IS-ENES3 Data Management Plan

Authors: Bryan Lawrence, Michael Lautenschlager, Sophie Valcke and the contributors to the initial IS-ENES3 data inventory.

Reviewers: Francesca Guglielmo, Sylvie Joussaume

Version	*Author*	*Date*	*Comment*
0.3	BNL	26/06/19	Released for review
0.4	FG, SJ	01/07/19	Reviewer comments
1.0	BNL	02/07/19	Final Version for D1.1

Summary

The IS-ENES3 data management plan is based on standard data management principles applied in the context of an infrastructure project supporting the earth system climate modelling community and downstream users. As the project consists of many different types of datasets, an inventory of datasets has been used to develop a set of data categories, and a separate plan is in place for each of those categories. In general, all data will be persisted in Zenodo, Github/Gitlab, or by one of the partner institutions (there are several designated data centres within the partner network), and in all cases, a curation policy will be in place. Some datasets will be entered into a sustainability register, and long-term plans developed for those data within the context of project sustainability.

Actions and Milestones

1. Establish an inventory of data expected to be produced during the project (DONE),
2. Use it to establish categories of data to be managed (DONE),
3. Develop a "sub-plan" for each category, which will represent the application of the project data management principles to datasets within that category (DONE), and
4. Annually update the inventory (June 2020, 2021, 2022), and re-assess the plan.

Introduction

The third project delivering the Infrastructure for the European Network for Earth System Modelling (IS-ENES3) has three main goals:

1. Pursue the integration of the Earth's climate modelling community and prepare the sustainability of its infrastructure.
2. Foster the common development of models and tools, and the efficient use of HPC.
3. Support the exploitation of model data by the Earth system science community, the climate change impact community and the climate service community.Pursue the integration.

This document outlines the data management plan for the IS-ENES3 project itself, and some principles for the longer-term sustainability of data within the shared infrastructure. It will be revised and updated at each reporting period.

The document consists of four main sections:

1. **Background**: consisting of a short summaries of the IS-ENES3 project itself and the objectives of data management.
2. **Scope**: an analysis of how data management should apply for IS-ENES3, and
3. **The Plan** itself: outlines an approach to understanding what data is the responsibility of the project, characteristics of the data from an initial inventory of data held, or expected to be held, and how data *management* will be undertaken within the project, and
4. **Sustainability Issues**: a short statement of those issues for sustainability which have arisen from this initial data management analysis.

Background

A brief introduction to IS-ENES3

The three objectives for IS-ENES can be characterised as "community integration", "support for HPC software and usage", and "support for data exploitation". A short summary of each follows:

Objective 1: IS-ENES3 will pursue the integration of the Earth's climate system modelling community and will prepare the sustainability of its infrastructure. It will:

1. Establish stronger ties with parties either underrepresented within, or relatively new to, the ENES community, in particular through training and schools (e.g. groups in Eastern Europe and the impact modelling community) (*grow community*);
2. Build on existing governance and communication activities to prepare a sustainable infrastructure by developing a robust stakeholder consensus on downstream user requirements (*build stakeholder interaction*);
3. Strengthen existing collaborations with third parties, and developing new relationships and synergies with other infrastructures, projects, initiatives, and communities, particularly those where IS-ENES will bring wider impact (*work with 3rd parties*);
4. Develop a strategy for the infrastructure needed by the ENES community in the next decade seeking opportunities for interoperability among infrastructures (*establish future strategic goals*).

Objective 2: IS-ENES3 will foster the common development of models and tools, and the efficient use of HPC. It will:

1. Create *sustainable communities* through the coordination of network and joint research activities;
2. Develop common, *sustainable model infrastructure*, including support for relatively *new software tools* evaluated in previous phases;
3. Where possible, *share development of common model components*, with a focus on developing a *common European sea ice model* and on improving the computational performance of the *European ocean platform*;
4. Promote the use of *new metrics for evaluating model computational performance* aimed at understanding the relative efficiency of codes and how best to use available HPC;
5. Provide *community leadership* and *knowledge exchange*, aimed at responding to the existing strategy, and *improving the exploitation of HPC both nationally and internationally* (e.g. ENES use of PRACE);

6. *Explore new avenues and tools*, such as Machine Learning (ML) and Artificial Intelligence (AI).

Objective 3: IS-ENES3 will support the exploitation of model data by the Earth system science community, the climate change impact community and the climate service community. It will:

1. *Maintain and develop* the European component of the global Earth System Grid Federation with the aim of supporting the 6th phase of the Coupled Model Intercomparison Project (CMIP6);
2. Develop a *new service* to ease multi-model data analytics;
3. Put the *infrastructure and governance* of key metadata and data standards (e.g. Climate Forecast conventions for NetCDF, documentation of models and simulations) onto a *sustainable footing*;
4. Raise the standard for Earth system model evaluation by *gathering more detailed understanding of user requirements*, by promoting standards for the science provenance, and by *developing* a state-of-the-art community European model evaluation framework;
5. Invest in the *operation and development* of the climate4impact platform and underlying services to enable *customised access to data, documentation*, and *information about model evaluation* to the climate impact community, climate service businesses and consultancies.

Basic Principles of Data Management

There are three core objectives for any project data management plan:

1. To ensure that third party data that the project needs is available to the project members,
2. To ensure that data which underpins project outcomes is available for scrutiny and re-use, and
3. To identify important datasets for long-term preservation, and put in place procedures for preserving *and* curating those data.

The first of these is often neglected, yet in terms of meeting success criteria and efficiency, it can be crucial.

The second principle is often summarised by requiring project data to be "Findable, Accessible, Interoperable, and Reusable", or FAIR. While FAIR currently gets a lot of attention, without addressing the third principle, that of longer-term curation, it is often a waste of time, insofar as data can only be FAIR if it still exists, and some institution is charged with (and funded) for curating the data to deliver on at least the FAR components (i.e. not address interoperability). This is reflected in the Commission's requirements for a data management plan, that it should include information on:

- the handling of research data during *and after* the end of the project
- what data will be collected, processed and/or generated
- which methodology & standards will be applied
- whether data will be shared/made open access and
- how data will be curated & preserved (including after the end of the project).

To that end, this data management plan devotes a section on sustainability, or to be more accurate, on how IS-ENES3 will identify datasets for longer-term preservation, and develop a plan for their sustainability.

FAIR Principles

The FAIR principles were introduced in Wilkinson (2016), and are repeated verbatim here:

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

Scope

In this section we discuss the scope of data management within the IS-ENES3 project by revisiting the background and extracting a general categorisation of the project in terms of data usage and production (top down) as well as developing an inventory of expected datasets (bottom up). Together these will define the scope of data to be managed under the auspices of this plan.

General Categorisation

The IS-ENES3 infrastructure project is delivered using three vehicles:

1. joint research activities,

2. networking, and
3. service delivery.

In general, we can think about the joint research activities as having the potential to "use and produce data", while the networking activities might be about "acquiring and using data", while the services should not involve the acquisition and production of data at all - beyond the production of metrics of usage (which themselves might be datasets). In practice these simple categorisations will likely break down, but they provide context for establishing the actual data management issues.

Another viewpoint of IS-ENES3 is that the project has four axes, being infrastructure to support:

1. community building,
2. scientific data delivery (not for data management per se),
3. model development and usage (by partners and third parties), and
4. the evaluation of model simulations (by partners third parties)

With respect to data delivery, it is important to understand that the bulk of the data infrastructure is about providing services to support *discovery and distribution* of simulation data from internationally coordinated numerical experiments. The choice of which data, and for how long it should be made available, is expressly not the responsibility of the IS-ENES3 project - but the responsibility of the modelling groups that provide the data.

From this perspective we might argue that the project itself could hold *no* scientific data, insofar as the scientific data inputs and outputs belong to the users of the infrastructure. However, it will be seen, the project will create, maintain, and develop a curation plan for some datasets, but not as many as might have been anticipated a priori. Many of these will be assigned to a partner to own ongoing data management. Any datasets which need to be managed by the project (as opposed to the partners) will be added to a *sustainability register*, being a list of datasets which need addressing in the context of the sustainability work package.

The project will clearly hold data about users and usage, and it will be clearly important to establish what part of that information is *"private"* and covered by the General Data Protection Regulations (GDPR) - and hence out of scope for the data management plan - and what part could or should be *public*, and treated as a dataset.

Inventory Based Categorisation

Experience tells us that general categorisations can hide interesting special cases, and so as input to the data management plan (and, as will be seen, as part of the ongoing data management), the project has developed an inventory of datasets.

The methodology for developing that inventory is outlined in the Appendix, here we extract the categories of data identified, and describe the accompanying data management issues. In some cases we will indicate the expectation that the project will be required to *directly manage the data*, what this means in practice will be discussed in "The Plan" below (but in all such cases, datasets will be added to the sustainability register). These categories highlight the very different sorts of datasets in play, and the necessity for a range of data management approaches.

Categories:

1. **Software**: For example, ESMValTool (a community diagnostic and performance metrics tool for routine evaluation of Earth System Models, see https://www.esmvaltool.org).
   - We consider *our* software products as data entities which we need to address in the context of a data management plan. In this context *our* means software products where we are funded to manage/direct/maintain the software itself.
   - We consider the maintenance of software versions *as data*, to be a key part of ensuring provenance of any data products they might produce. If the software does not produce data products, it could be argued that older versions do not need to be curated. Such data will be managed directly.
   - We do not consider software products to which we contribute, but managed by others (including by our partners, with other funding), as *our* products (e.g. NEMO, OAISIS-MCT, and XIOS are not IS-ENES products per se).
2. **Scientific Data produced by services**: For example, ESMValTool is being deployed as a service on the ESGF infrastructure deployed by IS-ENES3 (albeit on hardware owned by partner institutes), it will produce data products which represent evaluation of climate models. In practice their production is invoked by a third party.
   - These data products include not only binary data, but images.
   - It is yet to be decided whether these products should be considered as datasets for IS-ENES3 to be managed by IS-ENES3.
   - Whether or not IS-ENES3 manages such product data, the onus is still on IS-ENES3 to ensure that they have sufficient standardised metadata for downstream data management (by whoever).
3. (Input datasets): For example, those needed by the ESMValTool services.
   - It is expected that all such data should be held in a designated data centre, and have an existent plan for long-term data management. All input datasets will be assessed, and if a copy is not in an appropriate repository, a copy will be lodged with a partner designated data centre.
4. **Schema and Standards**: The CMIP Data Request schema, and accompanying documentation and tools, is an exemplar of an information standard which will underpin data production and management.
   - IS-ENES3 has a number of similar products. We will treat these as software, but recognise their importance to data management by others.
5. **Survey Data**: We will be surveying the community about the usage of, and requirements for, software and services. Such data will include personal information, but will also include aggregated information.
   - We will treat aggregated anonymous information as datasets. We will treat any raw data which includes personal data as “inappropriate for data management”, but manage it under the GDPR.
   - We will either expect one of the partner institutions to “own” the data management, or if none are suitable, we will manage it directly.
6. **Service Statistics**: Service usage statistics as well as user support related information is collected and stored, and used for service planning and for key performance indication.
   - We will treat aggregated anonymous information as datasets. We will treat any raw data which includes personal data as “inappropriate for data management”, but manage it under the GDPR.

- We will either expect one of the partner institutions to "own" the data management, or if none are suitable, we will manage it directly.

7. **Benchmarking Data**: Some of the tools and model development will involve the usage of benchmarking to produce evaluation data (for example, the performance and outcome quality). Such data is likely to underpin both deliverables and scientific papers.
   - Where these evaluations are carried out by individuals funded by IS-ENES3, we will treat these as IS-ENES3 data products.
   - We will either expect one of the partner institutions to "own" the data management, or if none are suitable, we will manage it directly.
8. **Model Documentation**: IS-ENES3 will acquire model documentation which will not exist elsewhere in a similar form. Such documentation will include scientific information and personal contact information.
   - The form of the documentation is actually information aggregated into documents. We will treat these as if they were formal publications, ensuring that any personal information exposed has been done so with consent.
   - The documents will be managed directly.

As part of the development of the inventory, it has been made clear that the actual scientific output of models will not be covered by this data management plan, being out of scope for the project.

The Plan

IS-ENES3 Data Management Principles

Generic basic data management principles have been outlined earlier. In this section we refine these principles for application within IS-ENES3, given the project scope.

IS-ENES3 should

1. Identify which data will be produced by the project (as opposed to distributed by the project), by establishing an inventory.
2. For each dataset in the inventory, either assign the responsibility for data management for that dataset to a partner organisation, or manage it directly.
3. Where datasets will be managed by a partner, record in the inventory a short description of how the partner expects to
   i. Deliver on the FAIR principles for that data, and
   ii. Manage the data long-term.
   iii. (Where the partner is a designated data centre in its own right, e.g. DKRZ or CEDA, it will be sufficient to record the application of FAIR as being "ingestion into a designated data centre".)
4. Where datasets will be managed by the project itself, manage them according the category principles discussed below, and enter them onto a *Sustainability Register* which will be managed by the project's Scientific Officer as part of WP2 (Sustainability). For each dataset the sustainability task will be expected to assess whether or not a long-term archive can be found for the data.
5. Regularly review this plan, in particular by updating the inventory.

Category Principles

Eight categories of data have been identified by the project: Software, Scientific Data from Services, Input Data, Schema and Standards, Survey Data, Service Statistics, Benchmarking Data, and Model Documentation. Of these, only input data, survey data, and benchmarking data conform to the usual definition of "Research Data", where strict application of data management principles is required, however, for each category we outline below a summary of how IS-ENES3 will manage that category of data.

1. **Software**: IS-ENES3 software products will be managed, distributed, and preserved on GitHub. Github delivers a complete implementation of FAIR for software. Should GitHub become unavailable during the project a similar service will be identified.
2. **Scientific Data from Services**. In the event that data products are deemed to be "owned" by IS-ENES3, as opposed to by the users of the service (and this has yet to be determined), the data will be initially available via that service (Findable, and Accessible, but possibly not Interoperable and Re-usable). Data products will be kept in the service environment, and added to the sustainability register, with the most likely expectation being that one of the partner data centres will be requested to curate the data (and deliver a complete FAIR implementation). To support all aspects of FAIR, all datasets will be accompanied with suitable information to support the production of discovery and browse metadata (as defined in Lawrence et al, 2009), and
    i. any binary data will be created using the NetCDF climate forecast conventions (cfconventions.org), and
    ii. any images will be accompanied by a manifest in an appropriate format (to be established on a case by case basis) describing image provenance.
3. **Schema and Standards** will be treated like software, by using GitHub for management and distribution. Important versions will be additionally published via Zenodo to provide an additional layer of accessibility and preservation.
4. **Input Data** is necessary for some of the services. Where such input data is "research data", if the data is not already held by a suitable scientific data repository, a copy will be deposited with a partner data centre, who will manage any issues around data access and licensing.
5. **Survey Data and Benchmark data** are expected to be relatively low in volume, and with no existent metadata standards will either be managed by a partner institution or deposited with Zenodo, using DOIs for identifiers, and text for accompanying metadata.
6. **Service statistics** will be published on an IS-ENES3 website, with accompanying textural metadata, and maintained for the duration of the project.
    - As these data are not "research" data per se, there is no clear scientific reason for preserving the data long-term, so that apart from mothballing the appropriate website pages, no long-term *curation* for these data will be established. However, such data will likely be persisted as part of service management.
7. **Model Documentation** (from the ES-DOC activity within IS-ENES3) will be published on websites, and stored in GitHub for long-term preservation. All documents will include a universal unique identifier (UUID), and both a schema identifier and a version identifier which conform to the ES-DOC meta-model (itself, to be published in the academic literature).

This data management plan takes cognisance of the fact that a number of the partner organisations are designated data centres, that is, they conform to accepted standards of data management practice, and have comprehensive ingestion procedures to ensure FAIR data.

- Where datasets cannot be "handed" over to a suitable partner organisation, the project will take responsibility, both by carrying out active data management during the project, and by registering such datasets on the "sustainability register".

Sustainability Implications

All datasets will fall into one of the following categories:

1. Held by a partner, who commits to sustaining the data in an appropriate manner, or
2. Held by a partner designated data centre, who by taking the data, will have entered it into a system which has appropriate sustainability procedures, or
3. Managed by the project itself, and entered onto the Sustainability Register.

Registration of a dataset onto the sustainability register indicates that the project believes the data should be long-term archived, but the mechanism/location/funding for such archival needs to be handled as an "infrastructure sustainability" issue.

It is expected that all datasets on the sustainability register can be further categorised as:

1. Already in a suitable persistence format with enough metadata for "hands off curation" (e.g. text documents deposited with Zenodo),
2. In a format which is suitable for long-term persistence, with suitable metadata, but likely to have "long-term curation costs" (e.g. they are high volume), for which longer-term funding is necessary.
3. Are not in a suitable state for curation, with the cost of adapting the data being unsupportable by the project itself.

For all these categories, the project will take a decision as to how to proceed as part of the sustainability plans being developed in WP2/NA1. These decisions will effectively form a first "re-appraise and consider disposal" step before long-term curation, and cannot be made until the bulk of the sustainability plan is in place.

Bibliography

Lawrence, B. N., Lowry, R., Miller, P., Snaith, H., & Woolf, A. (2009). Information in environmental data grids. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1890), 1003–1014. https://doi.org/10.1098/rsta.2008.0237

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Appendix - Establishing the Inventory

Inventory Questions:

Each task leader was asked to fill out an entry for any *"significant" "datasets"* created, or held by, the project. Task leaders will be regularly requested to review whether or not new datasets have been identified.

The following guidance was provided to define "significant" and "datasets":

- Significant - Would someone want or need to see this in order to substantiate project deliverables or publications? Would the people who provided it to us expect IS-ENES3 to be making it available? If in doubt, it's significant!"
- Dataset - Collection of data from one instrument, survey, or sub-project (including software).
  - (E.g. all of the ESGF model data will count as one dataset, but ESGF usage would count as another)
  - (In the case of software, consider only software which we develop and disseminate ourselves, i.e. you don't need to include contributions to third party software packages, including host institution models)."
  - We should consider "the set of data needed to underpin a deliverable" as a dataset if it the deliverable is based in any way on data per se.

The questions:

1. Dataset name
2. Contact person for this survey response
3. Description (up to 100 words)
4. Work package, task number, deliverable and/or milestone numbers which are relevant to this data
5. Is it personal data, covered by the GDPR?
6. Is it data owned by a third party, and we only have a copy?
7. Does this data have a license, or statement/assertion of IPR (e.g. copyright), if so, what is it?
8. Would anyone reasonable expect *us* to be making this data publicly available?
9. Would anyone reasonably expect to be able to request this data (from *us*), even if it is not public?
10. Is this data owned by us, but stored and distributed elsewhere (e.g. github)?
11. If *we* might need to distribute it, how might we do that?
12. What kind of data is this (numeric, text, imagery, software, survey data, model output …)?
13. Does it consist of mostly one sort of format, if so which (if a few dominant ones, please list the major ones)
14. (If we created the data) Are there any metadata or other conventions we should apply to this data?
15. (if third party) What metadata conventions does this conform to? (If any)
16. Do we need to keep this data after the project?
17. If so, will one of the partners do that as part of their normal business, or is it a sustainability issue?