

 www.egi.eu

 @EGI_eInfra

EGI Data Management requirements and plans

3rd Rucio Workshop

Andrea Manzi

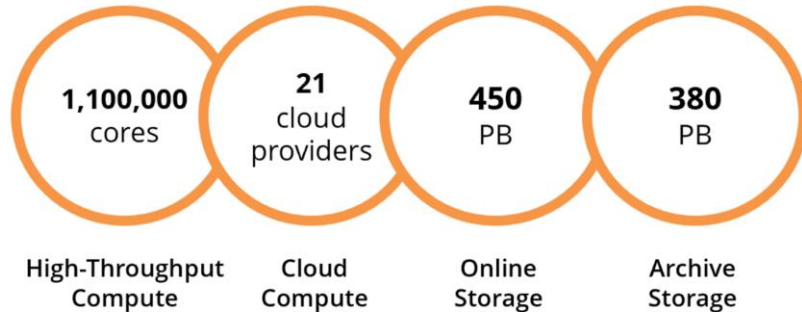
EGI Foundation



The work of the EGI Foundation
is partly funded by the European Commission
under H2020 Framework Programme

- EGI intro
- EGI Data Management requirements
 - Data management workshop
- EGI & Rucio
 - Evaluation & Requirements
- Future plans

EGI is a federation of computing and storage resource providers united by a mission to support research and development.



As of September 2019

Compute

- Cloud Compute**
Run virtual machines on demand with complete control over computing resources
- Cloud Container Compute**
Run Docker containers in a lightweight virtualised environment
- High-Throughput Compute**
Execute thousands of computational tasks to analyse large datasets
- Workload Manager**
Manage computing workloads in an efficient way

Storage and Data

- Online Storage**
Store, share and access your files and their metadata on a global scale
- Archive Storage**
Back-up your data for the long term and future use in a secure environment
- Data Transfer**
Transfer large sets of data from one place to another

Security

- Check-in**
Login with your own credentials

Training

- FitSM Training**
Learn how to manage IT services with a pragmatic and lightweight standard

Applications

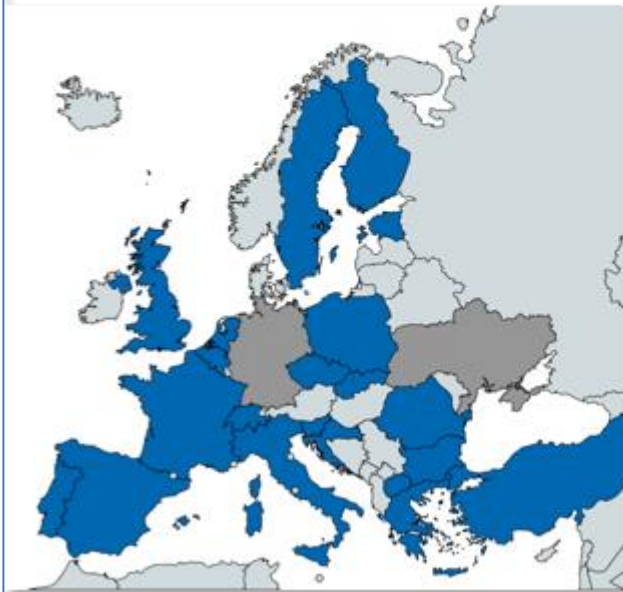
- Applications on Demand**
Use online applications for your data & compute intensive research
- Notebooks**
Create interactive documents with live code, visualisations and text

- ISO 27001 Training**
Learn how to manage and secure information assets

- Training Infrastructure**
Dedicated computing and storage for training and education

23 Countries and CERN

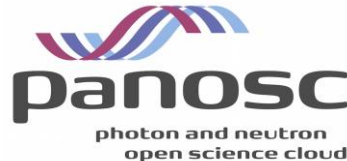
<https://www.egi.eu/about/egi-foundation/>







- Mobilise providers from the EGI Federation, EUDAT CDI, INDIGO-DataCloud and other major European research infrastructures to deliver a common catalogue of research data, services and software for research.



- <https://www.egi.eu/about/projects/>

EGI Service Portfolio: Data Management Services

- EGI is primarily a **compute** federation
- Data management to support the computing
 - Staging data, storage next to CPUs/GPUs
 - Archival is shifting out, primarily to community sites
- **Online** storage is
 - Grid storage for HTC Compute
 - Block and object storage for Cloud compute
- Data Transfer is **FTS**
 - OLA with CERN and STFC/RAL
- More data services in Alpha/Beta phase
 - **DataHub** based on OneData
 - Content distribution based on **CVMFS**

Data Management requirements

July's 2019 Workshop on Data Management

- Three entities **representing** a broad range of **solutions** and **services providers**



- Multiple **user communities**

- PaNOSC
- ASTRON
- Earth Observation JRC



- Goal: *proof-of-concept and pilot use cases to be implemented by XDC, ESCAPE and EGI in collaboration with the invited scientific communities*

<https://indico.egi.eu/indico/event/4698/>

- Research Infrastructures **generating** raw **data** at one or **few locations** (antennas, colliders, lasers,...)
 - Making data available for broader access through archives
 - Big storage and network capacities at one or more locations
 - Raw data transferred to these locations after filtering and calibration at the data source
 - Storage sites responsible for data archival, curation and sharing
- They need to
 - Manage transfers from acquisition locations to storage locations
 - Manage archiving and replication
 - Pre-stage data to computing sites
- Relevant communities and RIs
 - **EISCAT3D, ICOS-eLTER**

- RIs establishing a **federation** of **sites** providing **storage** and **computing**
 - Sites with computing resources for large-scale data analysis and analytics
 - Centrally provided and curated datasets and applications
- They need a Data **replication service** to copy community datasets **to sites**
 - Possible to use national providers
 - Selecting dataset and application
 - Using custom data
 - Performing data analysis/analytics
- Relevant communities and RIs
 - **ELIXIR, EPOS-ORFEUS, Earth Observation**

- **Remote archive** of a facility **experimental data**
 - perform analysis on data sets obtained during an experiment
 - **initial** transfer around **600TB**, **annual** volume from **300TB up to 10PB**
 - possibility to **retrieve** data **on-demand**

- **Remote Data analyses** using **Jupyter** notebook services of another provider
 - **compute** and **data** infrastructure are **distant**
 - a service to **host the results**

- User's data transfer to its home organization/home computer
 - The **volume** of data (up to 100TB) is **important**

- **EGI Rucio Testbed** under deployment
 - Based on Kubernetes
 - thanks to Rucio Helm charts
 - gaining experience on K8s deployments
 - PostgreSQL DB
 - Understanding and testing the various daemons
 - To be interfaced to CERN or STFC/RAL FTSS

- Access to **Rucio multi-VO instance @STFC/RAL**
 - Under implementation

- Selection of interested user communities to implement **pilots**



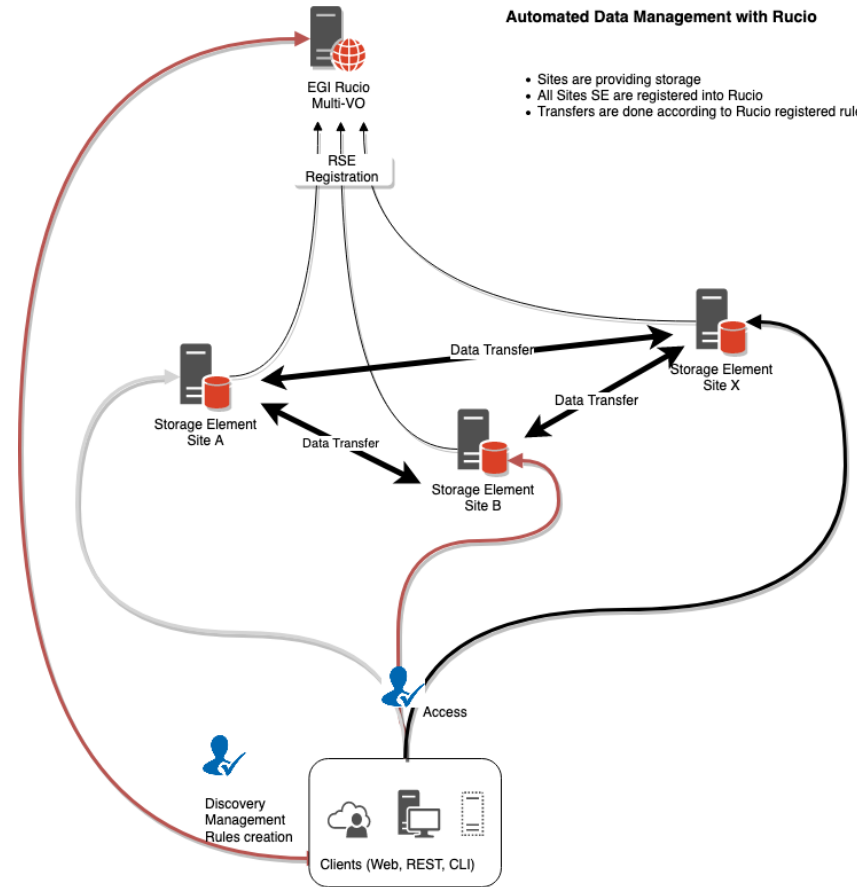
- **Multi-VO** support
- Integration with EGI AAI (**EGI Checkin**)
 - Follow up the activities for OIDC support (both in Rucio & FTS)
 - Token translation via Online CA (RAuth) for X509 access to storages?
- Integration with EGI Workflow Manager (**DIRAC**)
 - Usage of Rucio catalog in Dirac presented by Cedric in the context of Belle II
- Possibility to interface with **Globus** (already used by some communities, i.e. PaNOSC)
 - Ongoing development by BNL

- **Data Catalogue** and flexible metadata interface
- Access via Jupyter **Notebooks**, for instance via **Fuse**?
 - Presentation on a Fuse interface for RUCIO tomorrow
- **OneData** Integration (EGI Data-Hub)
 - This is more on the FTS side
- **PID** support
 - Via third-party services/tools
- Integration with **EGI ARGO** monitoring and **APEL** accounting
 - Plenty of Nagios probes for RUCIO available already

- **Rucio** as a one of the Data Management services for EGI
 - Part of the EGI Service Portfolio
 - Following well defined IT SM Procedures
- **Catch-all instance (Multi-VO)**
 - Operated by one of the member of the Federation
 - Possibly funded via the next EGI flagship EU project in 2021, together with some of the missing integration activities

Automated Data Management with Rucio

- Sites are providing storage
- All Sites SE are registered into Rucio
- Transfers are done according to Rucio registered rules





www.egi.eu



@EGI_eInfra

Thank you
for your attention.

Questions?



This work by the EGI Foundation
is licensed under a Creative Commons
Attribution 4.0 International License.



The work of the EGI Foundation
is partly funded by the European Commission
under H2020 Framework Programme