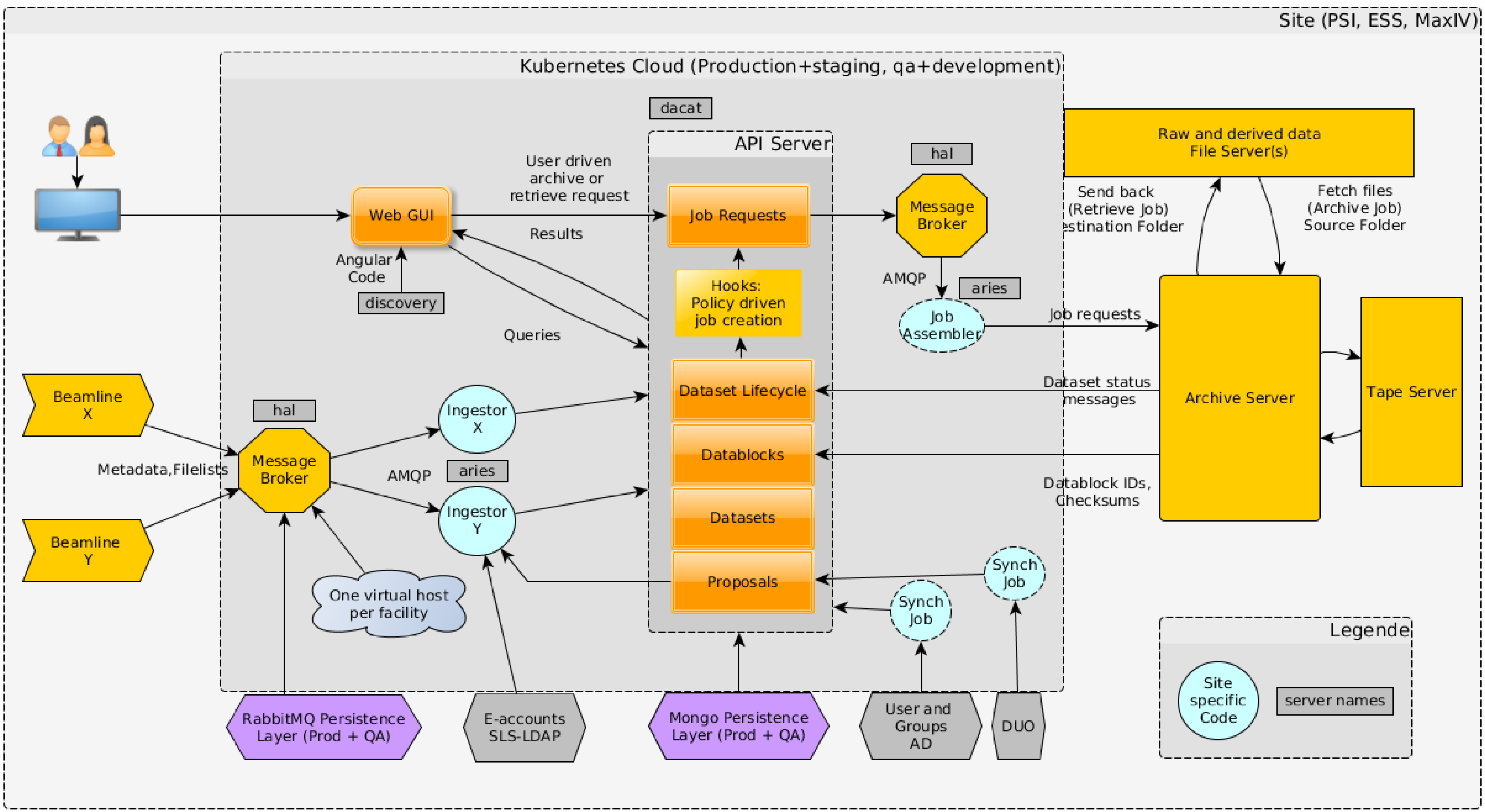
WIR SCHAFFEN WISSEN – HEUTE FÜR MORGEN

Luke Gorman :: Paul Scherrer Institut :: Photon Science Department

SciCat Data Catalog – Ingestion Experience

Ingestion Experience, November 6th 2018

Overall Data Catalog Architecture



Ingestion Methods

Automated ingestion at beamline (Raw Data)
This methods is setup once and integrates with the data acquisition process. The DAQ needs to call the beamline specific ingest script. Metadata can be extracted from the HDF5 raw data files.

Manual Ingestion (Raw and Derived Data):
This methods is triggered by the user. The user must prepare the metadata. Historic data is treated manually, initially with the help of the SciCat archive administrator.

Manual Ingestion

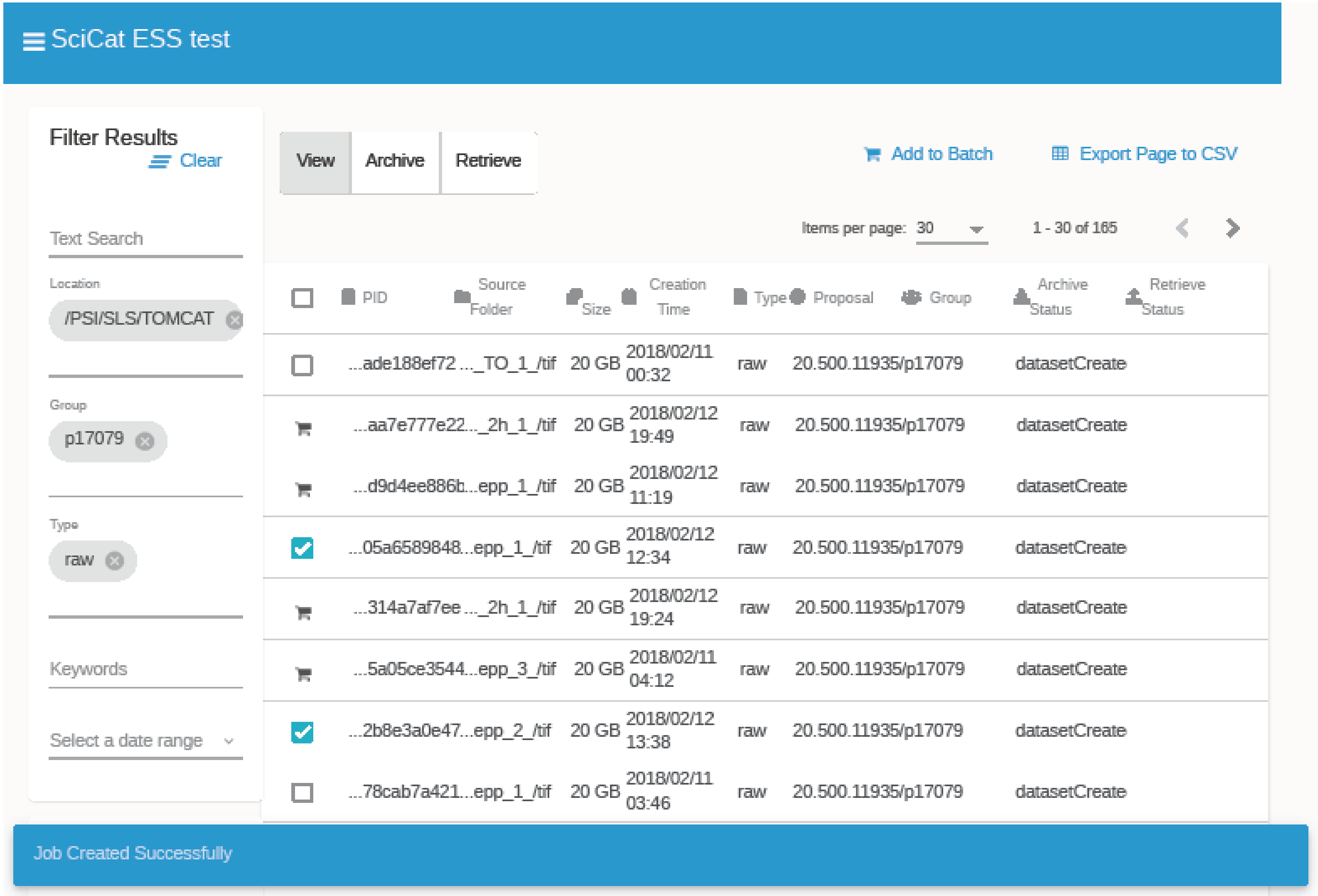
```
> datasetIngestor metadata.json

metadata.json:
{
"sourceFolder": "/some/folder/containing/the/derived/data",
"owner": "Thomas Meier",
"type": "derived",
"investigator": "thomas.meier@psi.ch",
.
"scientificMetadata": {
        "beamlineParameters": {
            "Monostripe": "Ru/C",
            "Ring current": {
                "v": 0.402246,
                "u": "A"
            },
            "Beam energy": {
                "v": 22595,
                "u": "eV"
            }
        .
```

Defining a Dataset

Data that is related:

- all the data collected in some time period/ during one measurement
- data with consistent environmental parameters (scientific metadata)
- all files of a dataset should usually be located in one folder
- up to the user to define
- think about retrieval when deciding on granularity of a dataset

After Ingestion - Selecting Datasets



SciCat ESS test

Filter Results

Clear

Text Search

Location: /PSI/SLS/TOMCAT ⊗

Group: p17079 ⊗

Type: raw ⊗

Keywords

Select a date range ⌄

View | Archive | Retrieve

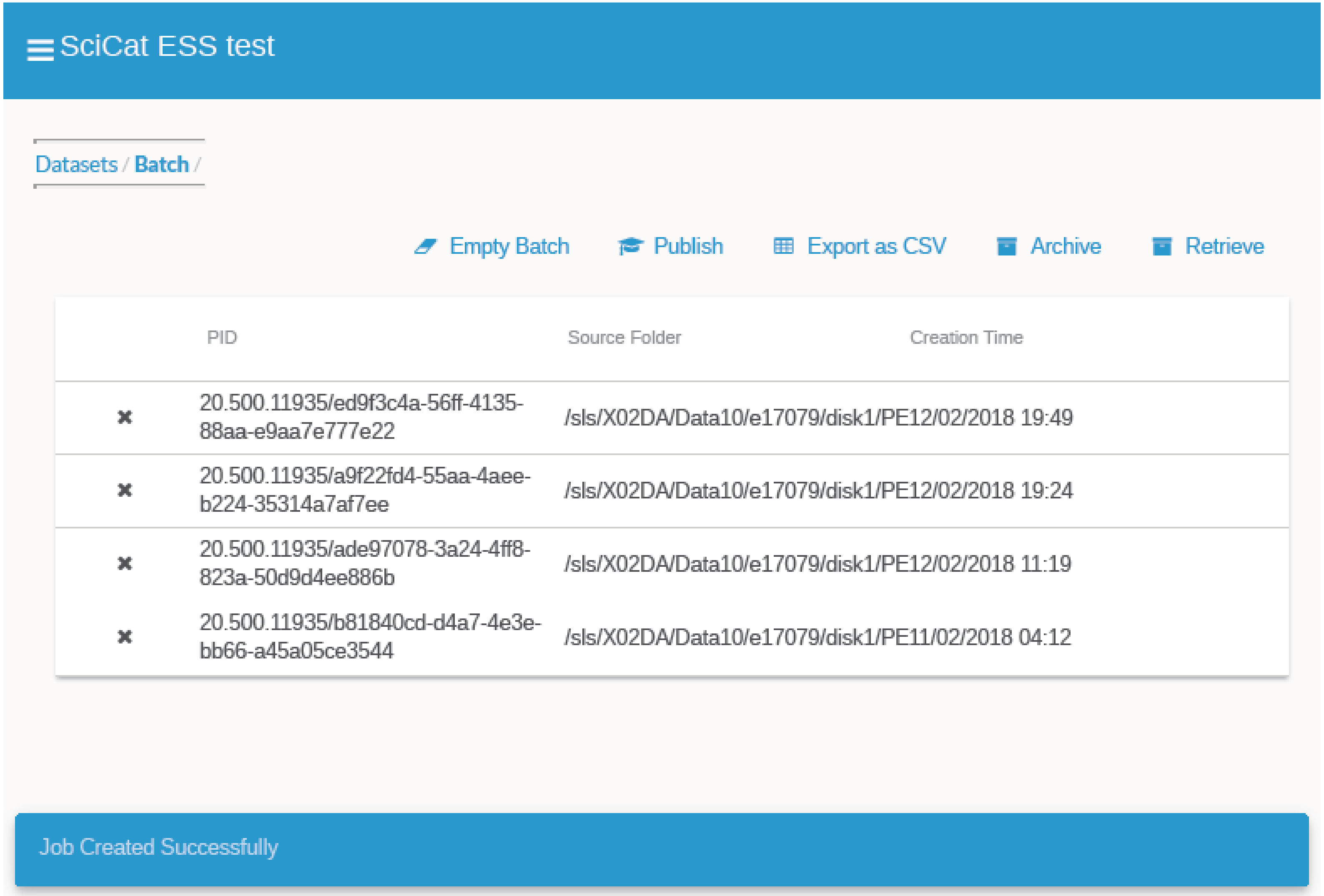
🛒 Add to Batch   ▦ Export Page to CSV

Items per page: 30 ▾   1 - 30 of 165   ‹ ›

☐	PID	Source Folder	Size	Creation Time	Type	Proposal	Group	Archive Status	Retrieve Status
☐	...ade188ef72	..._TO_1_/tif	20 GB	2018/02/11 00:32	raw	20.500.11935/p17079		datasetCreate	
🛒	...aa7e777e22	..._2h_1_/tif	20 GB	2018/02/12 19:49	raw	20.500.11935/p17079		datasetCreate	
🛒	...d9d4ee886b	...epp_1_/tif	20 GB	2018/02/12 11:19	raw	20.500.11935/p17079		datasetCreate	
☑	...05a6589848	...epp_1_/tif	20 GB	2018/02/12 12:34	raw	20.500.11935/p17079		datasetCreate	
🛒	...314a7af7ee	..._2h_1_/tif	20 GB	2018/02/12 19:24	raw	20.500.11935/p17079		datasetCreate	
🛒	...5a05ce3544	...epp_3_/tif	20 GB	2018/02/11 04:12	raw	20.500.11935/p17079		datasetCreate	
☑	...2b8e3a0e47	...epp_2_/tif	20 GB	2018/02/12 13:38	raw	20.500.11935/p17079		datasetCreate	
☐	...78cab7a421	...epp_1_/tif	20 GB	2018/02/11 03:46	raw	20.500.11935/p17079		datasetCreate	

Job Created Successfully

Executing Actions on Datasets



≡ SciCat ESS test

Datasets / **Batch** /

Empty Batch | Publish | Export as CSV | Archive | Retrieve

	PID	Source Folder	Creation Time
✖	20.500.11935/ed9f3c4a-56ff-4135-88aa-e9aa7e777e22	/sls/X02DA/Data10/e17079/disk1/PE	12/02/2018 19:49
✖	20.500.11935/a9f22fd4-55aa-4aee-b224-35314a7af7ee	/sls/X02DA/Data10/e17079/disk1/PE	12/02/2018 19:24
✖	20.500.11935/ade97078-3a24-4ff8-823a-50d9d4ee886b	/sls/X02DA/Data10/e17079/disk1/PE	12/02/2018 11:19
✖	20.500.11935/b81840cd-d4a7-4e3e-bb66-a45a05ce3544	/sls/X02DA/Data10/e17079/disk1/PE	11/02/2018 04:12

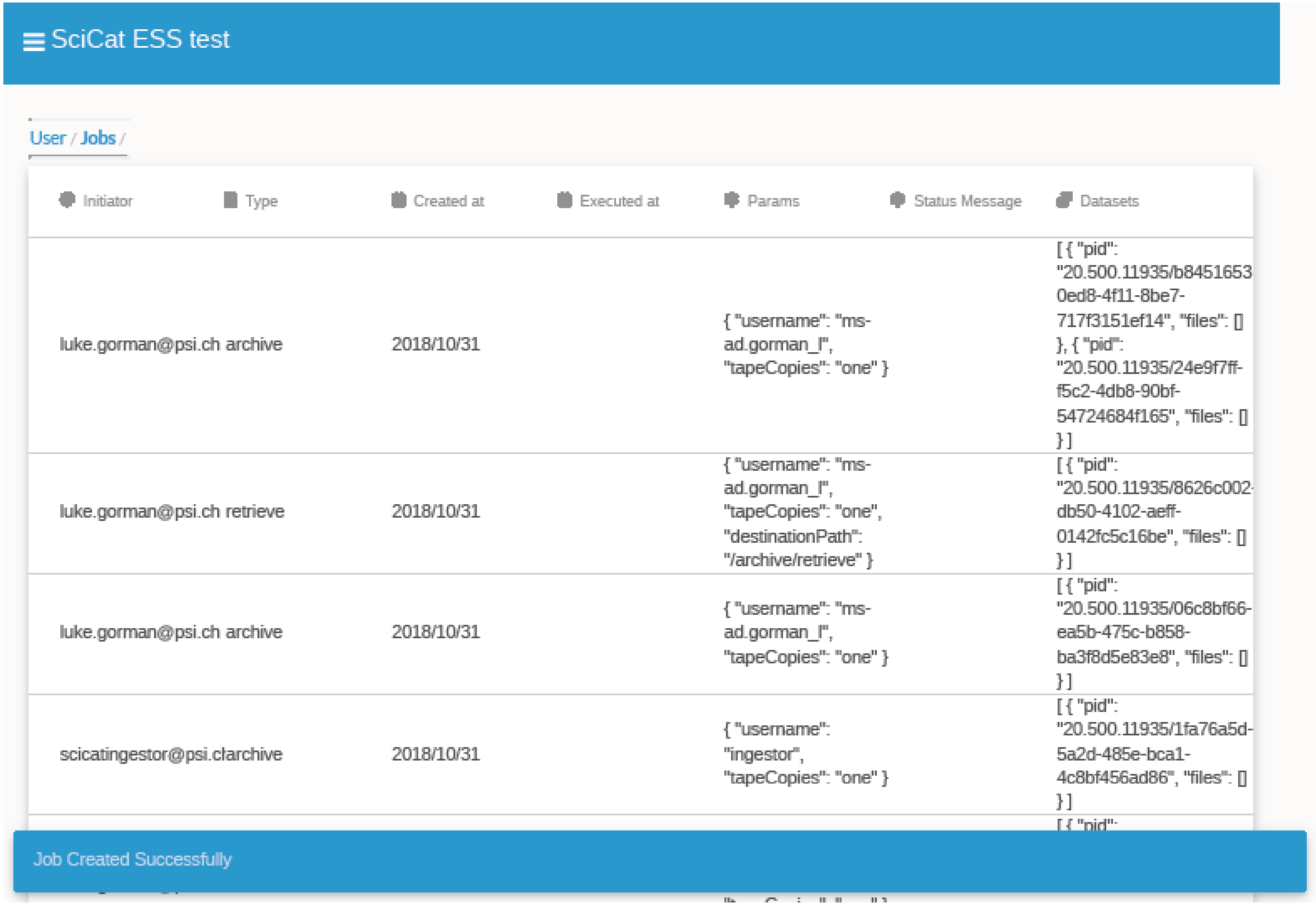
Job Created Successfully

Retrieve to Final Location

```
> datasetRetriever destinationPath
```

Only the user has write access to the destination path hence SciCat cannot move the data to this location.

Monitoring Archive/ Retrieve Jobs



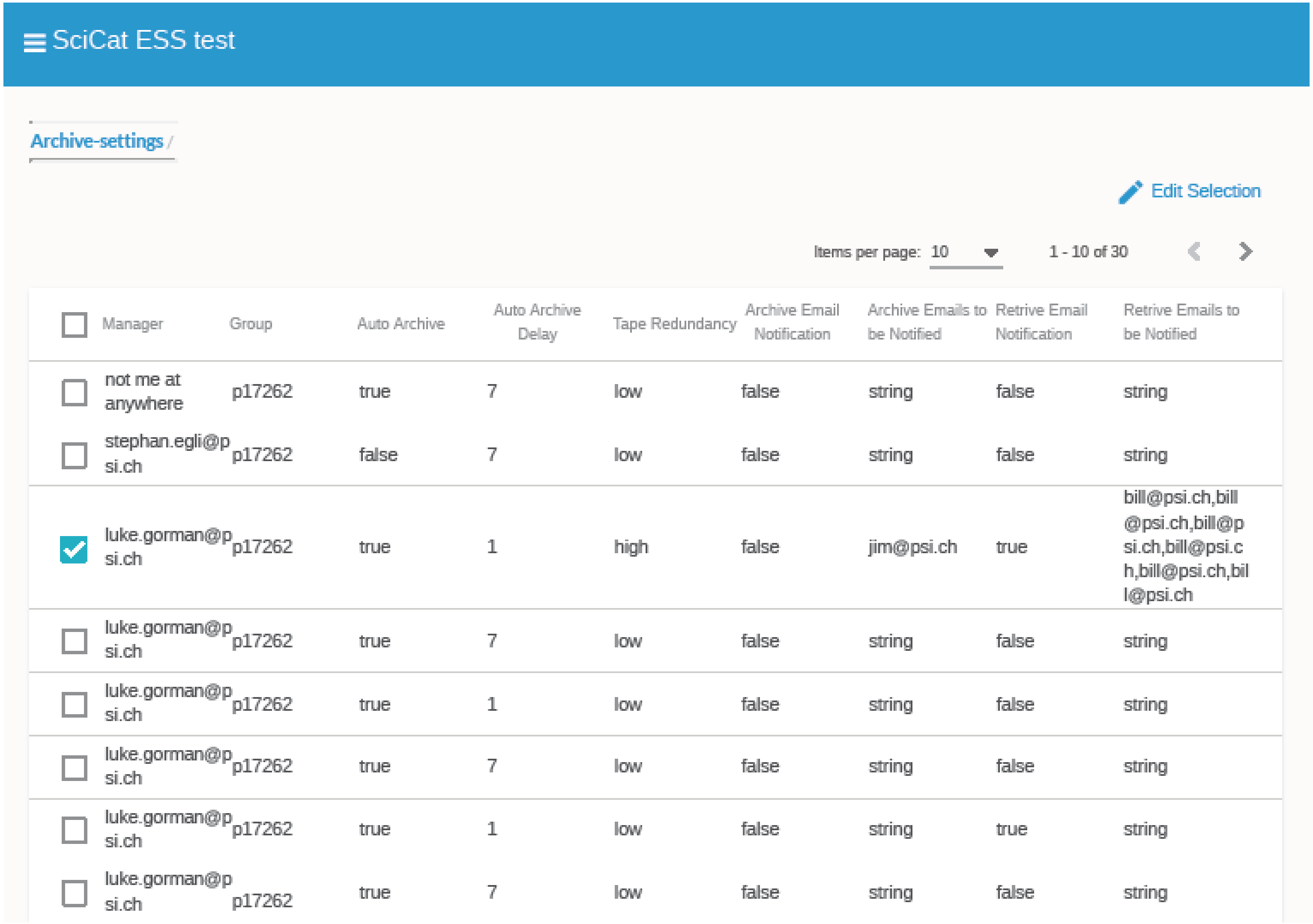
SciCat ESS test

User / Jobs /

Initiator	Type	Created at	Executed at	Params	Status Message	Datasets
luke.gorman@psi.ch	archive	2018/10/31		{ "username": "ms-ad.gorman_l", "tapeCopies": "one" }		[{ "pid": "20.500.11935/b8451653 0ed8-4f11-8be7-717f3151ef14", "files": [] }, { "pid": "20.500.11935/24e9f7ff-f5c2-4db8-90bf-54724684f165", "files": [] }]
luke.gorman@psi.ch	retrieve	2018/10/31		{ "username": "ms-ad.gorman_l", "tapeCopies": "one", "destinationPath": "/archive/retrieve" }		[{ "pid": "20.500.11935/8626c002-db50-4102-aeff-0142fc5c16be", "files": [] }]
luke.gorman@psi.ch	archive	2018/10/31		{ "username": "ms-ad.gorman_l", "tapeCopies": "one" }		[{ "pid": "20.500.11935/06c8bf66-ea5b-475c-b858-ba3f8d5e83e8", "files": [] }]
scicatingestor@psi.cl	archive	2018/10/31		{ "username": "ingestor", "tapeCopies": "one" }		[{ "pid": "20.500.11935/1fa76a5d-5a2d-485e-bca1-4c8bf456ad86", "files": [] }]
						[{ "pid":

Job Created Successfully

Configuring Group Policies



□	Manager	Group	Auto Archive	Auto Archive Delay	Tape Redundancy	Archive Email Notification	Archive Emails to be Notified	Retrive Email Notification	Retrive Emails to be Notified
□	not me at anywhere	p17262	true	7	low	false	string	false	string
□	stephan.egli@psi.ch	p17262	false	7	low	false	string	false	string
☑	luke.gorman@psi.ch	p17262	true	1	high	false	jim@psi.ch	true	bill@psi.ch,bill@psi.ch,bill@psi.ch,bill@psi.ch,bill@psi.ch,bill@psi.ch
□	luke.gorman@psi.ch	p17262	true	7	low	false	string	false	string
□	luke.gorman@psi.ch	p17262	true	1	low	false	string	false	string
□	luke.gorman@psi.ch	p17262	true	7	low	false	string	false	string
□	luke.gorman@psi.ch	p17262	true	1	low	false	string	true	string
□	luke.gorman@psi.ch	p17262	true	7	low	false	string	false	string

Problems/ Challenges

- Decision on what to archive:
  - collection of the metadata needs to be done at the beamline (needs local expertise). Some part of the metadata can however be augmented centrally e.g. linking to the proposal
  - what is the minimum meta data
  - what scientific metadata is available
- decision on when to archive (directly after taking data/ delayed after cleanup)
- wildly different DAQ systems and languages, tools, scripts at beamlines
- no unique folder structures and filename conventions between beamlines/ users
- support for all sorts of file types, not only HDF5 files
- Large historic datasets both in terms of overall size (36TB) and number of files (600k files) for a single dataset

Problems/ Challenges

- BL scientists have little time, but are willing to help getting things going
- some times compromise is needed: decision about separating ingest as two step option: without scientific metadata first, add scientific metadata later. Motivation born from need to archive fast because running out of disk space
- Invest time both in catalog and beamline side for initial discussion to explain the ingestion options. Ingest and retrieve Manual exists, but who reads manuals?
- Support for metadata creation.
- Q: how much data will come from existing sources and how much will be manually added by the beam line scientist? For the latter, additional metadata creation forms may be needed.