

Open and FAIR data initiative

With contributions from RIC, FS, Daphne, PUNCH, INFa, dCache dev and ops

FS & DAPHNE: Anton, Lisa A., Christoph R. and Linus

PUNCH: Bas (Basavaraja) and Hubert

dCache: Tigran, Christian and soon Mwai

InFa: Peter, Regina, Noel and Johannes

User Support: Frank S.

RIC: Sophie, Paul, Tim, Uwe, Franz and myself

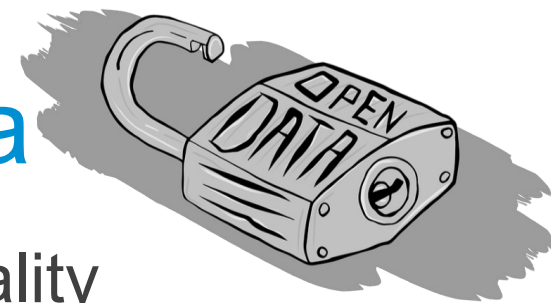
Presented by Patrick Fuhrmann, RIC

RIC : FAIR and Open Science experience

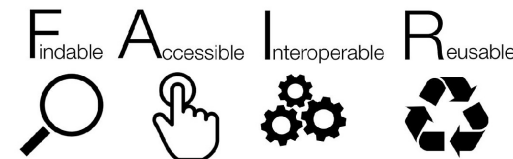
Working in the realm of FAIR and Open Science and data for years

- Anton, Sophie, Paul and myself coordinated a 6 Million Euro Project on FAIR data for Synchrotrons with 10 European Synchrotron Facilities
 - Created templates for DMP and Data Policies for Photon facilities
 - Worked on ontologies for metadata of beamline techniques.
- Funded the SciCat OAI-PMH protocol plug-in for harvesting
- Close connection to DAPHNE and FS on Open and FAIR Data.
- MoUs with LEAPS partners on FAIR data and the VISA portal.
- EOSC Beyond on FAIR data with NFDI head office (Y. Sure-Vetter)
- Our Open Science Network -> DESY must not become a FAIR Data island.
 - We represent DESY in the EOSC Association
 - Florian from L and myself represent DESY in the Open Science AK of the Helmholtz Open Science office
 - Paul represents DESY in the NIAC, defining Metadata structures in the NeXus file format.
 - FAIR Data from synchrotrons
 - Sophie organized a European Workshop in Paris for FAIR Photon data for AI
 - Sophie produced a [DMP promotional movie for Photon Science](#) for the EOSC.


How do we interpret Open and FAIR data



- Open data by itself is of no value unless it fulfills certain quality criteria!
- Those criteria became known as FAIR data principles starting 2014!
- However FAIR data of course doesn't need to be open, it just defines it's attributes and the level of reusability.
- One of 1000 definitions: Cambridge Crystallographic Data Center
 - The FAIR Data Principles set out criteria that enable the philosophy of openness to be realized in a tangible way through modern publishing practices and infrastructures that support current data science needs.
- The most important reasons for FAIR
 - Verifiable scientific results (enforced by journals)
 - Machine readability for ML and AI
- Current hot topic in the Helmholtz: The H-Foundation Models
 - We are involved with imaging and data management



What would be a minimum viable system for DESY ?

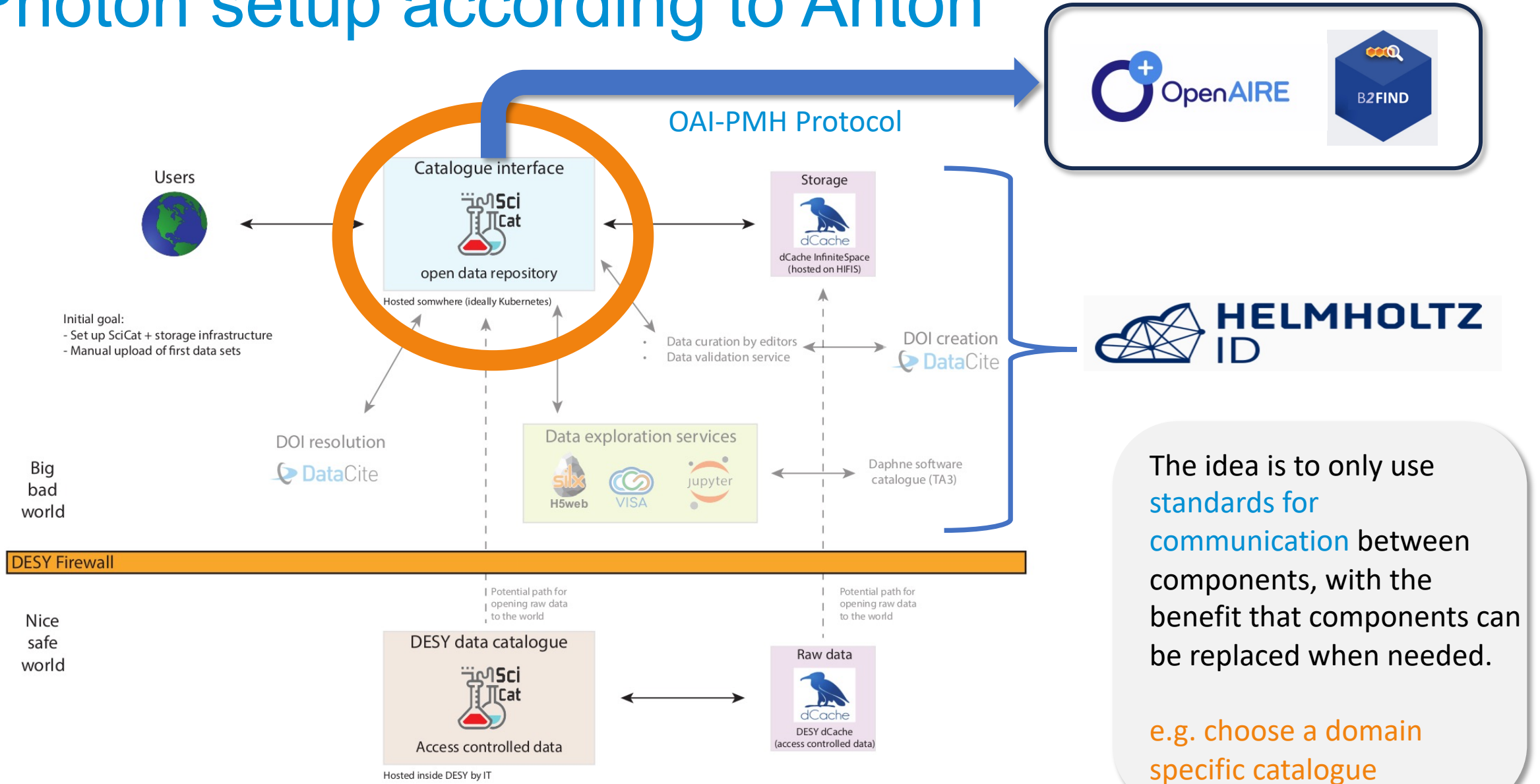
- Long term storage system with
 - Federated Authenticated and non-authenticated access
 - Standard protocol access (http, NFS)
 - Metadata catalogue with
 - Federated Authenticated and non-authenticated access
 - Mandatory core metadata fields
 - Optional domain specific metadata fields
 - OAI-PMH protocol for harvesting core metadata by high level catalogues
 - DOI minting service (See Linus's/Anton's position paper with L)
 - Open Science (Virtual Research) infrastructure
 - MoU on VISA with other Synchrotrons in Europe
- 
- Second Phase

Importance of proper metadata definitions

- Reminder : **Mandatory core metadata fields**
 - Core metadata already has been defined in prior activities and by responsible reference bodies. (e.g. Dublin Core, Data Cite V4.4)
- Reminder: **Optional domain specific metadata fields**
 - Those definitions need to be provided by the domain specific communities.
 - Like ExPaNDS/PaNOSC and DAPHNE (Proposal by Lisa et al.)
- **Parallel to Open Metadata**, additional information might be required
 - Beamline level
 - Facility level



Photon setup according to Anton

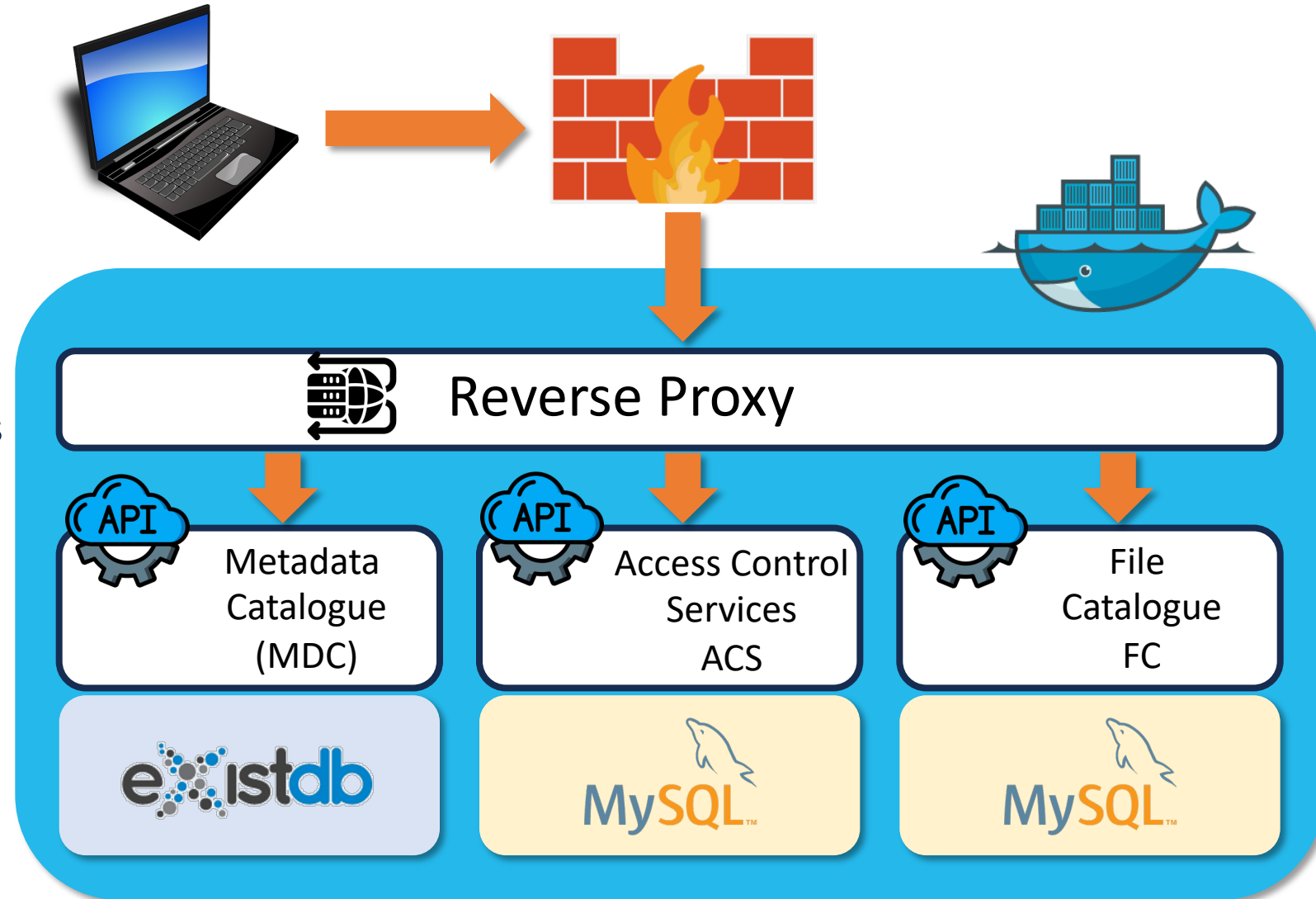


The idea is to only use **standards for communication** between components, with the benefit that components can be replaced when needed.

e.g. choose a domain specific catalogue

Other catalogue example: PUNCH: ILDG Metadata catalogue

- Configurable access policies
- Independent (and optional) use of MDC/FC/ACS functionalities, e.g.
 - only MDC or
 - only FC
 - also without ACS if write access is only done on server-side



What is the role of HIFIS?

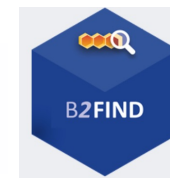
- HIFIS provides the **skillset** and the **financial capacity** to
 - offer the **Helmholtz ID system** according to the AARC blueprint
 - operate **core services**, e.g.
 - HIFIS Storage
 - compose **high level services**
 - FAIR Open Repository
 - Data Transfer services
 - provides service **redundancy across Helmholtz sites**.
- **HIFIS Support**: Single point of contact for all components.

What is the role of Information Fabrics @ IT

- InFa primarily works on the beamline SciCats
 - thanks to resources provided by DAPHNE
- InFa is part of the European SciCat development team and
- InFa will keep DESY communities updated on new software developments (SciCat technical and management meetings, ...)
- InFa is deploying and operating Open SciCat for RIC/HIFIS
- InFa is the sole contact for SciCat deployments and requirements analysis at DESY

Contact : scicat.service@desy.de

Envisioned Process

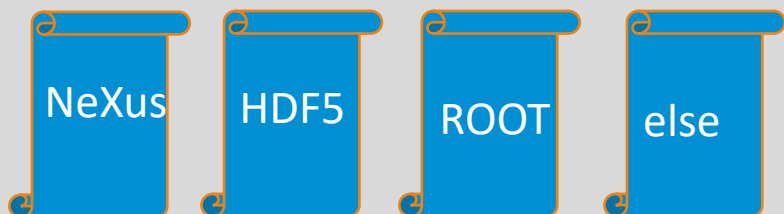


Staging Space

Additional
Manual
Metadata



Metadata Extraction (e.g. from NeXuS)



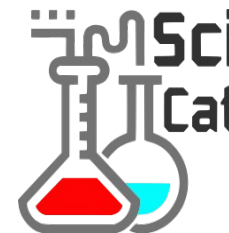
DATA

Publishing Process

Data Quality
verification

Metadata
Enrichment

Open Catalogue



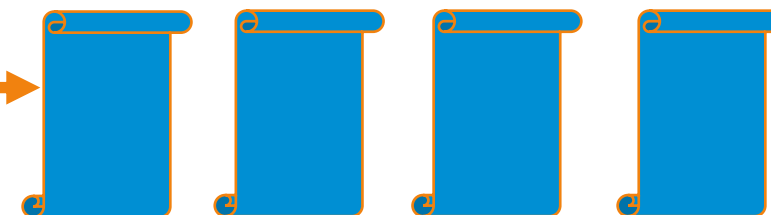
Metadata can still be
extended/modified.



Optional DOI minting

Persistent Space (Frozen)

Data is immutable and no
longer removable





Progress











HIFIS-Storage.desy.de




The storage drop box and the final storage space!

 Root desy public-data upload







Type	Name	Creation time	File location	Size
	daphne4nfdi	29/11/2023, 14:17:40	Disk	--
	it-ric	29/11/2023, 15:24:43	Disk	--
	punch4nfdi	29/11/2023, 14:18:05	Disk	--

Write access granted by Helmholtz VO membership!

The catalogue!



Search

PID

Text Search

benasque

Location

Group

Type

Keywords

Start Date – End Date

+

Add Condition

My Data

All Public Data

All

Archivable

Retrievable

Work In Progress

System Error

User Error

□

Name

Source Folder

□	Benasque	...
---	----------	-----

📄

General Information

Name	Benasque
Description	Sloan Digital Sky Survey
PID	undefined/904c64d5-cf00-4eb9-8999-d80c4ce34caa 📄
Type	derived
Creation Time	2022-09-25 22:35
Keywords	

👤

Creator Information

Owner	Patrick Fuhrmann
Investigator	patrick.fuhrmann@desy.de

+

Create Dataset



Name	
Benasque	

Related Documents

https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08311.JPG ,
 https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08312.JPG ,
 https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08313.JPG ,
 https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08314.JPG ,
 https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08315.JPG ,
 https://hifis-storage.desy.de/desy/open-data/upload/ric-it/DSC08316.JPG

General Information

Name	Benasque
Description	Sloan Digital Sky Survey
PID	undefined/904c64d5-cf00-4eb9-8999-d80c4ce34caa
Type	derived
Creation Time	2022-09-25 22:35
Keywords	

Creator Information

Owner	Patrick Fuhrmann
Investigator	patrick.fuhrmann@desy.de
Contact Email	patrick.fuhrmann@desy.de
Owner Group	it-ric
Access Groups	

File Information

Source Folder	/nfs
----------------------	------

Scientific Metadata

View
Edit

Battery-Level	67%
Lens-Info	24-240mm f/3.5-6.3
Lens-Model	FE 24-240mm F3.5-6.3 OSS
Field of View	54.4 deg

Tiny Demo (thanks to Tim)



Demo (Thanks to Tim for the preparation)

The preparation

Dataset Name

```
patrick@zitpocx39737-w scicat-demo % ls spain
```

```
total 450704
```

```
-rw-rw-r--@ 1 14417920 Dec 18 06:27 DSC08312.JPG
-rw-rw-r--@ 1 14123008 Dec 18 06:27 DSC08313.JPG
-rw-rw-r--@ 1 17596416 Dec 18 06:27 DSC08314.JPG
-rw-rw-r--@ 1 16384000 Dec 18 06:27 DSC08316.JPG
-rw-rw-r--@ 1 16416768 Dec 18 06:27 DSC08317.JPG
-rw-rw-r--@ 1 11894784 Dec 18 06:27 DSC08318.JPG
-rw-rw-r--@ 1 18284544 Dec 18 06:27 DSC08320.JPG
-rw-rw-r--@ 1 18317312 Dec 18 06:27 DSC08322.JPG
-rw-rw-r--@ 1 21889024 Dec 18 06:27 DSC08323.JPG
-rw-rw-r--@ 1 20578304 Dec 18 06:27 DSC08327.JPG
-rw-rw-r--@ 1 22872064 Dec 18 06:27 DSC08331.JPG
-rw-rw-r--@ 1 15564800 Dec 18 06:27 DSC08335.JPG
-rw-rw-r--@ 1 22413312 Dec 18 06:27 DSC08339.JPG
-rwxr-xr-x 1 1856 Dec 18 06:39 extract.sh
```

Data Files

```
patrick@zitpocx39737-w scicat-demo % spain/extract.sh
```

```
% spain/metadata.json
```

```
{
  "creationTime": "2023-12-18T06:27:01+01:00",
  "scientificMetadata": {
    "Battery-Level": "67%",
    "Lens-Info": "24-240mm f/3.5-6.3",
    "Lens-Model": "FE 24-240mm F3.5-6.3 OSS",
    "Field of View": "54.4 deg"
  },
  "inputDatasets": [
    "https://hifis-storage.desy.de:/desy/open-data/upload/ric-it/DSC08312.JPG",
    "https://hifis-storage.desy.de:/desy/open-data/upload/ric-it/metadata.json"
  ],
  "owner": "Patrick",
  "owner": "Fuhrmann",
  "ownerEmail": "patrick.fuhrmann@desy.de",
  "contactEmail": "patrick.fuhrmann@desy.de",
  "sourceFolder": "/nfs",
  "size": 0,
  "packedSize": 0,
  "type": "derived",
  "keywords": [],
  "description": "Pictures from Benasque",
  "datasetName": "spain",
  "isPublished": true,
  "ownerGroup": "it-ric",
  "accessGroups": [],
  "datasetlifecycle": {"archivable": true, "retrievable": false, "publishable": false},
  "techniques": [],
  "investigator": "patrick.fuhrmann@desy.de",
  "usedSoftware": ["curl", "SciCat", "dCache"]
}
```

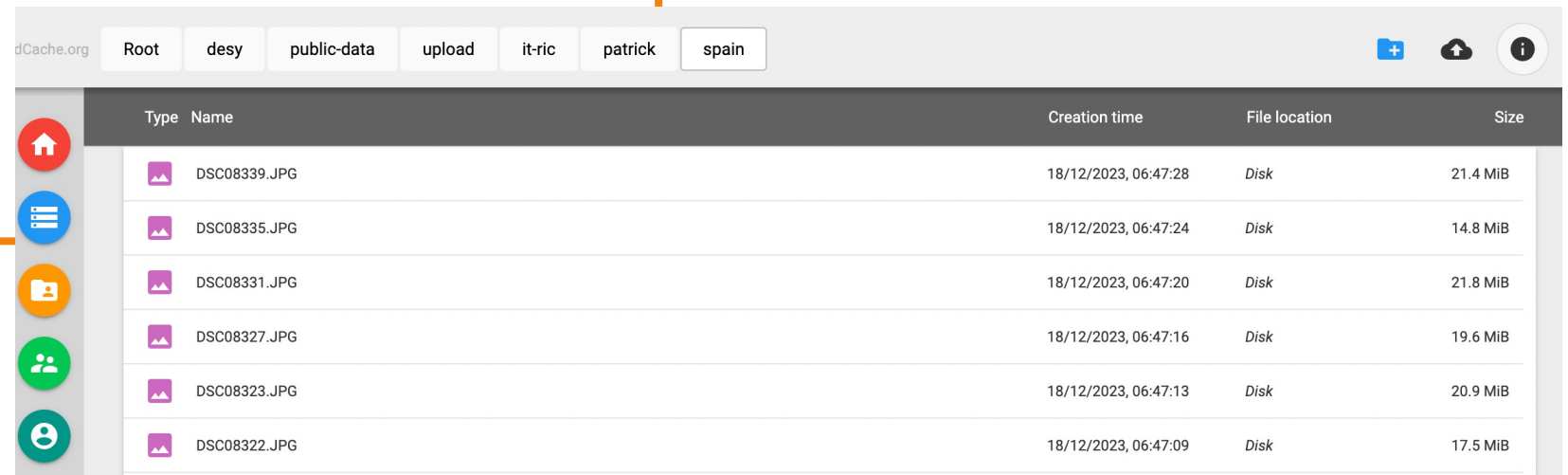

Demo (cont.)







The upload

```
patrick@zitpocx39737-w scicat-demo % ./pd_upload.sh spain
Uploading spain/DSC08312.JPG
Uploading spain/DSC08313.JPG
Uploading spain/DSC08314.JPG
Uploading spain/DSC08316.JPG
Uploading spain/DSC08317.JPG
Uploading spain/DSC08318.JPG
Uploading spain/DSC08320.JPG
Uploading spain/DSC08322.JPG
Uploading spain/DSC08323.JPG
Uploading spain/DSC08327.JPG
Uploading spain/DSC08331.JPG
Uploading spain/DSC08335.JPG
Uploading spain/DSC08339.JPG
```

Dataset Name

Data Files



Type	Name	Creation time	File location	Size
	DSC08339.JPG	18/12/2023, 06:47:28	Disk	21.4 MiB
	DSC08335.JPG	18/12/2023, 06:47:24	Disk	14.8 MiB
	DSC08331.JPG	18/12/2023, 06:47:20	Disk	21.8 MiB
	DSC08327.JPG	18/12/2023, 06:47:16	Disk	19.6 MiB
	DSC08323.JPG	18/12/2023, 06:47:13	Disk	20.9 MiB
	DSC08322.JPG	18/12/2023, 06:47:09	Disk	17.5 MiB

Demo (cont.)

The registration

```
patrick@zitpocx39737-w scicat-demo % ./pd_register.sh spain/metadata.json
```

```
"undefined/7cf02d08-4e90-4bbc-a21f-e182dbb6633a"
```

The check

```
patrick@zitpocx39737-w scicat-demo % ./pd_list.sh
```

```
"undefined/7cf02d08-4e90-4bbc-a21f-e182dbb6633a"
```

Scientific Metadata	
View	Edit
Search ×	
Battery-Level	67%
Lens-Info	24-240mm f/3.5-6.3
Lens-Model	FE 24-240mm F3.5-6.3 OSS
Field of View	54.4 deg

Datasets / undefined/7cf02d08-4e90-4bbc-a21f-e182dbb6633a /	
Details	Datafiles
Jupyter Hub	
General Information	
Name	spain
Description	Pictures from Benasque
PID	undefined/7cf02d08-4e90-4bbc-a21f-e182dbb6633a
Type	derived
Creation Time	2023-12-18 06:27
Keywords	
Creator Information	
Owner	Fuhrmann
Investigator	patrick.fuhrmann@desy.de
Contact Email	patrick.fuhrmann@desy.de
Owner Group	it-ric
Access Groups	

Thanks