

# EGI DataHub

Data as a Service – Distributed Data  
Management



---

[www.egi.eu](http://www.egi.eu)

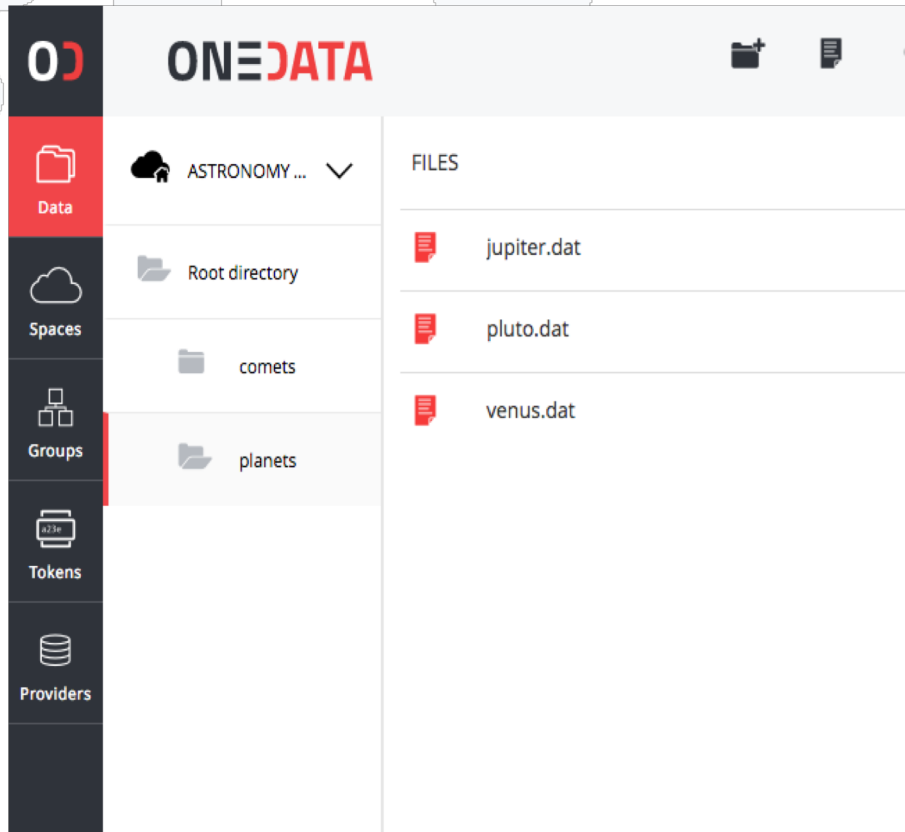
This work by EGI.eu is licensed under a  
[Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

- 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 a sharing service
- No easy way of **combining multiple datasets from different data providers**
- Users need to **access data locally** and from **compute resources**

# 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

# On the client side



**ONEDATA**

ASTRONOMY ...

FILES

- jupiter.dat
- pluto.dat
- venus.dat

```
[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

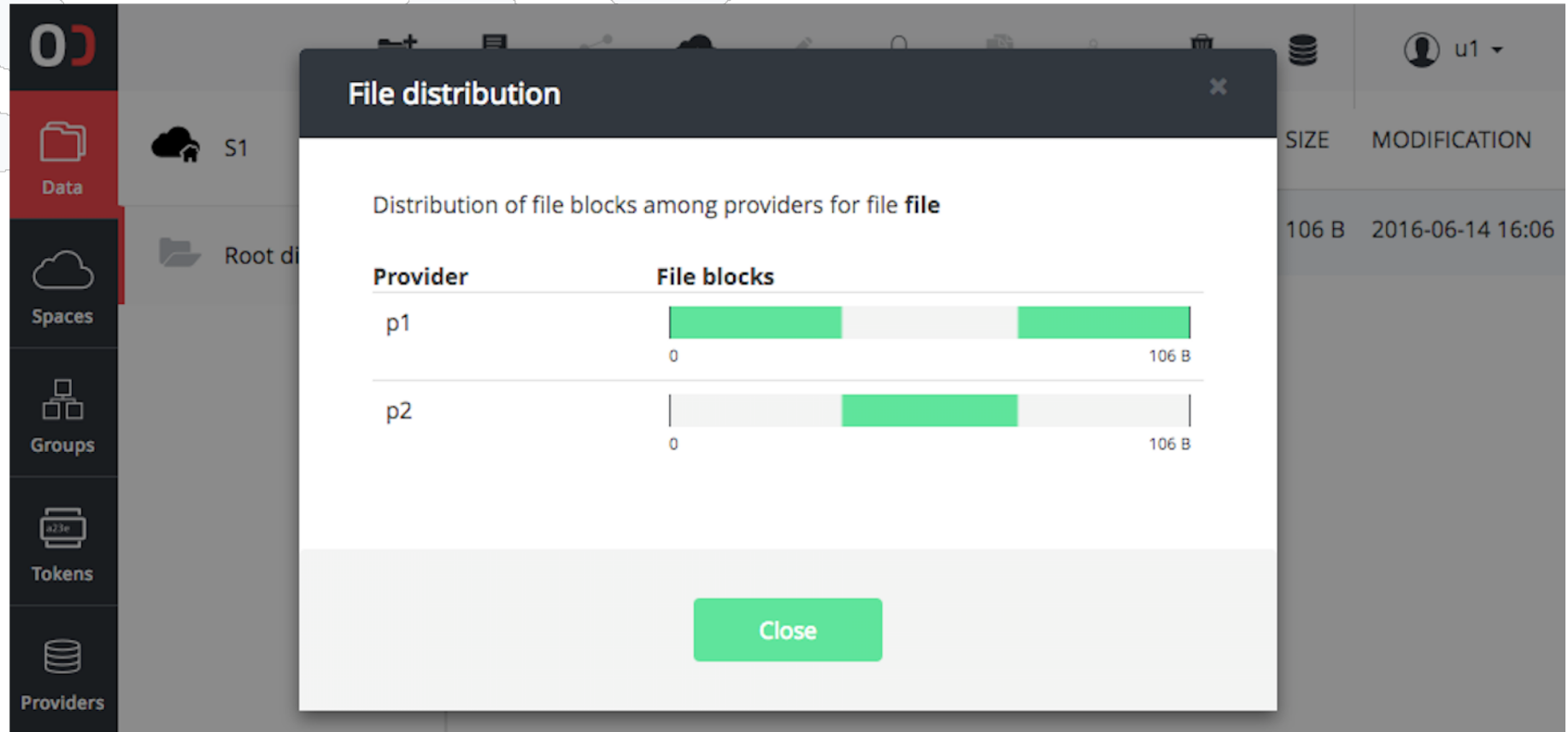
./Astronomy Datasets/comets:
total 0
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:20 enck.dat
-rw-r--r-- 1 1124656 1733762 10000000 Sep 26 19:19 halley.dat

./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

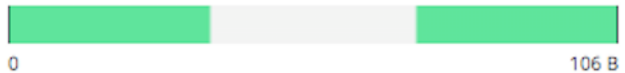

./Big Data Experiment:
total 0
-rw-r--r-- 1 1124656 1337123 10000000 Sep 26 19:08 cats_images.tgz
-rw-r--r-- 1 1124656 1337123 5000000 Sep 26 19:13 galaxies.img
-rw-r--r-- 1 1124656 1337123 5000000 Sep 26 19:14 spam_mails.tgz

./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]#
```

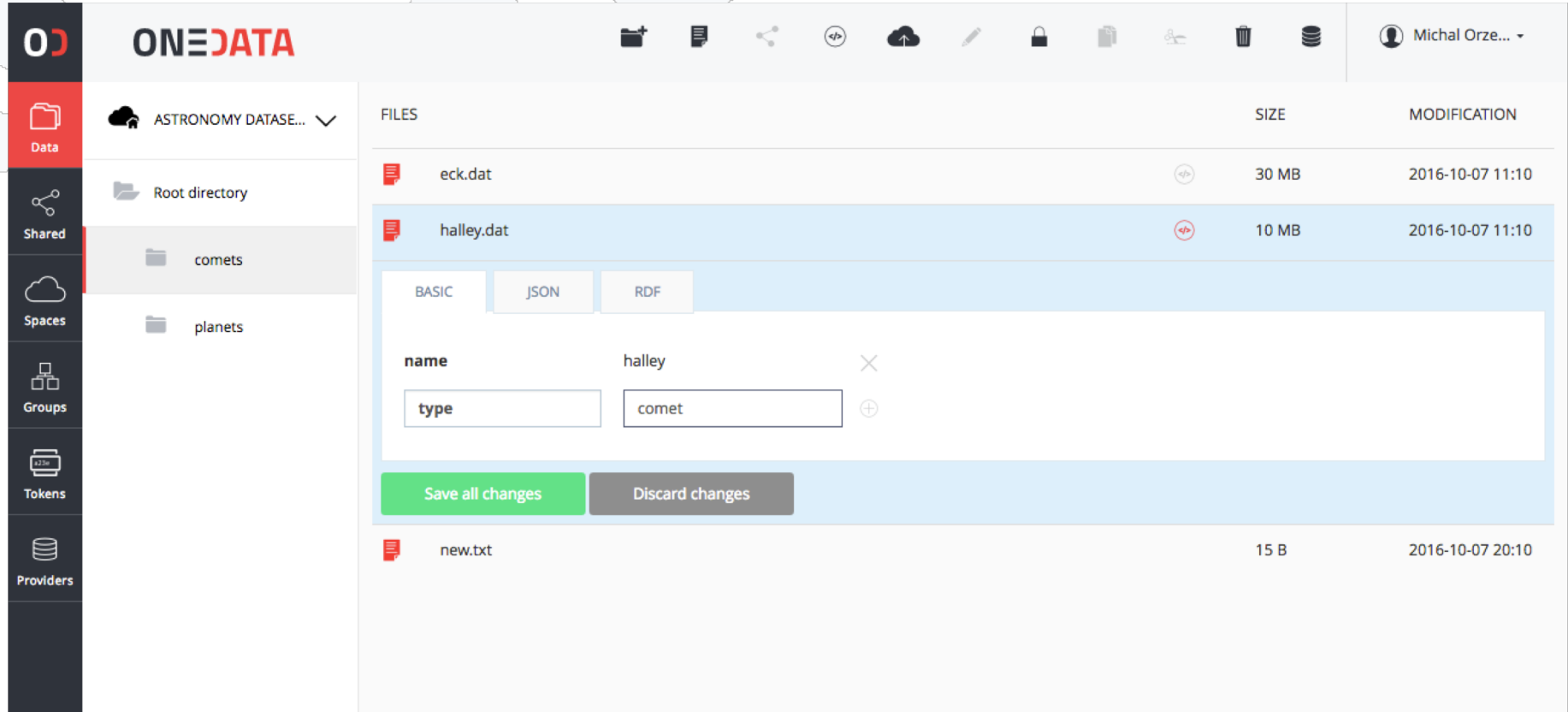




The screenshot shows a 'File distribution' dialog box overlaid on a cloud management interface. The dialog box title is 'File distribution' and it contains the text 'Distribution of file blocks among providers for file file'. Below this text is a table with two columns: 'Provider' and 'File blocks'. The table shows two providers, p1 and p2, each with a horizontal bar representing the distribution of file blocks. The total size of the file is 106 B. The background interface shows a sidebar with 'Data', 'Spaces', 'Groups', 'Tokens', and 'Providers' sections, and a main area with a table showing file details.

Provider	File blocks
p1	 0 106 B
p2	 0 106 B

Close



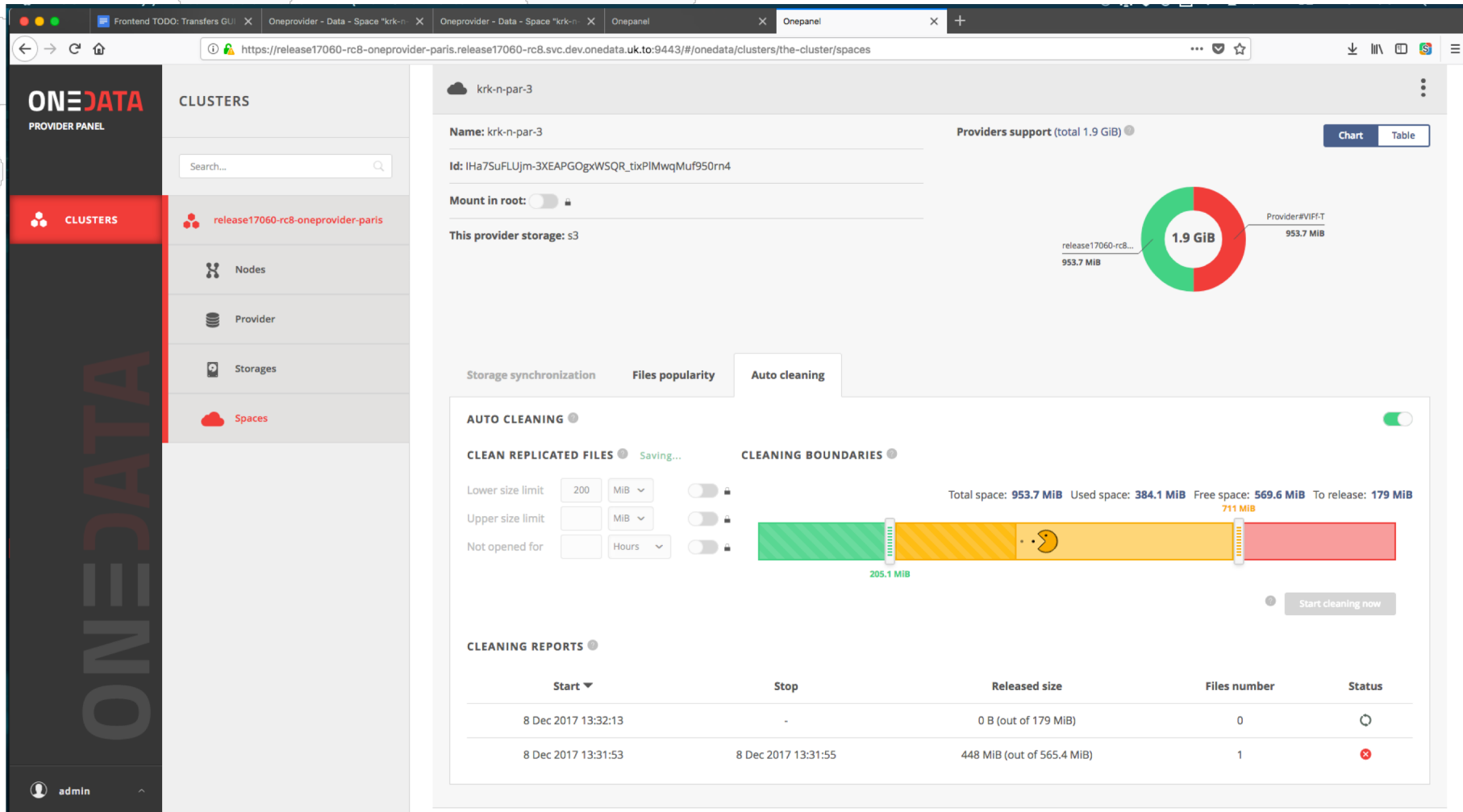
The screenshot shows the ONE DATA interface. On the left is a navigation sidebar with sections: Data, Shared, Spaces, Groups, Tokens, and Providers. The main area is titled 'ASTRONOMY DATASE...' and shows a file browser view. A table lists files with columns for FILE, SIZE, and MODIFICATION. The file 'halley.dat' is selected, and its metadata is displayed in a form below the table. The form has tabs for BASIC, JSON, and RDF. The 'name' field contains 'halley' and the 'type' field contains 'comet'. At the bottom of the form are 'Save all changes' and 'Discard changes' buttons.

FILE	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

Metadata form for 'halley.dat':

- Tabs: BASIC (selected), JSON, RDF
- name: halley
- type: comet
- Buttons: Save all changes, Discard changes

# File popularity and smart caching



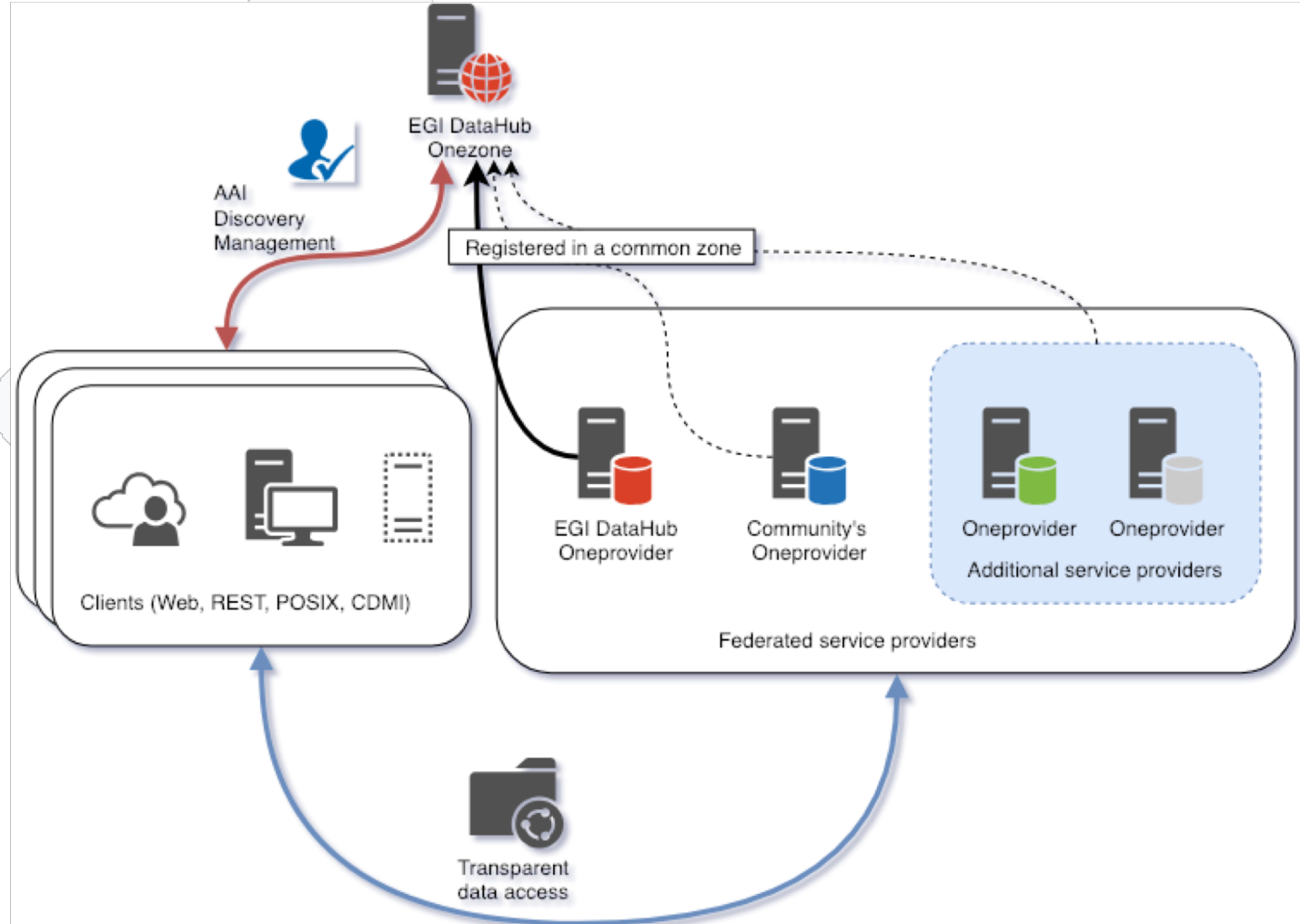
The screenshot shows the ONE DATA PROVIDER PANEL interface. The left sidebar contains navigation options: CLUSTERS, Nodes, Provider, Storages, and Spaces. The main content area displays details for a cluster named 'krk-n-par-3'. It includes a search bar, a search result for 'release17060-rc8-oneprovider-paris', and a 'Mount in root' toggle. A donut chart shows 'Providers support (total 1.9 GiB)' with segments for 'release17060-rc8...' (953.7 MIB) and 'Provider#VIFI-T' (953.7 MIB). Below this, there are tabs for 'Storage synchronization', 'Files popularity', and 'Auto cleaning'. The 'Auto cleaning' tab is active, showing 'AUTO CLEANING' settings (enabled), 'CLEAN REPLICATED FILES' (Saving...), and 'CLEANING BOUNDARIES'. A progress bar indicates 'Total space: 953.7 MiB', 'Used space: 384.1 MiB', 'Free space: 569.6 MiB', and 'To release: 179 MiB'. A 'Start cleaning now' button is present. A 'CLEANING REPORTS' table is shown at the bottom.

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	❌

# Multiple usage models

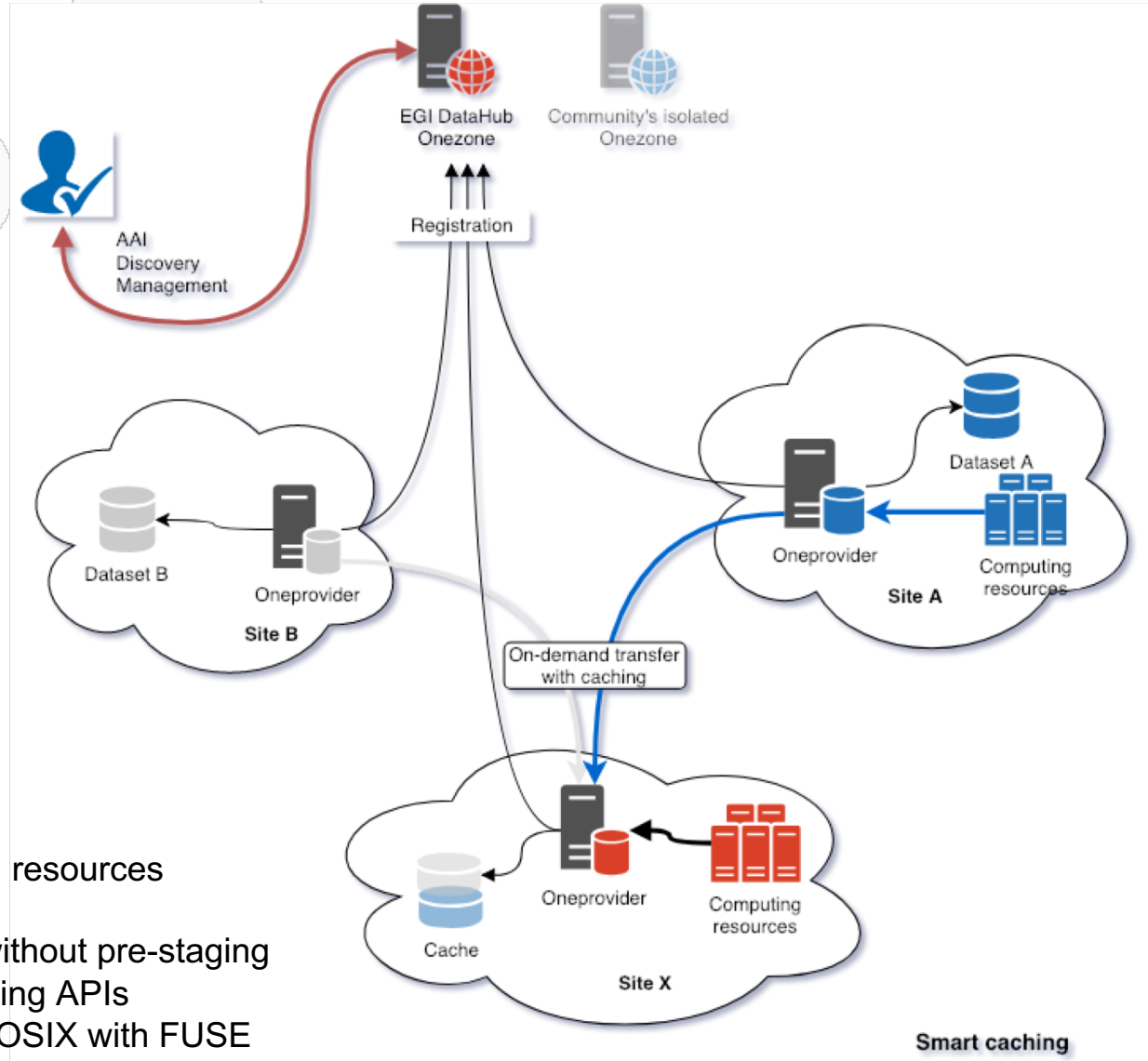
- Transparent data access service
- Doing smart caching of remote storage
- Federating data sources/providers
- Publishing datasets
- Notebooks with DataHub

# DataHub for transparent data access



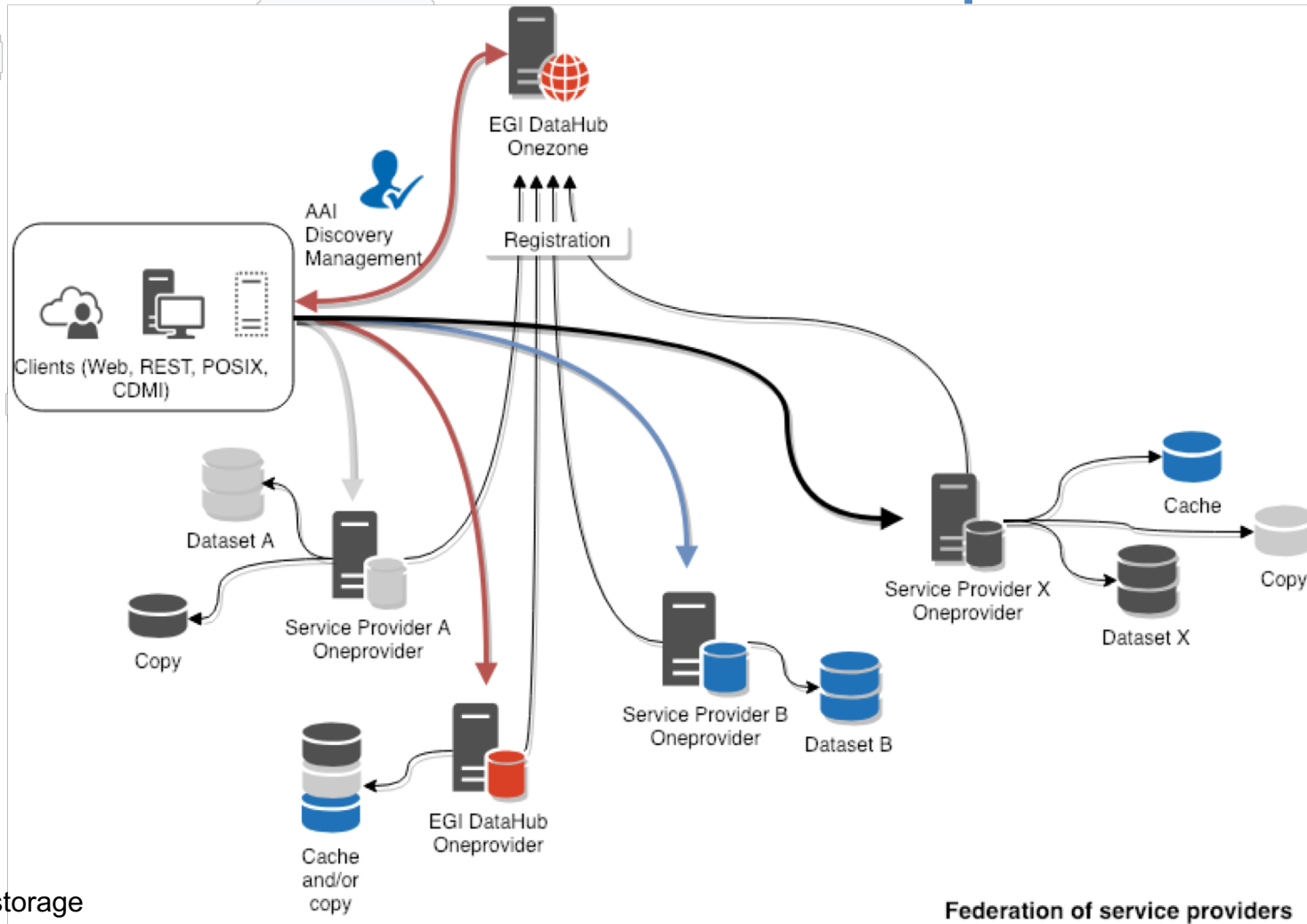
- Clients uses one or more providers to access data
- Data can be accessed over multiple protocols

# Smart caching of remote storage



- 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

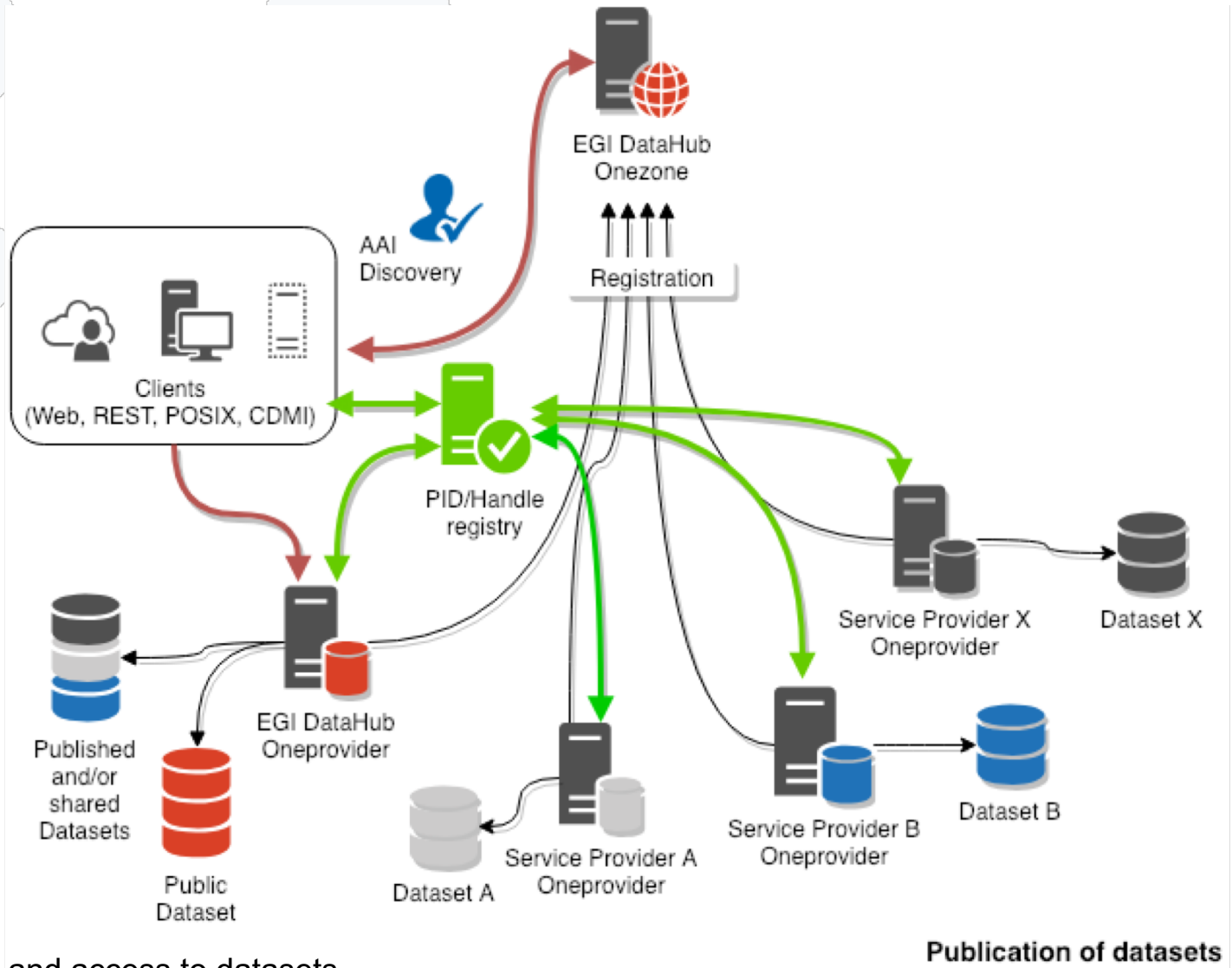
# Federation of service providers



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

# Publishing and discovery of datasets

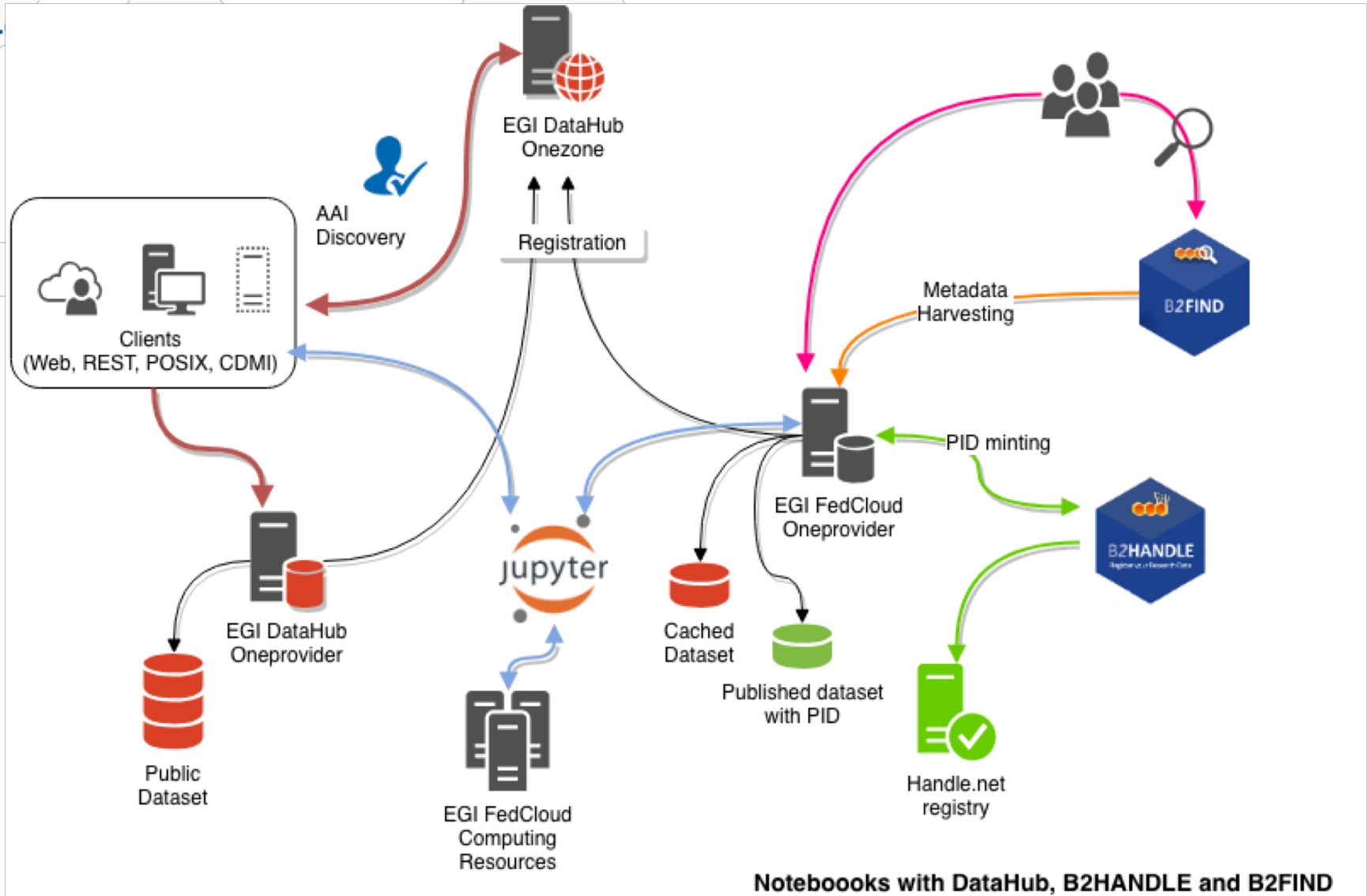


- PID minting
- Publishing, discovery and access to datasets

**Publication of datasets**



# Notebooks with DataHub



**Notebooks with DataHub, B2HANDLE and B2FIND**

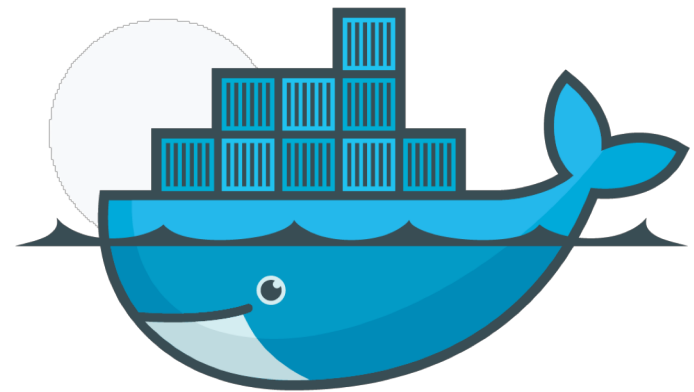
# Steps to use DataHub and Onedata

- 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

# Deploying Onedata

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

**ONEDATA**



**docker**

# 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

- EGI DataHub
  - <https://datahub.esi.eu/>
  - <https://community.esi.eu/c/egi-services/datahub>
  - <https://egi-datahub.readthedocs.io/>
  - [https://wiki.esi.eu/wiki/EGI\\_Federated\\_Data](https://wiki.esi.eu/wiki/EGI_Federated_Data)
- System requirements
  - [https://onedata.org/docs/doc/system\\_requirements.html](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>

# Thank you for your attention.

*Questions?*



---

[www.egi.eu](http://www.egi.eu)

This work by EGI.eu is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).