future weeks and months and a decision support tool for the renewable industry.

More information: www.s2s4e.eu

1 Introduction

1.1 Purpose of the document

The objective of this document is to define policies and technical solutions about the data collected and generated during the project, across the different work packages, the data being scientific or not (personal data, results from surveys, etc.).
The data presented in this document will be categorized in two.

The first category is the **climate data**, that go from the raw model and observations data (described more thoroughly below) to the final plots of essential climate variables presented in the website of the project.

The second category is the **personal data** and refers to names, emails, addresses of the users and the answers given during the surveys.

1.2 Applicable and reference documents

The applicable documents are listed in the table below:

Id	Deliverables	WP
D8.1	POPD - Requirement No. 1: Information sheet of users	WP8
D8.2	POPD - Requirement No.2: Guideline for personal data management	WP8

2 Glossary/Definitions

Reanalysis: systematic approach to produce datasets for climate monitoring and research. Reanalyses are created via a data assimilation scheme and model(s) which ingest a certain set of observations. The resulting dataset will be referred to as “reanalysis” in this document.

Grib format: GRIB (GRIdded Binary or General Regularly-distributed Information in Binary form) is a concise data format commonly used in meteorology to store historical and forecast weather data. It is standardized by the World Meteorological Organization's Commission for Basic Systems, known under number GRIB FM 92-IX, described in WMO Manual on Codes No.306. Currently there are three versions of GRIB. Version 0 was used to a limited extent by projects such as TOGA, and is no longer in operational use. The first edition (current sub-version is 2) is used operationally worldwide by most meteorological centres, for Numerical Weather Prediction output (NWP)

NetCDF format: NetCDF (Network Common Data Form) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. NetCDF is commonly used to store and distribute scientific data. The latest version of the NetCDF format is NetCDF 4 (also known as NetCDF enhanced, introduced in 2008), but NetCDF 3 (NetCDF classic) is also still widely used.

3 Data collection

3.1 Type of data collected

3.1.1 Original data collected

Due to the characteristics of the activities carried out within the S2S4E project, the data managed are classified in two types: scientific and personal data. Both types will be managed according to the open data policy recommended by the European Commission and the General Protection data Regulation (GDPR) respectively.

It is expected not only to collect data for the correct development of the research activities, but also to create different types of data as outcomes of the project.

3.1.1.1 Climate data

During the project, the main source of climate data used will be a collection of model outputs, reanalysis, and observations.

For model outputs, the data used will be from ECMWF[3] and MeteoFrance[4] system 5 (SEAS5) downloaded from the Copernicus Climate Change Service (C3S), NCEP CFS version 2[5], the National MultiModel Ensemble (NMME[6]), and POAMA from the Australian Bureau of Meteorology (BOM[7]). These files will be downloaded in their native format (grib for ECMWF, netCDF for the other ones). They will be used by work packages 4 (S2S climate predictions) and 5 (Operational Climate Service: Decision Support Tool (DST) implementation).

[3] https://software.ecmwf.int/wiki/display/FCST/Implementation+of+Seasonal+Forecast+SEAS5

[4] https://www.umr-cnrm.fr/IMG/pdf/system6-technical.pdf

[5] https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/climate-forecast-system-version2-cfsv2

[6]https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/

[7] http://opendap.bom.gov.au:8080/thredds/catalogs/bmrc-poama-catalog.html

The reanalysis data will be used by WP 3 (Observational datasets), 4 and 5 and will consist of the: ERA5[8], ERA-Interim[9] data (from ECMWF), NCEP/NCAR reanalysis[10], MERRA-2[11] and the Japanese ReAnalysis data (JRA55[12]). They will be downloaded in their original grib format and processed afterwards.

The observational data will be used by work package 3. The exact datasets used will be determined further in the project, but in a first time, the data from wind "tall towers" will be used. This data is a collection of individual observations on wind farms spread over the world, coming into different formats that can be grib, NetCDF or csv.

A more detailed table about the climate data that will be used within the project is available in this live document:

This table (https://tinyurl.com/ydes27le) presents the different datasets, with links, their spatial and temporal resolutions, their time coverage, available variables, number of members, etc.

3.1.1.2 Personal data

Personal data managed in this project are only related to professional information (name, position, firm, e-mail and phone number). Besides professional information, pictures, videos and audio recordings could be gathered always with the informed consent of the people involved.

This information will be collected by work packages 2 (Definition of user needs and the role of S2S forecasts in decision-making processes), 5 (Operational Climate Service: Decision Support Tool Implementation , 6 (Positioning, exploitation and business models), 7 (Dissemination, communication and user engagement) and 8 (Ethics requirements).

Personal data management is described in deliverables *D8.1 Requirement No.1: Information sheet of users* and *D8.2 Guideline for personal data management.*

Although not directly considered Personal data, all the knowledge associated to

3.1.1.3 Other projects outcomes reports

Knowledge and experience from other projects might be the third type of data collected by the project (besides climate and personal data). The collection of this data will be done by WP 2, 3, 4, 5 and 6 and might be used to inform the work of these WPs and can be used in some deliverables. The format of this data is usually in the form of deliverables, reports or technical notes in pdf format.

[8] https://climate.copernicus.eu/era5-public-release-2010-2016

[9] https://www.ecmwf.int/en/forecasts/datasets/archive-datasets/reanalysis-datasets/era-interim

[10] https://climatedataguide.ucar.edu/climate-data/ncep-reanalysis-r2

[11] https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/

[12] http://jra.kishou.go.jp/JRA-55/index_en.html

3.1.2 Data created during the project

During the project, from the "raw/original" data mentioned in the previous section, some indicators, tailored products, summaries and forecasts outlooks will be created by the different work packages. This data will mainly consist in netCDF files following the CF[13] conventions. Work packages 2 to 7 will be involved in this process of data creation.

The interaction with energy users and other stakeholders will gather the personal data of the people with whom the project interacts for practical logistics. Besides collecting personal data, other other information obtained with this interaction with users will mainly consist in the detailed feedback received through participatory activities. The format of exchange of this data will be determined later in the project according to the collection method, and will be updated in the next versions of the Data Management Plan. Personal data together with the user feedback gathered by the DST will be stored in a secure environment at the BSC.

3.2 Data collection method

To collect the data mentioned above, different technical solutions will be used.

For the user's feedback and personal data, the information will be gathered through participatory activities held along the project (surveys, on-line forms, interviews, workshops, events and face-to-face meetings).

For the climate data, the technical infrastructure and solutions will be defined and presented in D5.4 (Architecture Design Document and Interface Design Document). Mainly, this will consist in ssh connections to the remote data servers and retrievals through ftp or the webapi from the ECMWF.

4 Documentation and Metadata

4.1 Climate data standards and metadata

Regarding climate data standards, a guideline during the project is to comply to the data standards established in the geospatial data community (like CF conventions). The data will be formatted in NetCDF following the INSPIRE Directive standards[14]. As the proposal aims at using

[13] http://cfconventions.org/

[14] https://inspire.ec.europa.eu/

data from existing platforms (Copernicus, etc.) the original data standards will be kept for these data and adapted to the needs of S2S4E for the post-processed data.

Regarding personal data, the documentation and methodology will be further defined in *WP8 Ethics requirements* and reported in the WP's specific deliverables and milestones, as well as in the next versions of the DMP.

4.2 Git branching strategy

In order to keep track of the different processes that led to the generation of the different data during the project, the source codes of the different softwares used by the DST will be kept and managed with a version control system (i.e.: git server hosted at BSC) and a well-defined branching strategy according to the Agile method development based on steps called *sprints*.

This strategy is based on five active branches along the project:

- The master branch contains the last stable sprint version.
- The daily branch contains all feature developments that are done during the current Sprint but not fully validated.
- The common branch contains all common developments shared between multiple features.
- The qualif (qualification) branch contains all features that are done and validated.
- The prod (production) branch is the main branch for production, it contains the version currently in production.

Each feature developed during the sprint is done in a specific branch named as the user story. Usually, it's a fork of the *master* branch.

When the feature is done (development, unit tests, integration tests), it is merged into *daily* branch to be validated.

When the feature is successfully validated, it is merged into the *qualif* branch.

A feature could be a fork of *qualif* branch if it needs another feature developed during the sprint.

Sometimes, a feature development requires to add or modify *common* projects or parent pom files. In this case, either the development is done in the user story branch and reported into the *common* branch, or it is directly done into the common branch.

Once this change have been pushed into the remote common branch, the developer has to "*ring the bell*" that warns all developers to merge the common branch into their own branch.

Some days before the end of the sprint, a validation is done with all features merged into *qualif* branch to test the integrity of the software. During this phase, developers should react quickly to correct any raised bugs to ensure a stable version. Every step of the processing will be tested and released only when successfully tested. Technical details of the different checks will be given in D5.4 (DST Architecture Design Document and Interface Design Document).

At the end of the sprint, the *qualif* branch is merged into the *master* branch that represents the next stable release.

The unfinished features are reported to the next sprint.

This *master* release could be merged into the *prod* branch that is deployed into production environment.

5 Ethics and Legal Compliance

The main ethical aspects that can affect research activities of the project are linked to participatory methods and are related to a) personal data protection, b) data confidentiality and c) informed consent. These issues are better detailed in deliverables D8.1 Information users sheet and D8.2 Guidelines for personal data management. For any report on results from participatory activities, the main ethical aspects will be reported in a specific section of the report indicating procedures and solutions (consent forms, anonymization, encryption of answers if necessary, etc.).

All Intellectual Property Rights (IPR) issues will be organized, managed and discussed within the Innovation Management Board (IMB) (to know more about this board, see deliverable D1.4 Composition and terms of reference of the Innovation Management Board (IMB) and External Advisory Board (EAB)).

Original licences from climate data will be preserved. Software developed during the project will have identified owners and licences. Further details will be given in the next version of the DMP when all the different components of the DST have been selected.

6 Storage and Backup

During the development phase of the project, some data samples will be downloaded individually by the different partners from the original data sources and stored in their individual data storages.

For the operational part of the DST, the climate data (files and databases) will be stored at BSC datacentre (both data and metadata) and at SMHI for the hydrological forecast. The data will be stored at BSC on a shared GPFS file system managed by the Operations department from

BSC. The storage is mounted in GPFS Native RAID ensuring reliability, availability, data protection and recovery.

7 Selection and Preservation

The results of the participatory activities will be preserved as well as the algorithms used and the software (in gitlab). Depending on the data, an expiration time will be set. These expiration dates will be defined later during the project when decisions have been made on all the datasets used by the different WPs.

8 Data Sharing

Following the EC recommendations on data sharing, the data generated within S2S4E (energy indicators, derived variables, diagnostics, etc.) will be freely (with registration) accessible through the DST (Open data policy) following the same policies as the input data from the sources Copernicus Climate Change, NMME, S2S Project and other sources.

The results of the participatory activities will be shared through the project wiki or more secure channels according to the requested confidentiality level of the information provided by the participants.

9 Responsibilities and Resources

The BSC, as project coordinator, will be responsible for the data management of the project, as well as Capgemini, The Climate Data Factory and Cicero.

Regarding personal data, each institution in the consortium is responsible for the management of the data following the requirements of the General Data Protection Regulation as explained in *D8.1 POPD - Requirement No. 1: Information sheet of users* and *D8.2. POPD - Requirement No.2: Guideline for personal data management.*

The following table shows the total efforts foreseen for the DMP and its updates along the project lifetime. It corresponds to task 5.1 Data Management Plan - definition of data protocols and formats.

Table 1: Costs assigned to data management in S2S4E

	PM	Total costs
BSC	3	13,500.00€
TCDF	1	6,600.00€
Capgemini	3	22,875.00€
TOTAL	**7**	**42,975.00 €€**