EUROPEAN COMMISSION
Community research

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK
FOR EARTH SYSTEM MODELLING

IS-ENES2 DELIVERABLE (D -N°: 5.1)

ENES Data Service Infrastructure requirements and recommendations

File name: {IS-ENES2_D5_1.pdf}

Reporting period: e.g. *01/04/2013 – 30/09/2014*

Author(s): Martin Juckes, Michael Lautenshlager, Sébastien Denvil

Reviewer(s): Wim Som de Cerff,
Sandro Fiore

Release date for review: ***19/03/2015***

Final date of issue: ***18/01/2016***

Revision table			
Version	**Date**	**Name**	**Comments**
0.5	1/10/2014	Draft	For consultation between authors
0.6	24/10/2014	2nd draft	For consultation with WGCM Infrastructure Panel
0.7	16/03/2015	3rd draft	Including metadata requirements
1.0draft	19/03/2015	Final draft	Submitted for final review
1.1	18/01/2016	Revised	Revised with responses to reviews.

Abstract

The data service infrastructure links together existing resources within and beyond Europe. This document reviews institutional, procedural and technical requirements.

Project co-funded by the European Commission's Seventh Framework Programme (FP7; 2007-2013) under the grant agreement n°312979		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants including the Commission Services	
RE	Restricted to a group specified by the partners of the **IS-ENES2** project	
CO	Confidential, only for partners of the **IS-ENES2** project	

CAPACITIES

Table of contents

Executive Summary

The emergence since 2011 of a distributed archive services creates a new set of requirements for system operators. A set of high level requirements has been identified. Identification of these requirements will help to organise discussion on detailed technical requirements. The requirements are organised into sections on administrative requirements, which relate to institutional commitments to appropriate levels of support, operational requirements and technical requirements. These three categories reflect the fact that we are interested both in the operation of the system and the specification of the system. The effective operation of the federated system depends on the cooperation between institutions. Poor performance at one institution can dominate the user experience and render the entire federation dysfunctional. The administrative requirements relate to the procedures in place to deal with possible poor performance at an institution.

42 high-level requirements and recommendations have been identified. A review of the status-quo finds that the current distributed archive is lacking in many areas, despite the significant progress that has been made in recent years. At the institutional level there is an agreement among the IS-ENES2 project partners, but no progress to establishing a global framework which can go beyond the "best-effort" approach used for the archives of the Coupled Model Inter-comparison Project, Phase 5 (CMIP5 – Taylor et al., 2012)[1]. Discussion of system requirements takes place in a broad range of meetings and working groups, so it is hard to gain an overview of which priorities are being tackled[2].

[1] The WGCM Infrastructure Panel established in 2014 has some remit in this area, and took a further step forward in March 2015 with a decision to set up an operations management team with representatives from all institutions expecting to host CMIP6 data.

[2] Some progress on developing and ESGF Governance model has been made during preparation of this document.

1. Preambule

This is a review of ENES Data Service Infrastructure (ENES-DSI) requirements from a stakeholder perspective. It will cover 3 key stakeholder groups: the archive centres, the data providers and the associated service providers. It is clear and widely accepted that the ENES-DSI needs to join up services between federated archives within and beyond the European Research Area. The approach adopted uses community standards and software packages to create a set of common services which allow transparent access to distributed data resources. This is a new mode of operation which has already delivered considerable benefits to user communities but also raised new problems for all stakeholders.

The document has been prepared through discussion among the IS-ENES2 project partners. It should inform development priorities and deployment procedures.

This document expresses infrastructure requirements from a range of user perspectives, including the infrastructure operators: the ENES-DSI is considered as a component of a wider infrastructures supporting the scientific community, and the operators of the wider infrastructure are key users of the ENES-DSI. Earth System Grid Federation (ESGF) (Williams et al., 2015) is a core component of the ENES-DSI, and many, but not all, of the requirements below will translate into ESGF requirements.

The requirements are expressed in the RFC 2119 terminology[3]: "MUST" indicates an absolute requirements, while "SHOULD" indicates a recommendation which may be waived after careful consideration of the options. Following the RFC guidelines, "MUST" is used sparingly. RFC2119 leaves the mechanism for waiver of a recommendation open: this ambiguity is resolved, in the context of this document, by requirement R1.02 below.

[3] https://www.ietf.org/rfc/rfc2119.txt

2. Requirements

2.1 The federation of participating institutions

ESGF is, as the name suggests, a federation of institutions. The distributed archive run by the federation creates a unified archive which can be greater than the sum of the parts, but which can also be degraded by weak links. Maintenance of effective services at any site depends on an appropriate degree of responsiveness in other sites to resolve issues.

R1.01: Institutional contacts

All participating institutions MUST provide contact details for a technical manager of the system, a contact for scientific enquiries about the data, and of a project or program manager for administrative coordination across data nodes. These contact details should be accessible to all participating institutions.

R1.02 Reporting non-compliance

Failure to comply with any requirements listed here MUST be reported to federation partners.

R1.03 Institutional responsibilities

Participating institutions SHOULD commit to abiding by the protocols and data publication standards of any activities or projects for which they are publishing data.

R1.04 Search catalogue terms of use

The ESGF Application Programming Interface (API) makes it possible for 3rd party portals and software to browse the catalogue. Terms of use SHOULD be established which oblige 3rd parties providing discovery services to users to include appropriate information on provenance and support.

R1.05 Dialogue with system operators

There SHOULD be a mechanism for groups and individuals operating the system to provide feedback – distinct from system development forums.

R1.06 Emergency response

The SHOULD be a mechanism for dealing with major service interruptions arising from security breaches.

2.2 The operational archive

ESGF is also an operational archive, providing continuous access to data from multiple institutions. The operational archive is what the users see. This section includes requirements which go beyond technical software issues, though in many cases their resolution is dependent on development of suitable tools as well as on institutional agreements.

R2.01: User support

Users SHOULD have ready access to help, including a help desk with a query management system.

R2.02: Register of known problems with the data

It SHOULD be possible for users to easily find documentation of known problems and to easily know whether a given dataset belong to the latest version or not.

R2.03: Ability to comment on data collections

Users SHOULD be able to comment on data in a way which is accessible to other users.

R2.04: Links to model and experiment documentation

It SHOULD be easy to navigate to model and experiment documentation associated with all datasets.

R2.05: Quality Control

Quality control requirements may vary between different activities, but users will expect a reasonable level of consistency in the data. Quality control protocols SHOULD be available to users. Appendix 1 provides a draft list of format compliance concepts: within each concept there may be project dependent specifications, but having a set of organising concepts should help to clarify what users can expect and how specifications are established and verified.

R2.06: Security against friendly fire

When a catalogue entry is clearly erroneous (e.g. multiple references to the same file, files with overlapping time periods, files with names which are inconsistent with the dataset identifier, files of zero size, files which are consistently not accessible) it SHOULD be possible to filter it from search results (i.e. it should be possible for the operators of an ESGF Index Node index node to filter out datasets which they have identified as being clearly erroneous). Better understanding of institutional responsibilities may reduce the need for this, but the need is unlikely to go away completely. This needs both software and an agreed protocol.

R2.07 Version control

Data SHOULD be published with version control to enable tracking of changes.

R2.08 Clear data format requirements

Data format requirements SHOULD be clear and testable.

R2.09 Citable data: long term archive

Data which is placed in a long term archive SHOULD be assigned a DataCite DOI with suitable guarantees and documentation (metadata).

R2.10 Citable data: from publication

It SHOULD be possible to cite data from publication, even though not all data will end up in the long term archive.

R2.11 Segregation of archive from data share activities

As ESGF provides a flexible software platform allowing institutions to federate anything from data sharing services to full archive services including appropriate documentation and user support. There is a need to segregate these different service configurations to avoid confusion between the differently managed data collections. It SHOULD be possible to isolate services which are providing a true archive service from more limited data share activities: preventing the intrusion of data collections without linked documentation into archive index nodes.

R2.12 Failover services

Failover services SHOULD be in place to ensure that the operation of the archive as a whole can continue when any element is removed.

2.3 The ESGF software

This section deals with the requirements directly relevant to the ESGF software package. Fulfilment of these requirements is thus the responsibility of the software developers.

R3.01 Robust publication process

Publication processes MUST be able to support publication of large volumes of data in a timely manner and follow publication constraints (e.g. version control).

R3.02 Access control

Access control MUST support both browser and scripted access.

R3.03 OPeNDAP support

The OPeNDAP data transport architecture and protocol MUST be supported.

R3.04 Bulk download of browser-generated data requests

Users SHOULD be able to select data and download in bulk.

R3.05 Server-side processing

User SHOULD be able to perform federation based (server side) data processing for data reduction and transformation.

R3.06 Synchronisation of local data collections

Users SHOULD be able to keep a local collection of data synchronised with the archive contents.

R3.07 Efficient Search

Use of a large and complex archive relies on effective support for data discovery. Search queried SHOULD return results in under one second.

R3.08 Service status

It SHOULD be possible for tools (such as download scripts) to determine the status of key service elements in order to give the user intelligible error codes.

R3.09 System monitoring

Reports of system usage, giving details of activity at individual centres and across the archive are required. This MUST be automatic and robust.

R3.10 Robust installation and upgrade process

Installation and upgrade SHOULD be able to run smoothly and reliably.

R3.11APIs for access by client tools

The system SHOULD have standardised Application Programming Interfaces to support 3rd party software tools.

R3.12 Processing applied to browser-generated data requests

Users SHOULD be able to perform federation based (server side) data processing for data reduction and transformation.

R3.13 Verifiable data publication version management

The conformance of a data publisher to the version management policy SHOULD be testable.

R3.14 Single view

It SHOULD be possible to search across all appropriate catalogue (e.g. not having to know which project to search in first).

2.4 The ES-DOC software ecosystem

These specifications relate to the Earth System Documentation (ES-DOC) service (Lawrence et al., 2012), which provides a central repository for technical documentation.

R4.01 Service status

It MUST be possible for tools (such as web user interface) to determine the status of key service elements in order to give the user intelligible error codes.

R4.02 System monitoring

Reports of system usage, giving details of activity at individual centres and across the archive are required. This MUST be automatic and robust.

R4.03 Robust publication process (including creation)

Common Information Model (CIM) documents publication processes MUST be able to support the publication of all type of CIM documents in a programmatic manner (automatically) and through an intuitive web user interface.

R4.04 Access control

Access control when required (creating CIM documents) MUST support both browser and scripted access.

R4.05 To view CIM documents

Users SHOULD be able to access every CIM document type and to visualize it.

R4.06 To compare CIM documents

Users SHOULD be able to compare CIM documents having the same type.

R4.07 Efficient Search

Use of a complex metadata repository relies on effective support for CIM documents discovery. A flexible search service SHOULD be provided.

R4.08 APIs for access by client tools

The system SHOULD have standardised Application Programming Interfaces to support 3rd party software tools.

R4.09 Forcing description

Forcing datasets used by models are often poorly described. ES-DOC SHOULD provide a standard way of describing such datasets.

R4.10 Link to model documentation articles

It SHOULD be made easy to populate the CIM at the same time that a model documentation article is being written.

3. State of system against each requirement

Requirement	**Status (March 2015)**
R1.01: Institutional contacts	Generally only have a technical contact.
R1.02: Reporting non-compliance	New
R1.03 Institutional responsibilities	Adherence to CMIP5 standards and procedures was weak – partly due to lack of appropriate tools.
R1.04 Search catalogue terms of use	New
R1.05 Dialogue with operators	Mainly through developers.
R1.06 Emergency Response	New
R2.01: User support	In place, but not clear if system will scale.
R2.02: Register of known problems with the data	A skeleton service was in place for CMIP5, but far from complete and not well integrated with ESGF services.
R2.03: Ability to comment on data collections	New
R2.04: Links to model and experiment documentation	The system for linking to documentation is in place, but work-flow for generation and maintenance of documentation is unclear. Some rationalisation in progress through ES-DOC.
R2.05: Quality Control	Not cleanly implemented for CMIP5
R2.06: Security against friendly fire	New
R2.07 Version control	Basic system in place, but not uniformly implemented. No systematic recording of purpose of new versions or means of differencing versions.
R2.08 Clear data format requirements	Requirements for CMIP5 expressed in multiple documents. Some rationalisation in progress through WIP.
R2.09 Citable data: long term	In place. Mapping of ESGF vocabularies to DataCite

archive	DOI vocabularies could be revisited including metadata elements in data formats.
R2.10 Citable data: from publication	Not currently available.
R2.11 Segregation of archive	New
R2.12 Failover services	New
R3.01 Robust publication process	To do.
R3.02 Access control	In place, but difficult to use for some users.
R3.03 OPeNDAP	In place for disk archives. Not supported for offline data.
R3.04 Bulk download and processing of browser-generated data requests	In place.
R3.05 Processing of browser-generated data requests	In place. The authorisation process causes some users problems.
R3.06 Synchronisation of local collection	Available in the form of synda (formerly synchro-data)[4]. Need to ensure that this capability is maintained and supported.
R3.07 Efficient Search	In place. But large and complex queries (such as those needed to grant R3.02) can cause some issues in term of performance.
R3.08 Service status	New
R3.09 System monitoring	More details in IS-ENES2 Milestone 111 (Monitoring system and dashboard design): in progress.
R3.10 Robust installation and upgrade process	Progress being made.
R3.11 APIs for access by client tools	Good APIs for data discovery in place.
R3.12 Processing	Server side data processing currently only available in a

[4] http://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synda

	limited form. Resource management is an issue.
R3.13 Verifiable data publication version management	New
R3.14 Single view	In place.
R4.01 ES-DOC Service Status	Partly in place. Must be enhance to have a broader coverage.
R4.02 ES-DOC Service monitoring	To do.
R4.03 Robust creation and publication processes	Scriptable creation tools and web UI exists. They need to be enhanced and adapted to streamline and ease the creation/publication process.
R4.04 Access control	To do.
R4.05 View CIM documents	In place.
R4.06 Compare CIM documents	In place for most of the document types. Comparing and contrasting simulations is a high priority.
R4.07 Search over CIM documents	Good APIs for CIM documents discovery in place.
R4.08 ES-DOC APIs	Good APIs are available to view, create, update and compare CIM documents
R4.09 Forcing dataset descriptions	Partly in place based on previous project like CMIP5. This need to be rework in close collaboration with data producer.
R4.10 Link to model documentation articles	To do.

4. Conclusion

The creation of a federated system relies on provision of uniform services with high availability at each participating institution. Some of the requirements were clear in advance, others have become clear as the system has evolved. This list is intended to serve as a reference point for discussions. Many of these requirements will need more detailed analysis to be carried out by specialist teams.

5. Appendix 1: Quality control: compliance test draft glossary

This appendix lists a series of compliance checks, with identifiers, titles (in italics) and a short description. This list will need to be extended for CMIP6. In some cases the formulation is shaped by ad hoc requirements specified for CORDEX which should be managed in a more generic way through well-structured MIP tables for CMIP6.

variable_ncattribute_mipvalues
Consistency of variable NetCDF attributes with MIP tables.
Test of the consistency of the NetCDF attributes of a variable with those specified in a MIP tables entry. The attributes tested will generally include the standard_name, long_name, units, cell_methods, _FillValue and missingvalue. The last 3 have special treatment. The "cell_methods" attribute has a compound structure. The test should parse the structure to evaluate whether the information specified in the MIP tables is given correctly. Inclusion of additional information in the "cell_methods" attribute should be allowed. The two attributes "_FillValue" and "missingvalue" provide the same information. "_FillValue" is given special treatment by the NetCDF libraries, "missingvalue" is included for compatibility with earlier standards. They should be both present or both omitted (for some MIPs it may be specified that they should always be present) and if present, they should have the same value (generally specified in the MIP requirements).

global_ncattribute_present
Presence of required NetCDF global attributes
Test that required global NetCDF attributes are present in the data file.

variable_ncattribute_present
Presence of required NetCDF variable attributes
Test that required NetCDF variable attributes are present in the data file.

variable_in_group
The data variable is a valid member of the indicated collection of variables
Test that the data variable is in a list of variables associated with a specified collection: in the context of MIPs this means checking that the variable is present in the relevant MIP table.

variable_type

Data type of variable

Verify that the data variable is of the appropriate data type (e.g. single precision or double precision).

global_ncattribute_cv

Global NetCDF attribute controlled vocabularies

Verify that NetCDF global attribute values are consistent with controlled vocabulary constraints.

filename_filemetadata_consistency

Consistency of NetCDF global attributes and variable name with file name.

Verify the consistency between the file name and the internal metadata of the file.

exception

Compliance checker failed to run as designed

An exception error code indicates that the software crashed while running checks. Details should be in the processing log files.

parse_filename

Parse the file name into component elements

File names will generally consist of a sequence of elements separated by a special character. This test checks that the correct number of elements are present (or that the number of elements is in the correct range).

parse_filename_timerange

Parse the time range specified in the file name

If the file name contains a time range, this test will check that the given element has the correct syntax (usually "start-end", where "start" and "end" are strings such as "19900101" or "199001").

filename_timerange_length

Verify the number of characters used to specify the time range

Verify that the number of characters used to specify the time range fits the requirements. This will be 6 if the time range is specified to the nearest month, 8 if it is specified to the nearest day.

time_attributes

Verify that the time variable is present (if needed) with the appropriate attributes

Check that a "time" variable is present and has appropriate attributes, including units and, if required, bounds. The bounds attribute is required if the "cell_methods" attribute on the data variable specifies that the data is not instantaneous.

pressure_levels

Check attributes, bounds and values of pressure levels

Check attributes, bounds and values of pressure levels. Where data is interpolated to pressure levels, the MIP data request generally defines the levels required.

height_levels

Check attributes, bounds and values of height levels

Check properties of the height vertical coordinate

grid_mapping

Check the grid_mapping attributes

Check the attributes specifying the grid in the grid_mapping variable. The usage of the "grid_mapping" variable is defined in the etCDF CF Convention.

rotated_latlon_attributes

Check the attributes of the rotated latitude and longitude coordinate variables

Check the attributes of the rotated latitude and longitude coordinate variables. This test checks variable attributes (e.g. long_name, standard_name, units, and axis attributes) and the type of the variable.

rotated_latlon_domain

Check the domain specified by rotated latitude and longitude coordinate variables

Check the domain specified by rotated latitude and longitude coordinate variables. There may be some tolerance specified, rather than requiring an exact match.

regular_grid_attributes

Check the attributes of the latitude and longitude coordinate variables

Check the attributes of the latitude and longitude coordinate variables.

regular_grid_domain

Check the domain specified by latitude and longitude coordinate variables

Check the domain specified by latitude and longitude coordinate variables. This may also include a check on the grid spacing.

filename_timerange_value

Check the time range specified in the file name

Check that the time range specified in the file name is consistent with the data request (e.g. some variables should be in blocks of 10 years, starting on January 1st in the 1st year of a decade).

6. Glossary

API: Application Programming Interface;
CIM: Common Information Model (data model for ES-DOC content);
DOI: Digital Object Identifier;
CMIP5: Coupled Model Intercomparison Project, Phase 5;
CORDEX: Coordinated Regional Downscaling Experiment;
ENES: European Network for Earth System Modelling
ES-DOC: Earth System Documentation (es-doc.org);
ESGF: Earth System Grid Federation;
MIP: Model Inter-comparison Project;
OPeNDAP: Open-source Project for a Network Data Access Protocol, http://www.opendap.org/
RFC: A Request for Comments (RFC) is a type of publication from the Internet Engineering Task Force (IETF) and the Internet Society, the principal technical development and standards-setting bodies for the Internet;
Synda: Data synchronisation tool for ESGF archives: http://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synda;
WCRP: World Climate Research Programme;
WGCM: Working Group on Coupled Models (a WCRP working group).

7. References

Lawrence, B.N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R.W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M.-P., Murphy, S., Pascoe, C., Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A. and Valcke, S. (2012) *Describing Earth system simulations with the Metafor CIM.* Geoscientific Model Development, 5 (6). pp. 1493-1500. ISSN 1991-9603 doi: 10.5194/gmd-5-1493-2012;

Taylor, K.E., R.J. Stouffer, G.A. Meehl: An Overview of CMIP5 and the experiment design." Bull. Amer. Meteor. Soc., **93**, 485-498, doi:10.1175/BAMS-D-11-00094.1, 2012M

Williams, Dean N.; Lautenschlager, Michael; Balaji, Venkatramani; Cinquini, Luca; DeLuca, Cecilia; Denvil, Sebastien; Duffy, Daniel; Evans, Ben; Ferraro, Robert; Juckes, Martin; Trenham, Claire; "Strategic Roadmap for the earth system grid federation", Big Data, 2015 IEEE International Conference Proceedings: pages 2182-2190, Santa Clara, CA, USA, October 29 2015 – November 1 2015, DOI: 10.1109/BigData.2015.7364005.