

DATA CURATION @



EUROPEAN
SPALLATION
SOURCE

Gareth Murphy

Experiment Control & Data Curation

2018-07-03



brightness

What is data curation?

- Data curation - organizing, integrating data and metadata, presenting and publishing, preserving and archiving
- Latin *cura animarum* - cure of souls



Powerful pulsed neutron source

- 17 Partner countries
- Construction work in progress

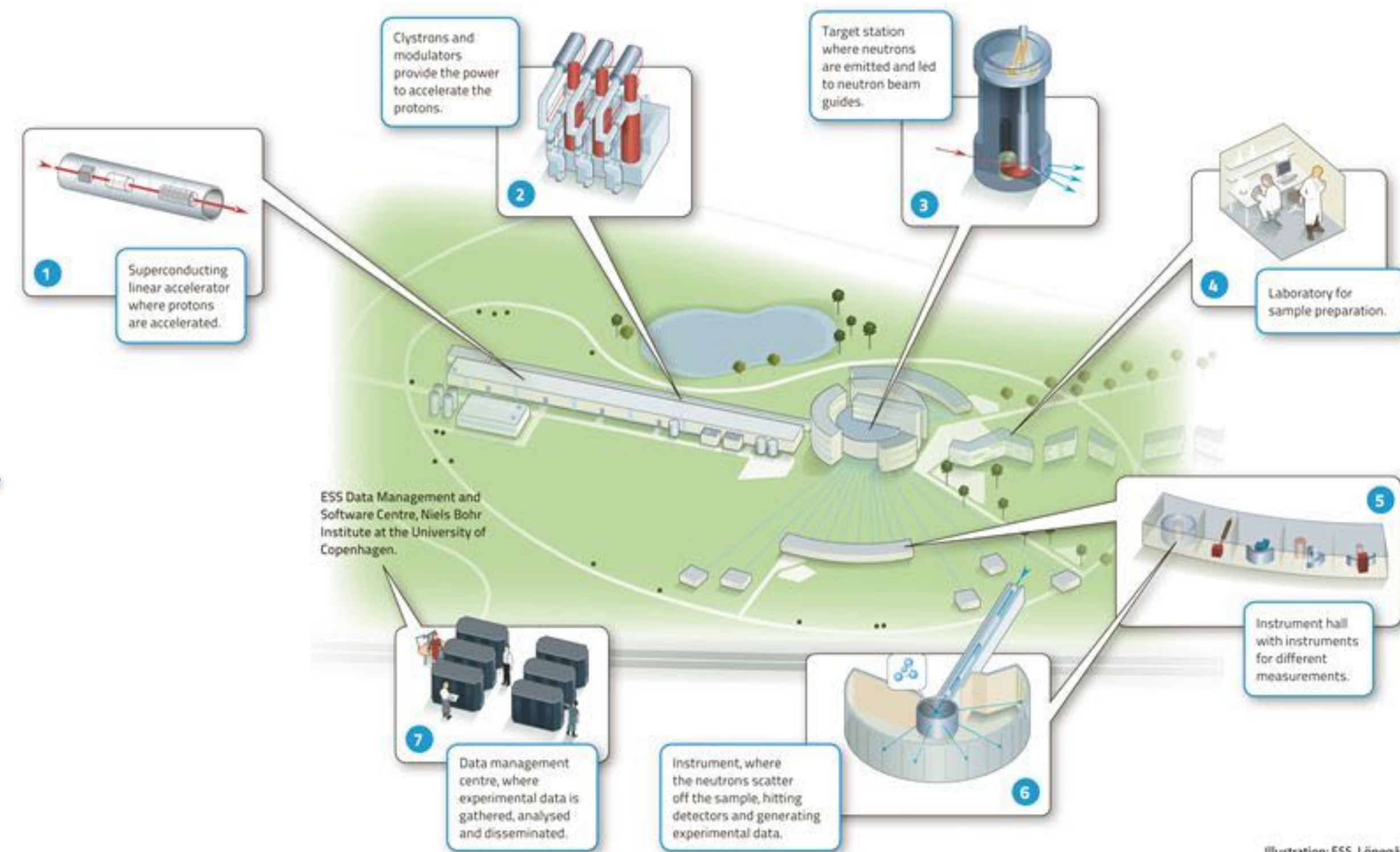
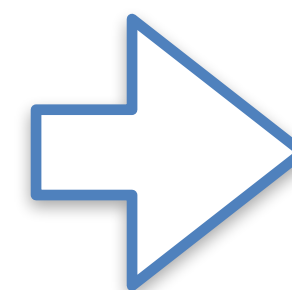
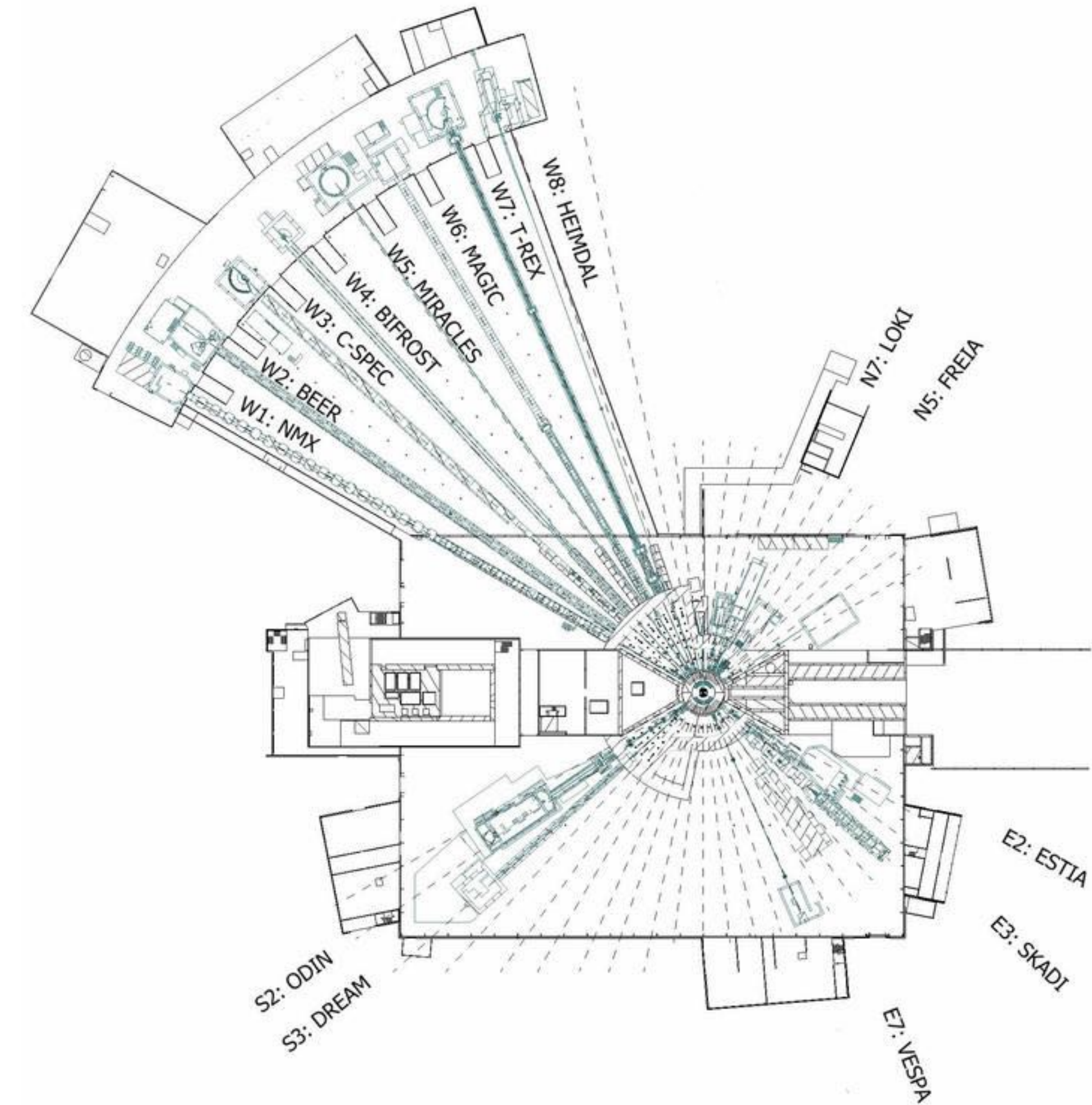
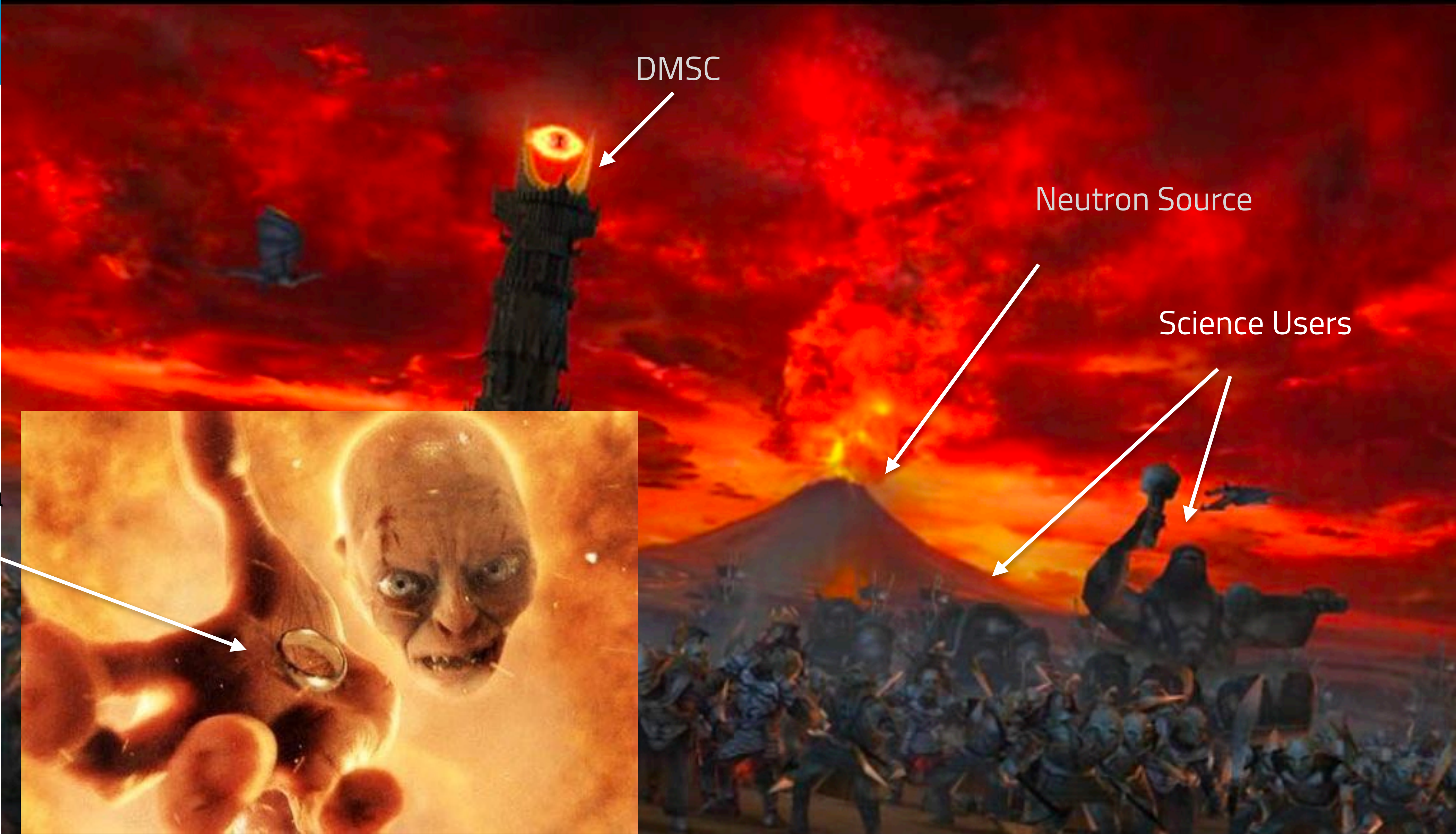


Illustration: ESS, Lönegård

- 15 instruments/beamlines
- Imaging, spectroscopy, diffraction
- Each instrument has different data requirements
- Traditionally, communities have had different data types, formats, analysis and reduction methods, standards - problem for data management
- By standardizing across instruments, we can make this process simpler and quicker



Lund 2025 ...



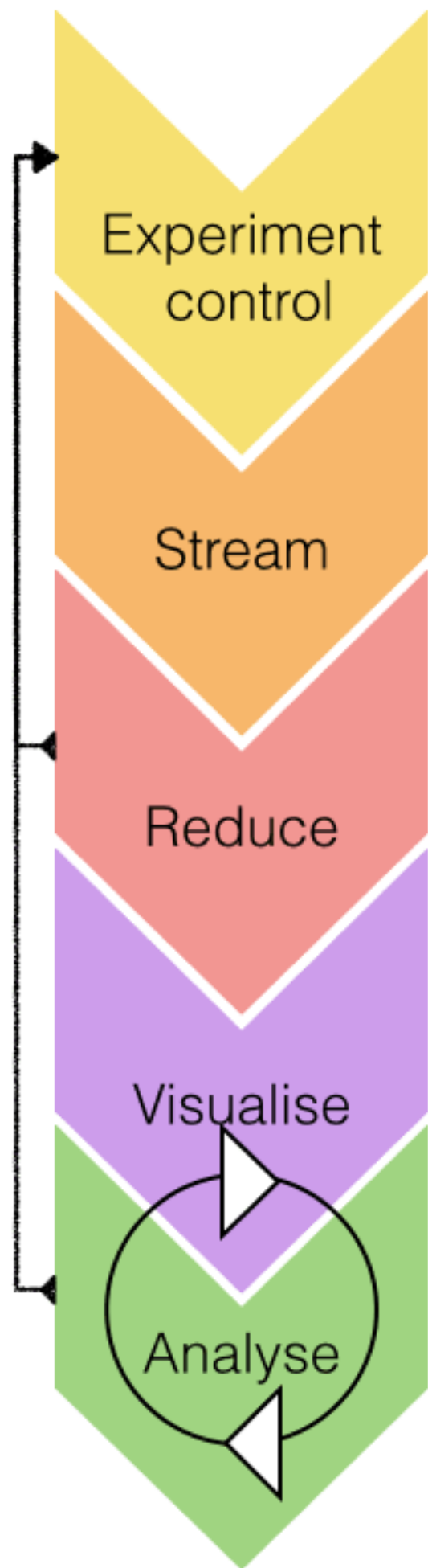
DMSC

Neutron Source

Science Users

Metadata

Data Management & Software Centre (DMSC)

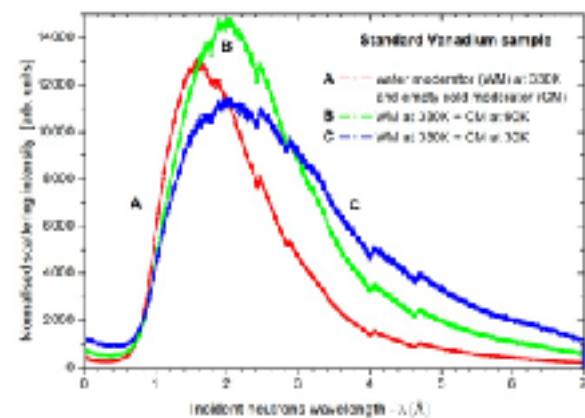


- DMSC - one team to rule the data
- Create uniform file writer for every beam line
- Connect data acquisition to data reduction and analysis
- Create/acquire metadata and send to data catalogue
- Owner + ORCID, time, wavelength, license, type



Raw, reduced and derived data

- Raw data - unprocessed data at full resolution, with communications artifacts removed (e.g. frame headers)
- Reduced - transformed and corrected from instrument units to physical units,
- Derived data - images, plots, statistics
- NASA define several processing levels raw = level 0, reduced = level 1, derived = level 2
- How to manage all this data?



- SciCat: Manages the **metadata** of raw and derived data which is taken at experiment facilities
- **administrative** metadata : data steward, data management lifecycle, file details, size etc
- **scientific** metadata: describing the sample, beamline and experiment parameters relevant for the users data analysis
- SciCat was developed at PSI as in-kind contribution to ESS



SciCat supports unstructured metadata

- Scientific needs can change between proposal writing time and experiment time
- Not all parameters are known until an experiment begins/ends
- As well as “known unknowns”, metadata allows for “unknown unknowns”

SciCat dashboard

SciCat ESS test

localhost:4200/datasets/10.17199%2FBRIGHTNESS%2FSONDE0018

SciCat ESS test

Datasets / 10.17199/BRIGHTNESS/SONDE0018 /

Details Datafiles Attachments Admin

PID	10.17199/BRIGHTNESS/SONDE0018
Owner	Ramsey Al Jebali
Contact Email	ramsey.aljebali@esss.se
Source Folder	/users/detector/experiments/sonde/IFE_june_2018/data/temp/t50
Size	651 MB
Creation Time	Jun 9, 2018
Type	base
Version	2.8.1
Owner Group	ess
Principal Investigator	Ramsey Al Jebali
Creation Location	SONDE

Scientific Metadata

```
{  
  "id": 3  
}
```

SciCat ESS test

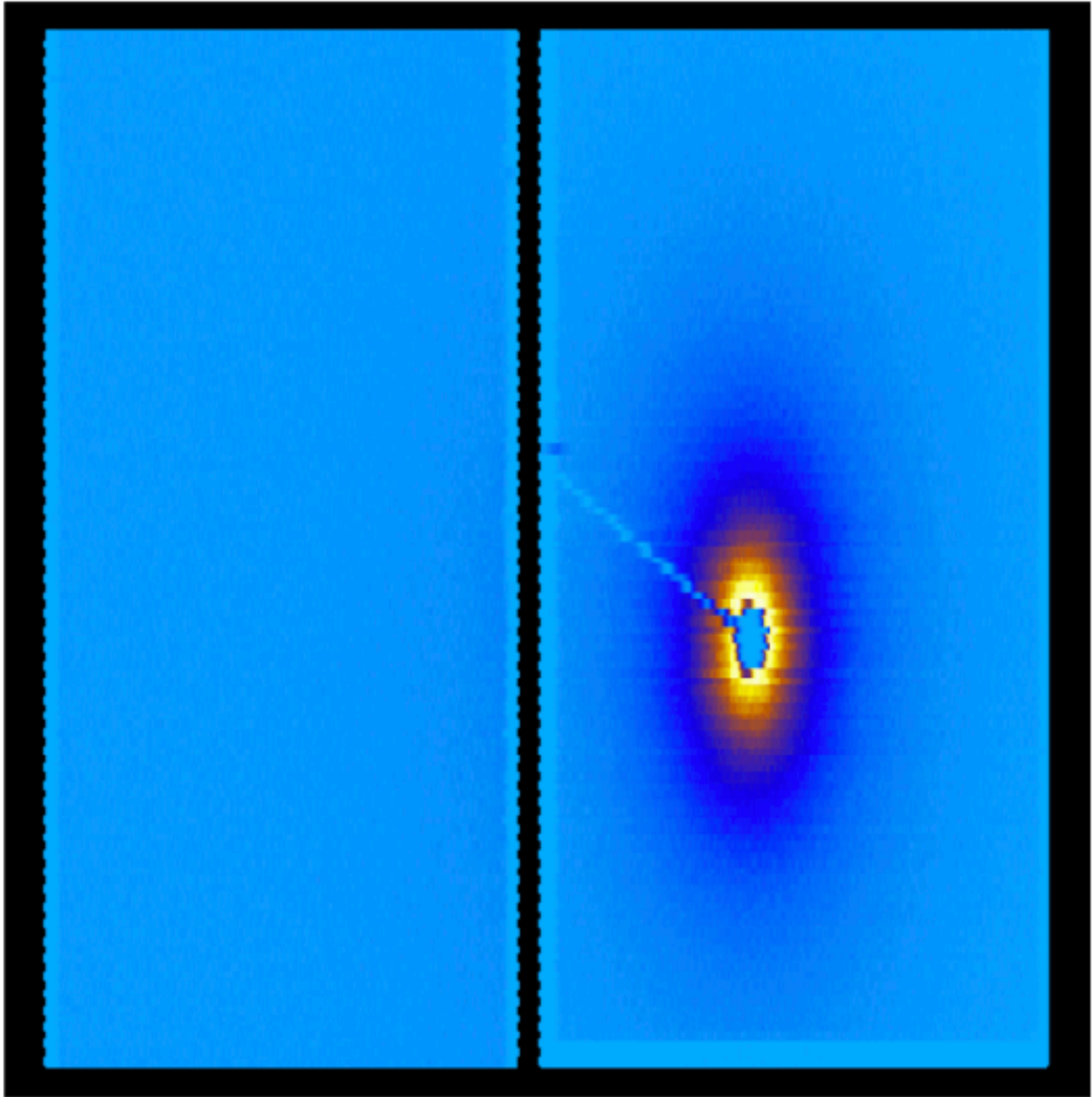
localhost:4200/datasets/10.17199%2FBRIGHTNESS%2FSONDE0018

SciCat ESS test

Datasets / 10.17199/BRIGHTNESS/SONDE0018 /

Details Datafiles Attachments Admin Export to CSV

Attachments



SciCat Team

ESS



Gareth Murphy



Lottie Greenwood

MAXIV



Hannes Petri

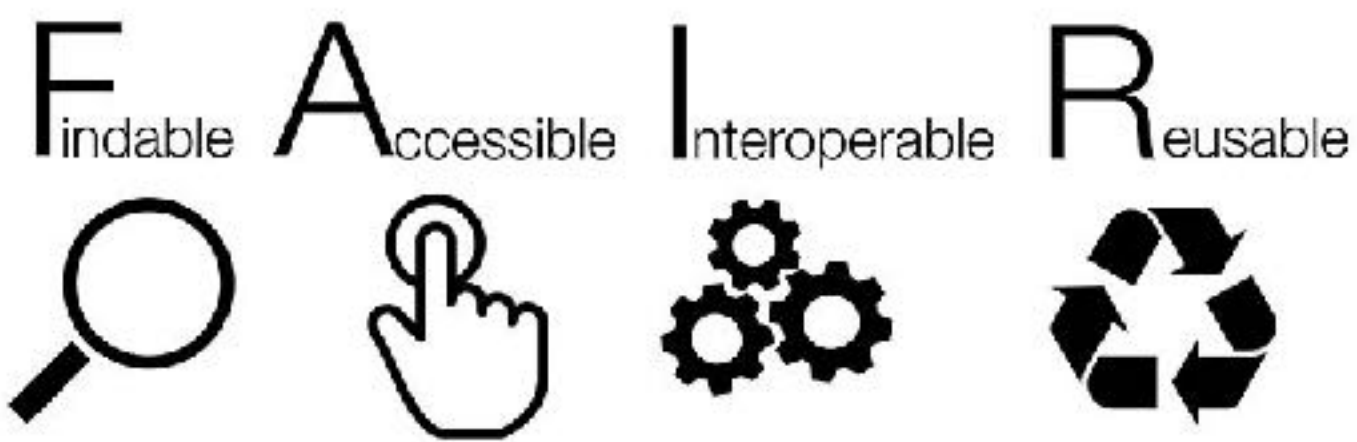
PSI



Stephan Egli



Luke Gorman



To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

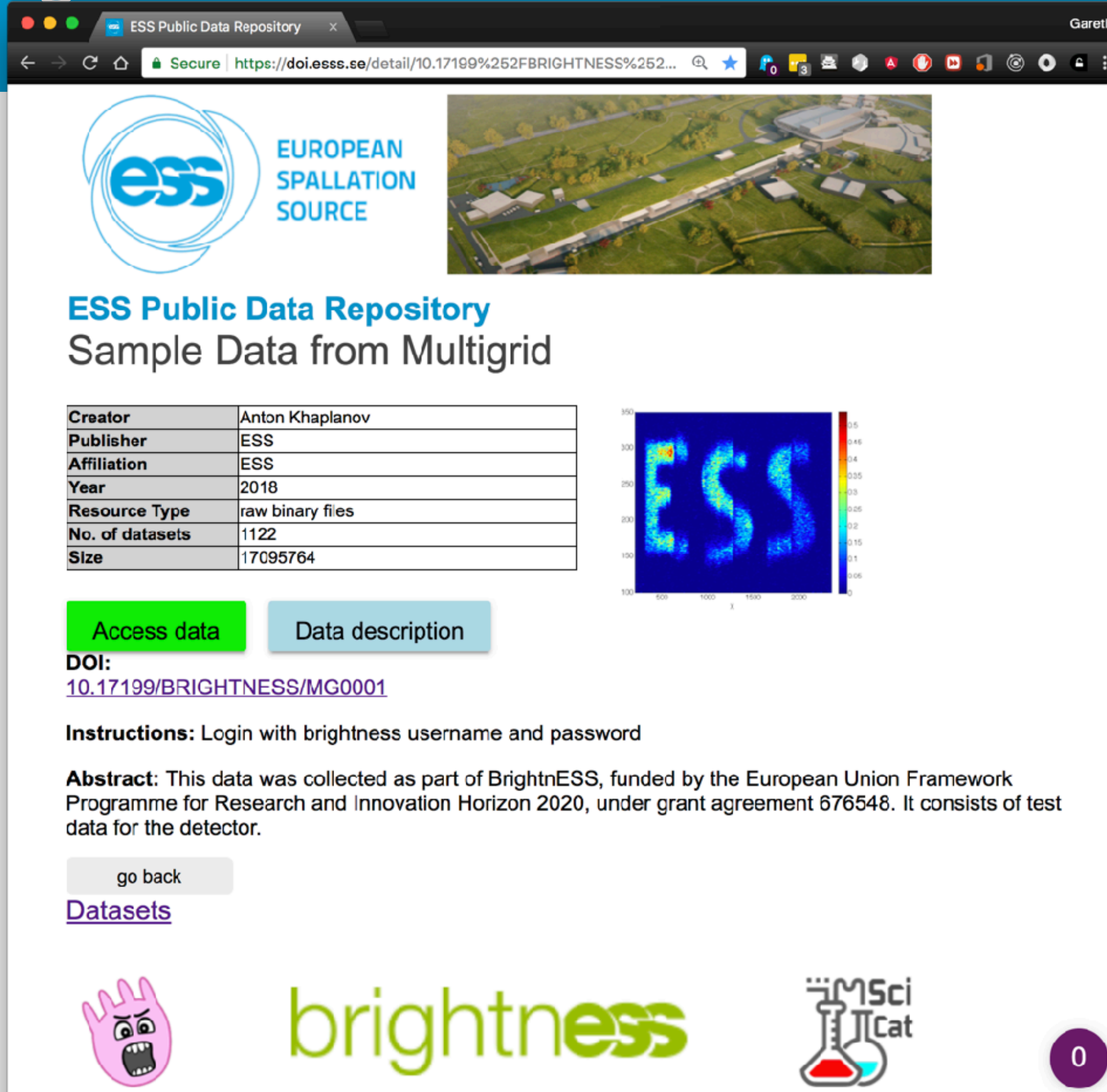
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards

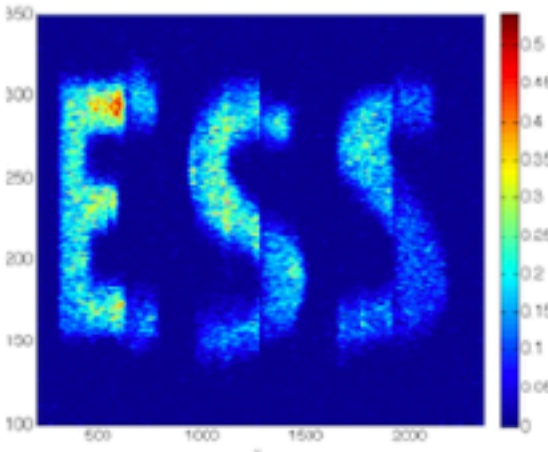


Landing page server



The screenshot shows a web browser window with the URL <https://doi.esss.se/detail/10.17199%252FBRIGHTNESS%252...>. The page features the ESS logo and an aerial view of the facility. The main title is "ESS Public Data Repository Sample Data from Multigrid". A metadata table is displayed, followed by "Access data" and "Data description" buttons. The DOI is [10.17199/BRIGHTNESS/MG0001](https://doi.org/10.17199/BRIGHTNESS/MG0001). Instructions and an abstract are provided, along with a "go back" button and a "Datasets" link. The footer contains logos for a pink hand character, "brightness", "SciCat", and a purple circle with the number "0".

Creator	Anton Khaplanov
Publisher	ESS
Affiliation	ESS
Year	2018
Resource Type	raw binary files
No. of datasets	1122
Size	17095764



[Access data](#) [Data description](#)





DOI:
[10.17199/BRIGHTNESS/MG0001](https://doi.org/10.17199/BRIGHTNESS/MG0001)

Instructions: Login with brightness username and password

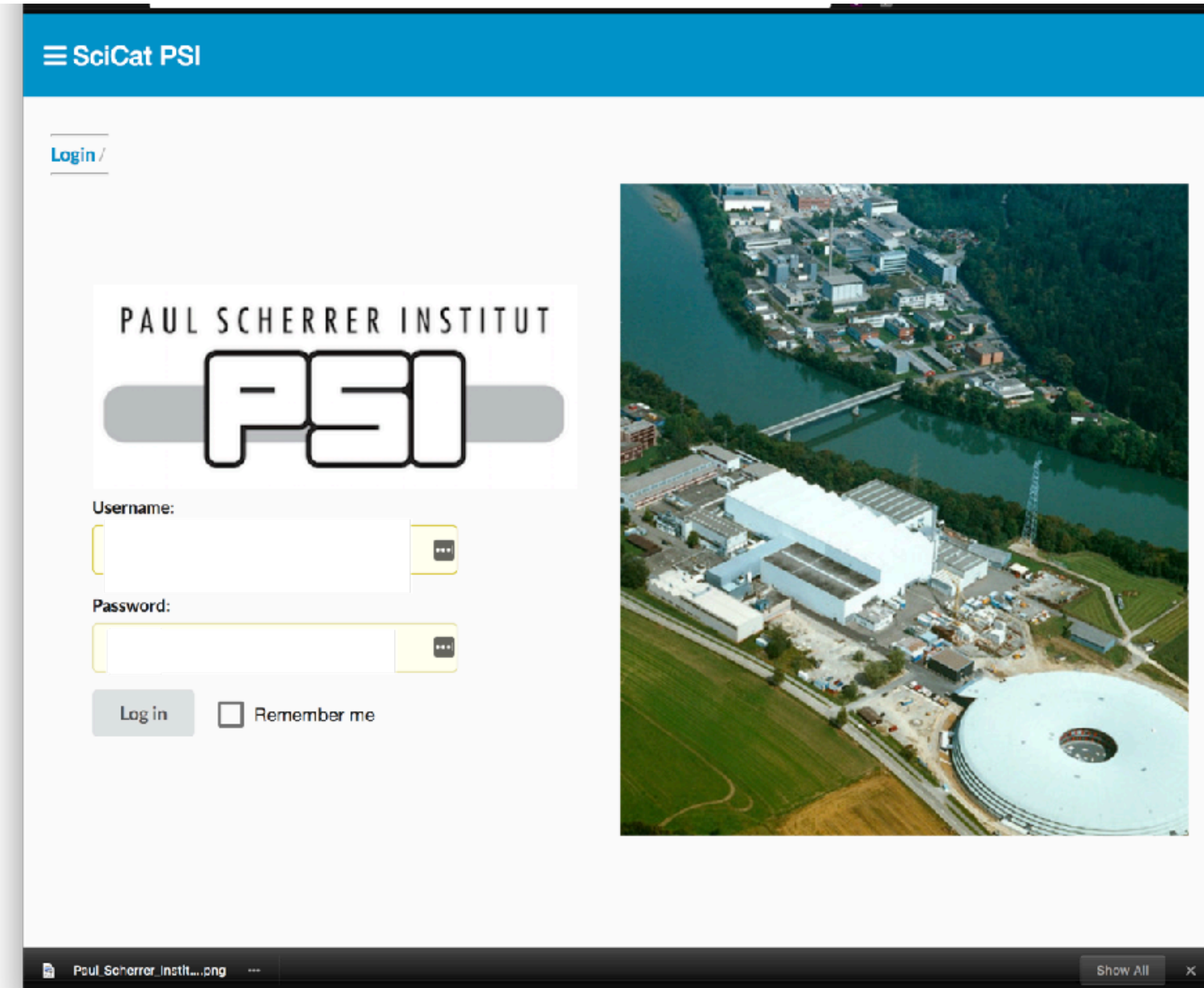
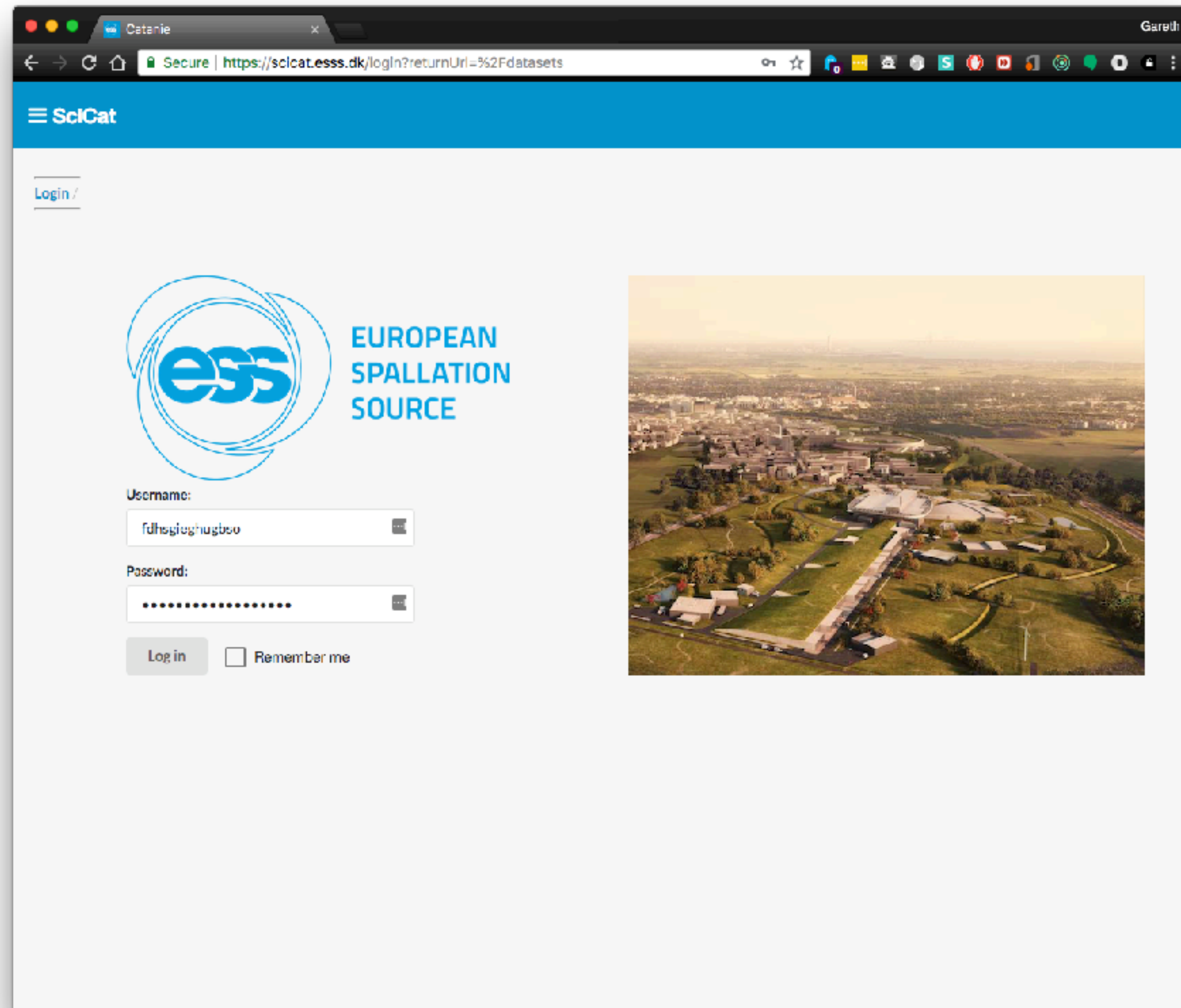
Abstract: This data was collected as part of BrightnESS, funded by the European Union Framework Programme for Research and Innovation Horizon 2020, under grant agreement 676548. It consists of test data for the detector.

[go back](#)

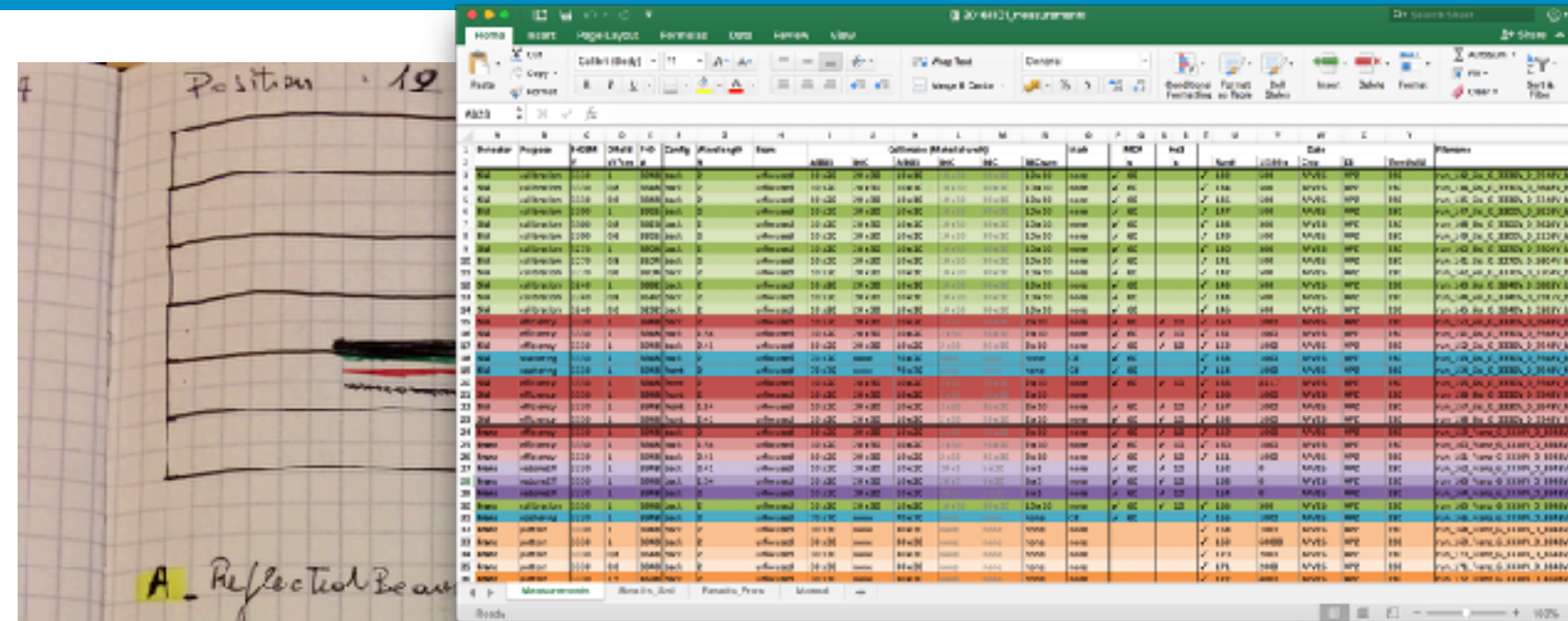
[Datasets](#)



Deployed at ESS, PSI and MAXIV



Capturing metadata at the beam line



Position 19

Beamline	Start	Stop	Current	Total	Time	Notes
A	167.05	166.0	167.05	166.0	10 min	Reflected Beam
B	167.05	166.0	167.05	166.0	10 min	Direct Beam
C	167.05	166.0	167.05	166.0	10 min	Direct Beam

A - Reflected Beam

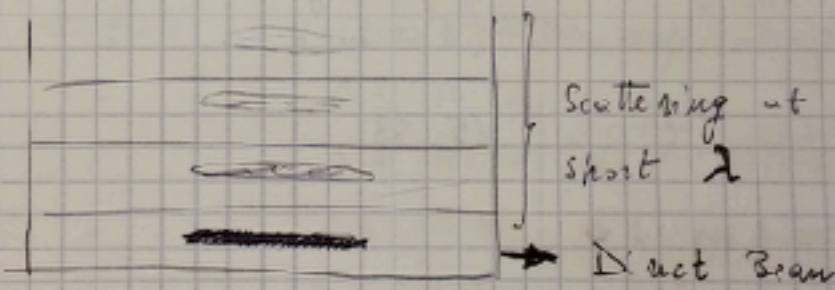
10 min. each

Direct Beam (Position = -16)

B - Direct Beam : Measurements without any sample
2 minutes (43 files)

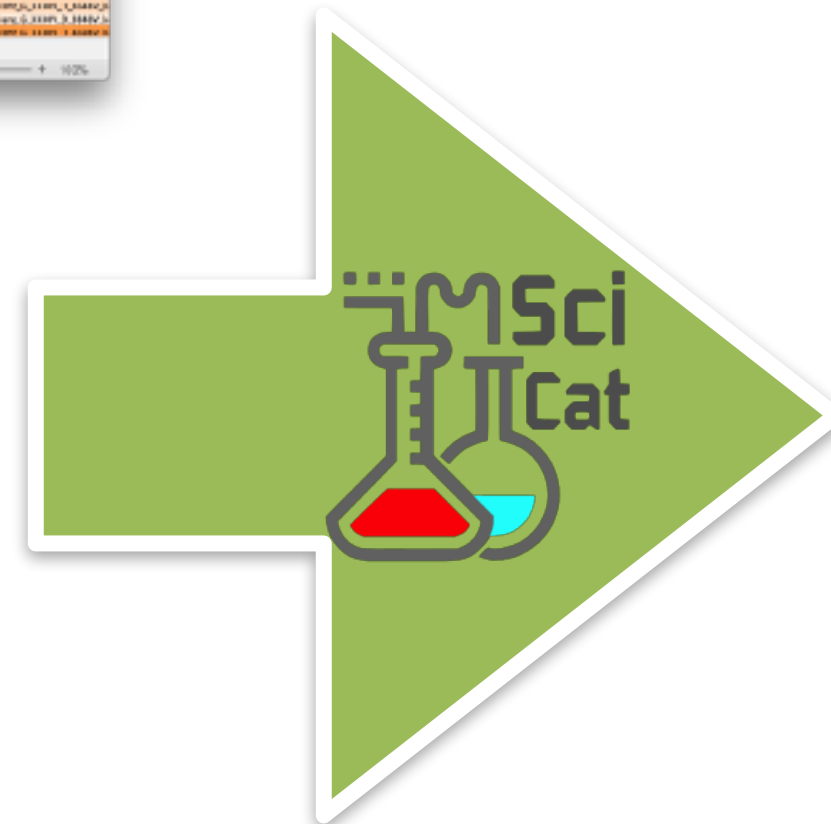
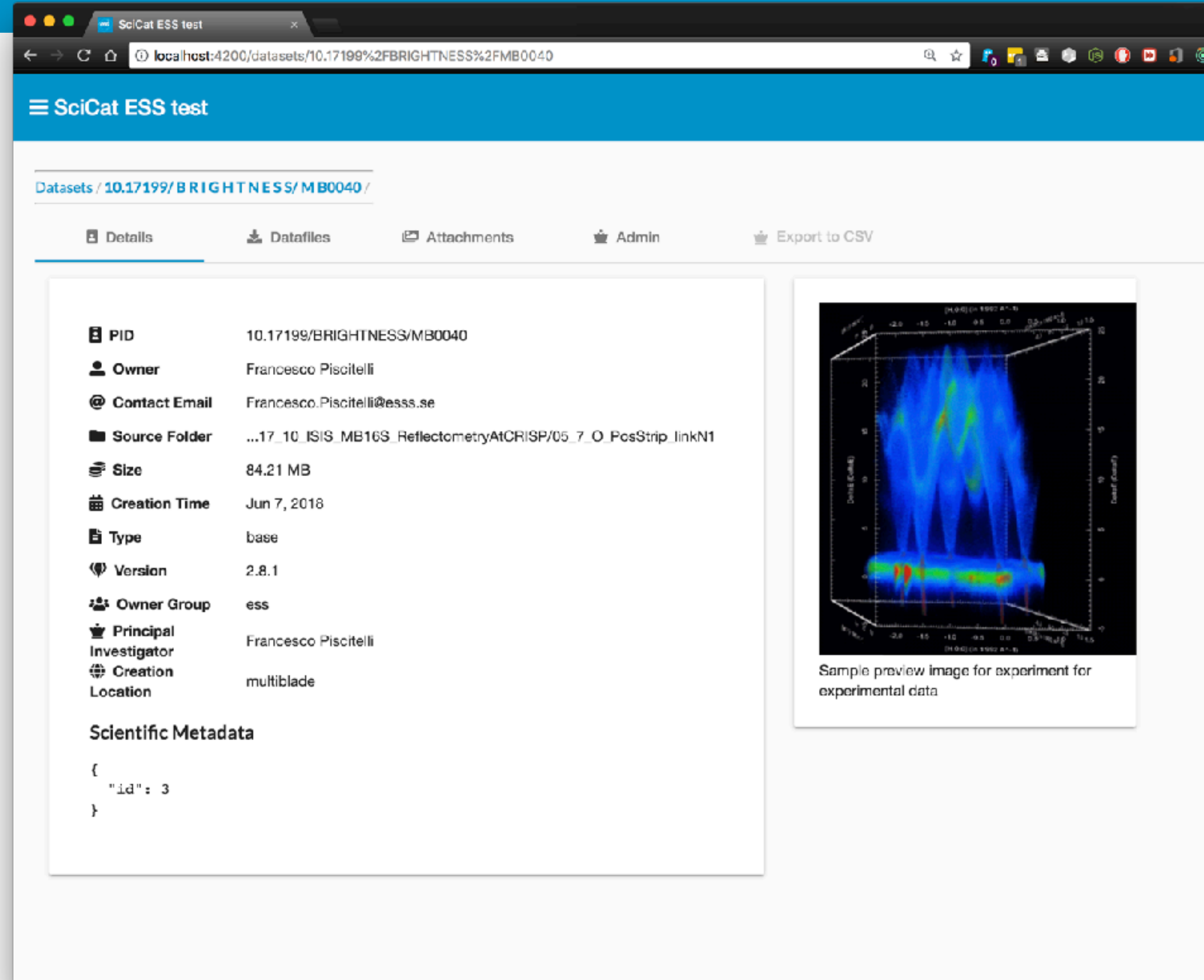
START (167.05 / 0.18) Current / Total (µA)
STOP (166 / 0.14) 36 min

B : Direct Beam measurements



C - Direct Beam : 5 mm Al plate in front of the window.
2 minutes

C : try thickness (167 / 2.82) START 10 min (5 files)
Al window (166.6 / 25.91) STOP
(166.0 / 1.63) START
(166 / 25.18) STOP 10 min

SciCat ESS test

localhost:4200/datasets/10.17199%2FBRIGHTNESS%2FMB0040

SciCat ESS test

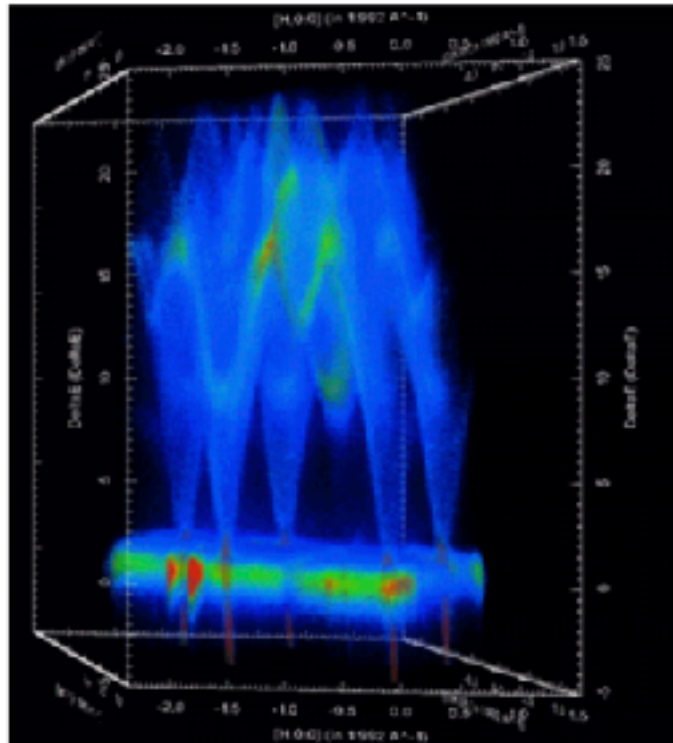
Datasets / 10.17199/BRIGHTNESS/MB0040/

Details | Datafiles | Attachments | Admin | Export to CSV

- PID**: 10.17199/BRIGHTNESS/MB0040
- Owner**: Francesco Piscitelli
- Contact Email**: Francesco.Piscitelli@ess.se
- Source Folder**: ...17_10_ISIS_MB16S_ReflectometryAtCRISP/05_7_O_PcsStrip_linkN1
- Size**: 84.21 MB
- Creation Time**: Jun 7, 2018
- Type**: base
- Version**: 2.8.1
- Owner Group**: ess
- Principal Investigator**: Francesco Piscitelli
- Creation Location**: multiblade

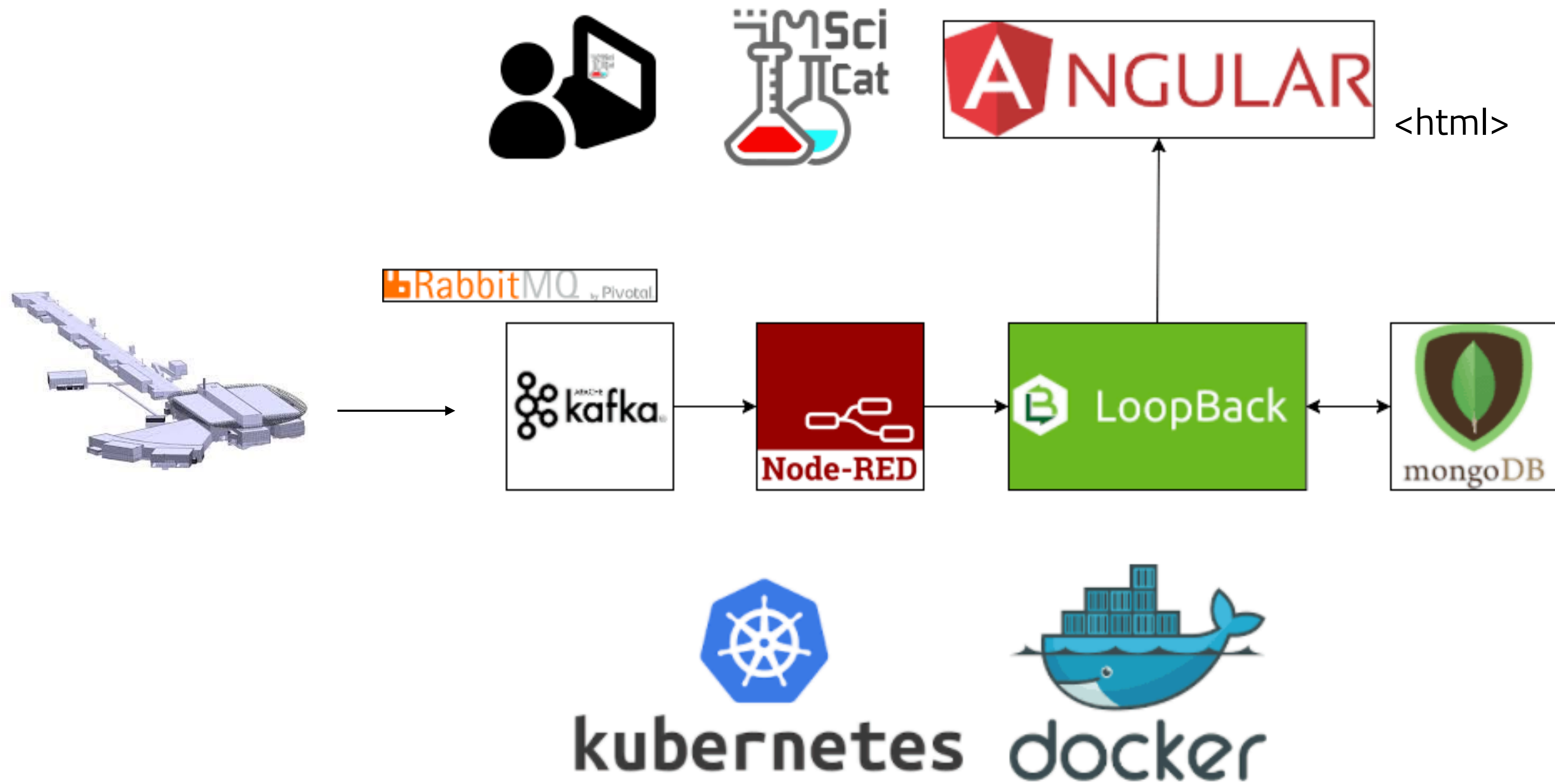
Scientific Metadata

```
{
  "id": 3
}
```

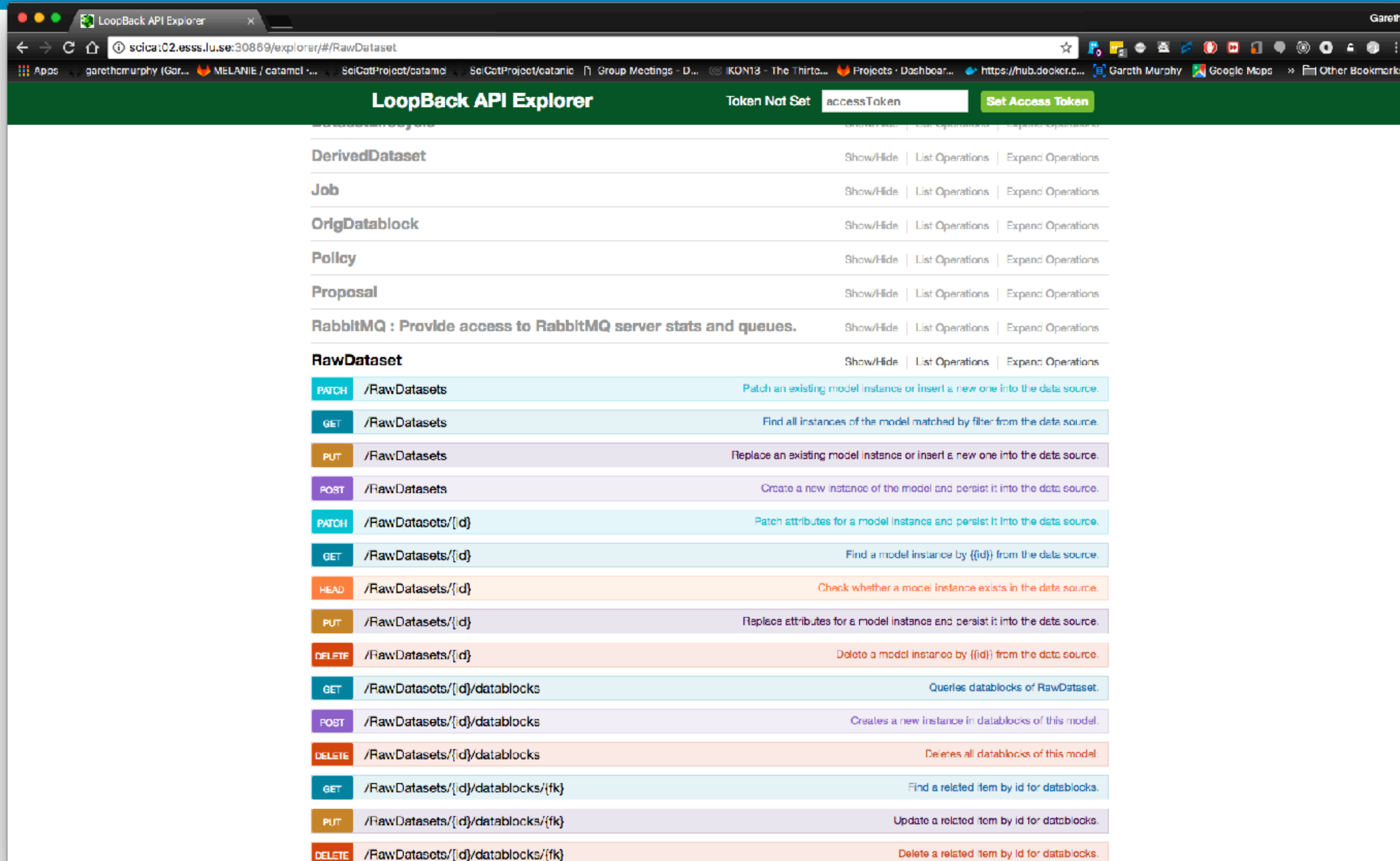


Sample preview image for experiment for experimental data

SciCat Architecture



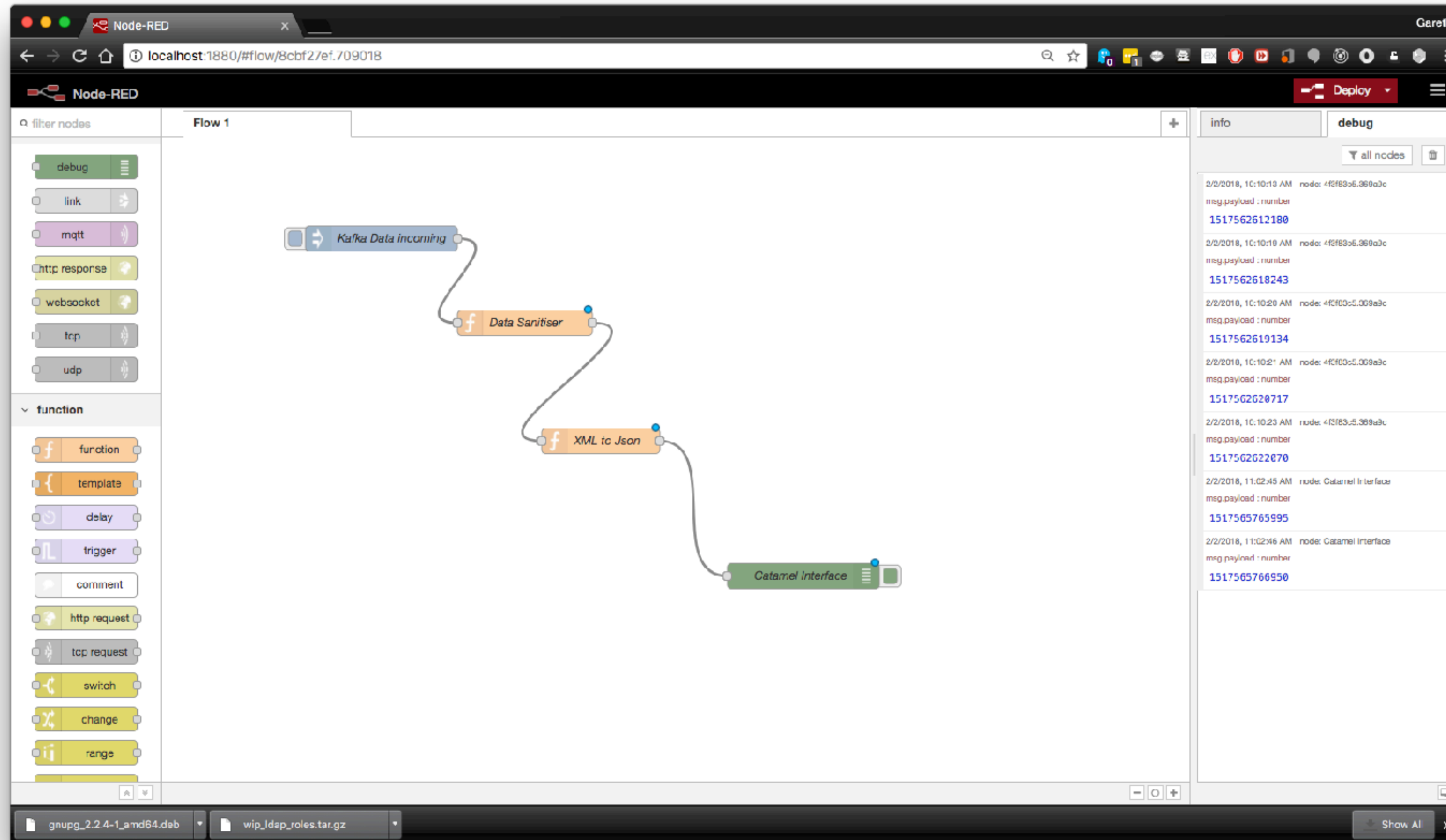
Loopback API Framework



The screenshot displays the LoopBack API Explorer interface. At the top, the title "LoopBack API Explorer" is visible, along with a "Token Not Set" indicator and a "Set Access Token" button. The main content area lists several models: DerivedDataset, Job, OrigDatablock, Policy, Proposal, RabbitMQ (with a description: "Provide access to RabbitMQ server stats and queues."), and RawDataset. The RawDataset model is expanded to show a list of API endpoints with their respective HTTP methods and descriptions.

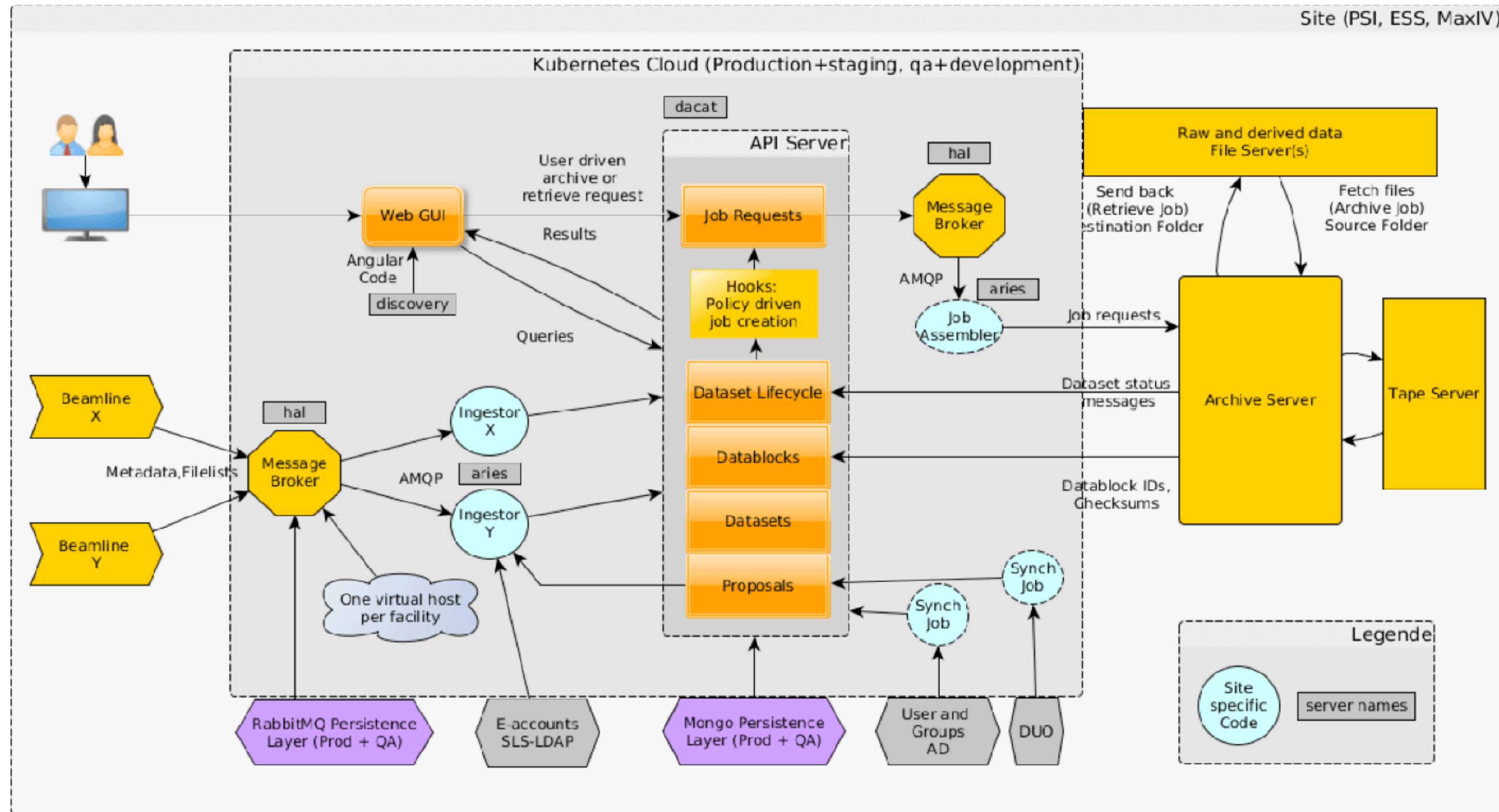
Method	Endpoint	Description
PATCH	/RawDatasets	Patch an existing model instance or insert a new one into the data source.
GET	/RawDatasets	Find all instances of the model matched by filter from the data source.
PUT	/RawDatasets	Replace an existing model instance or insert a new one into the data source.
POST	/RawDatasets	Create a new instance of the model and persist it into the data source.
PATCH	/RawDatasets/{id}	Patch attributes for a model instance and persist it into the data source.
GET	/RawDatasets/{id}	Find a model instance by {{id}} from the data source.
HEAD	/RawDatasets/{id}	Check whether a model instance exists in the data source.
PUT	/RawDatasets/{id}	Replace attributes for a model instance and persist it into the data source.
DELETE	/RawDatasets/{id}	Delete a model instance by {{id}} from the data source.
GET	/RawDatasets/{id}/datablocks	Queries datablocks of RawDataset.
POST	/RawDatasets/{id}/datablocks	Creates a new instance in datablocks of this model.
DELETE	/RawDatasets/{id}/datablocks	Deletes all datablocks of this model.
GET	/RawDatasets/{id}/datablocks/{fk}	Find a related item by id for datablocks.
PUT	/RawDatasets/{id}/datablocks/{fk}	Update a related item by id for datablocks.
DELETE	/RawDatasets/{id}/datablocks/{fk}	Delete a related item by id for datablocks.

Node-Red flow editor



The screenshot shows the Node-RED web interface in a browser window. The address bar shows the URL `localhost:1880/#flow/8cbf2/ef.709018`. The interface is divided into several sections:

- Left Panel (Nodes):** A sidebar with a search bar and a list of nodes. The 'function' category is expanded, showing nodes like 'function', 'template', 'delay', 'trigger', 'comment', 'http request', 'tcp request', 'switch', 'change', and 'range'.
- Flow Editor (Center):** A workspace titled 'Flow 1' containing a flow with four nodes connected in sequence:
 - Kafka Data incoming:** A blue node that receives data from Kafka.
 - Data Sanitiser:** An orange function node that processes the incoming data.
 - XML to Json:** An orange function node that converts XML data to JSON.
 - Catamel interface:** A green node that outputs the processed data.
- Right Panel (Info/Debug):** A panel with tabs for 'info' and 'debug'. The 'debug' tab is active, showing a list of messages with their timestamps, node IDs, and payloads. The payloads are numbers: 1517562612180, 1517562618243, 1517562619134, 1517562620717, 1517562622070, 1517565765995, and 1517565766950.



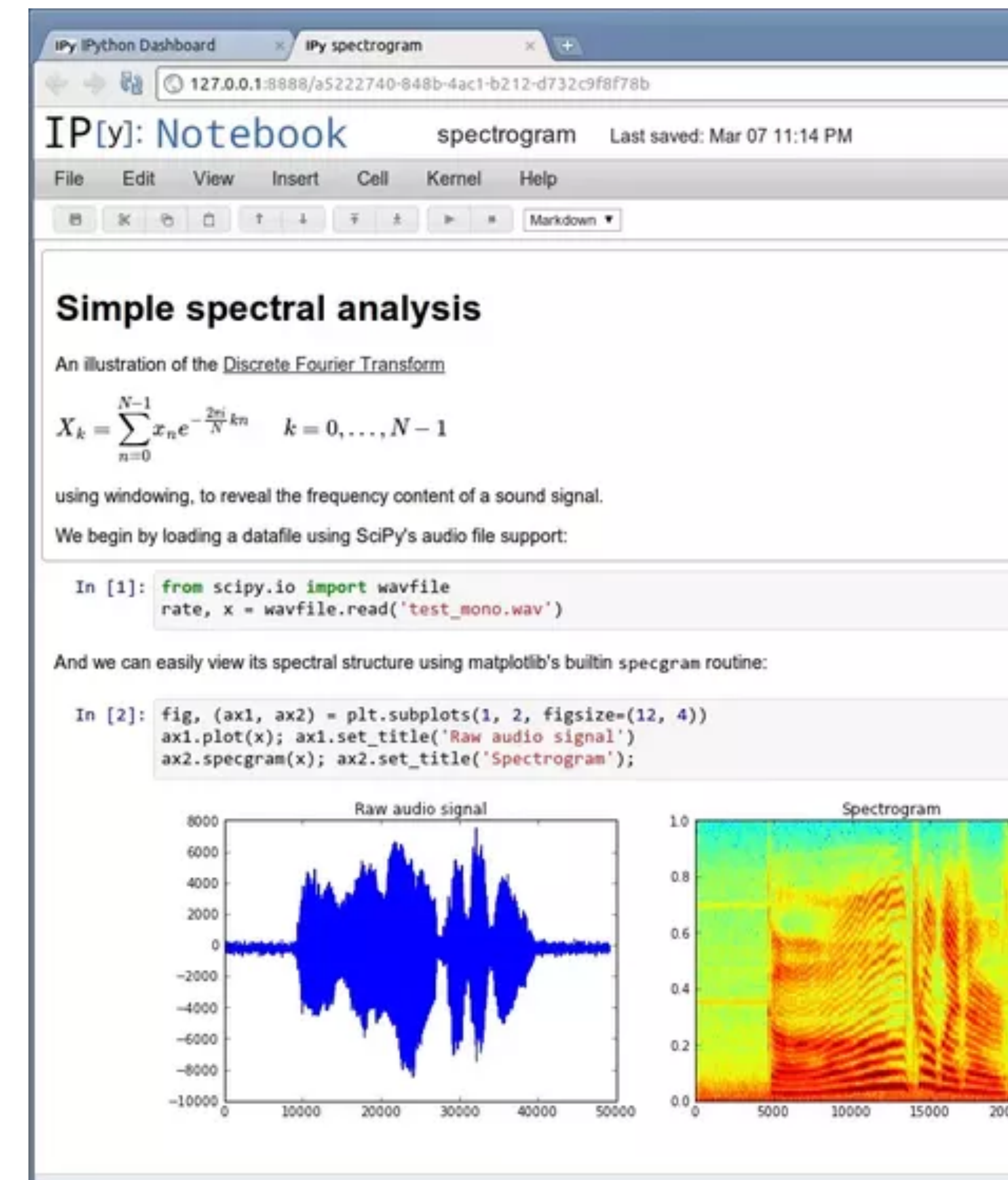
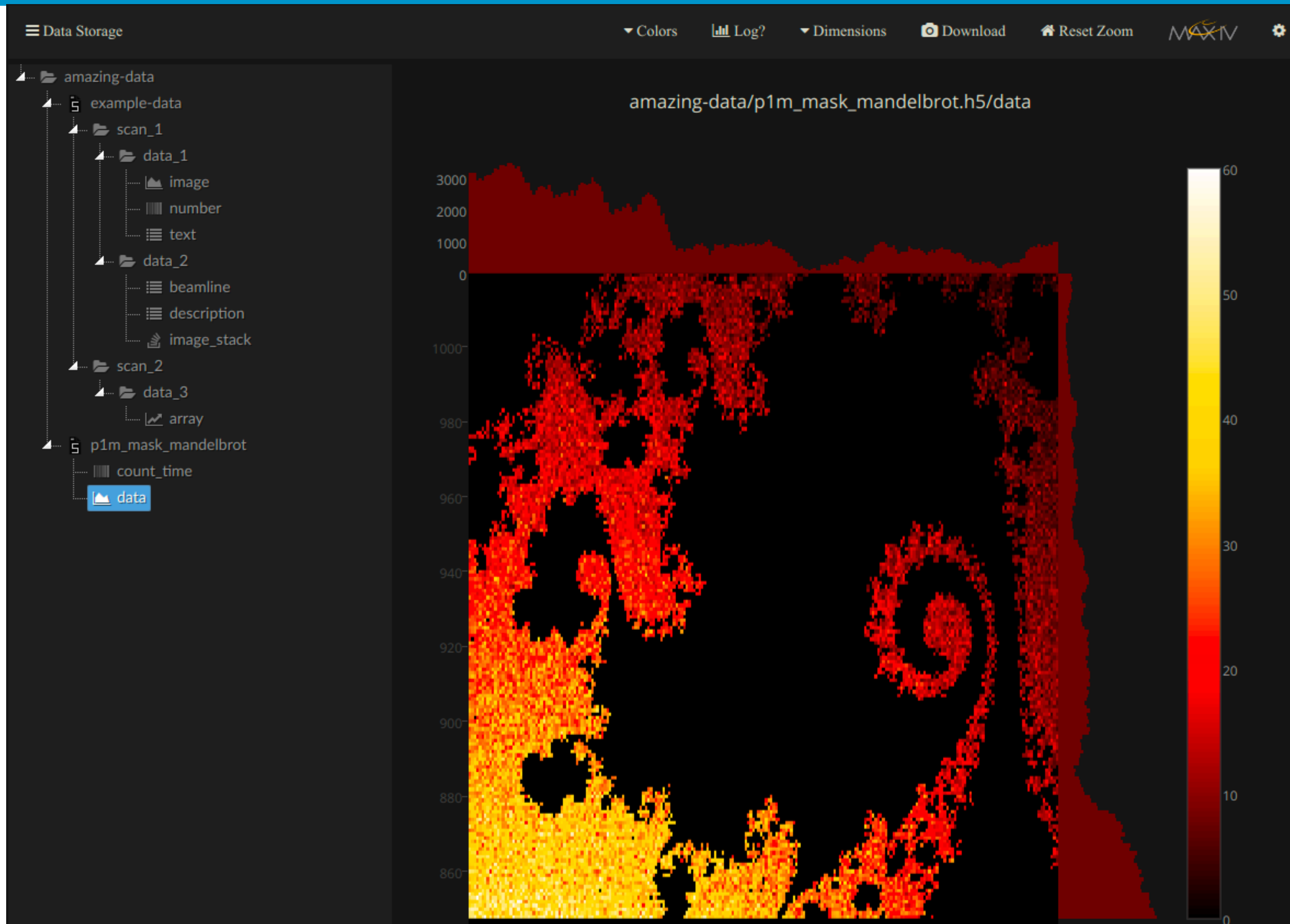
- Digital Object Identifier (DOI) must connect to accessible landing page, which displays metadata
- <https://doi.org/10.17199/BRIGHTNESS/SONDE0001>
- Landing page server
- Users should be able to make their data public and acquire a DOI and landing page

Photon and Neutron Open Science Cloud (PANOSC)






- FAIR - PaNOSC will comply with the FAIR principles in the following ways:
- Findable - all data will have a DOI, rich metadata, common api for federated search
- Accessible - api will support open protocol, metadata accessible even without data
- Inter-operable - metadata to follow community standards (Nexus), register metadata
- Reusable - follow community standardise metadata, clear licence (CC-BY)

Future additions



Thank you



- Client app so less strain on server - most of load on your browser   
- Kubernetes allows autoscaling - if we have enough CPUs

Updating microservices

- MongoDB, Loopback slow release cycle
- Kubernetes quarterly - kubectl upgrade
- Angular 6 month cycle - ng upgrade

- Angular, supported by Google
- Loopback, supported by IBM

Researcher persistent identifier



- ORCID
- Can uniquely identify researcher using instruments
- Can follow data use and citations
- Data creator/steward can be identified uniquely

The screenshot shows the ORCID search results page. The browser address bar shows the URL: <https://orcid.org/orcid-search/quick-search/?searchQuery='european%20spalla...'>. The page header includes the ORCID logo and navigation links: FOR RESEARCHERS, FOR ORGANIZATIONS, ABOUT, HELP, and SIGN IN. Below the header, it states "4,963,023 ORCID iDs and counting. See more...". The search results are displayed as a table with the following columns: ORCID ID, First/given name, Last/family name, Other names, and Affiliations.

ORCID ID	First/given name	Last/family name	Other names	Affiliations
https://orcid.org/0000-0002-0078-0372	Nikolaos	Gazis		European Spallation Source
https://orcid.org/0000-0003-3797-0476	Konstantin	Batkov		European Spallation Source
https://orcid.org/0000-0001-5371-9199	Emanuele	Laface		European Spallation Source
https://orcid.org/0000-0003-0175-179X	Javier	Cereijo Garcia		European Spallation Source, European Spallation Source AB, University of Vigo, University of A Coruña, University of Santiago de Compostela
https://orcid.org/0000-0002-0206-0387	Susan	Everett	S. M. Everett	Oak Ridge National Laboratory, European Spallation Source ERIC, University of Tennessee
https://orcid.org/0000-0002-2109-1226	Mads	Bertelsen		
https://orcid.org/0000-0003-1875-4700	Chung-Chuan	Lai		European Spallation Source ESS AB, Linköping University, National Tsing-Hua University
https://orcid.org/0000-0001-8688-4238	Masatoshi	Arai		European Spallation Source ESS AB, Tohoku University
https://orcid.org/0000-0001-5434-3728	Morten	Sales		
https://orcid.org/0000-0002-7015-1053	Mats	Lindroos		Lunds Universitet, European Spallation Source ESS AB, European Organization for Nuclear Research, Chalmers tekniska högskola
https://orcid.org/0000-0001-8287-0269	Zoe	Fisher		European Spallation Source ESS AB, Los Alamos National Laboratory, University of Florida, University of Stellenbosch