Federated data usage statistics in the Earth System Grid federation

A.Nuzzo, **M.Mirto, P. Nassisi, S. Fiore, G. Aloisio**
CMCC Foundation
(Euro Mediterranean Center on Climate Change)

6th Annual ESGF Conference
Dec 6-9 2016, Washington, DC

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

cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici

Outline

- *Goals and main tasks*
- *Architecture in the large – single node level*
- *Architecture in the large – federation level*
- *Federation protocol*
- *New Dashboard-UI module*

Goals and main tasks

The main goal of the DWT was to provide a ***distributed and scalable monitoring framework*** responsible for:

- capturing usage metrics, system status and aggregated information at the single site level and at the federated level
- providing the user with a user friendly interface including widget showing aggregated statistics and monitoring information.

The Dashboard system faces this important challenge through two main components:

ESGF Dashboard
(back-end engine)

ESGF Dashboard-UI
(front-end layer)

Architecture in the large – Single node level

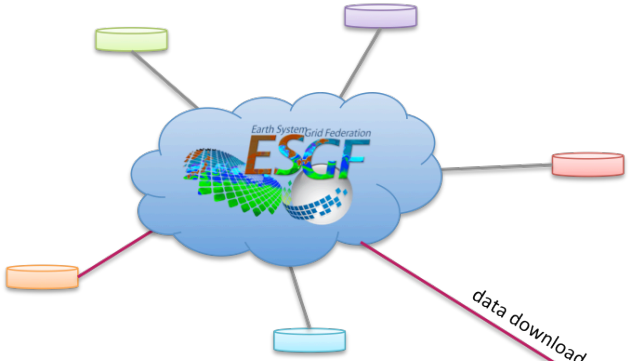
ESGF DATA NODE

ESGF node manager filter

esgf_dashboard
ESGCET

DASHBOARD_QUEUE

ID	url_path	duration	size	timestamp	success	processed
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...



Features:
- Extended set of statistics
- Fine grain level
- Project specific views
- More scalable design

ETL

Multi Tier Database
1 Datawarehouse +
A set of data marts

METADATA

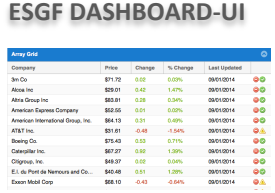
ESGF dashboard

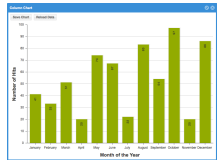
getMetadata(url_path)

METADATA

SOLR

ESGF DASHBOARD-UI

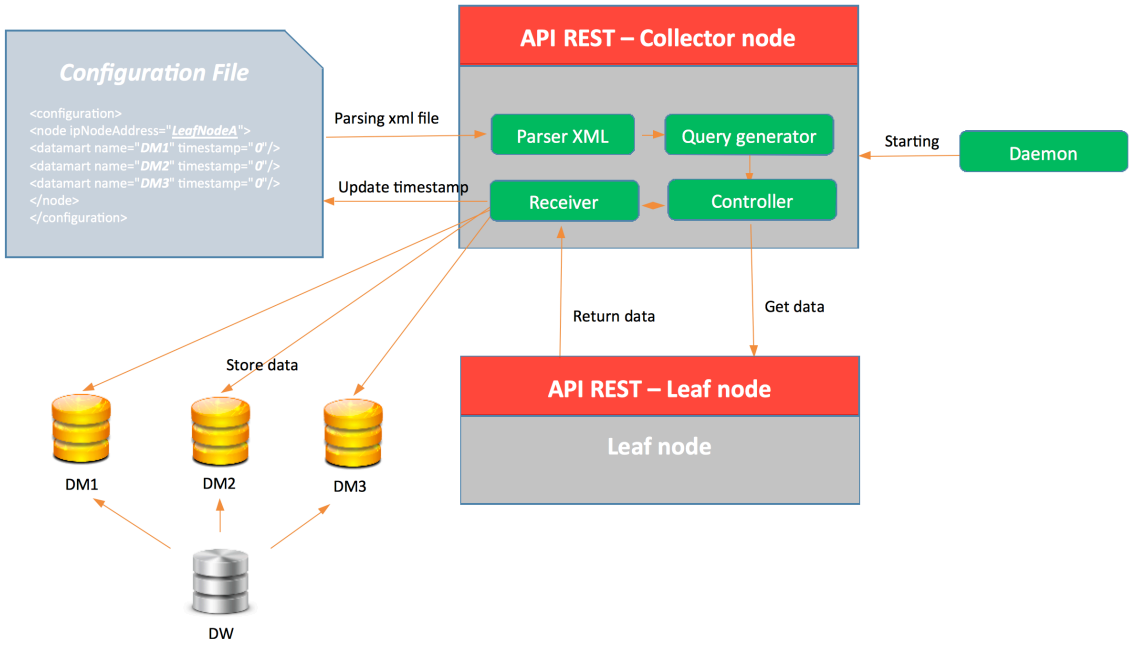




Architecture in the large – Federation level

Till now, the federated statistics have been collected by manually executing a set of different queries on the various data nodes and importing the results into a single database.

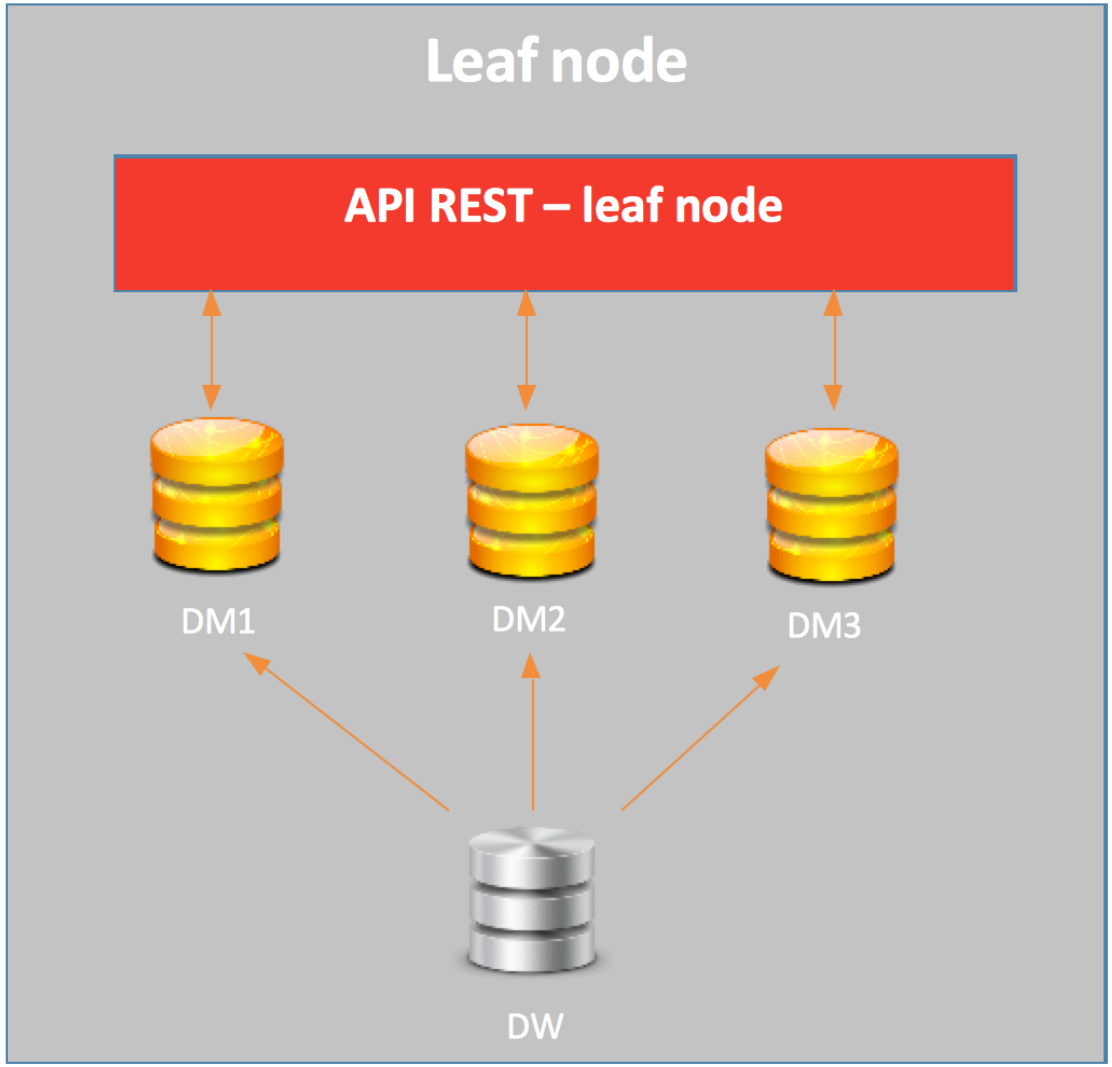
- The federated protocol is based on a hierarchical view of the system
- Two kinds of nodes:
  - Collector node
  - Leaf node



Architecture in the large – Federation level

Leaf node

- Dashboard back-end engine
  - A data warehouse storing all the data related to the downloads
  - A set of data marts containing specific statistics information
- A set of RESTful API providing the collector node the possibility to access data marts and getting the statistics.



Architecture in the large – Federation level

Collector node

The collector node has a more complex structure because, in addition to making its information available to the collector through the RESTful API, is in charge to query its leaf nodes.

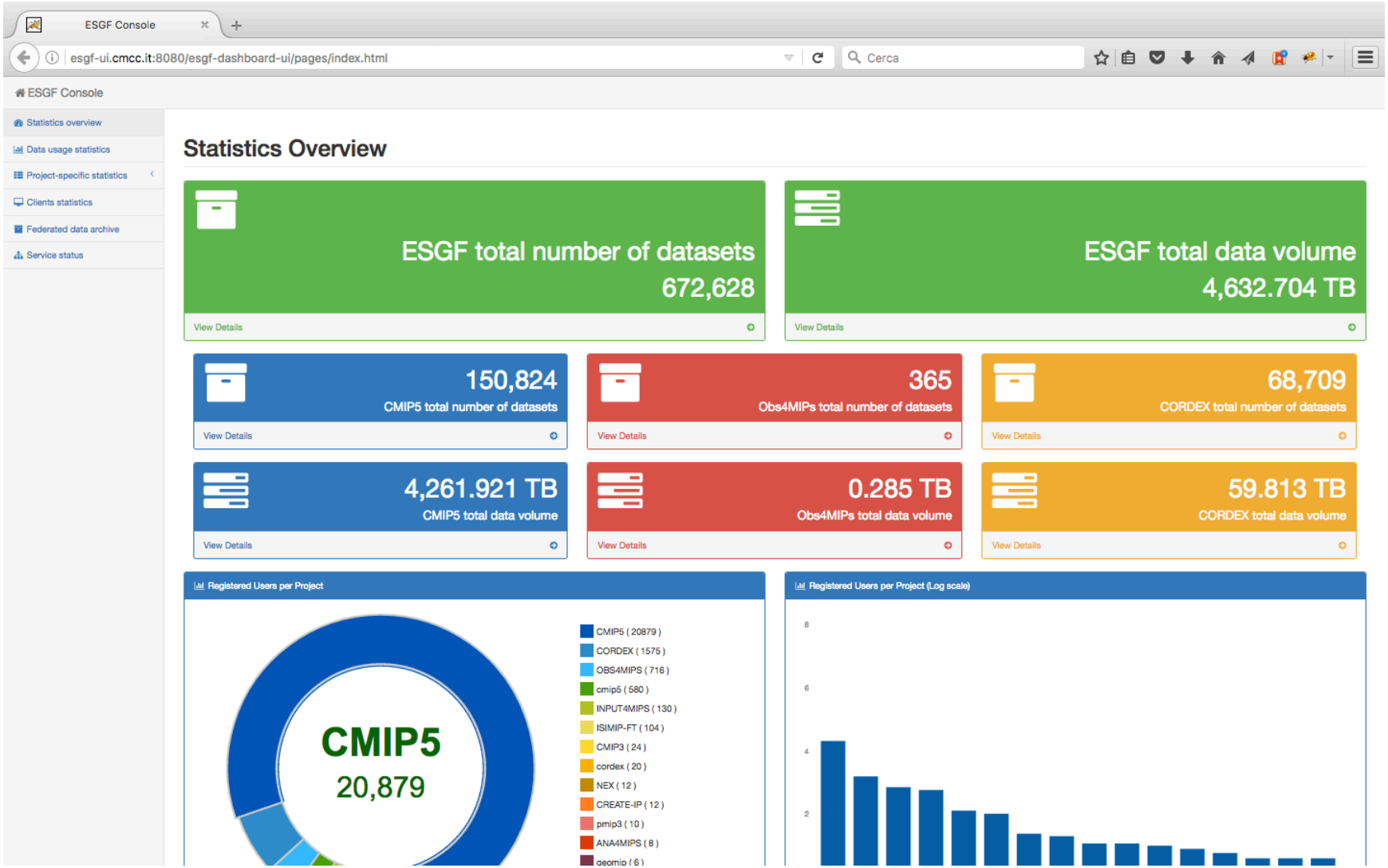
The collector node is composed by:

- ❖ data warehouse and data marts
- ❖ RESTful API
- ❖ xml configuration file
- ❖ federation component

A first prototype of such protocol has been successfully installed and tested on four sites: CMCC, DKRZ, NASA/JPL, PCMDI

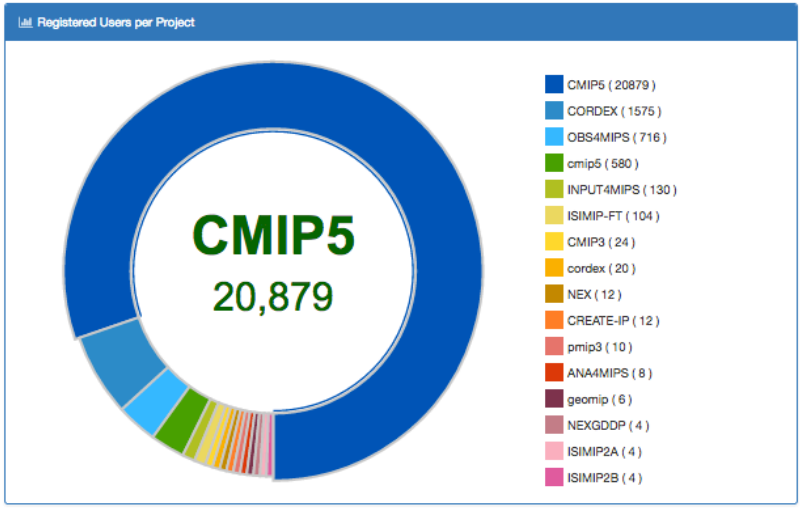
```
– <configuration>
  – <node ipNodeAddress="esgf-fedtest.dkrz.de" port="8080">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  – <node ipNodeAddress="esgf-data.jpl.nasa.gov" port="0">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  – <node ipNodeAddress="pcmdi11.llnl.gov" port="8080">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  </configuration>
```

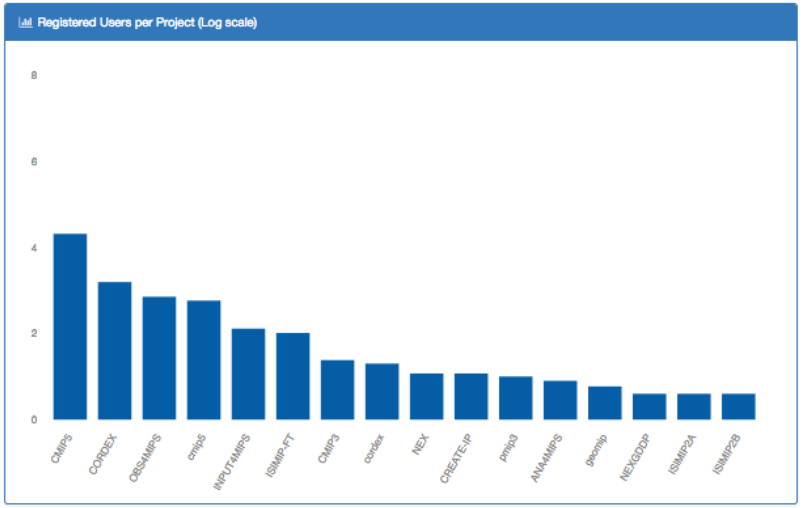
New Dashboard-UI – Statistics Overview

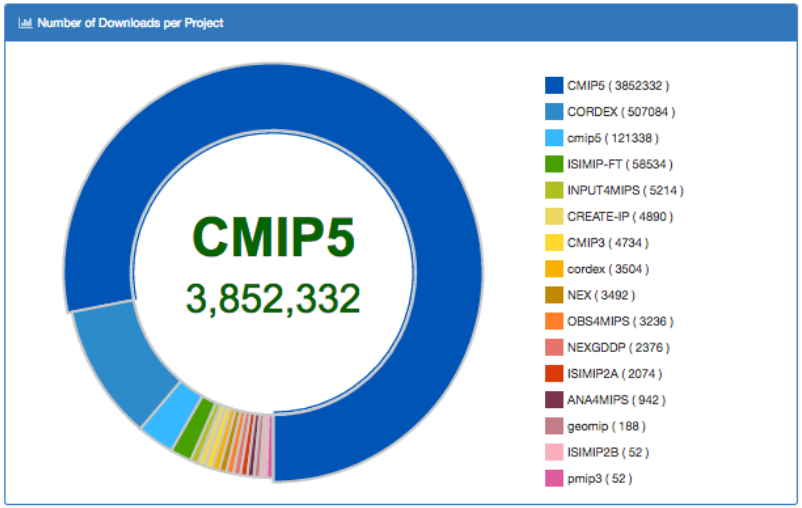


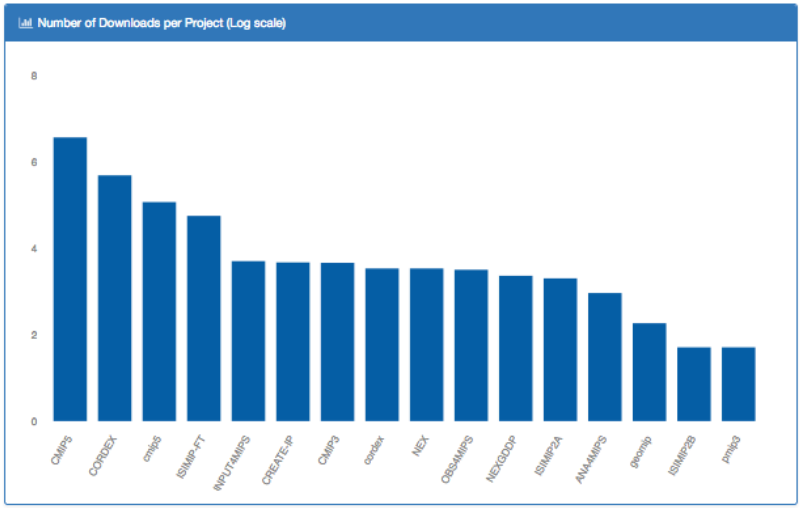
New Dashboard-UI – Statistics Overview

Registered Users and Number of Downloads per Project

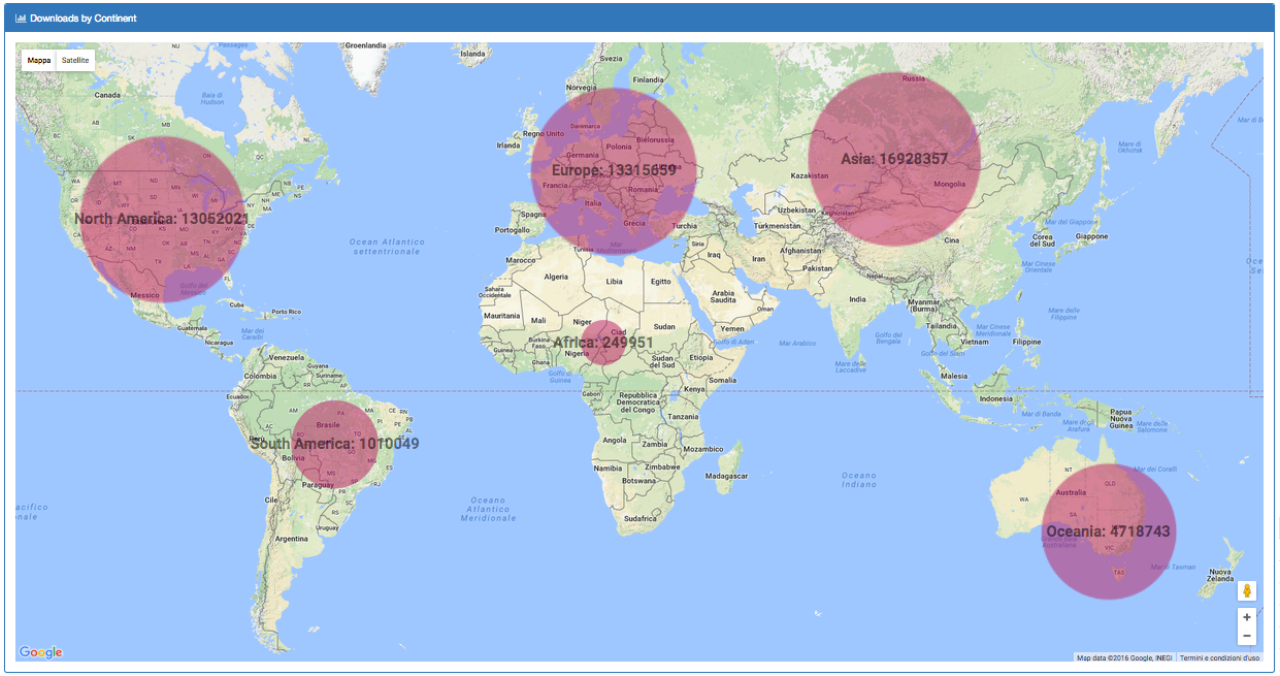








New Dashboard-UI



Number of Downloads by Continent and Countries

Belize	9
Puerto Rico	1
Nicaragua	1

Egypt	713
Ethiopia	575
Ghana	376
Morocco	239
Benin	217

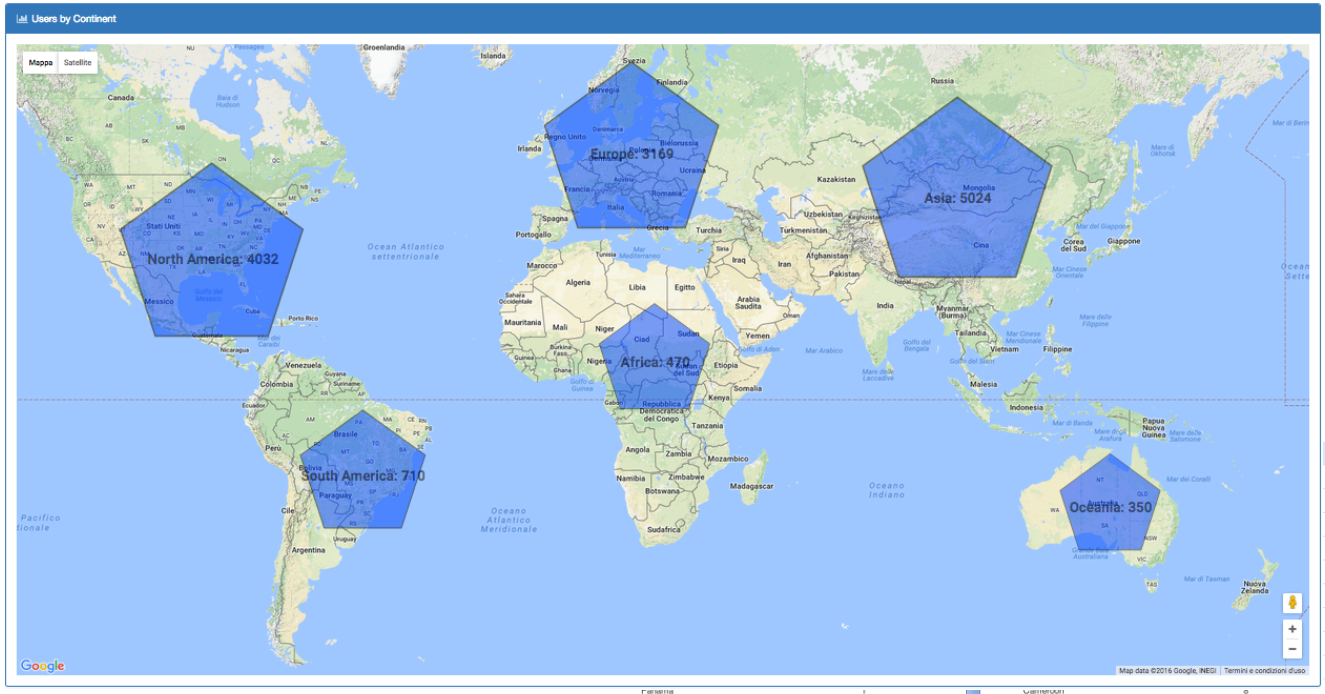
Europe	6881487
Germany	3214626
Spain	1514863
Switzerland	659415
United Kingdom	555524
France	367639
Netherlands	107846
Norway	80835
Italy	67388
Sweden	56446
Portugal	51221
Greece	25523

South America	519605
Chile	407411
Brazil	60832
Colombia	37777
Argentina	8367
Peru	4943
Ecuador	210
Bolivia	33
Suriname	19
Venezuela	11
Uruguay	2

Asia	9328428
China	6938391
Japan	1333097
Korea, South	348387
Iran	313300
Thailand	138416
India	105774
Taiwan	59041
Hong Kong	31600
United Arab Emirates	14544
Singapore	14208
Turkey	9631

Oceania	2373828
Australia	2157734
New Zealand	216087
New Caledonia	7

New Dashboard-UI



Number of Users by Continent and Countries

Nicaragua	1

Burkina Faso	7

Europe	1585
Germany	508
France	188
Italy	121
Spain	109
Netherlands	83
Russian Federation	79
Norway	75
Sweden	64
Switzerland	47
Belgium	37
Denmark	36

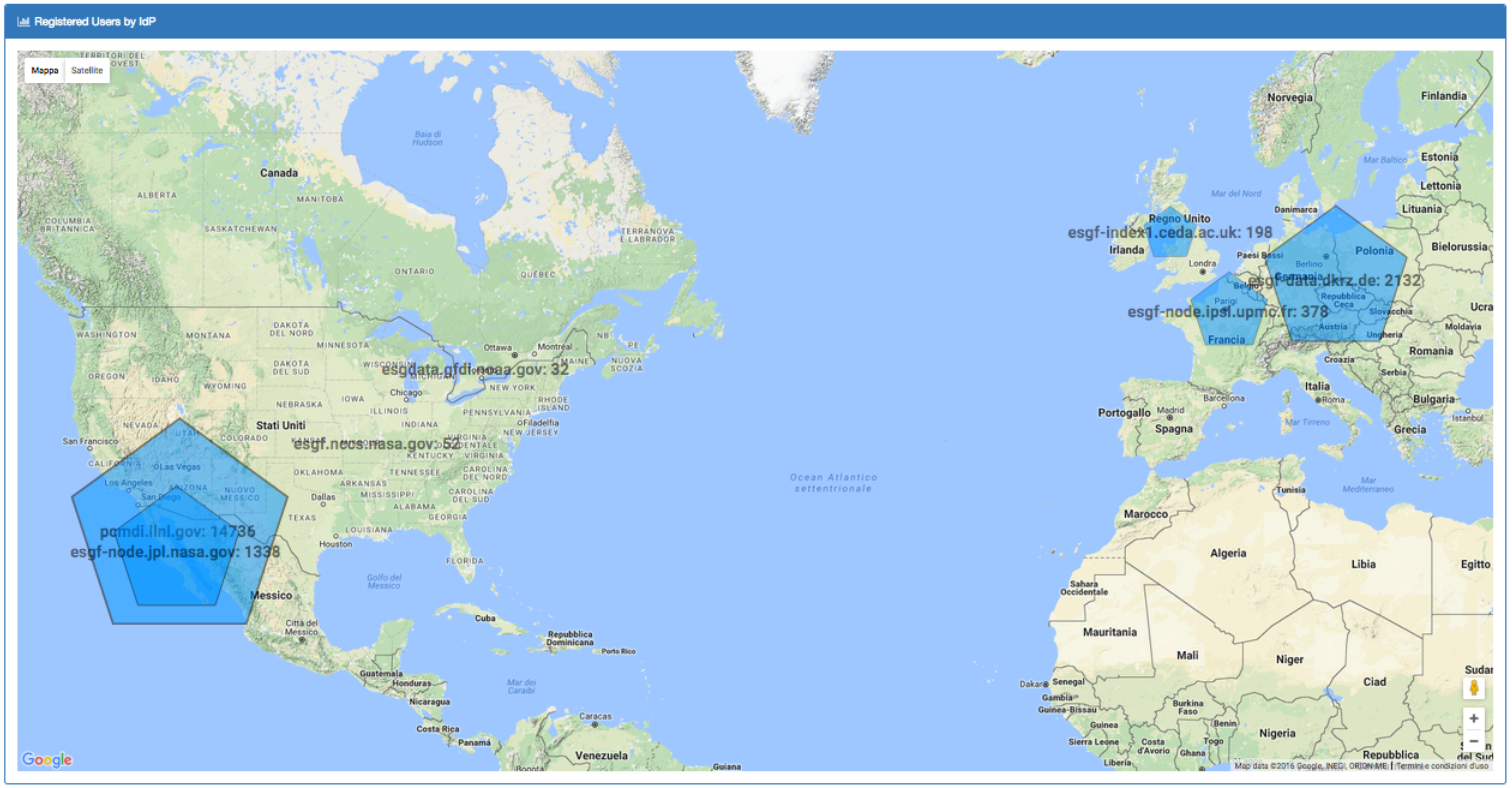
South America	356
Brazil	154
Colombia	71
Chile	56
Argentina	39
Peru	20
Bolivia	6
Suriname	3
Venezuela	3
Ecuador	2
Paraguay	1
Uruguay	1

Asia	2515
China	1084
India	460
Japan	217
Iran	171
Korea, South	135
Thailand	97
Taiwan	46
Indonesia	40
Pakistan	37
Malaysia	29
Israel	21

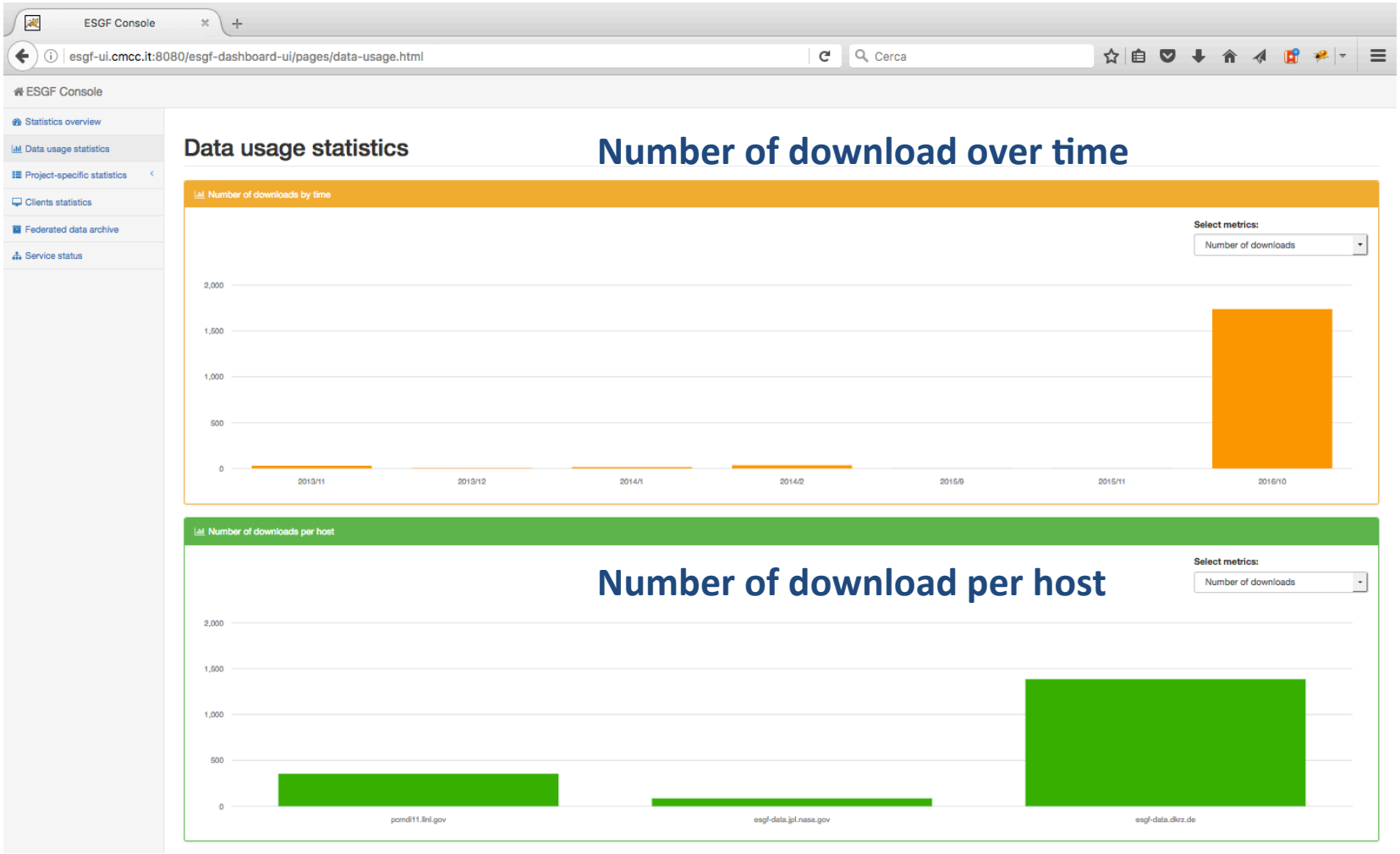
Oceania	175
Australia	154
New Zealand	19
Norfolk Island	1
New Caledonia	1

New Dashboard-UI

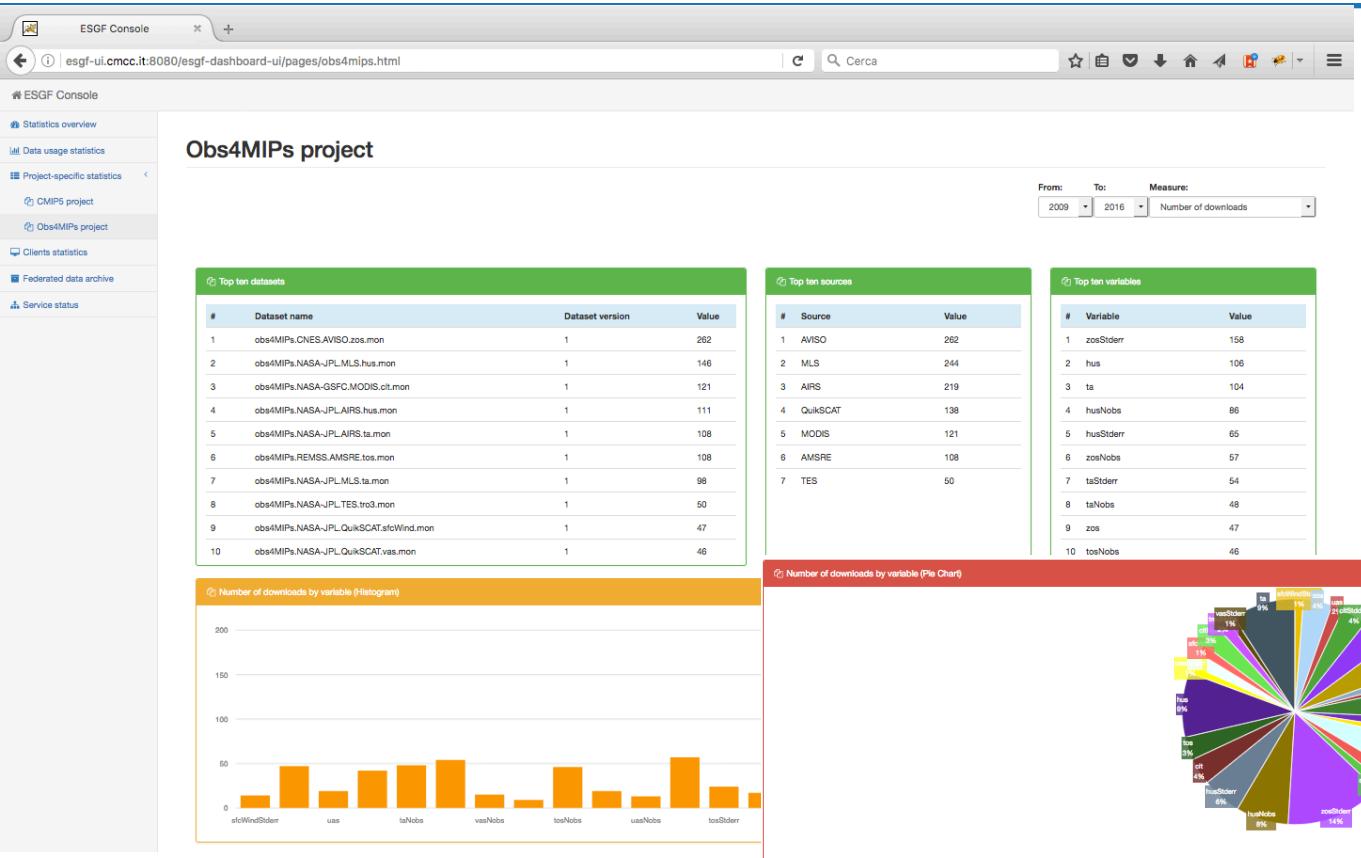
Number of Registered Users by IdPs



New Dashboard-UI – Data Usage Statistics section



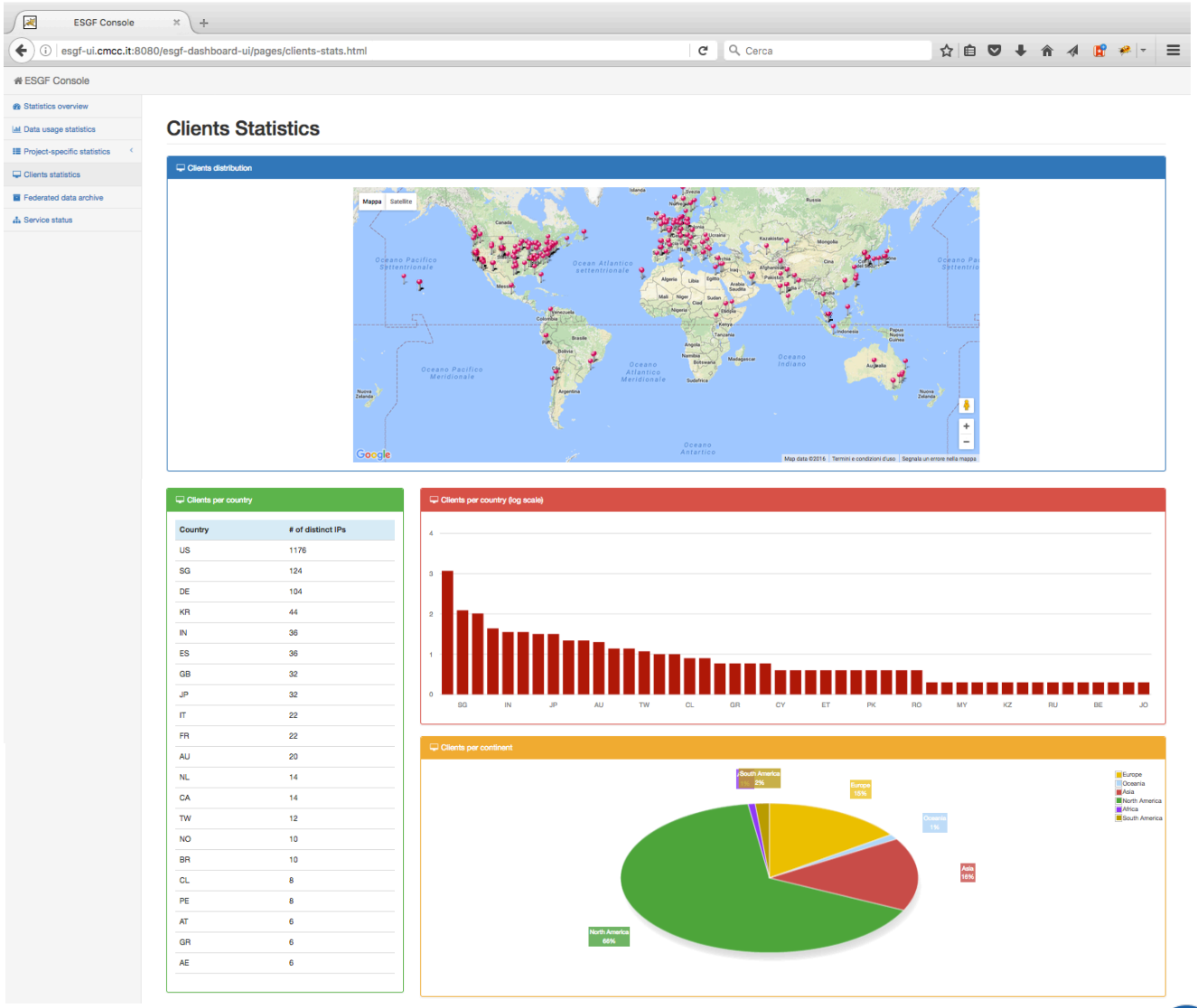
New Dashboard-UI – Project specific section



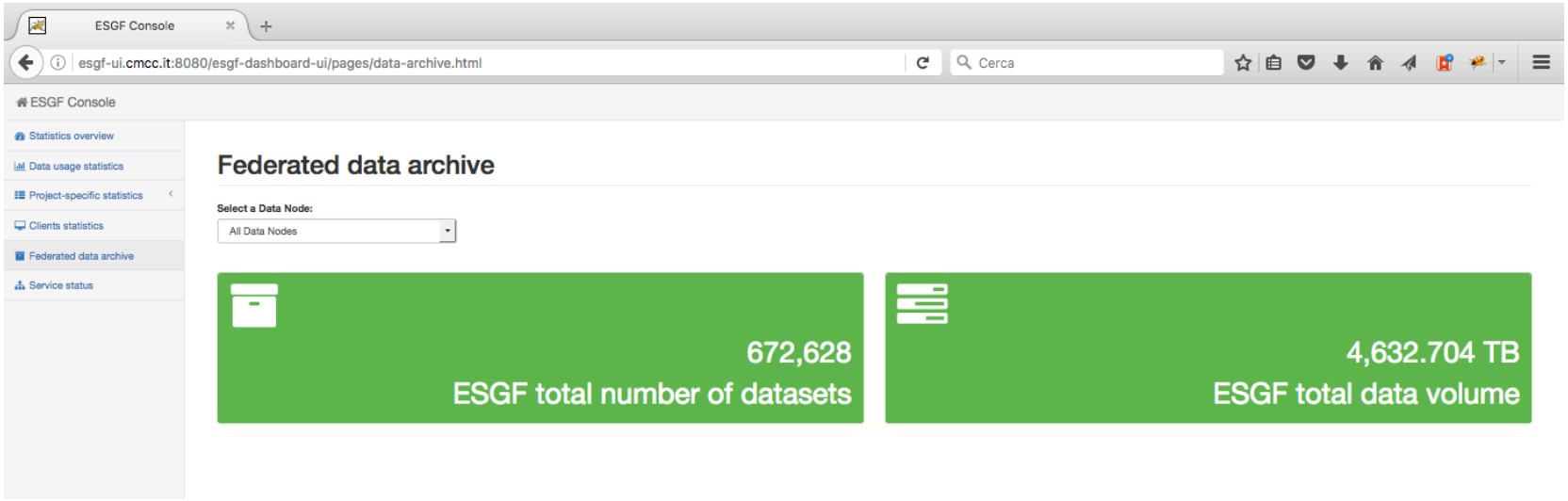
Obs4MIP4 project

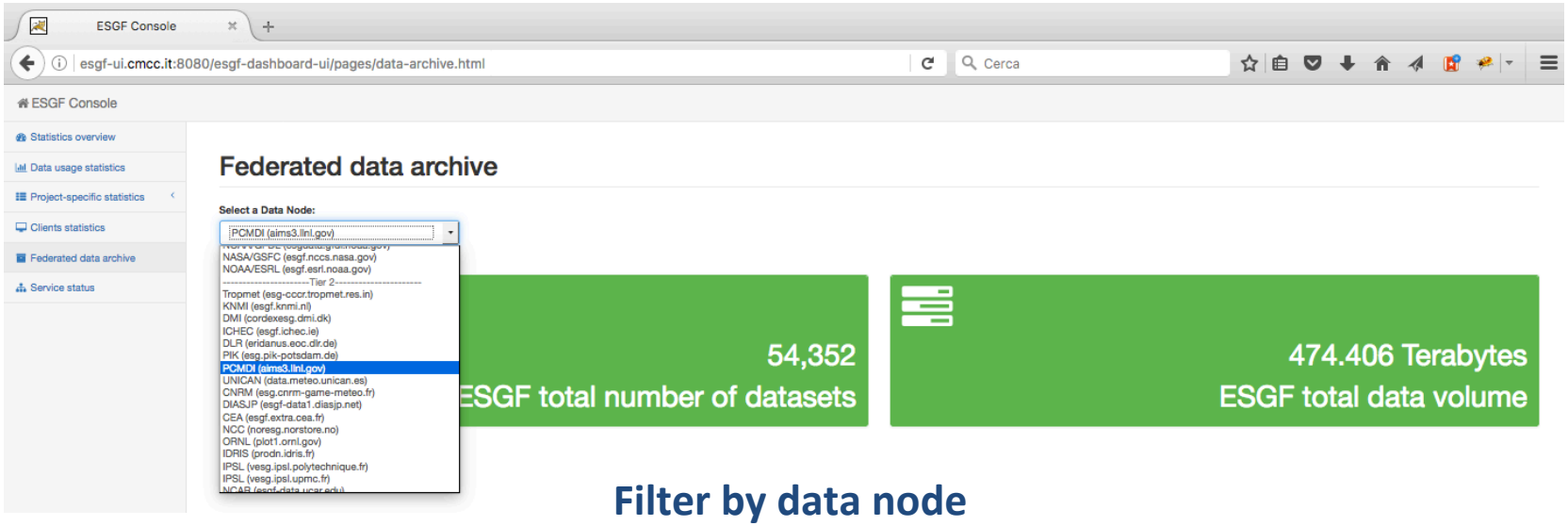
New Dashboard-UI - Client statistics section

Number of users who made a download



New Dashboard-UI – Federated Data Archive section





Filter by data node

New Dashboard-UI – Federated Data Archive section

Total number of datasets and related size for each Model and Modeling Institute for CMIP5 project (data obtained by SOLR module).

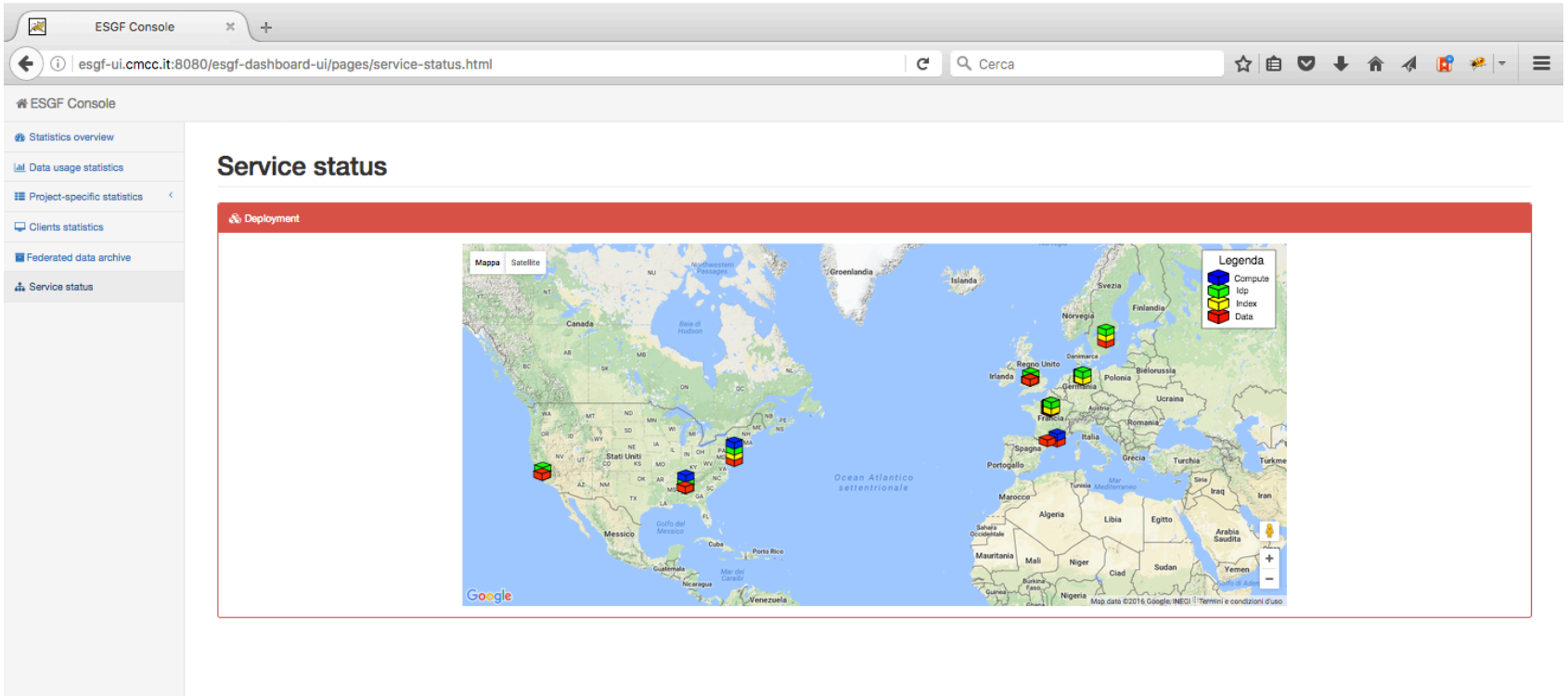
Published CMIP5 data per Model

#	Model	# of datasets	Size (TB)
1	ACCESS1.0	300	35.476
2	ACCESS1.3	316	33.825
3	BCC-CSM1.1	5,101	41.844
4	BCC-CSM1.1(m)	403	3,610.014
5	BNU-ESM	513	19.93
6	CCSM4	4,956	166.347
7	CESM1(BGC)	309	22.716
8	CESM1(CAM5)	378	31.279
9	CESM1(CAM5.1-FV2)	56	2.928
10	CESM1(FASTCHEM)	51	3.269
11	CESM1(WACCM)	155	3.873
12	CFSv2-2011	2,644	34.793
13	CMCC-CESM	87	1.254
14	CMCC-CM	992	144.024
15	CMCC-CMS	110	4.227
16	CNRM-CM5	3,390	135.5
17	CNRM-CM5-2	263	11.909
18	CSIRO-Mk3.6.0	3,120	57.593
19	CSIRO-Mk3L-1-2	26	0.03
20	CanAM4	184	8.027
21	CanCM4	19,118	23.317
22	CanESM2	2,577	39.844

Published CMIP5 data per Institute

#	Modeling institute	# of datasets	Size (TB)
1	BCC	5,504	97.992
2	BNU	513	19.93
3	CCCMA	21,879	71.187
4	CMCC	1,189	149.504
5	CNRM-CERFACS	3,653	147.409
6	COLA-CFS	1,189	7.953
7	CSIRO-BOM	616	69.301
8	CSIRO-QCCCE	3,120	57.593
9	FIO	230	6.694
10	ICHEC	3,620	134.946
11	INM	486	21.402
12	INPE	24	7.957
13	IPSL	10,757	699.208
14	LASG-CESS	1,553	40.014
15	LASG-IAP	418	7.597
16	MIROC	16,791	823.919
17	MOHC	24,720	148.004
18	MPI-M	11,655	195.248
19	MRI	7,271	404.758
20	NASA-GISS	7,681	227.295
21	NASA-GMAO	2,520	8.631
22	NCAR	4,956	166.347

New Dashboard-UI – Service status section



Deployment distribution

Thank you

cmcc
Centro Euro-Mediterraneo
sui Cambiamenti Climatici