Leveraging and Easing End Users' Climate Data Access by Interfacing Infrastructures

IN33E-0889

Abstract 340841

Christian Pagé (Cerfacs, France)

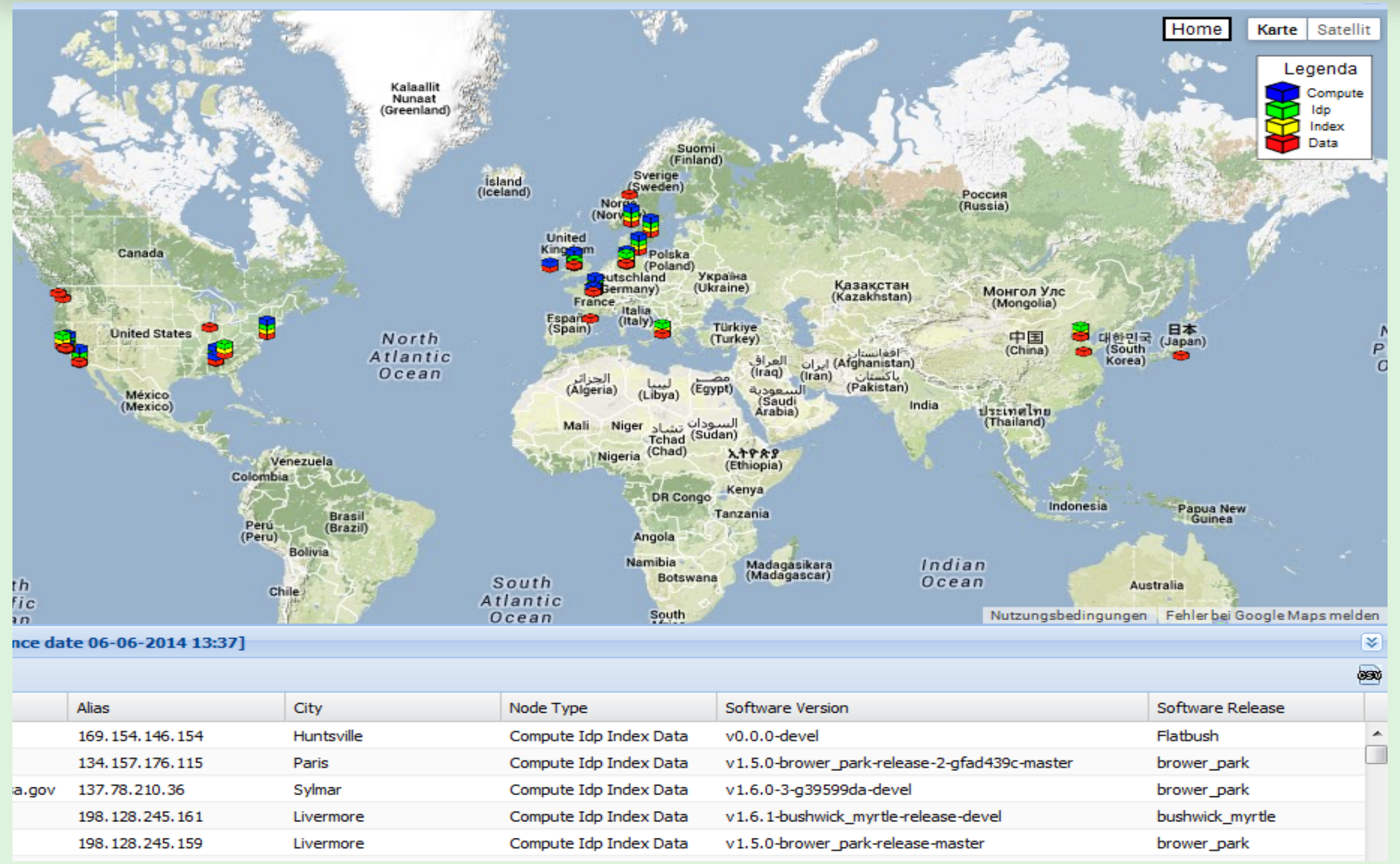
christian.page@cerfacs.fr http://linkedin.com/in/pagechristian https://www.researchgate.net/profile/Christian_Page http://cerfacs.fr/~page

CECI, Université de Toulouse, CNRS, Cerfacs, Toulouse, France

Wim Som de Cerff & Maarten Plieger & Alessandro Spinuso (KNMI, Netherlands)

Iraklis Angelos Klampanos & Vangelis Karkaletsis (NCSR Demokritos, Greece) Malcolm Atkinson (University of Edinburgh, UK)

I New Challenges for Science



nce date 06-06-2014 13:37]

	Alias	City	Node Type	Software Version	Software Release
	169.154.146.154	Huntsville	Compute Idp Index Data	v0.0.0-devel	Flatbush
	134.157.176.115	Paris	Compute Idp Index Data	v1.5.0-brower_park-release-2-gfad439c-master	brower_park
a.gov	137.78.210.36	Sylmar	Compute Idp Index Data	v1.6.0-3-g39599da-devel	brower_park
	198.128.245.161	Livermore	Compute Idp Index Data	v1.6.1-bushwick_myrtle-release-devel	bushwick_myrtle
	198.128.245.159	Livermore	Compute Idp Index Data	v1.5.0-brower_park-release-master	brower_park

FIG 1: Earth System Grid Federation (ESGF) of Servers

- The Earth System Grid Federation is a Peer2Peer International Network of Servers
- Currently, the *Download & Analyze* workflow is no longer sustainable
- European Data Nodes are managed by IS-ENES (Infrastructure for the European Network of Earth System Modelling)

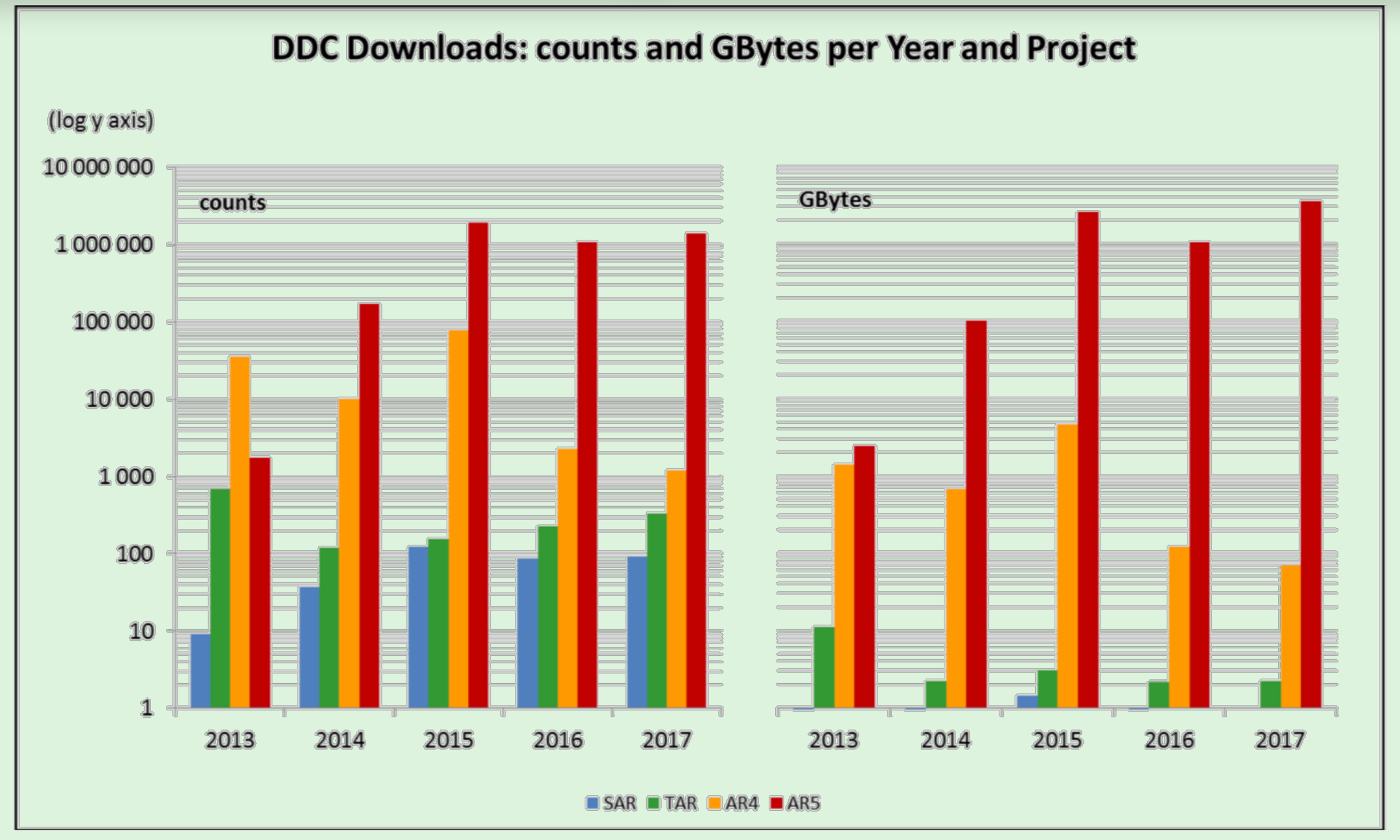


FIG 2: DKRZ ESGF Data Node Data Volume (GB) Downloaded for SAR, TAR, AR4 and AR5

- Increase in Data Production is significant
- Moving data will be soon much more expensive than computing time
- A change of paradigm

	CMIP5	CMIP6	CMIP7
Year	2012	2017	2022
Power factor	1	30	1000
Npp	200	357	647
Resolution [km]	100	56	31
Number of mesh points [millions]	3,2	18,1	108,4
Ensemble size	120	214	388
Number of variables	800	1068	1439
Interval of 3-dimensional output (hours)	6	4	3
Years simulated	90000	120170	161898
Storage density	0,00002	0,00002	0,00002
Distributed Archive Size (Pb)	**3,19**	**86,05**	**2260,20**

FIG 3: Climate Model Intercomparison Projects (CMIP) Archive Size (PB)

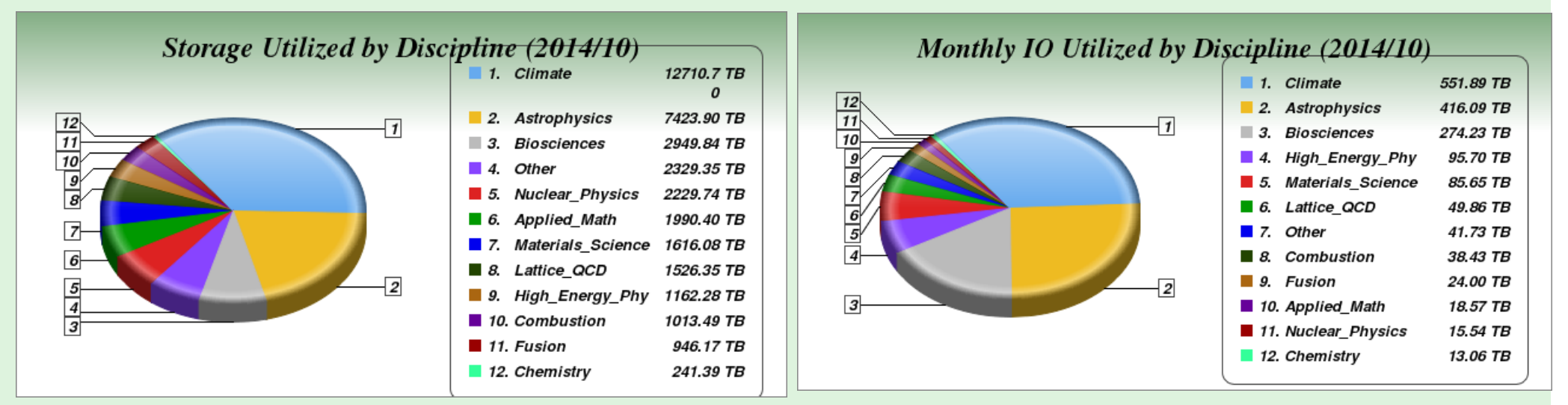


FIG 3: National Energy Research Scientific Computing Center (NERSC) Storage and I/O by Discipline

- Climate has large needs for Storage and I/O
- Large and heterogeneous communities of users

II C4I IS-ENES Tailored UI

- https://climate4impact.eu
- Developed and managed by IS-ENES
- Platform for researchers to explore climate data and perform analysis
- Not only UI, but also Standard Services (WPS, WCS, ...)
- Connects to ESGF web services
- Tailored for end-users
- Supports on-demand data processing and statistical downscaling
- Now containerized version
  - docker & docker-compose
- Compatible with CMIP6, CORDEX, etc.

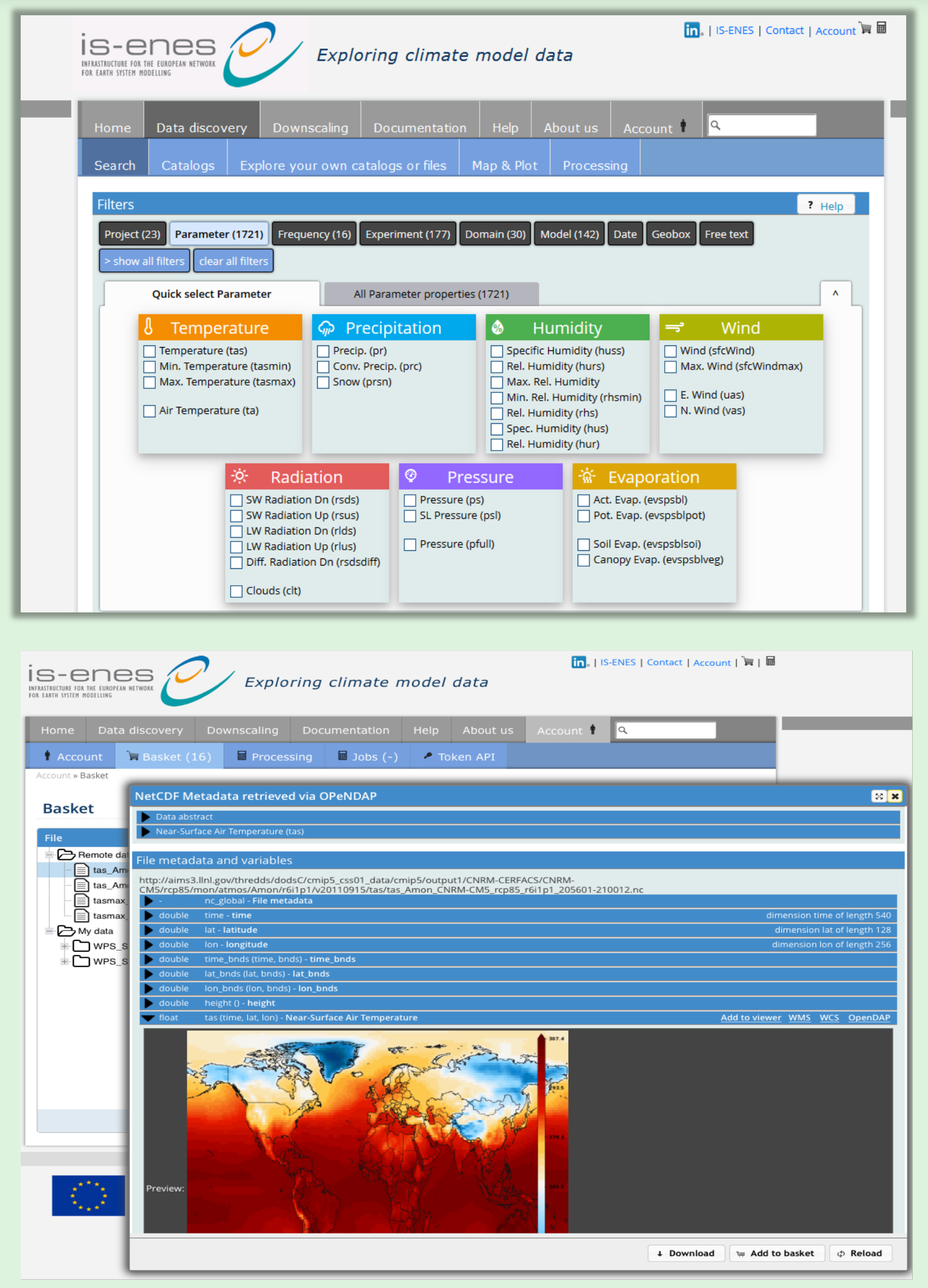


FIG 4: C4I Faceted Search and Interface

III DARE Platform

- dispel4py: High-level streaming dataflow specification API/library. Automated collection of lineage.
- S-ProvFlow: Reproducibility as a Service. Based on W3C-PROV.
- Exareme: Large-scale dataflow processing on the cloud.
- Semagrow: Semantics and linked-data.
- BigDataEurope: Big Data Analytics

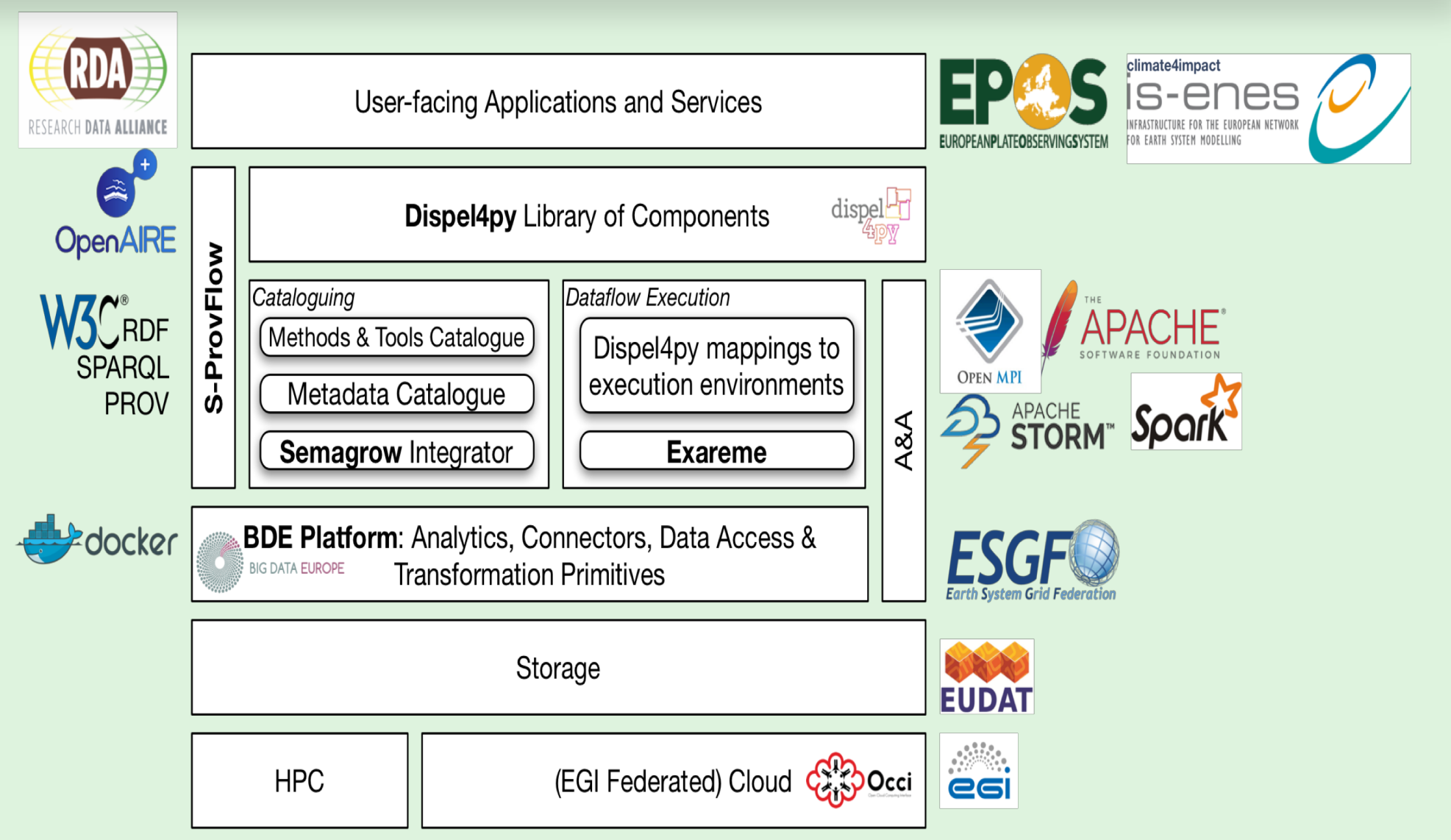

User-facing Applications and Services
Dispel4py Library of Components
S-ProvFlow
Cataloguing
Methods & Tools Catalogue
Metadata Catalogue
Semagrow Integrator
Dataflow Execution
Dispel4py mappings to execution environments
Exareme
A&A
BDE Platform: Analytics, Connectors, Data Access & Transformation Primitives
Storage
HPC
(EGI Federated) Cloud
RDF
SPARQL
PROV

FIG 5: DARE provisional architecture stack indicating related platforms, infrastructures and technologies

IV Mapping Communities' Needs

- User Stories Approach
  - Bottom-Up
- Translating a User Story to
  - Feature
  - Capability
- Leads to specification of Components

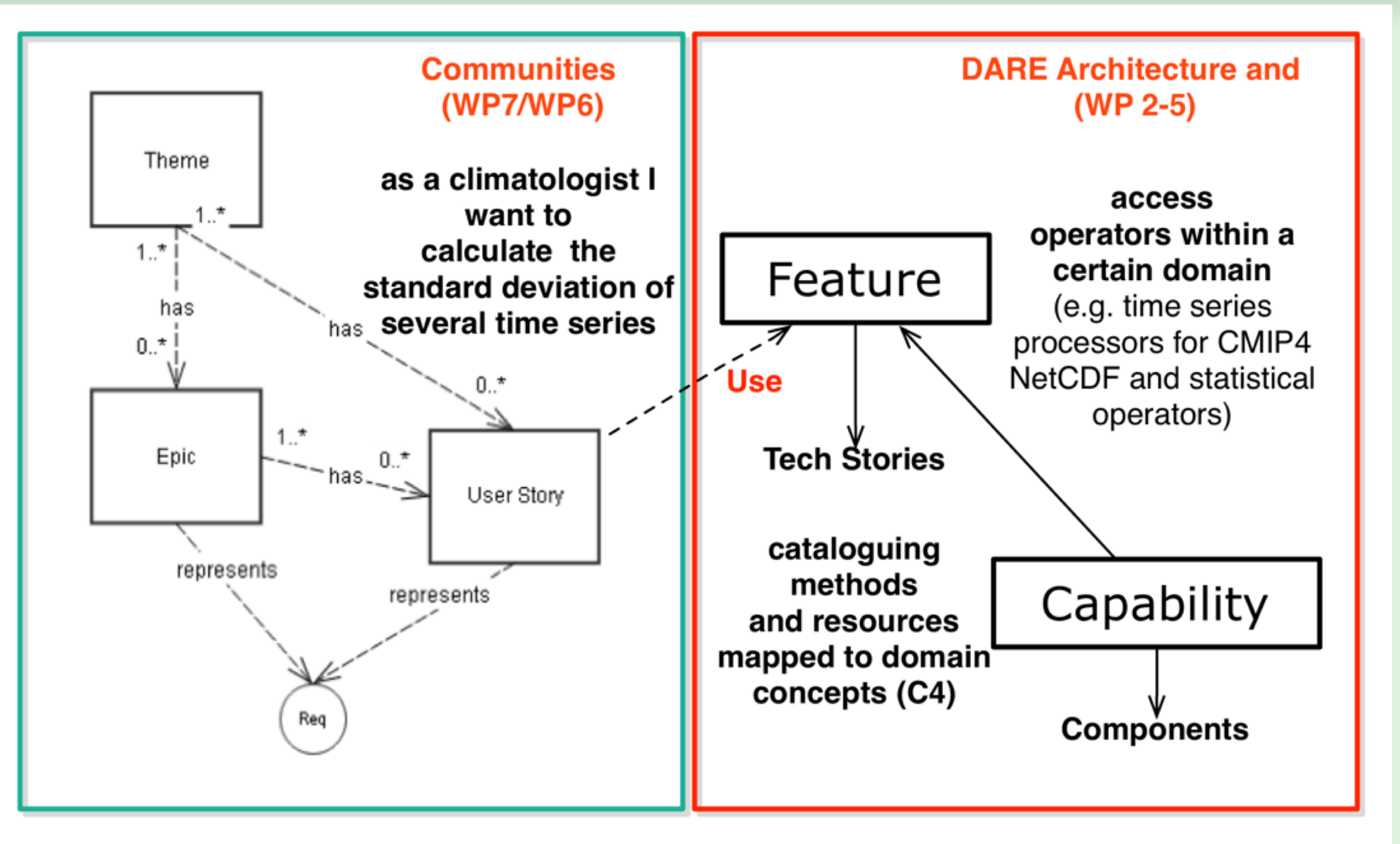

Communities (WP7/WP6)
Theme
Epic
User Story
Req
has
represents
1..*
0..*
as a climatologist I want to calculate the standard deviation of several time series
DARE Architecture and (WP 2-5)
Use
Feature
Tech Stories
access operators within a certain domain (e.g. time series processors for CMIP4 NetCDF and statistical operators)
cataloguing methods and resources mapped to domain concepts (C4)
Capability
Components

FIG 6: Mapping User Stories to Technical Architecture

V IS-ENES Climate Domain Pilot

- Supports a very large number of Use Cases
- Integration of several components
- icclim Climate Indices backend
- Interfaces to external sources
- Major DARE contributions
  - Provenance/Lineage
  - Seamless integration to high-performance data analytics platform

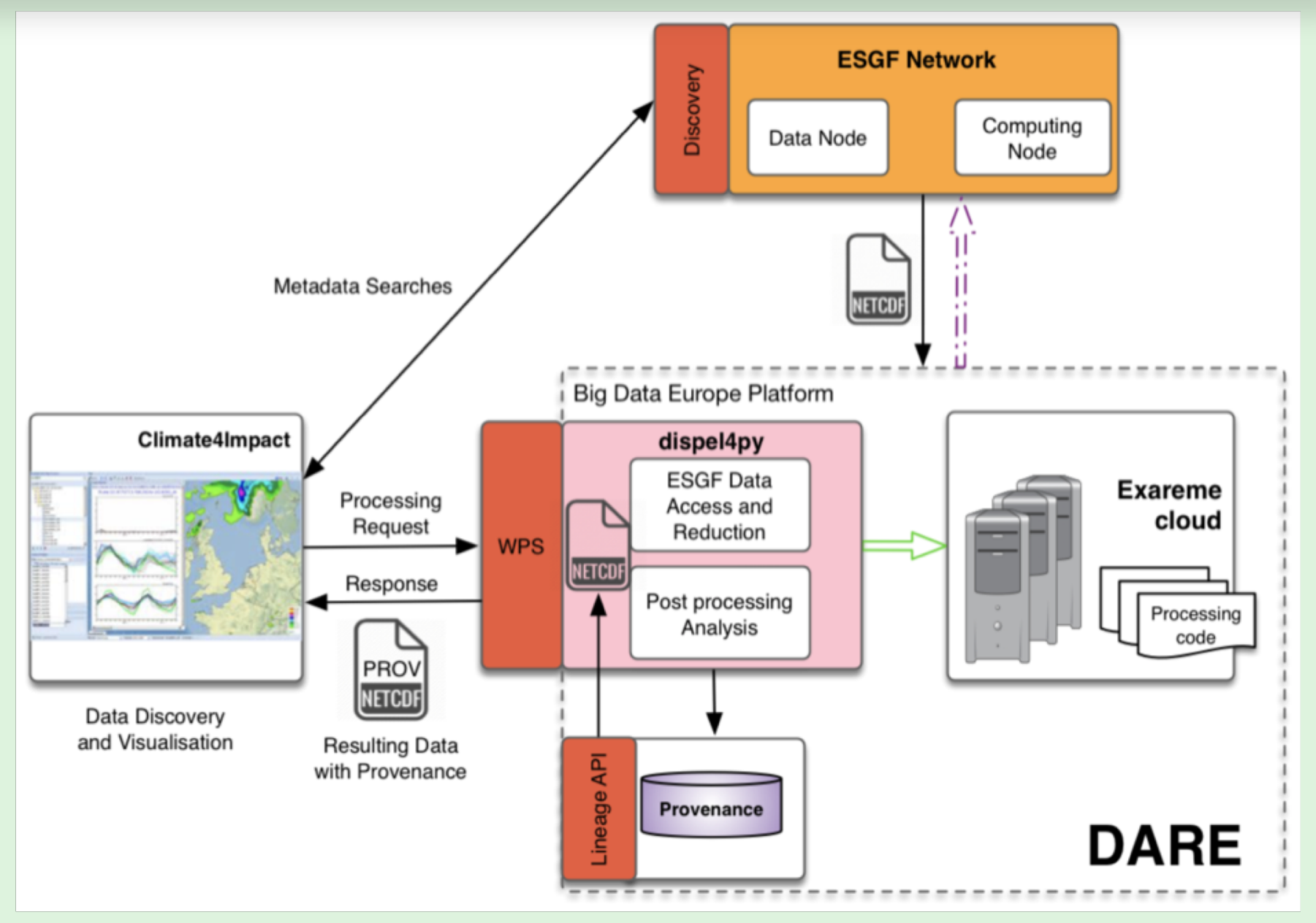


FIG 7: IS-ENES Climate Domain Pilot Schematics

VI Tracking Provenance & Lineage

- Deployed using docker-compose
- Data captured in a document-oriented database based on JSON representation (MongoDB)
- S-ProvFlow collection of methods exposed through OpenApi2.0 API abstraction layer
- Visualization Tool: Provenance Relationships

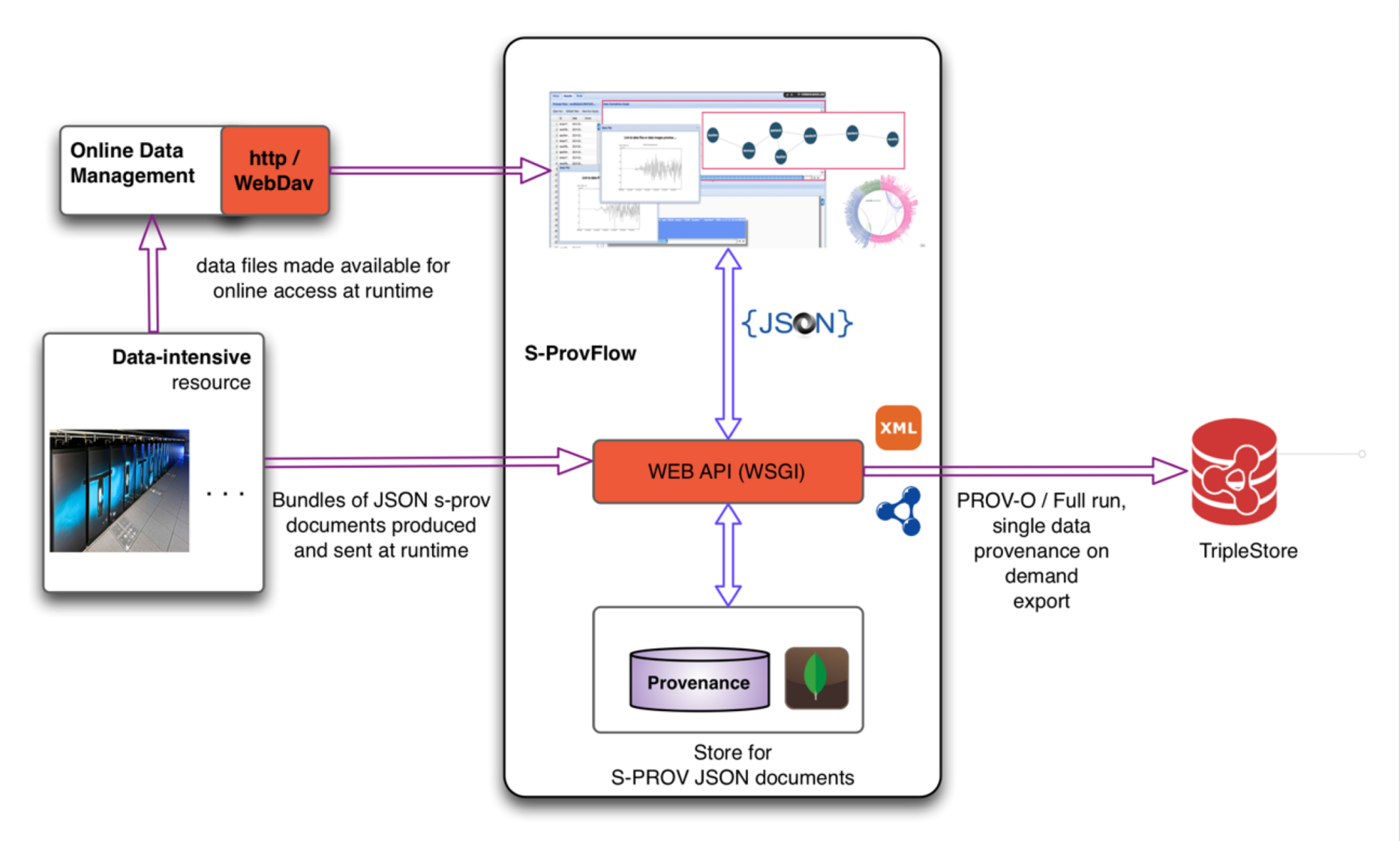

Online Data Management
http / WebDav
data files made available for online access at runtime
Data-intensive resource
Bundles of JSON s-prov documents produced and sent at runtime
S-ProvFlow
{JSON}
XML
WEB API (WSGI)
Provenance
Store for S-PROV JSON documents
PROV-O / Full run, single data provenance on demand export
TripleStore

FIG 8: Schematic architecture exploiting the S-ProvFlow system for acquisition, visualisation, data access and provenance export services

VII Future Work

- Finalize the DARE Architecture (Dec 2018)
- Implement dispel4py Processing Elements for both Domain Pilots
- Develop the OGC WPS interface to dispel4py
- Evaluate where Exareme can be useful within some Domain Pilots Use Cases
- Write interfaces from dispel4py to/from ESGF
  - Both for the data nodes as well as for the computing nodes
- Interface to external services (see Fig. 5, e.g. EUDAT B2 Services)
- Develop the Provenance Model and Semantics for Climate Data and Processing
- Perform first execution tests on the DARE testbed
- Collaborate with H2020/IS-ENES3 to:
  - Design the C4I UI to deal with complete workflows

Get the Poster here!!

Some References

H2020/DARE Project: http://project-dare.eu

IS-ENES C4I: https://climate4impact.eu

EUDAT CDI: https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi

icclim: https://github.com/cerfacs-globc/icclim

https://tinyurl.com/AGU-DARE

The DARE project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 777413