www.egi.eu
@EGI_eInfra

EGI Data Management requirements

Feedback from the EGI communities

Baptiste Grenier
EGI Foundation

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

Overview

- The EGI Federation
- Data Management-related EGI use-cases
  - Application and data distribution platform
  - Application and data replication
  - User workflows over distributed data and compute
- Rucio to the rescue?

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

The EGI Federation

A European e-infrastructure

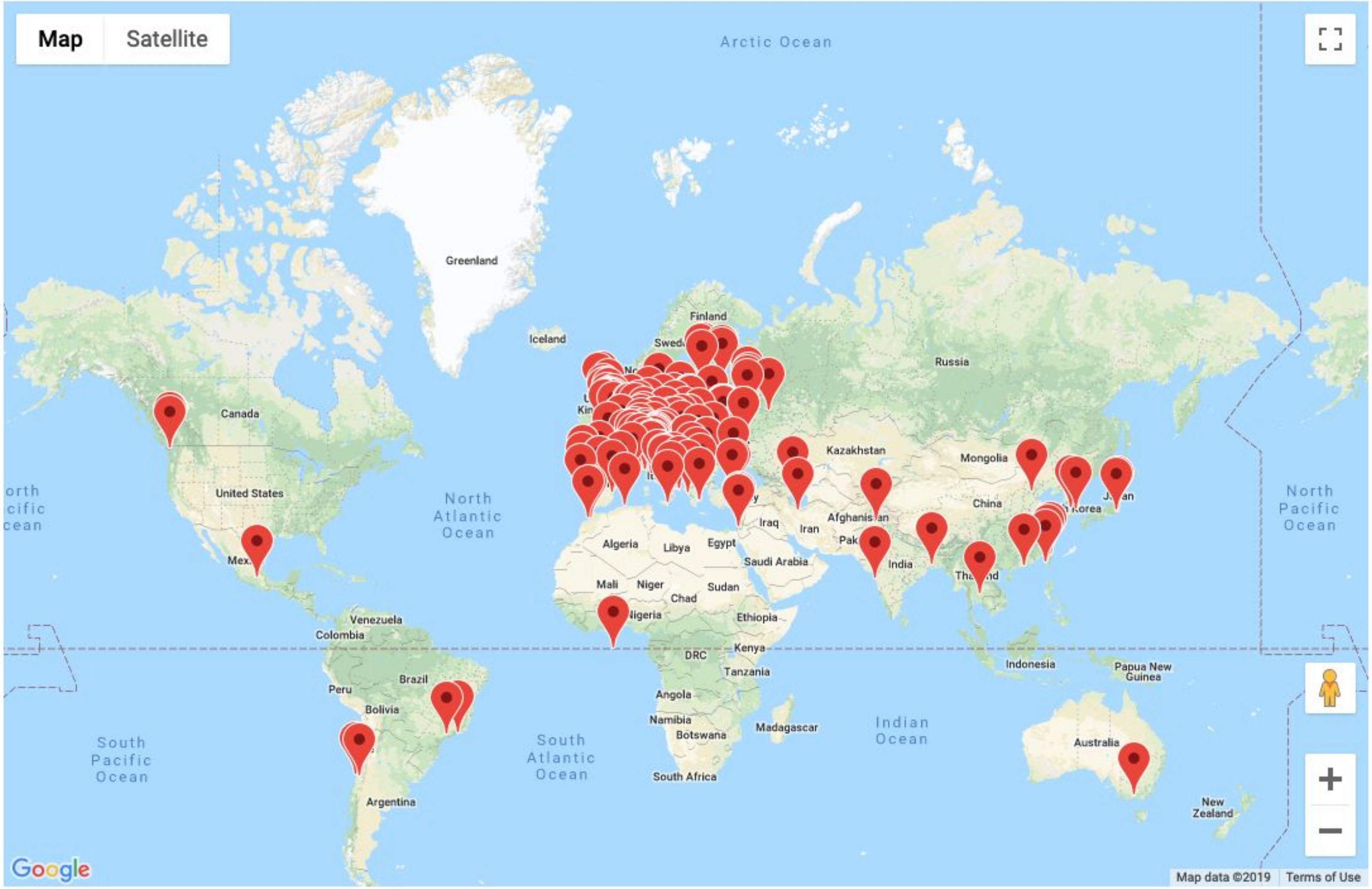
- **23** Countries
- **1** EIRO: **CERN**

www.egi.eu/about/egi-foundation/

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

EGI Federation: sites information

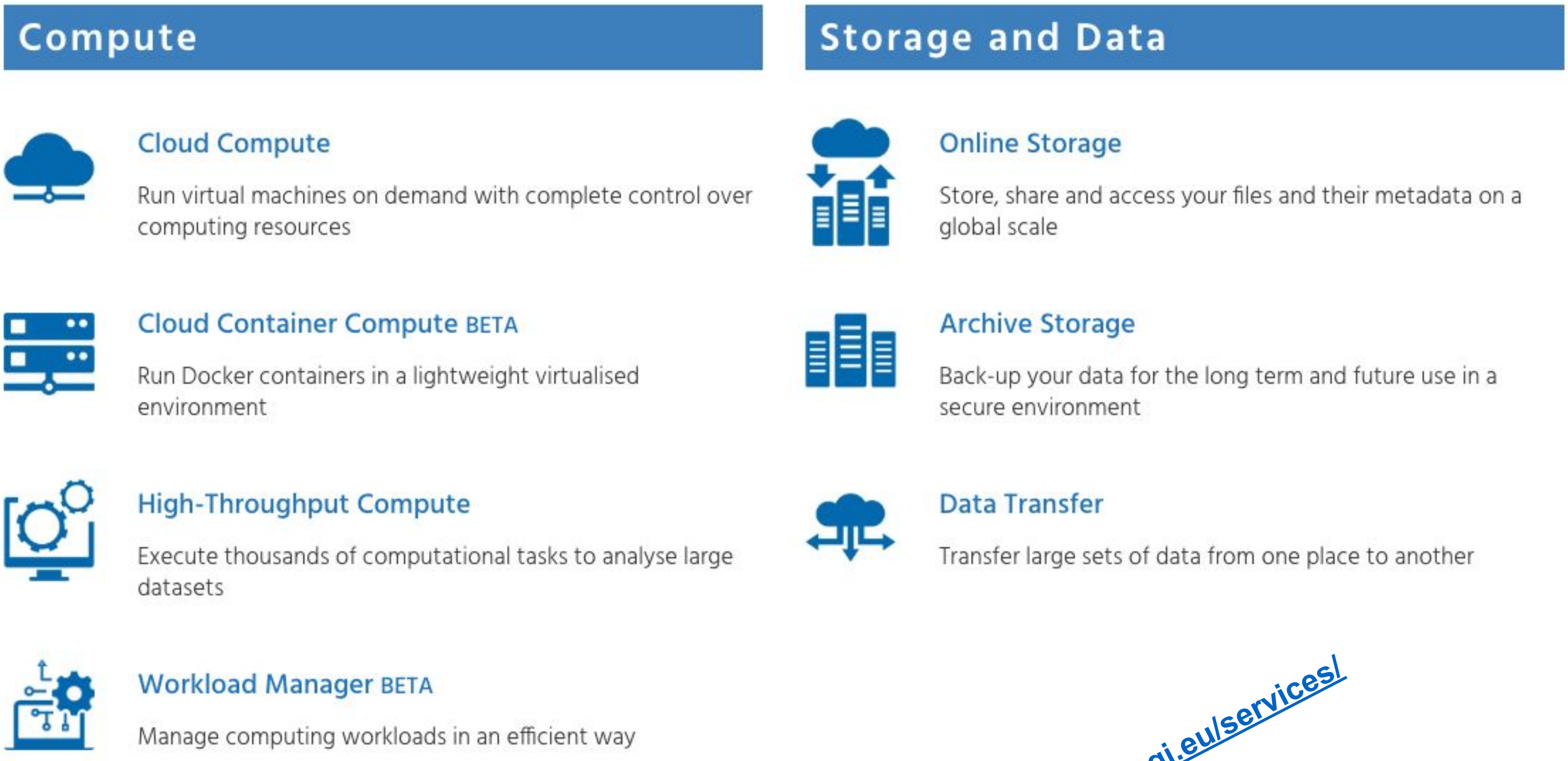
https://operations-portal.egi.eu/vapor/resources/GL2Map



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

About the EGI Federation

EGI Service Portfolio



Compute

Cloud Compute

Run virtual machines on demand with complete control over computing resources

Cloud Container Compute BETA

Run Docker containers in a lightweight virtualised environment

High-Throughput Compute

Execute thousands of computational tasks to analyse large datasets

Workload Manager BETA

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

https://www.egi.eu/services/

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

About the EGI Federation

EGI Service Portfolio: focus on data 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, to EUDAT,...

- Online storage is
  - Grid storage for HTC Compute
  - Block and object storage for Cloud compute
- Data Transfer is FTS

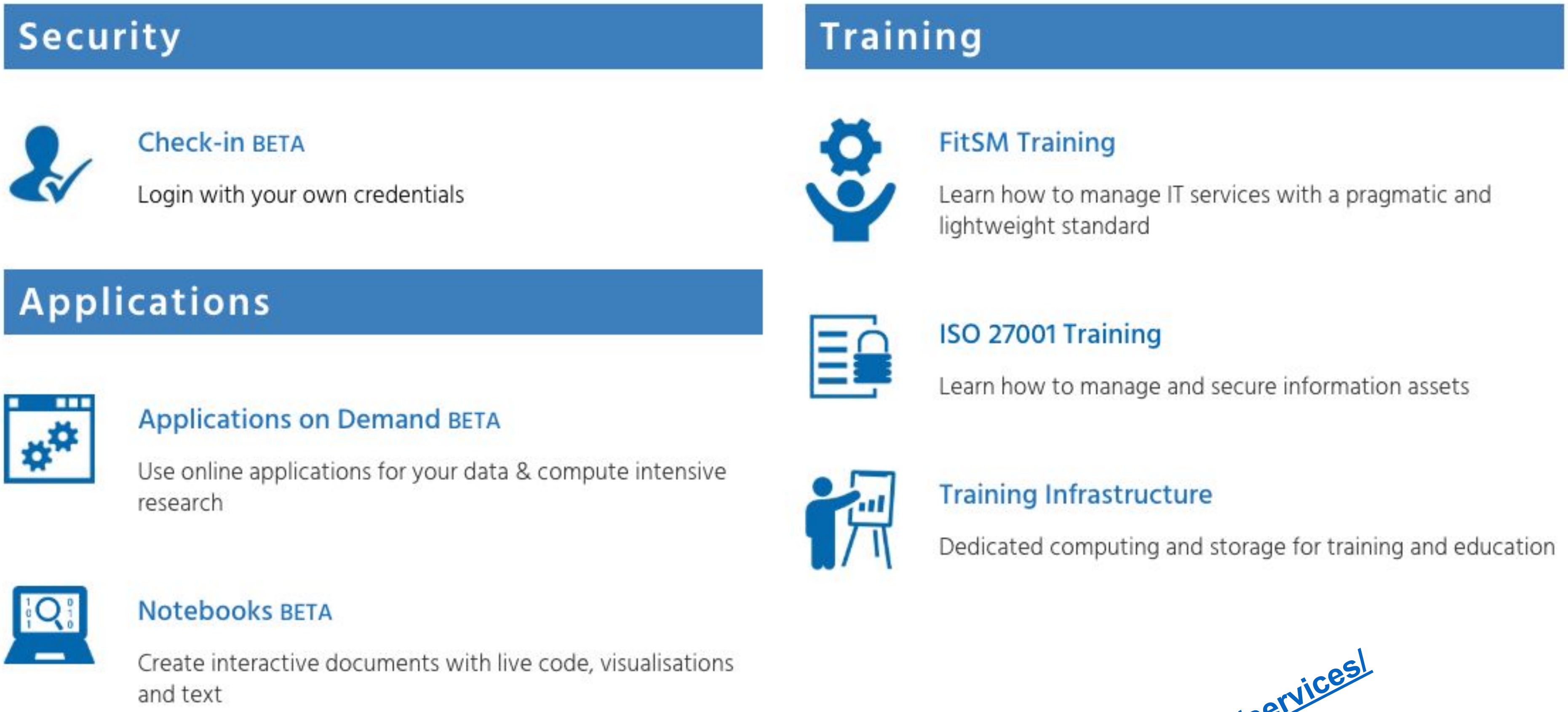
- More data services in the pipeline
  - DataHub based on OneData
  - Content distribution based on CVMFS

- Open to introduce additional services or implementations based on new technologies - such as RUCIO

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

About the EGI Federation

EGI Service Portfolio



Security

Check-in BETA

Login with your own credentials

Applications

Applications on Demand BETA

Use online applications for your data & compute intensive research

Notebooks BETA

Create interactive documents with live code, visualisations and text

Training

FitSM Training

Learn how to manage IT services with a pragmatic and lightweight standard

ISO 27001 Training

Learn how to manage and secure information assets

Training Infrastructure

Dedicated computing and storage for training and education

https://www.egi.eu/services/

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

About the EGI Federation

Internal Services for the EGI Federation

- Security
  - Check-in (COmanage, RCauth)
  - Attribute Management
- Operations
  - Marketplace
  - Accounting (APEL)
  - Collaboration Tools
  - Configuration Database (GOCDB)
  - Operational Tools (Operations Portal, Vapor)
  - Helpdesk (GGUS)
  - Service Monitoring (ARGO)
  - Validated Software and Repository (UMD, CMD)
- Coordination
  - Operations Coordination and Support
  - Community Coordination
  - ITSM Coordination
  - Technical Coordination
  - Strategy and Policy development
  - Project Management and Planning
  - Security Coordination
  - Communications

https://www.egi.eu/internal-services/

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

The EGI Federation

2018 figures

- 4.4 Billion CPU core wall time delivered in 2018
  - 1 million computing cores
  - 356 PB disk & 380 PB tape storage
- 1170 open access publications
- +41 new international projects
- 31 large scale ESFRI projects/landmarks supported

- EGI Conference - 2019/05/06-08 - Amsterdam Science Park
  - https://indico.egi.eu/indico/event/4431/overview
  -

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

About the EGI Federation

Collaborating globally and supporting the EOSC

- Together with **EUDAT** and **INDIGO** DataCloud, **EGI Foundation** is at the core of the **EOSC-hub** project
- Liaising with services providers to integrate their services
- Supporting user communities / RIs to leverage available services
- Liaising with other projects
- Liaising with other e-infrastructures

Africa and Arabia:
Council for Scientific and Industrial Research, South Africa

Latin America:
Universida de Federal do Rio de Janeiro

India:
Centre for Development of Advanced Computing

China:
Institute of HEP,
Chinese Academy of Sciences

USA

Canada

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

About the European Open Science Cloud

- The **EOSC** is a vision set up by the European Commission to give Europe a global lead in scientific data infrastructures

- EOSC will offer a virtual environment with open and seamless services for storage, management, analysis and re-use of research data

- The entry point for the EOSC is the **EOSC Portal**: http://www.eosc-portal.eu

- EOSC services are available through the **EOSC Marketplace**: https://marketplace.eosc-portal.eu/

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

The EGI Foundation

Supporting the coordination of the Federation

- EGI Foundation
  - A **non profit Foundation** established in Amsterdam since **2010**
  - **A legal entity** to **represent** the EGI Federation
    - In EU-funded projects, international bodies, collaborations
  - Supporting the **coordination** of the EGI **Federation**
  - Steered by the **EGI Council** including all the EGI Federation participants
  - Day to day running **supervised** by the **Executive Board**
    - **members** of the EGI **Council** appointed for two-years terms
  - Managed following **FitSM** practices, certified **ISO 9001** and **ISO 20000**
  - And even offering you a nice job opportunity: https://www.egi.eu/about/jobs/

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

A few EGI Use Cases

Related to data management

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

Application and data distribution platform

- 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 transfered to these locations after filtering and calibration at the data source
  - Storage sites responsible for data archival, curation and sharing

- **Centrally** provided **applications platforms** can make "reference applications" easily accessible

- Rucio could be used for
  - managing transfers from acquisition locations to storage locations
  - managing archiving and replication
  - pre-staging data to computing sites

- Relevant communities and RIs
  - EISCAT_3D, ICOS-eLTER,...

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

Application and data replication

- 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

- Data **replication service** to copy community datasets **to sites**
  - Possible to use national providers
    - Maximise the usage of national fundings and lower access cost
    - Using community/national AAI systems
  - Selecting dataset and application
  - Using custom data
  - Performing data analysis/analytics

- Application replication service to deploy community-specific or reference applications to sites

- Rucio could be used for
  - Data replication service

- Relevant communities and RIs
  - ELIXIR, EPOS-ORFEUS, Earth Observation,...

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

User workflows over distributed data and compute

- **Long running communities** producing **data** and **applications** at **several locations** with no or **light coordination**
- A rich set of data and applications accumulated
  - In diverse formats and with different metadata descriptions

- Challenge
  - Making all the data and applications accessible and combinable across the community respecting local policies, restrictions and limitations of provider sites
- Need for an agreed data and metadata format, and agreed APIs and some central catalogue where files could be discovered

- Rucio could help to
  - catalogue multiple storage endpoints and datasets
  - expose some metadata
  - access files using client or APIs

- Relevant communities and RIs
  - Fusion, Disaster Mitigation, Radio astronomy,...

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

EGI needs VS Rucio

From a Rucio boeotian

Requirements	Status	Comments
AAI integration (OIDC and X509 through Check-in)	?	Initial (?) OIDC support. TTS/RCauth?
Replica and transfer management over an heterogeneous and distributed infrastructure	✓	Of course!
Discovery of data with a catalogue	✓	Possibility to have a “public” catalogue?
Metadata management	?	Is actual support flexible enough?
Local access to data	✓	Rucio client. (What about AAI?)
Integration with other tools (APIs)	✓	API and clients. (What about AAI?)
Combining multiple datasets from different providers	?	
Interaction with DIRAC	?	https://github.com/rucio/rucio/issues/1808 ?
Integration with EGI Notebooks	✓	API and clients. (What about AAI?)
Onedata integration	☒	OIDC and REST API?
PID support	?	Using third party solution?

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

How to go forward?

- Evaluation: is a production-grade test instance provided by Rucio team?

- Adoption: Integration with EGI infrastructure
  - Check-in and RCauth (AAI)
  - Integration with other relevant EGI Services (DIRAC, Notebooks,…)
  - Procedure for production service in EGI
    - Creating a Service Design and Transition Package (SDTP)
      - ➢ PROC19, Accounting (APEL), Monitoring (ARGO), Support (GGUS), UMD,..
  - EGI Strategic and Innovation Fund (SIF) to support parts of the effort?

§

- How EGI could propose the service
  - Supporting user communities in expressing their needs
  - Assessing applicability of available solutions
  - Putting in place pilot test and providing support
  - Linking with application developers
  - Catch-all and Community-specific instances

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

Thank you
for your attention.

Questions?

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

www.egi.eu
@EGI_eInfra