RDA plenary Meeting 2020
BoF - Data Movement:
What Infrastructure Fabrics are Required?

ENES climate data infrastructure:

Data Movement

Stephan Kindermann
DKRZ

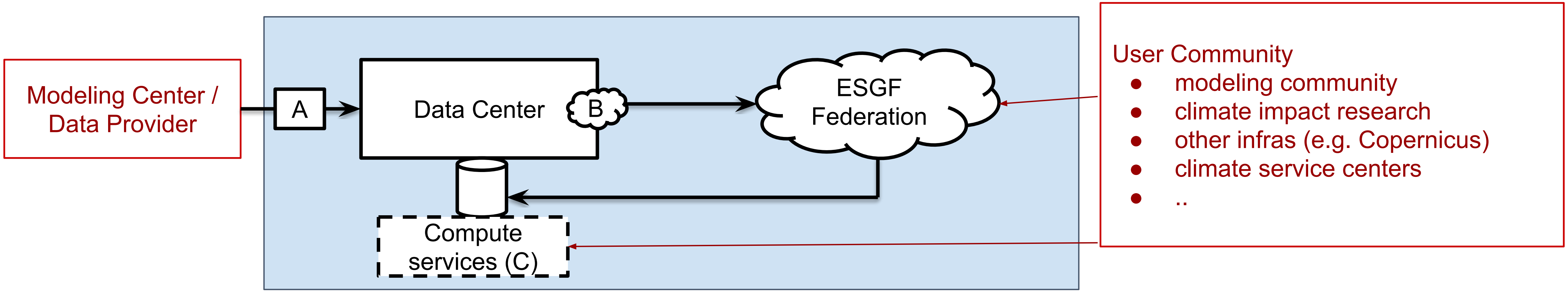
European Network for Earth System Modelling (ENES)

The IS-ENES3 project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824084

The ENES Climate Data Infrastructure (CDI)

The ENES CDI:

- **A: Data centers integrating services in a federation (ESGF based)**
  - > 30 data node installations worldwide (~ 50% in Europe), > 25 PByte of data
  - tier 1 data nodes host large replica data pools
    - optimized data replication routes among tier 1 sites (globus)



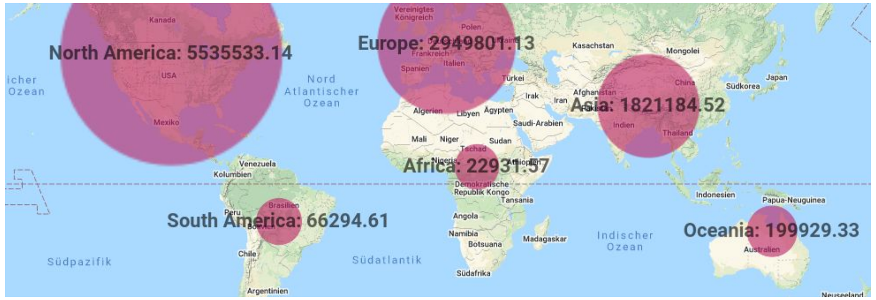
User Community
- modeling community
- climate impact research
- other infras (e.g. Copernicus)
- climate service centers
- ..

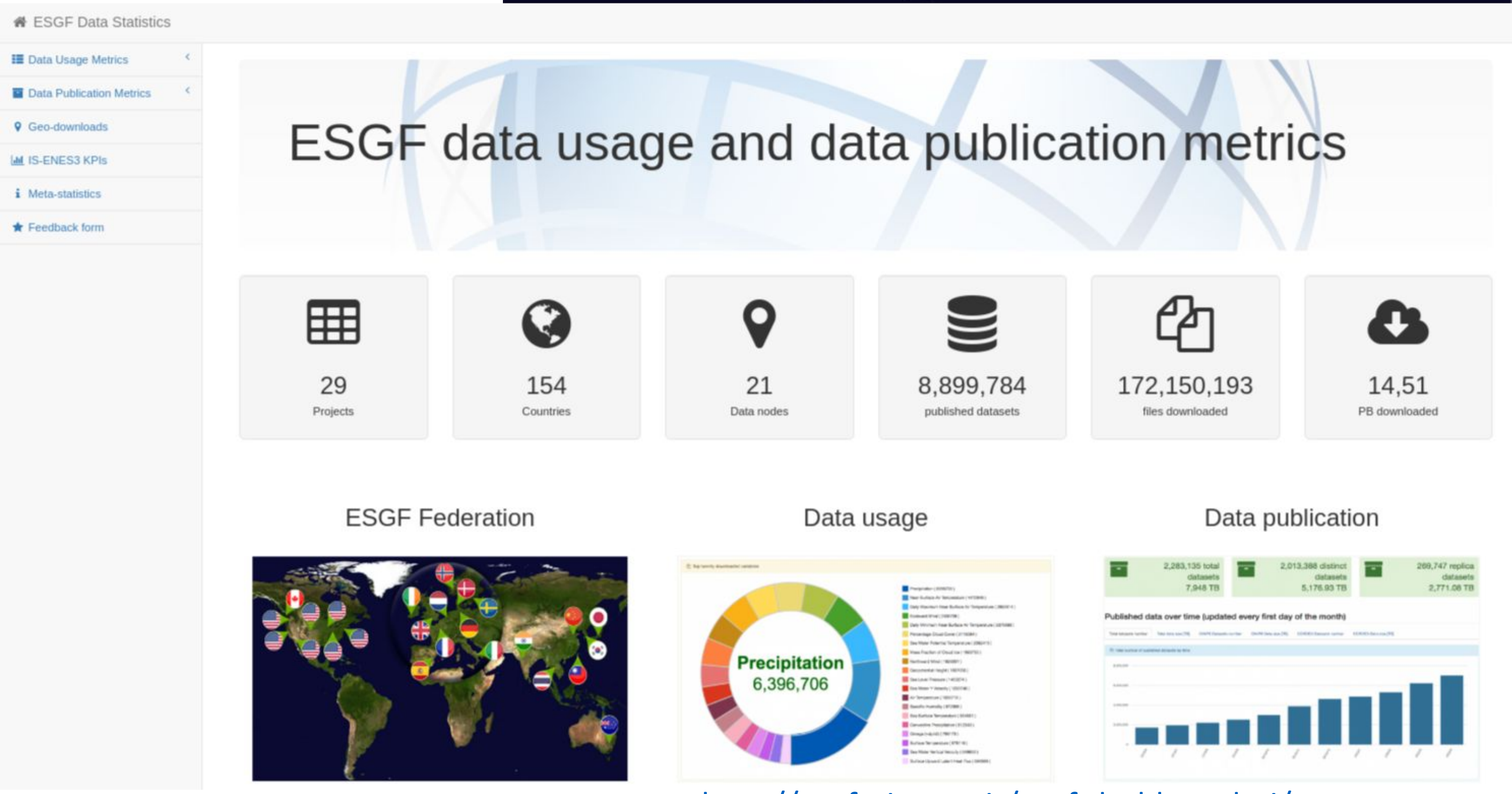
- **A: Metadata and Data standards**
- **C: Enabling data analysis activities near to the data**

The ENES CDI: objectives

Support data distribution for internationally coordinated climate model intercomparison projects (CMIP, currently CMIP6)

- Operational infrastructure
- 1000s of users downloading 100s of TBytes per month (mostly netcdf data)
- Tier1 centers manage large data pools and coordinate data replication (Petabyte range)
- Data near processing support at Tier1 sites





ESGF Data Statistics

- Data Usage Metrics
- Data Publication Metrics
- Geo-downloads
- IS-ENES3 KPIs
- Meta-statistics
- Feedback form

ESGF data usage and data publication metrics

29	154	21	8,899,784	172,150,193	14,51
Projects	Countries	Data nodes	published datasets	files downloaded	PB downloaded

ESGF Federation

Data usage

Precipitation
6,396,706

Data publication

2,283,135 total datasets 7,948 TB

2,013,388 distinct datasets 5,176.93 TB

269,747 replica datasets 2,771.08 TB

Published data over time (updated every first day of the month)

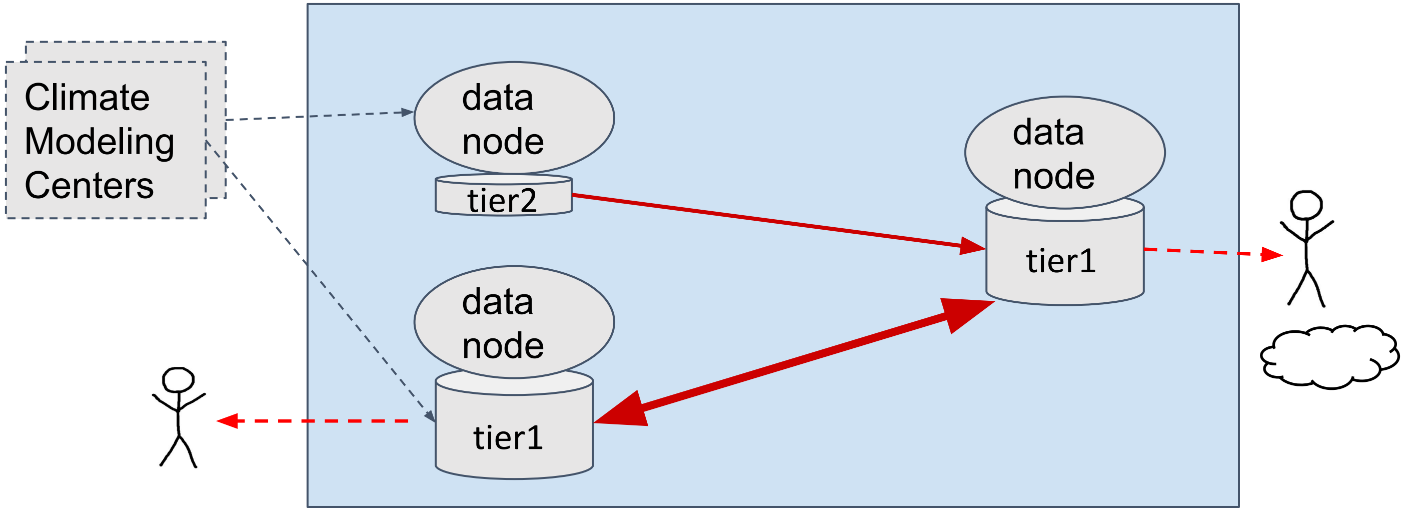
http://esgf-ui.cmcc.it/esgf-dashboard-ui/

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

Climate Data Movement Infrastructure Fabric

tier2 to tier1 data movement:

- for load balancing, resilience, long term preservation ...
- http and globus based
- common replication manger SW agent at tier1 sites
  - integrated in site specific workflows



tier1 to tier1 data movement:

- for load balancing, resilience, data analysis support, ..
- globus usage, optimized (and monitored) communication routes (disk to disk …)
- replica consistency (versioning, withdrawal of data,..)

Data placement:

- user → data manager → …
- managed multi-PByte data pools
- Overall replication strategy
  + site specific user requests

Data scheduling

- workflow managers at tier1 sites
- Overall global “namespace”
- declarative spec of “which data”

Data integrity

- actionable PIDs for all data sets (versioning + replica info)
- errata service
- all modifications to overall data collection via a common “data publisher component” by data managers

Climate Data Movement Infrastructure Fabric: Summary

- **data transport backbone not the key bottleneck**
  - perfsonar monitoring, coordination with nw providers and broadband optimizations at tier1 sites
  - institutional “last mile” to disk optimizations required a lot of effort (only partly on the technical side ..)

- **data transport operational issues**
  - no overall globus support (especially at best effort base managed tier2 sites)
  - management of data integrity not fully automatic and relies on “good behavior” of site admins

- **data management organisational issues**
  - community specific tooling to manage data replication
    - declarative specification of data collection to be transfered ⟶ file list ⟶ transfer manager (using status database) ⟶ downloading ⟶data publication as “replica”
  - maintaining overall data integrity (data retraction, versioning across sites, …) only partly automatized
  - yet overall we managed to transfer 10s of PBytes of CMIP6 data in 2020

- current infrastructure is posix file based (netcdf format mostly), usage of institutional and commercial clouds will have significant impact in the future (cloud native storage formats e.g. Zarr, cloud native repl. support)

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

The Consortium

Coordinated by **CNRS-IPSL**, the IS-ENES3 project gathers **22 partners** in **11 countries**

cnrs with cea SORBONNE UNIVERSITÉ
Met Office
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
DKRZ DEUTSCHES KLIMARECHENZENTRUM
BSC
CERFACS
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Koninklijk Nederlands Meteorologisch Instituut
Ministerie van Infrastructuur en Waterstaat
SMHI
UK Research and Innovation
UC UNIVERSIDAD DE CANTABRIA
DLR
netherlands eScience
Norwegian Meteorological Institute
METEO FRANCE Toujours un temps d'avance
MANCHESTER 1824 The University of Manchester
ΔΗΜΟΚΡΙΤΟΣ
WAGENINGEN UNIVERSITY & RESEARCH
CHARLES UNIVERSITY
NORCE
LiU LINKÖPING UNIVERSITY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

Our website
https://is.enes.org/

Follow us on Twitter !
@ISENES_RI

Contact us at
is-enes@ipsl.fr

YouTube Find our videos on our channel !

zenodo Join our ZENODO community !