www.egi.eu
@EGI_eInfra

EGI DataHub

Data as a Service – Distributed Data Management

Baptiste Grenier

EGI Foundation

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

Motivation

- Putting up a (scalable) distributed data infrastructure needs specific expertise, resources and knowledge
- No easy way to discover and transfer data
- No easy way of making data (publicly) accessible without transferring it to a sharing service
- No easy way of combining multiple datasets from different data providers
- Users need to access data locally and from compute resources

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

EGI DataHub: components and concepts

- **EGI DataHub**: a **Onedata Onezone**, the **federation** and **authentication** service. SSO with all the connected storage providers (Oneprovider) through **EGI Check-in**

- **Oneprovider**: **data management** component deployed in the data centres, **provisioning data** and **managing transfers**. A default one is operated for EGI by CYFRONET.

- **Space**: a **virtual volume** where **users organize data**. A space is supported by one or multiple Oneproviders

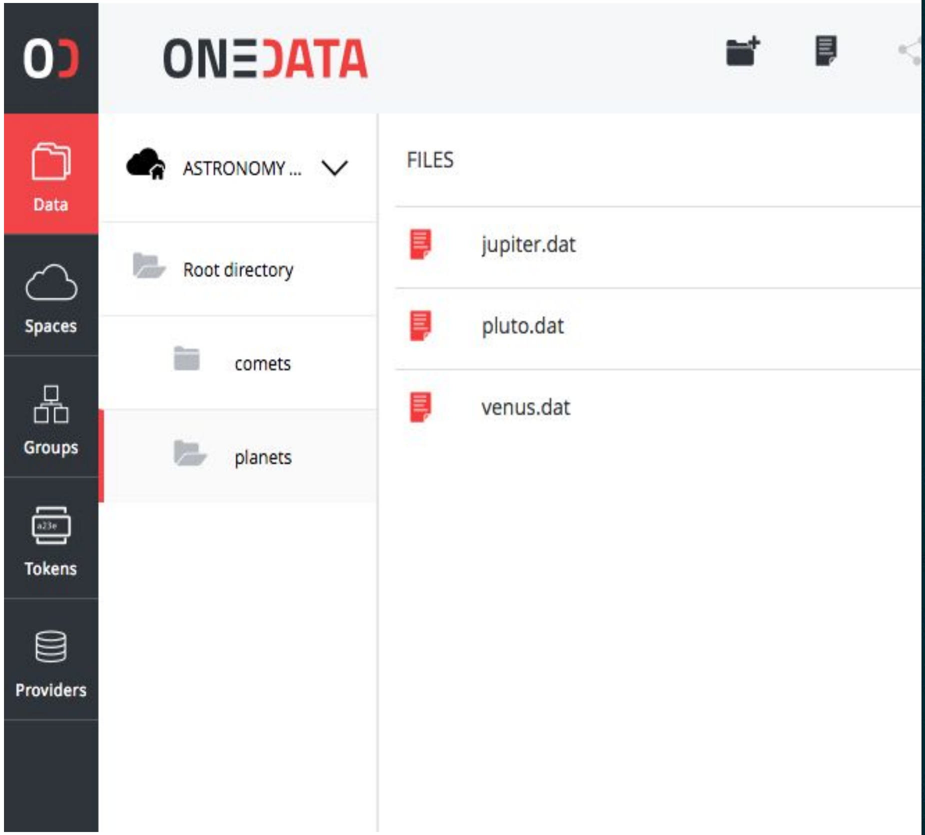
- **Oneclient**: a client providing access to the spaces through a FUSE mount point (**local POSIX access**)

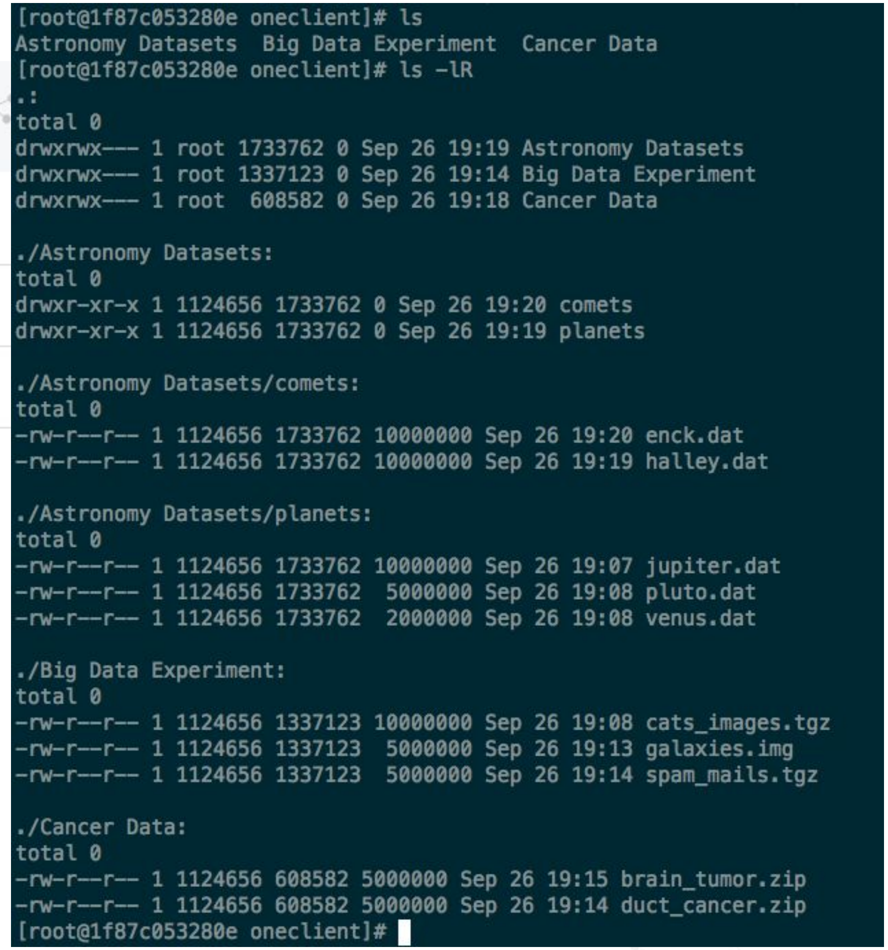
- **Web interfaces** and **APIs** are also available

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

On the client side

Web interface and Oneclient on the CLI



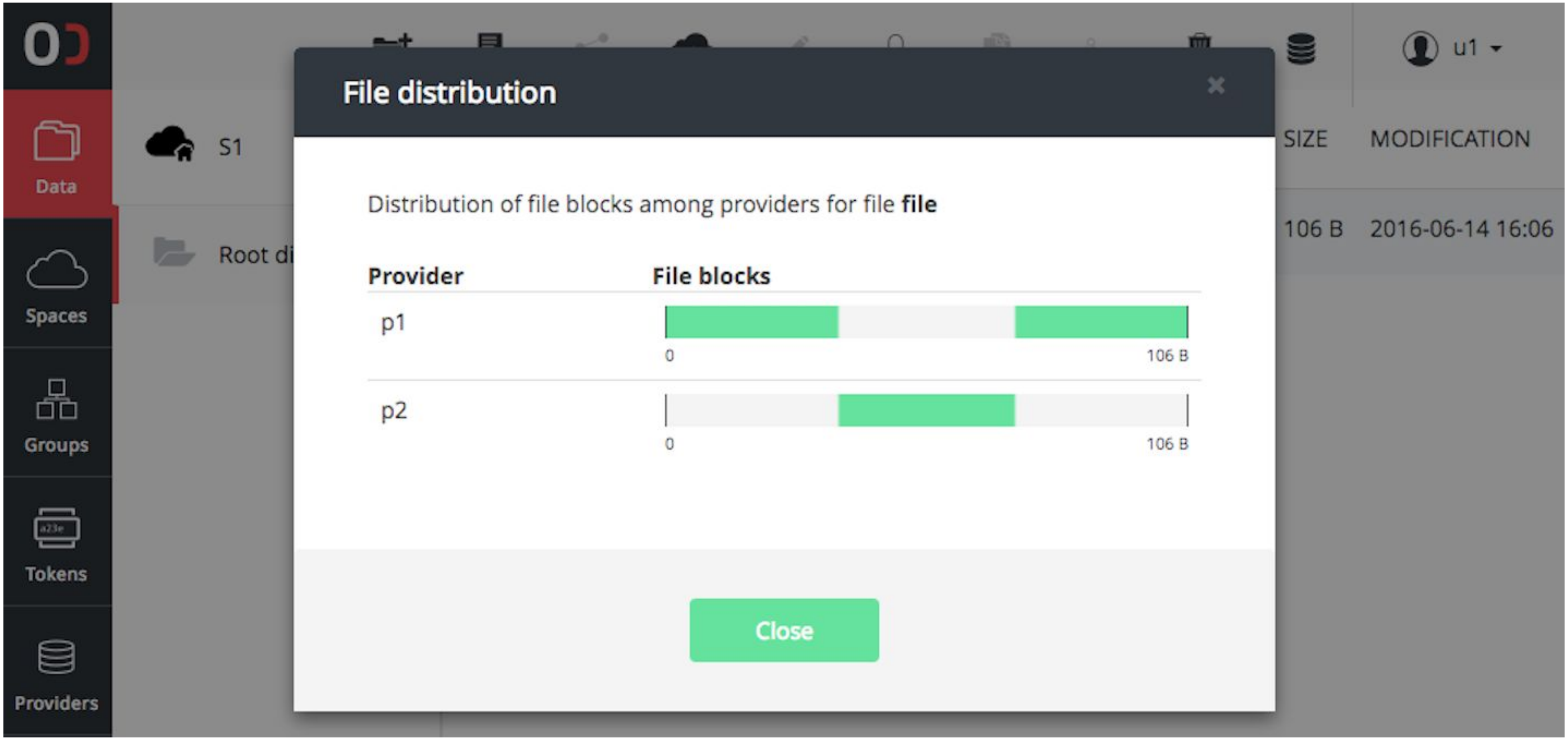


```
[root@1f87c053280e oneclient]# ls
Astronomy Datasets  Big Data Experiment  Cancer Data
[root@1f87c053280e oneclient]# ls -lR
.:
total 0
drwxrwx--- 1 root 1733762 0 Sep 26 19:19 Astronomy Datasets
drwxrwx--- 1 root 1337123 0 Sep 26 19:14 Big Data Experiment
drwxrwx--- 1 root  608582 0 Sep 26 19:18 Cancer Data

./Astronomy Datasets:
total 0
drwxr-xr-x 1 1124656 1733762 0 Sep 26 19:20 comets
drwxr-xr-x 1 1124656 1733762 0 Sep 26 19:19 planets

./Astronomy Datasets/comets:
total 0
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:20 enck.dat
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:19 halley.dat

./Astronomy Datasets/planets:
total 0
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:07 jupiter.dat
-rw-r--r-- 1 1124656 1733762  5000000 Sep 26 19:08 pluto.dat
-rw-r--r-- 1 1124656 1733762  2000000 Sep 26 19:08 venus.dat

./Big Data Experiment:
total 0
-rw-r--r-- 1 1124656 1337123 10000000 Sep 26 19:08 cats_images.tgz
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:13 galaxies.img
-rw-r--r-- 1 1124656 1337123  5000000 Sep 26 19:14 spam_mails.tgz

./Cancer Data:
total 0
-rw-r--r-- 1 1124656 608582 5000000 Sep 26 19:15 brain_tumor.zip
-rw-r--r-- 1 1124656 608582 5000000 Sep 26 19:14 duct_cancer.zip
[root@1f87c053280e oneclient]# 
```

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

Replica management

File distribution across providers



File distribution

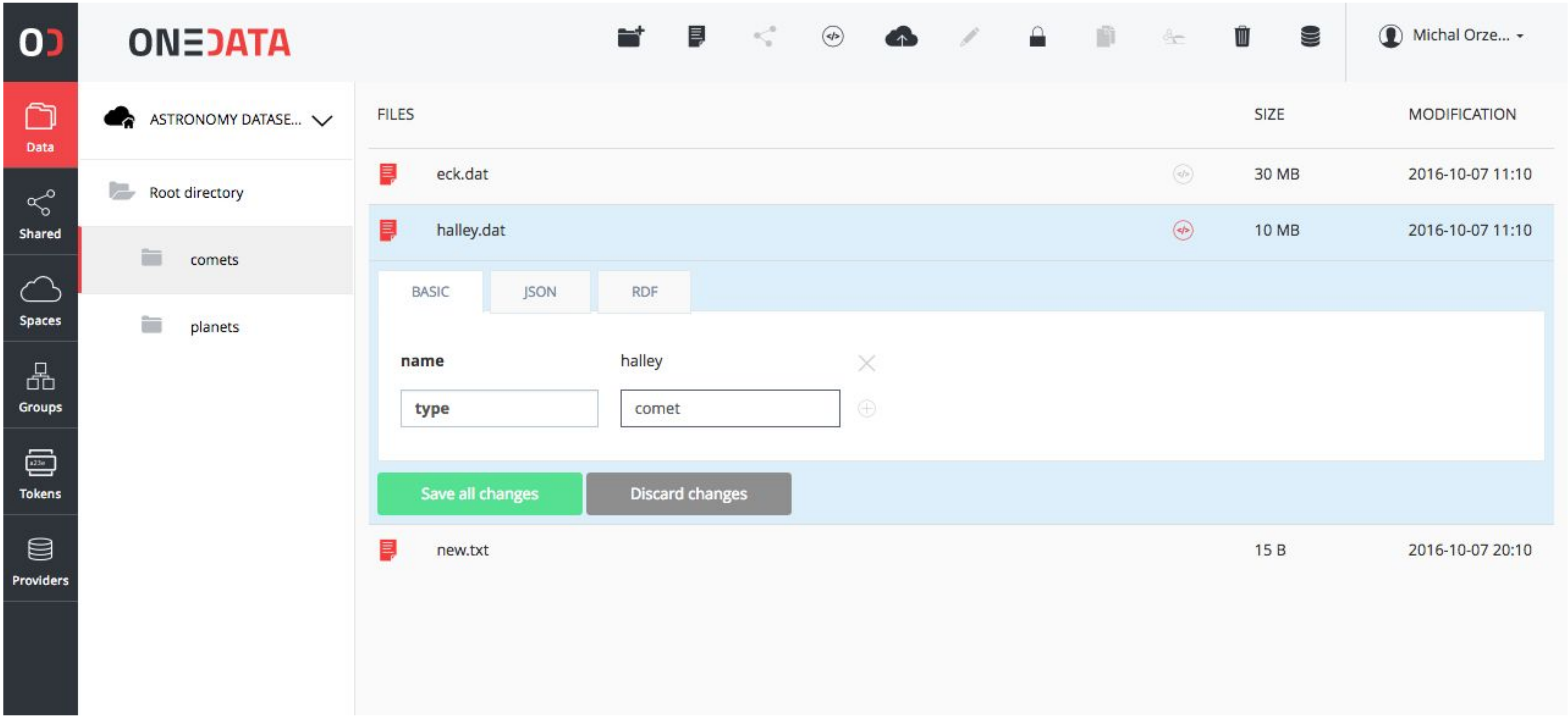
Distribution of file blocks among providers for file **file**

Provider	File blocks
p1	0 … 106 B
p2	0 … 106 B

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

Metadata management

Attaching metadata to files

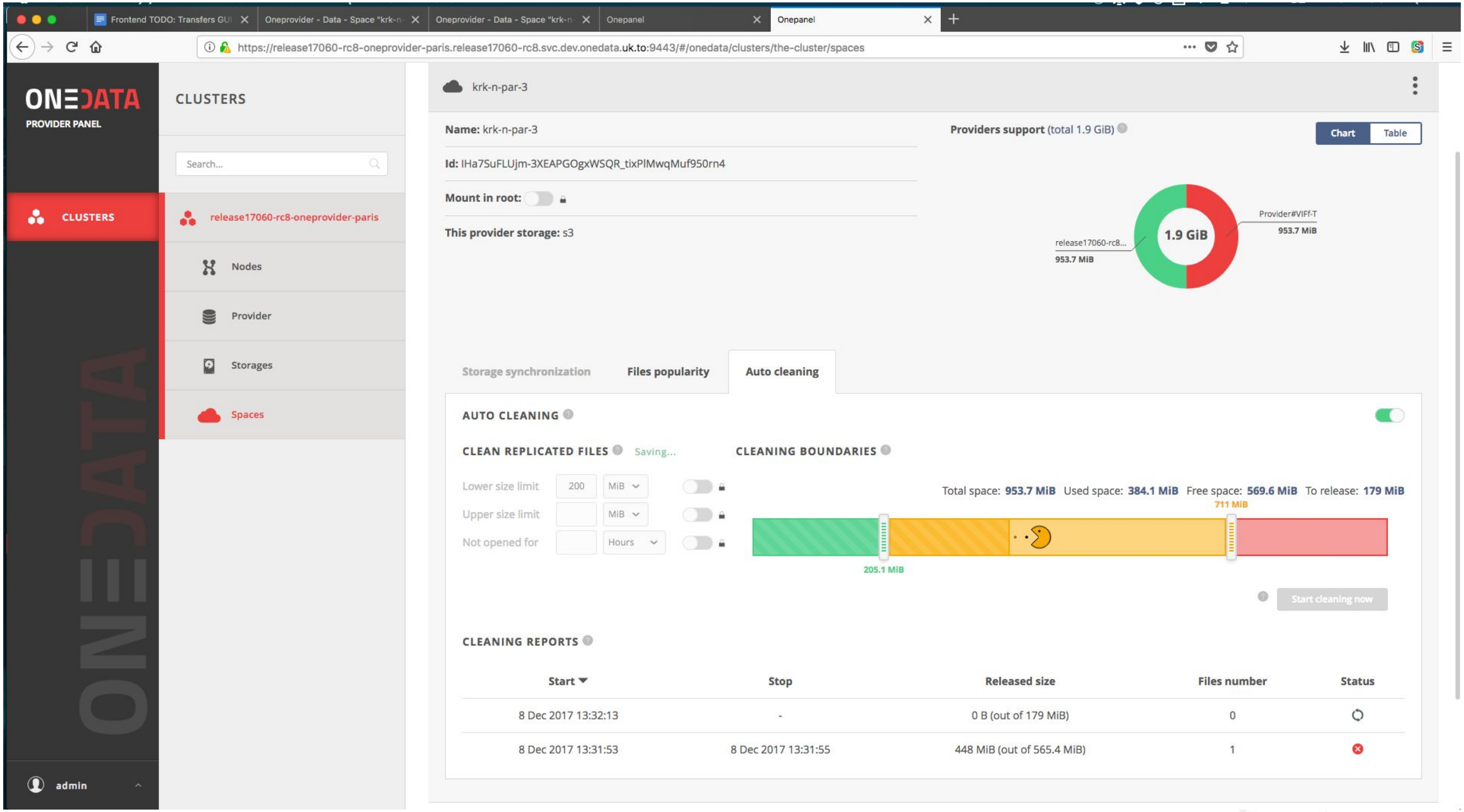


FILES	SIZE	MODIFICATION
eck.dat	30 MB	2016-10-07 11:10
halley.dat	10 MB	2016-10-07 11:10
new.txt	15 B	2016-10-07 20:10

name	halley
type	comet

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

File popularity and smart caching



Name: krk-n-par-3

Id: IHa7SuFLUjm-3XEAPGOgxWSQR_tixPlMwqMuf950rn4

Mount in root:

This provider storage: s3

Providers support (total 1.9 GiB)

AUTO CLEANING

CLEAN REPLICATED FILES

Lower size limit 200 MiB

Upper size limit MiB

Not opened for Hours

CLEANING BOUNDARIES

Total space: **953.7 MiB** Used space: **384.1 MiB** Free space: **569.6 MiB** To release: **179 MiB**

CLEANING REPORTS

Start	Stop	Released size	Files number	Status
8 Dec 2017 13:32:13	-	0 B (out of 179 MiB)	0	○
8 Dec 2017 13:31:53	8 Dec 2017 13:31:55	448 MiB (out of 565.4 MiB)	1	☒

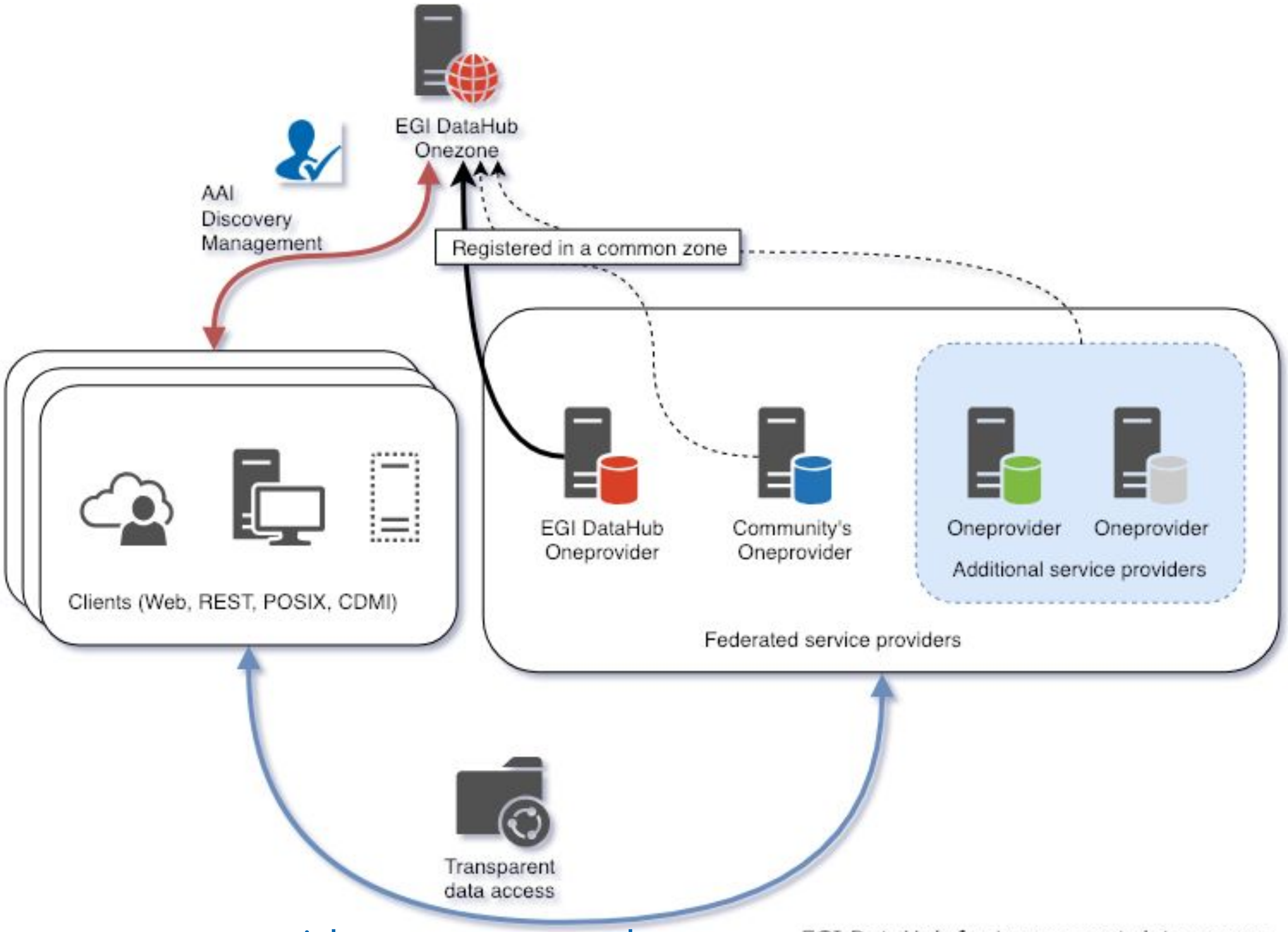
The work of the EGI Foundation
is partly funded by the European Commission
under H2020 Framework Programme

Multiple usage models

- Transparent data access service
- Doing smart caching of remote storage
- Federating data sources/providers
- Publishing datasets
- Notebooks with DataHub

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

DataHub for transparent data access

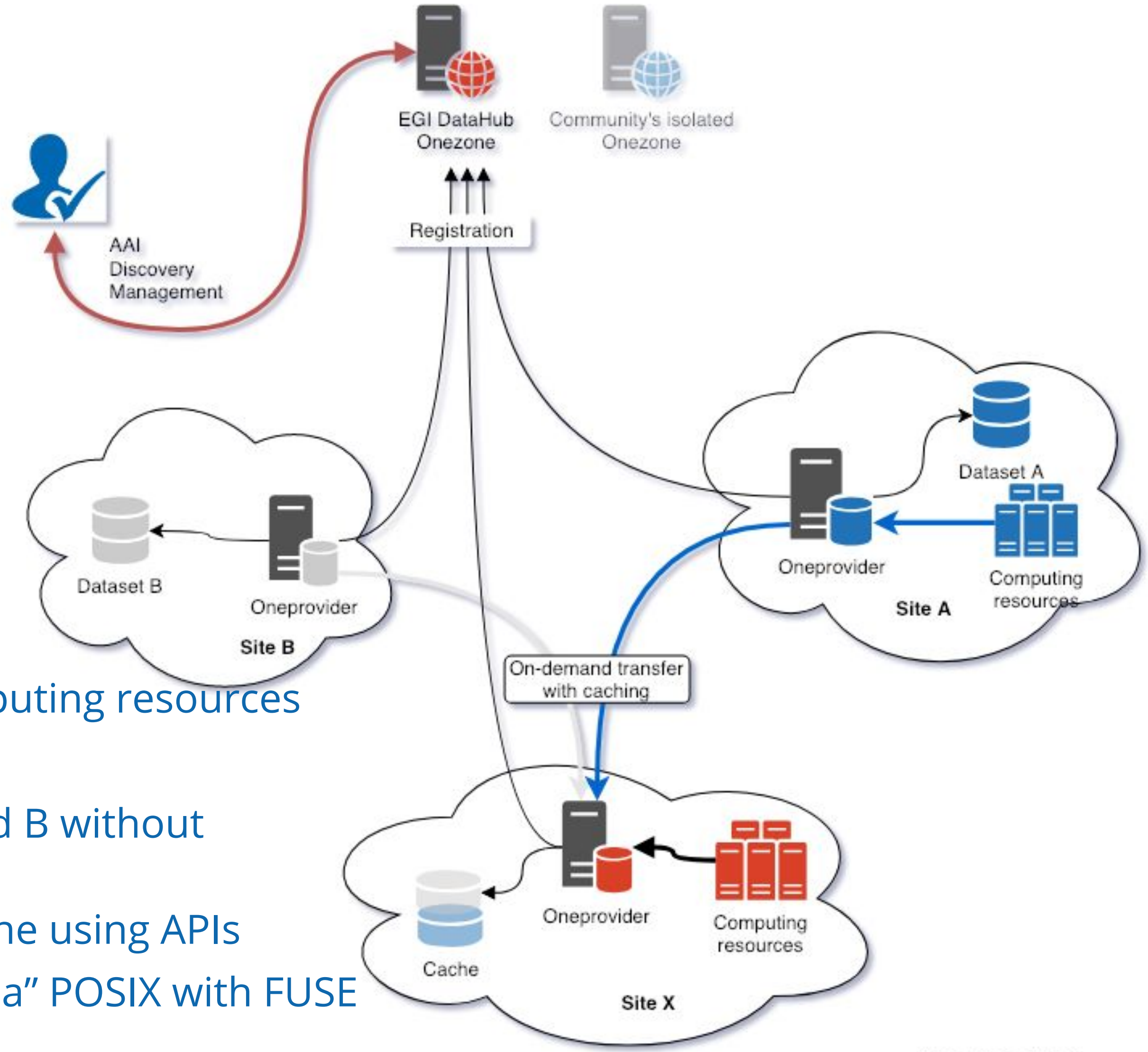


EGI DataHub for transparent data access

- Clients uses one or more providers to access data
- Data can be accessed over multiple protocols

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

Smart caching of remote storage

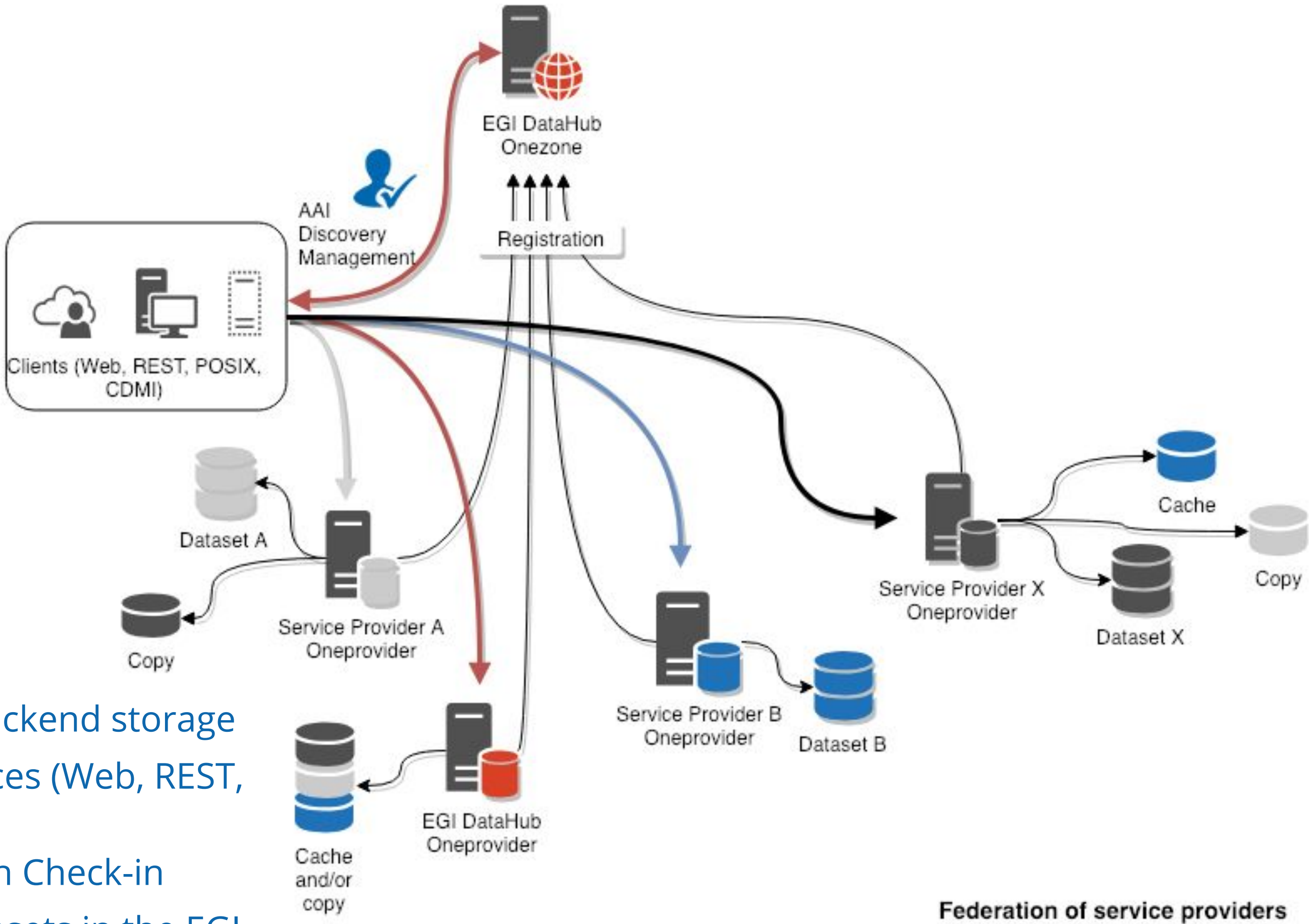


Smart caching

- Site A hosts data and computing resources
- Site B hosts only data
- Site X uses data from A and B without pre-staging
- Pre-staging can also be done using APIs
- Data is accessed locally "à la" POSIX with FUSE

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

Federation of service providers

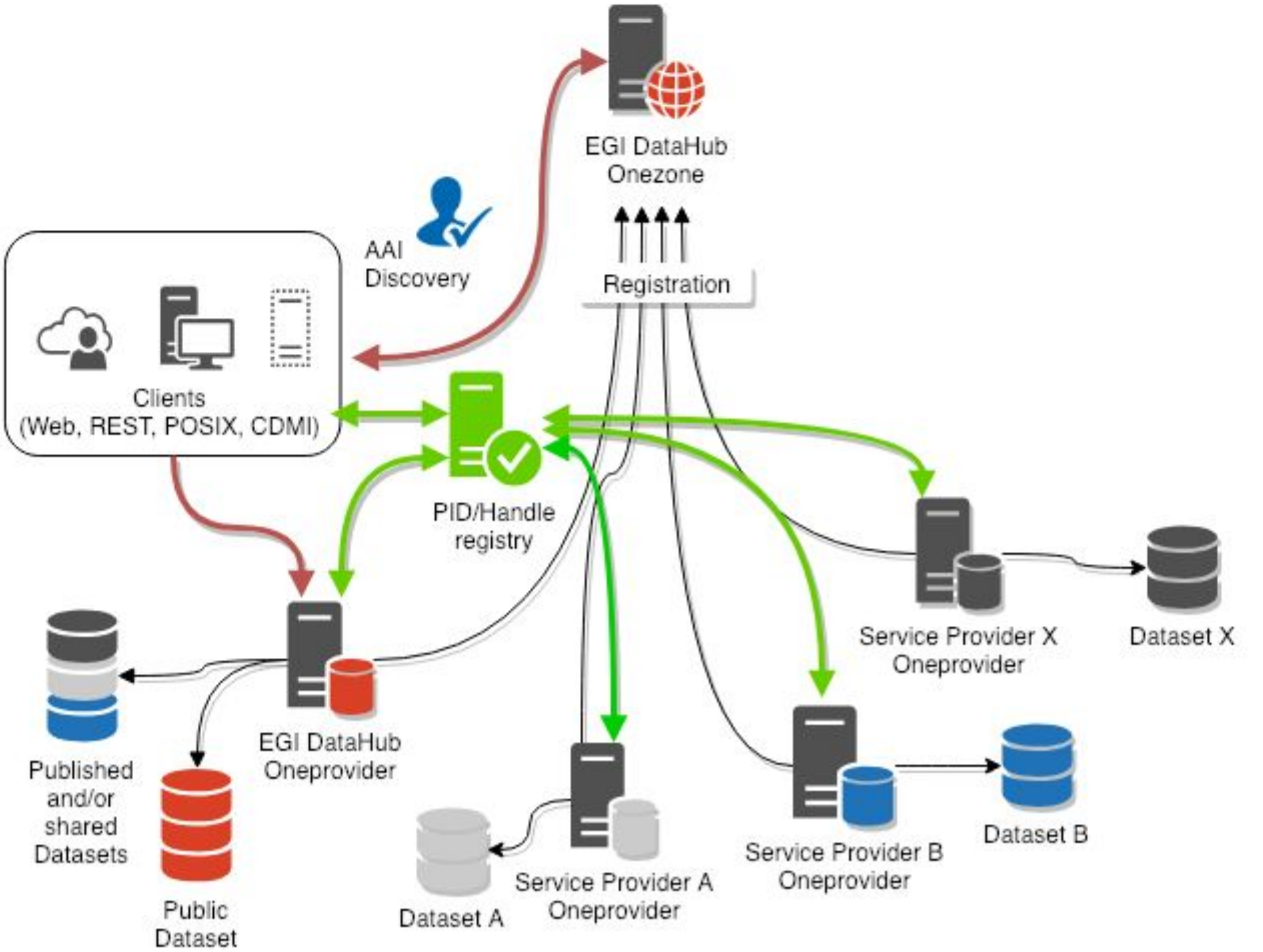


Federation of service providers

- Heterogenous backend storage
- Common interfaces (Web, REST, POSIX, CDMI)
- Common AAI with Check-in
- Discovery of Datasets in the EGI DataHub

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

Publishing and discovery of datasets



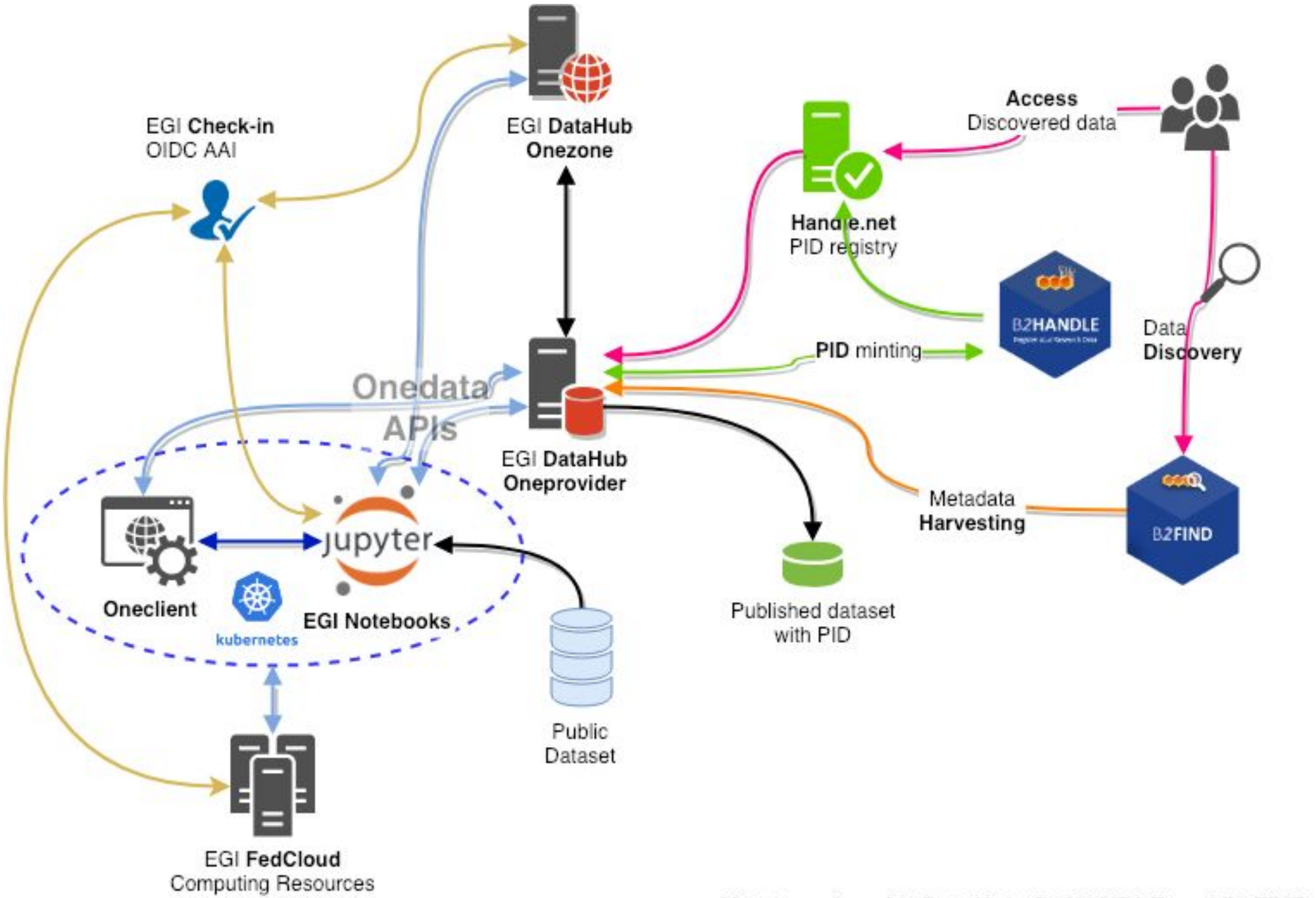
EGI DataHub Onezone
AAI Discovery
Registration
Clients (Web, REST, POSIX, CDMI)
PID/Handle registry
Service Provider X Oneprovider
Dataset X
EGI DataHub Oneprovider
Published and/or shared Datasets
Public Dataset
Dataset A
Service Provider A Oneprovider
Service Provider B Oneprovider
Dataset B

- PID minting
- Publishing, discovery and access to datasets

Publication of datasets

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

Notebooks with DataHub



Noteboooks with DataHub, B2HANDLE and B2FIND

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

Steps to use DataHub and Onedata

- Collecting and analysing dataset specificities
  - Number of files
  - Size of files
- Preparing a pilot
  - Designing and validating usage model
  - Integrating Onedata with existing resources
- Validating the pilot
- Deploying a production setup
  - Ensuring hardware requirements are sufficient
    - RAM, CPU, Disk, Network,…
    - Storage backend

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

Deploying Onedata

- Preferred model: using docker containers
  - Using docker-compose
  - Packages for Ubuntu 16.04 and CentOS 7 also available

ONEDATA

docker

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

Requirements for production

- Powerful-enough Oneprovider
  - RAM: 32GB
  - CPU: 8 vCPU
  - Disk: 50GB SSD
  - To be adjusted for the dataset and usage scenario

- For high IOPS
  - High-performance backend storage (CEPH)
  - Low latency network

- POSIX mounting
  - Oneprovider close to the Oneclient

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

Links

- EGI DataHub
  - https://datahub.egi.eu/
  - https://community.egi.eu/c/egi-services/datahub
  - https://egi-datahub.readthedocs.io/
  - https://wiki.egi.eu/wiki/EGI_Federated_Data

- System requirements
  - https://onedata.org/docs/doc/system_requirements.html

- Official Onedata documentation
  - https://onedata.org
  - https://onedata.org/#/home/documentation
  - Getting started
    - https://github.com/onedata/getting-started
  - Source code: https://github.com/onedata

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