EGI Open Data Platform how can DPM sites contribute ?

Peter Solagna – EGI Foundation

peter.solagna@egi.eu

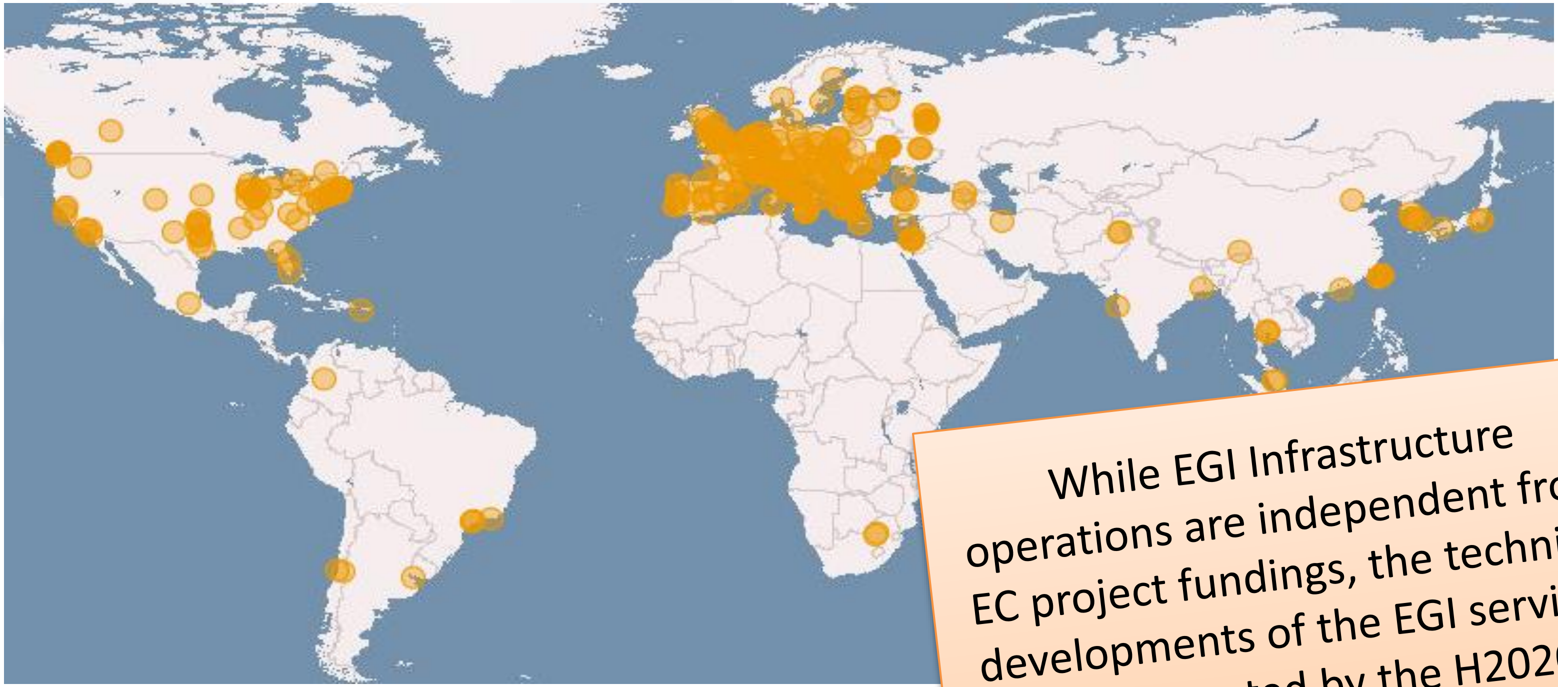
EGI-Engage is co-funded by the Horizon 2020 Framework Programme
of the European Union under grant number 654142

Summary

- Overview of the EGI Open Data Platform
- Use cases
- How can DPM and DynaFed be part of the EGI federated data solutions?

Overview of the EGI Open Data Platform

EGI Infrastructure



While EGI Infrastructure operations are independent from EC project fundings, the technical developments of the EGI services are supported by the H2020 project EGI-Engage.

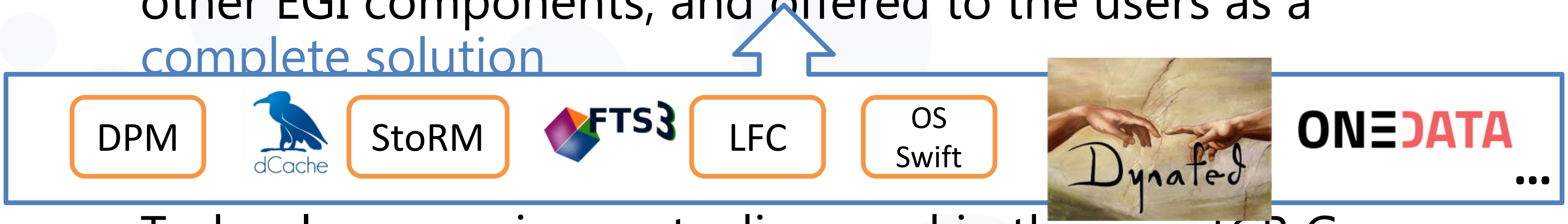
One of the EGI Engage goals is to develop the DataHub prototype.

- Distributed, federated storage and compute facilities
- Compute platforms (Grid, Cloud)
- Virtual Research Environments
- > 200 user research projects

- 300 reso

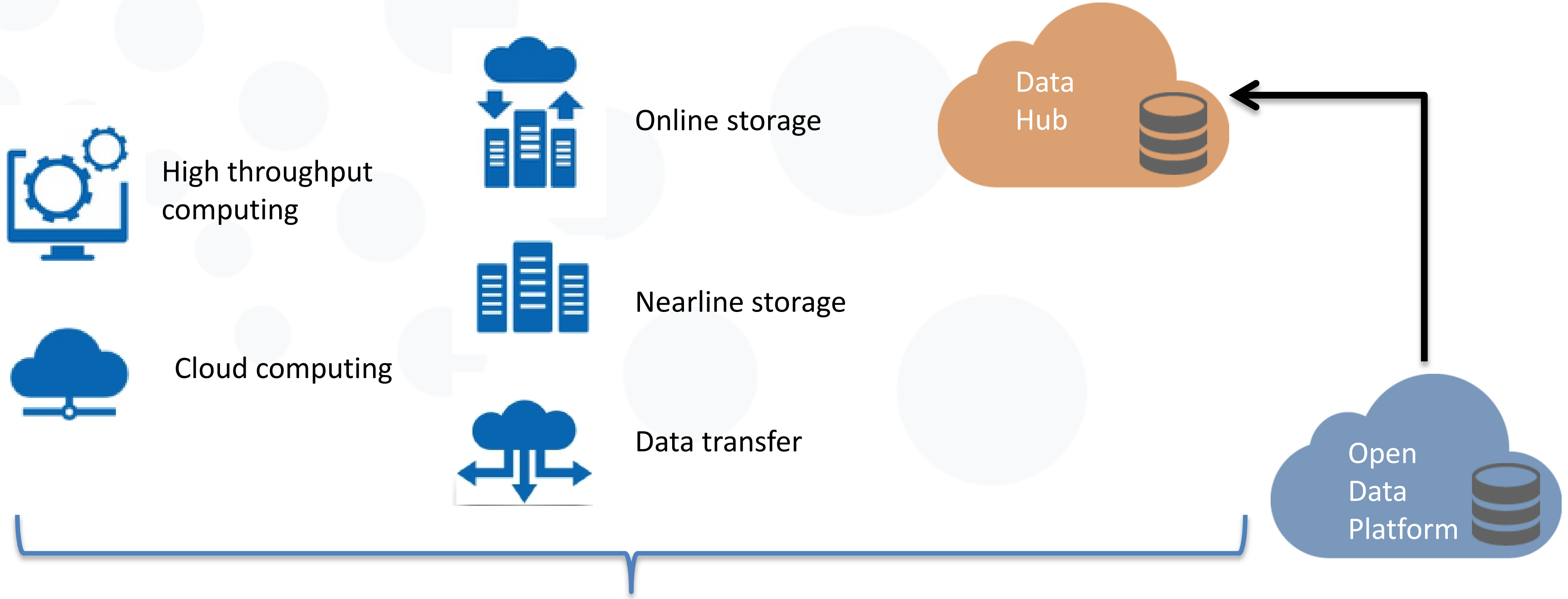
What are EGI’s ODP and Data Hub?

- Open Data Platform (ODP)
  - Set of technologies for data management, integrated with other EGI components, and offered to the users as a complete solution



  - Technology requirements discussed in the new TCB Group, which will include data technology providers representatives: DPM representatives welcome!
- Data Hub
  - One -centrally deployed- instance of (a subset of) the ODP services
  - Open datasets replicated on EGI resources to be accessed and exploited on other EGI services

EGI portfolio



The components of these services are provided by the resource centres and centrally

ODP- Community requirements collection

REQ1. Publication of open research data based on policies

REQ2. Make large data sets available without downloading them completely

REQ3. Enabling complex metadata queries

REQ4. Integration of the open data access data management with community portals

REQ5. Data identification, linking and citation

REQ6. Enabling sharing of data between researchers under certain conditions

REQ7. Sharing and accessing data across federations

REQ9. Data provenance

EGI-Engage

ODP – Requirements for the technologies

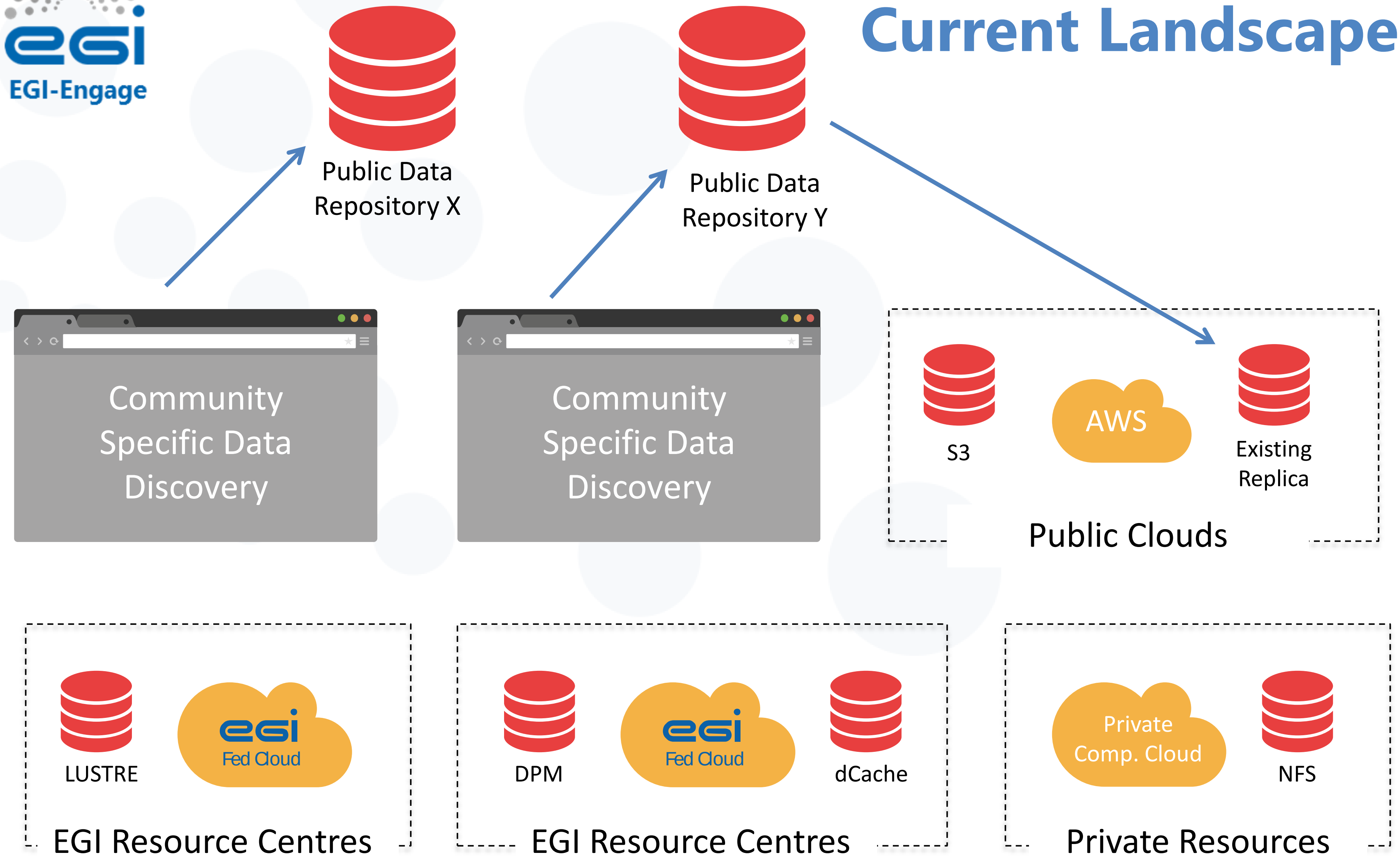
- **simplify access** and processing of data for EGI users
- **integrate and virtualize** existing EGI data storage solutions
- **optimize access** to open data provided by both external and internal EGI open data providers
- provide means for **tracking** open data usage **statistics**
- **data as a service** solution
- **persistent data identification (DOI)**
- **lower barrier** for EGI users in publishing their data

Not all these requirements are a must-have for every single technology solution, but can be achievable through a combination of multiple technologies

EGI DataHub Role

- Unified access to *reference* scientific data of public interest.
- Host *experimental-long term* or *temporary* scientific data and enable easy access to it by appropriate scientific applications.
- Distributed platform for managing replicas of publicly available data collection available on EGI Infrastructure

Current Landscape



Public Data Repository X

Public Data Repository Y

Community Specific Data Discovery

Community Specific Data Discovery

S3

AWS

Existing Replica

Public Clouds

LUSTRE

Fed Cloud

EGI Resource Centres

DPM

Fed Cloud

dCache

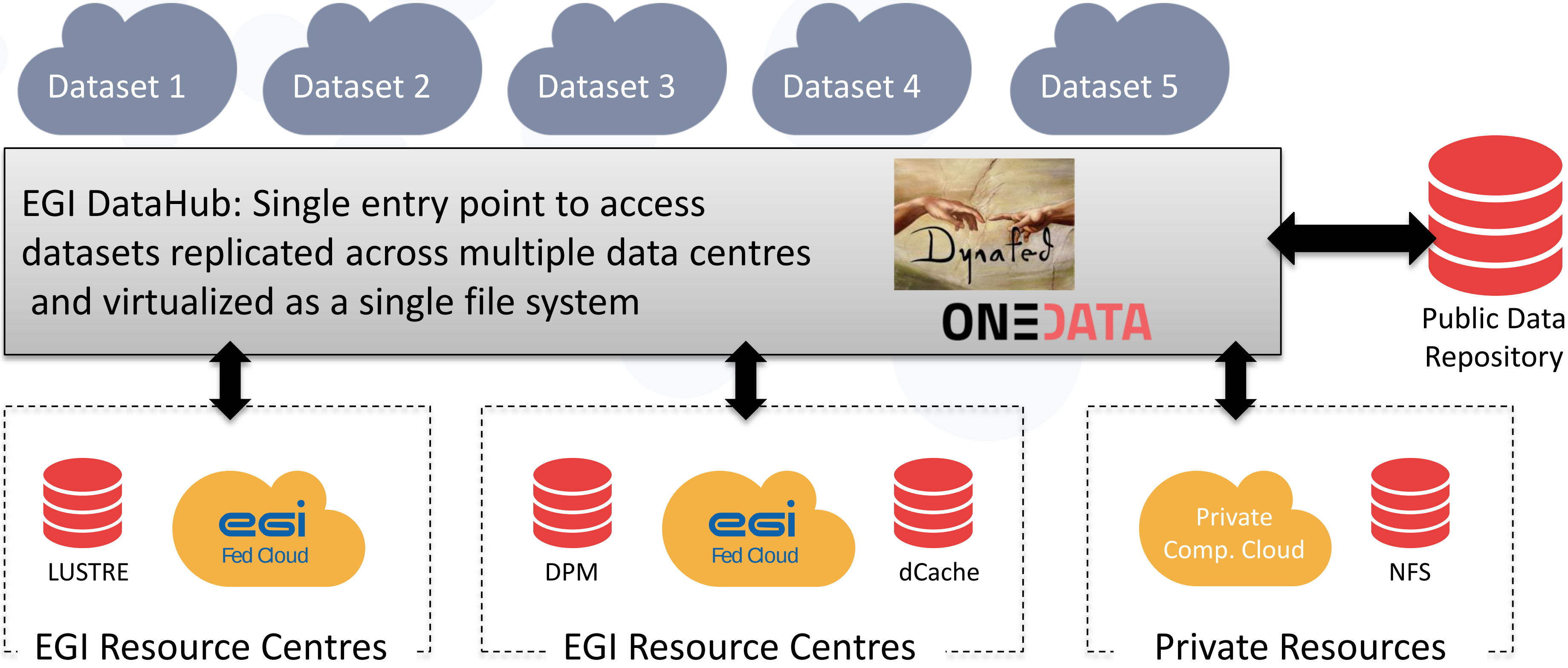
EGI Resource Centres

Private Comp. Cloud

NFS

Private Resources

New landscape with DataHub



EGI DataHub: Single entry point to access datasets replicated across multiple data centres and virtualized as a single file system

DataHub Processes / Use Cases

- EGI.eu collects interest for data collections that should be available in DataHub
- EGI.eu finds Resource Centres willing to support replica of the collections to improve access performances
- EGI.eu and Data Providers negotiate technical means of replication data collection to the Resource Centers
- Data Providers inform users about EGI replicas
- Data Collections discoverable in AppDB
- Users “link” collections of their interest in personal data hub space trough Data Providers JS widget or AppDB
- Users can access data collections via POSIX virtual file system or HTTP requests on all EGI resources
- Users can share their private resources in the same way as public collections

The use cases
(the data)

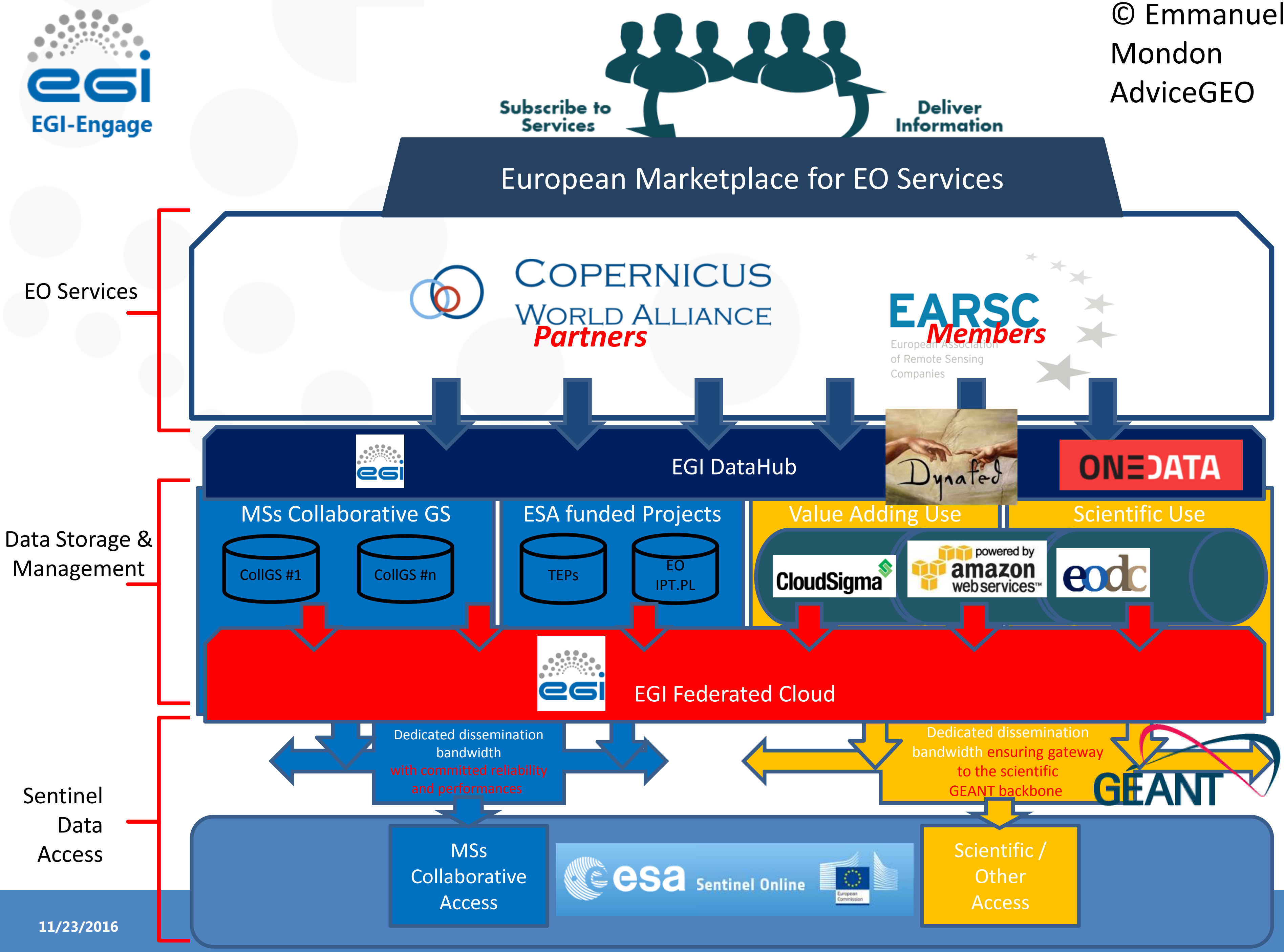
GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.
CHRIS

Copernicus: Sentinel Data

- Earth observation data produced by the Sentinel satellites
- Huge datasets, 10GB per file, EGI could replicate a subset of these datasets to support the exploitation in EGI computing resources. Order of: 10s of TBs
- Datasets are open and free to access
  - https://scihub.copernicus.eu/

EGI-Engage

© Emmanuel Mondon AdviceGEO



Subscribe to Services

Deliver Information

European Marketplace for EO Services

EO Services

COPERNICUS WORLD ALLIANCE *Partners*

EARSC *Members*

European Association of Remote Sensing Companies

EGI DataHub

Dynafed

ONEDATA

Data Storage & Management

MSs Collaborative GS

CollGS #1

CollGS #n

ESA funded Projects

TEPs

EO IPT.PL

Value Adding Use

Scientific Use

CloudSigma

powered by amazon webservices

eodc

EGI Federated Cloud

Sentinel Data Access

Dedicated dissemination bandwidth with committed reliability and performances

Dedicated dissemination bandwidth ensuring gateway to the scientific GEANT backbone

GEANT

MSs Collaborative Access

esa Sentinel Online

Scientific / Other Access

Copernicus: Sentinel Data

Exploitation

- Data from Earth Observation correlated with other scientific datasets for providing information directly integrated into business workflows

- Through the EGI Data Hub provide integrated access to computing and storage resources needed for data access and exploitation and provide the tools to manage the datasets in a distributed environment

- European-wide long-term storage for Copernicus data ensuring its high-availability for Researchers and for serving SMEs / Industries operating the DIAS platforms and marketplace guaranteeing their long-term availability and usability and realising economy of scale.
  - To be demonstrate this in a hybrid cloud pilot during 2017.

- Timeline: proof of concept/pilot during 2017, production 2018 onward

iMarine: fishery and aquaculture data

The Data

- Big number of diverse datasets exploitable for marine ecosystem, fishery and aquaculture use cases
  - Diverse licensing restrictions
- 50+ data providers
- 20,000+ temporal datasets
- 50,000+ spatial datasets

Exploitation

- Potentially 2700 users worldwide
- 50+ VRE for the data analysis
- iMarine use cases are already running on EGI resources, but with data accessed remotely
- Timeline: pilots during 2017, exploitation 2018 onward

Lifewatch: Algae bloom datasets

The Data

- Data from CdP Reservoir. Raw – Curated – Processed/Derived
- Real Time monitoring ~5GB. Model Data ~20GB for each 3D model.
- Metadata standards employed: EML
- Available in Pilot tests: Storage, transfer for processing, AAI integration

Exploitation

- It will be used in the LifeWatch Environment. Lifewatch community will use EGI FedCloud
- Timeline: 2017 onward

Other use cases

- ENES
  - Climate Model Intercomparison Data Analysis case study
  - Large input datasets O(1PB), large outputs as well
  - Exploitation using big data analysis tools on the cloud
- ELIXIR
  - Reference genomics datasets
- EPOS and ICOS
  - Seismology and Atmosphere data
- HBP
  - Brain atlas visualization

EGI-Engage

How can DPM, DPM sites, and DynaFed be part of the federated data solutions offered to the EGI users?

DynaFed: HTTP access to data

- For many use cases the DynaFed option ticks all the boxes
- The WLCG demonstrator is having good results with already demanding use cases
- Near-to zero effort required for the sites, leveraging on the existing Storage Elements
- A non-WLCG demonstrator could have equally good results for non-HEP use cases
- Enable access to local computing resources (on the same site) but also to close computing resources

For other specific requirements: OneData federation

- OneData is a software stack for distributed data management developed externally to EGI
  - Can federated multiple storage service types with different interfaces
  - Flexible access control mechanisms
  - Enables also POSIX access (on top of other interfaces)
  - Integrates metadata support and data discovery using PUID
- Still a pre-production solution in EGI, plan to have a production ready deployment for Q1 2017
- It would be interesting to test access to DPM SEs, perhaps through the WebDAV interface

Conclusions

- EGI Open Data Platform builds on top of the existing mature storage services and technologies provided by the sites
- General-purpose federation layer is necessary to support new (open) data use cases
- DPM services are an important subset of the storage resources available in EGI and federating them would have a big impact in the EGI ODP offer

Contacts

- Me: peter.solagna@egi.eu
- ODP and TCB on data management primary contact: Matthew Viljonen matthew.viljoen@egi.eu

- Next steps:
  - how to proceed?
  - Where is the best occasion to discuss an EGI demonstrator?

Thank you for your attention.

Questions?

www.egi.eu