General requirements for IS-ENES2

On request we are happy to support and provide advice on basis of our running environment (ESGF).

Context of general requirements in IS-ENES2

Completed questionnaire can be found at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/62/attachments/377/download?force=1&disposition=attachment

IS-ENES runs a distributed, federated data infrastructure based on few (3-4) main data centres and various associated smaller ones.

The requirements information provided to ENVRI+ refers to the climate modeling community, to two data dissemination systems (ESGF for project run time; LTA as long term archiving), to CMIP5 as climate modelling data project 2010-2015 and CMIP6 2016-2021.

Summary of IS-ENES2 general requirements

Use case 1: Data producers submit model data results to IS-ENES data nodes. Data is quality-checked and published in IS-ENES/ESGF data infrastructure. All data items are uniquely identified. Data is long term archived. Data aggregates (experiment level) are assigned DOIs. DOIs are used by end users in scientific publications. DOI-assigned data aggregates are published in various Metadata Catalogues e.g. in world data centers for climate.

Use case 2: End user of IS-ENES data infrastructure encounters problems (technical or scientific). End user contacts IS-ENES/ESGF user support (organized in first/second level support, second level support internationally distributed). General problems are documented in FAQ site.

Use case 3: End user wants to process large amounts of data. Three possibilities to do this:

> A) Download and process at home institute. This is supported via a bulk data download and synchronization tool for IS-ENES/ESGF sites.
>
> B) Contact a large IS-ENES/ESGF site who already has the required data available (replicated from other sites) and process there (personal interaction necessary to get account and permission at the site). This is supported by the user support service.
>
> C) Contact a web processing service or a portal providing (parts of) the requested analysis functionalities. This supported by the IS-ENES climate4impact portal (https://climate4impact.eu/) as well as by IS-ENES/ESGF web processing services (not yet fully in production)

Some more detailed IS-ENES use cases were submitted to the RDA (Research Data Alliance) Data Fabric interest group as well as Data Repository interest group and are available at:

https://rd-alliance.org/enes-data-federation-use-case.html and https://rd-alliance.org/climate-data-analytics-use-case.html .

Detailed requirements

Data is generated by climate modeling groups (as well as by some climate observational studies, relevant for climate model intercomparison projects). Data is post-processed according to the standards and agreements of the intercomparison project (e.g. CMIP, CORDEX). Data is ingested at IS-ENES/ESGF data nodes and quality-controlled (check intercomparison project conventions and standards). As a next step, data is published to the IS-ENES/ESGF data infrastructure. Publication makes metadata available and searchable and data accessible via IS-ENES portals (as well as via APIs). Important data products are replicated to dedicated long-term archival centers. There, additional quality checks are run as a pre-requisite for DOI assignment and availability for DOI based data citation.

The post-processing of the data according to the standards and conventions of intercomparison projects is supported by community tool (CMOR). The infrastructure is based on a large international open source community (The Earth System Grid Federation, ESGF), developing the individual components (security, catalogues, data access services, portal parts etc.). The computational environments are more heterogeneous and organized locally at sites according to site-specific constraints. Some computational facilities are integrated as part of the ESGF nodes and portals (simple sub-setting and visualization) or IS-ENES portals interfacing with the IS-ENES data infrastructure (e.g. the climate4impact portal). Larger computational services are exposed via Web Processing Services –this part is not yet in production and needs technical developments as well as future organizational/policy agreements.

Responsibilities of users who involved in this use cases are as follows:

Data producers:

> Deliver data (and metadata) according to the rules and regulations of the corresponding Model Intercomparison Project (defining a kind of data management plan).
>
> Inform data publishers about new versions and versioning related information.

Data publishers:

> Publish data according to defined “best practices” agreed upon in the data federation.
>
> Provide contact information in case of operational problems at the site.
>
> Inform federation about operational issues (down times etc.).

Data users:

> Provide citation information in published work based on the data.

Data from IS-ENES is replicated to EUDAT for data curation purposes (long term archival). IS-ENES data is harvested by EUDAT metadata catalogue (B2Find). Integration of other EUDAT services is foreseen (B2Drop, ..) to support cross-community data usage.

Mostly model data generated to enable Model Intercomparison Projects: e.g. CMIP5, CORDEX. Also some observational data used for intercomparison analysis activities: e.g. obs4Mips. The diversity will grow during the next phase of intercomparison projects currently starting (CMIP6).

All the components of the IS-ENES/ESGF data infrastructure are based on an international open source effort, called Earth System Grid Federation (ESGF). All the software is open source (https://github.com/esgf). Also the activities to provide future data near processing functionalities are organized in open source projects (see e.g. https://github.com/bird-house with documentation on http://birdhouse.readthedocs.org/en/latest/ as well as the climate4impact WPS activities).

No computing resources can bring to ENVRI+, except for testing&prototyping.

No (as n/a) for access to IS-ENES instrumentation/detectors or lab equipments.

On request IS-ENES is happy to support and provide advice on basis of our running environment (ESGF).

No access to related scholarly publications.

IS-ENES supports a website with information on our RI: https://is.enes.org

Through participation to ENVRIplus, IS-ENES aims better understanding of interdisciplinary use cases and end user requirements. A look on practices beyond the horizon of IS-ENES community. Example:. sharing of data management best practices.

IS-ENES expects ENVRI+ to provide Service and Data catalogues for comparison of IS-ENES model data to other data (e.g., observations).

Community specific standards for data formatting and access are used including, netcdf-CF (climate and forecast conventions), OpenDAP data access protocol, Thredds. Metadata is also (partially) exposed as ISO 19139 conforming documents.

Specific software are used by IS-ENES including, Federated Solr/Lucene indices to provide consistent data search across IS-ENES portals. Thredds servers for data access (developed by unidata). Globus GridFTP for large data transfers.

No plan for changing the standards in use.

The software is in continuous evolution especially because of security incidents in the past. Work in progress concerns among others a better automatic installation. The software is composed of a galaxy of components. Depending on requirements (scientific or operational) we have opportunities to make evolution on some components.

Because of the problems to provide stable operational procedures across an internationally distributed data federation (supported via different (local) funding streams) an operations team was formed to support CMIP6 data management in the data federation. This team will define best practices and supervise the operational data management activities at the sites. Please refer to Operations Team terms of reference document: https://docs.google.com/document/d/1oRWqxtWWEfsucTVhk0G3bMqHC0BL4dJwADrOG8Ukj-g/edit

Areas need to be improved in IS-ENES including,

- Be able to share best practices as fast as new nodes integrate the RI federation.
- Data near processing functionality has to be provided to A) reduce the download volumes from sites and B) support end users a means to be able to work with a worldwide-distributed climate data archive in the Petabyte range.

Data management plan of IS-ENES can be refer to e.g. CORDEX data management plan, CMIP6 data management preparation documents. WDCC documents.

For security and access, the IS-ENES data infrastructure supports single sign on across multiple portals as well as authorization based on membership to various “projects”.

CORDEX data are in general available for both commercial and research purposes. Some modelling centres decided to restrict the use of their data to “non-commercial research and educational purposes.” https://github.com/IS-ENES-Data/cordex/blob/9fa582a72c38ad13738885c1aeadc764bc3700fa/CORDEX_register.xlsx

The access to CMIP5 data is unrestricted except for the data from Japanese modeling centres, which are subject to similar restrictions as above: http://cmip-pcmdi.llnl.gov/cmip5/availability.html

Handling Volume and distribution of data (Multi-Petabyte range): Replication, Versioning. Providing related information for data products (provenance, user comments, usage, detailed scientific descriptions needed for usage).

Topic 4 (processing) is particular in interests.

Formalities (who & when)

Go-between	Yin Chen
RI representative	Sylvie Joussaume <sylvie.joussaume@lsce.ipsl.fr> Francesca Guglielmo <francesca.guglielmo@lsce.ipsl.fr>
Period of requirements collection	Oct -Nov 2015
Status	Completed


questions_IS-ENES_04_11.doc