

PIDs, Petabytes and Neutrons

Gareth Murphy

Data Curation Scientist

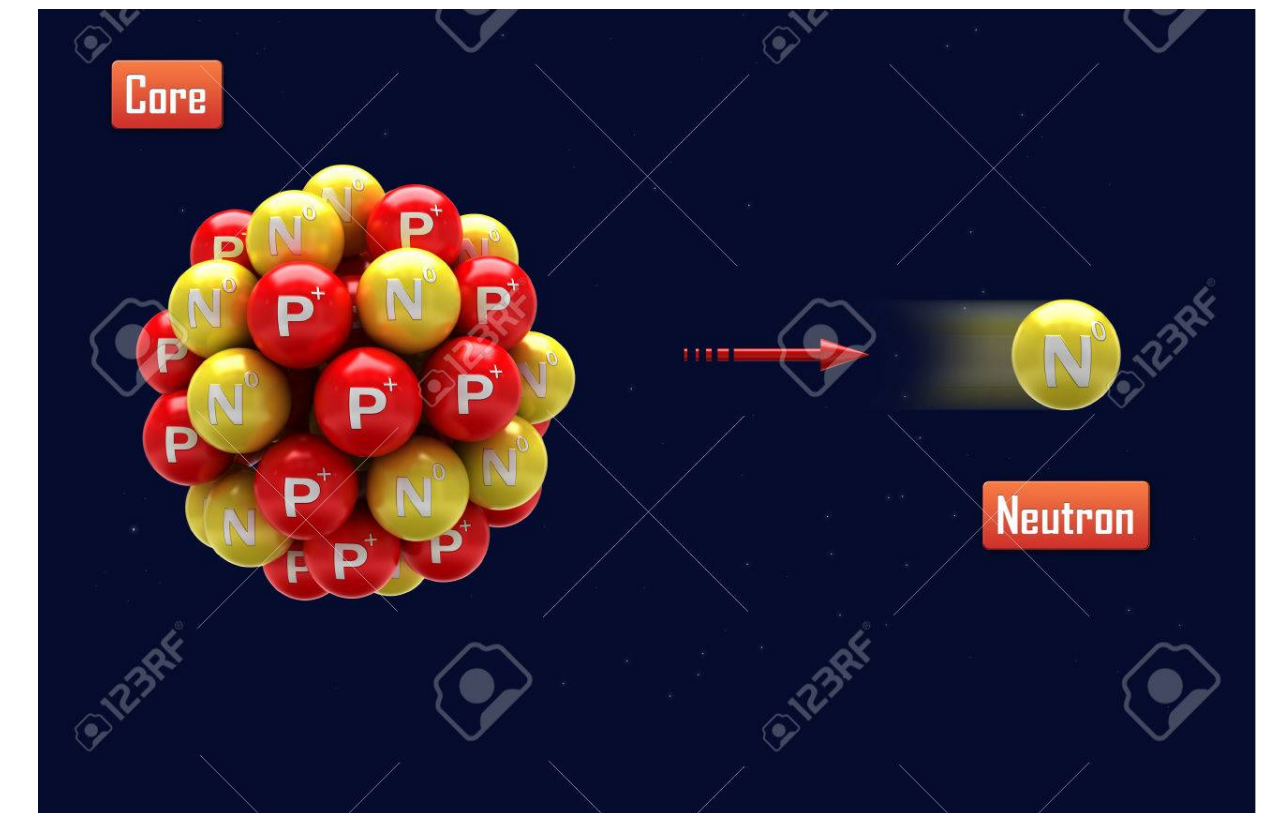
European Spallation Source

 orcid.org/0000-0002-2785-3674

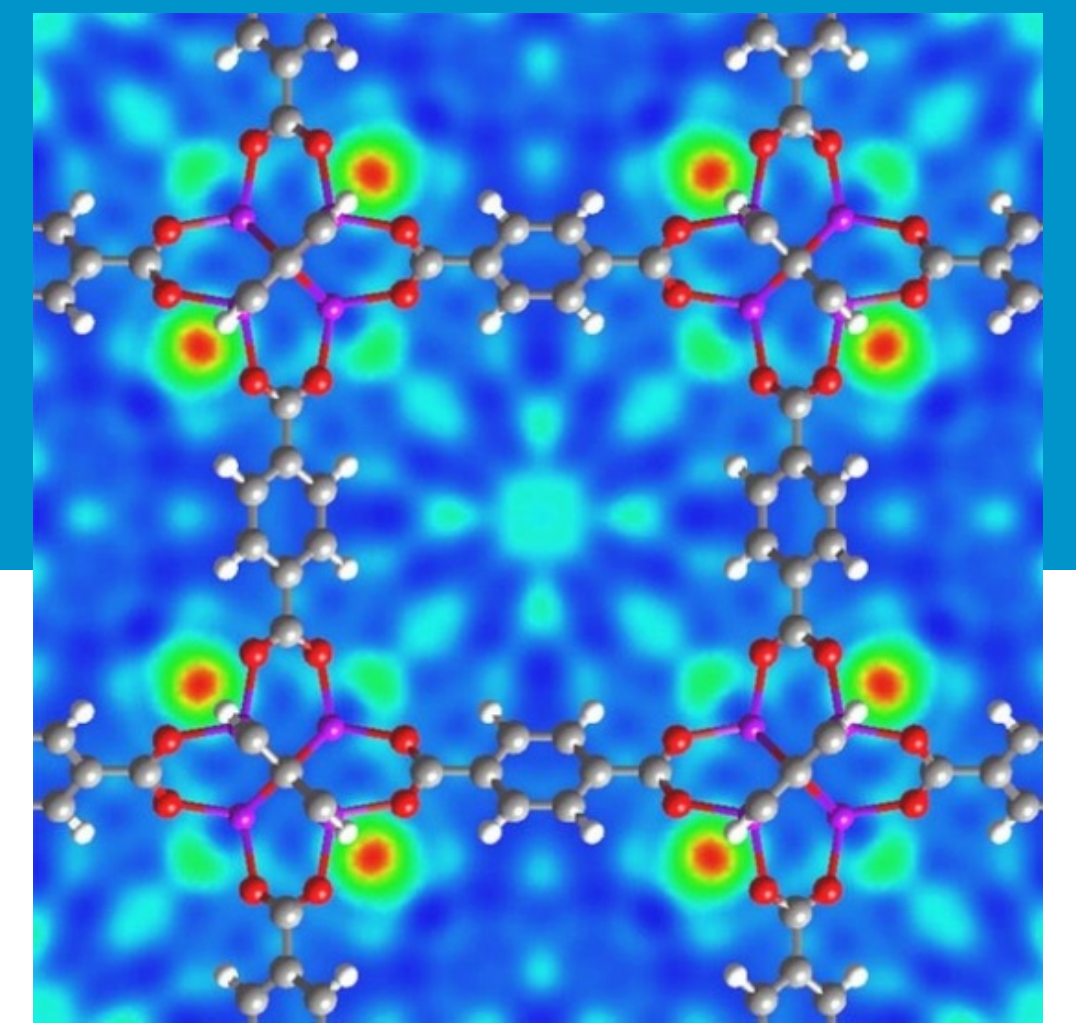


A few definitions

- *Spallation* - break-up of nucleus from the Middle English word *spall* meaning fragment
- *Data curation* - taking care of data - from Medieval Latin *cura animarum* - care of souls

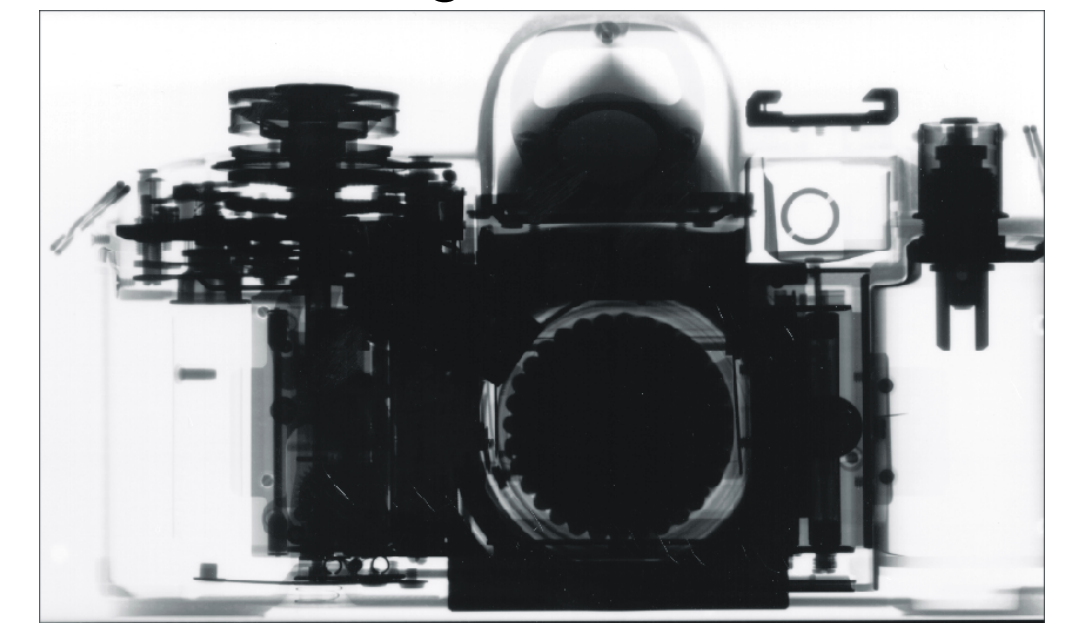


The European Spallation Source

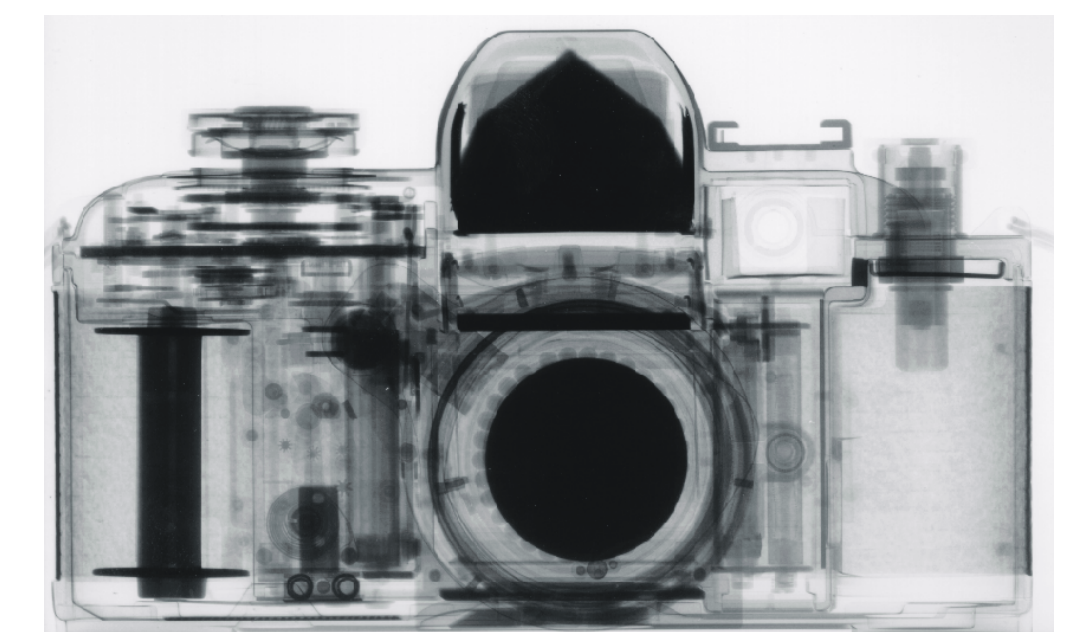


Neutron scattering of hydrogen in a metal organic framework

- An accelerator-based neutron source being built in Lund, southern Sweden
 - Material and life sciences research
- A collaboration of 15 European nations
 - Construction budget about 1860 million Euro
- Targeted to be the world's most powerful neutron source
 - 5 MW beam power, 2.5 GeV proton energy, 14 Hz repetition rate, 2.86 ms pulse@50 mA beam current
 - 22 neutron beam lines in construction budget
- First neutrons in 2020, full configuration in 2025



X-Ray Image



Neutron radiograph

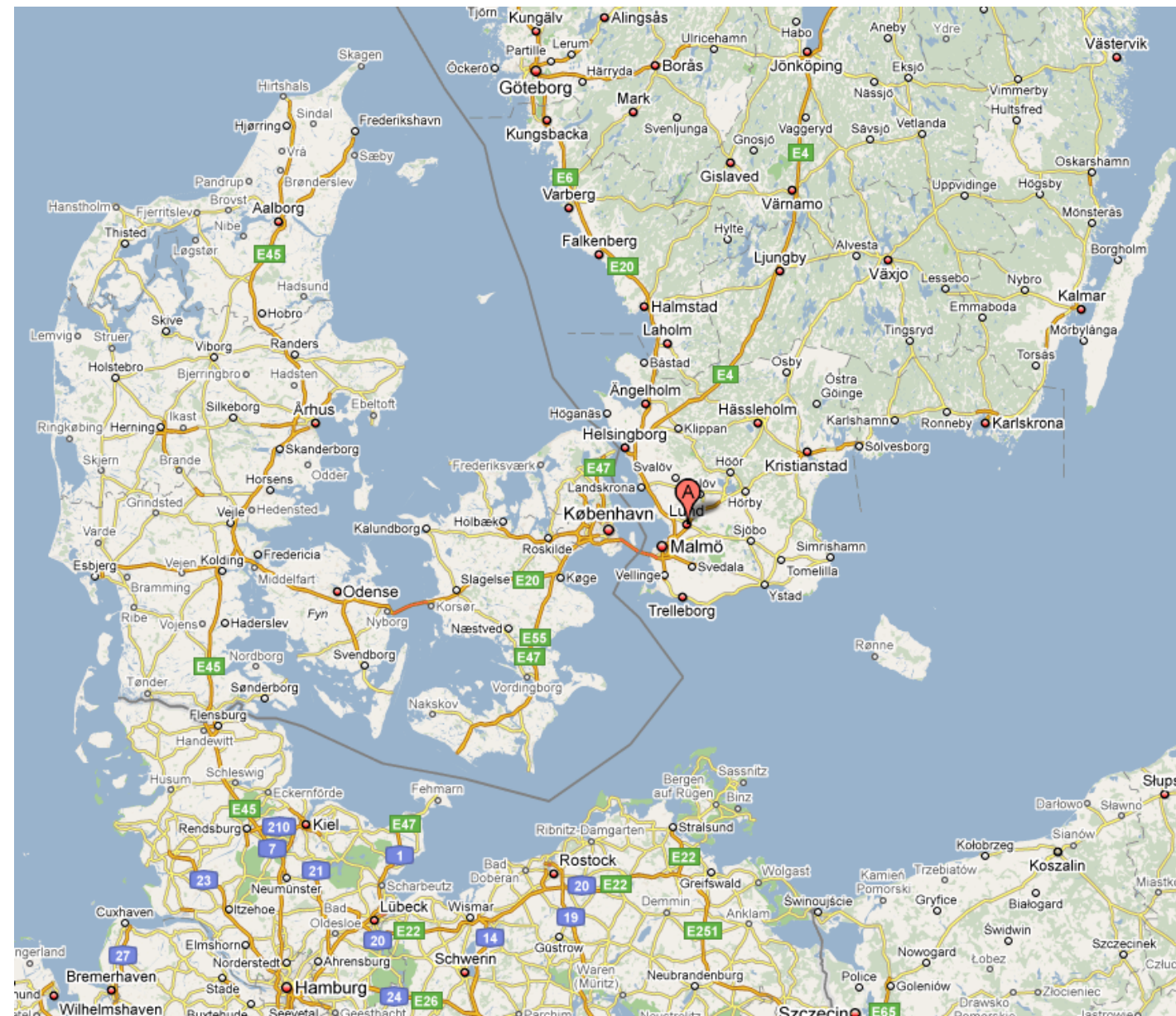
13 member states + 2 observers

- 471 people
- 48 nationalities
- Currently hiring ... need two more to break 50!

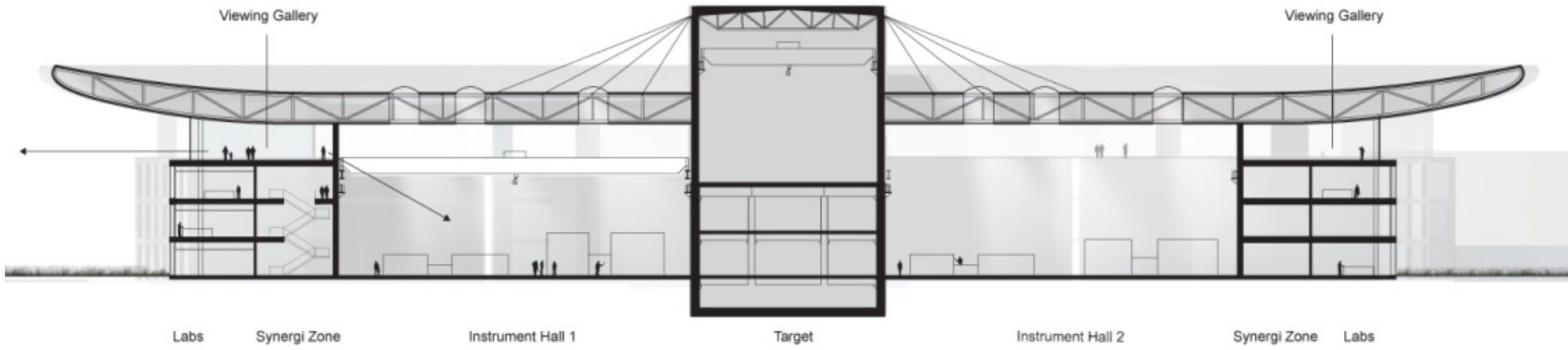
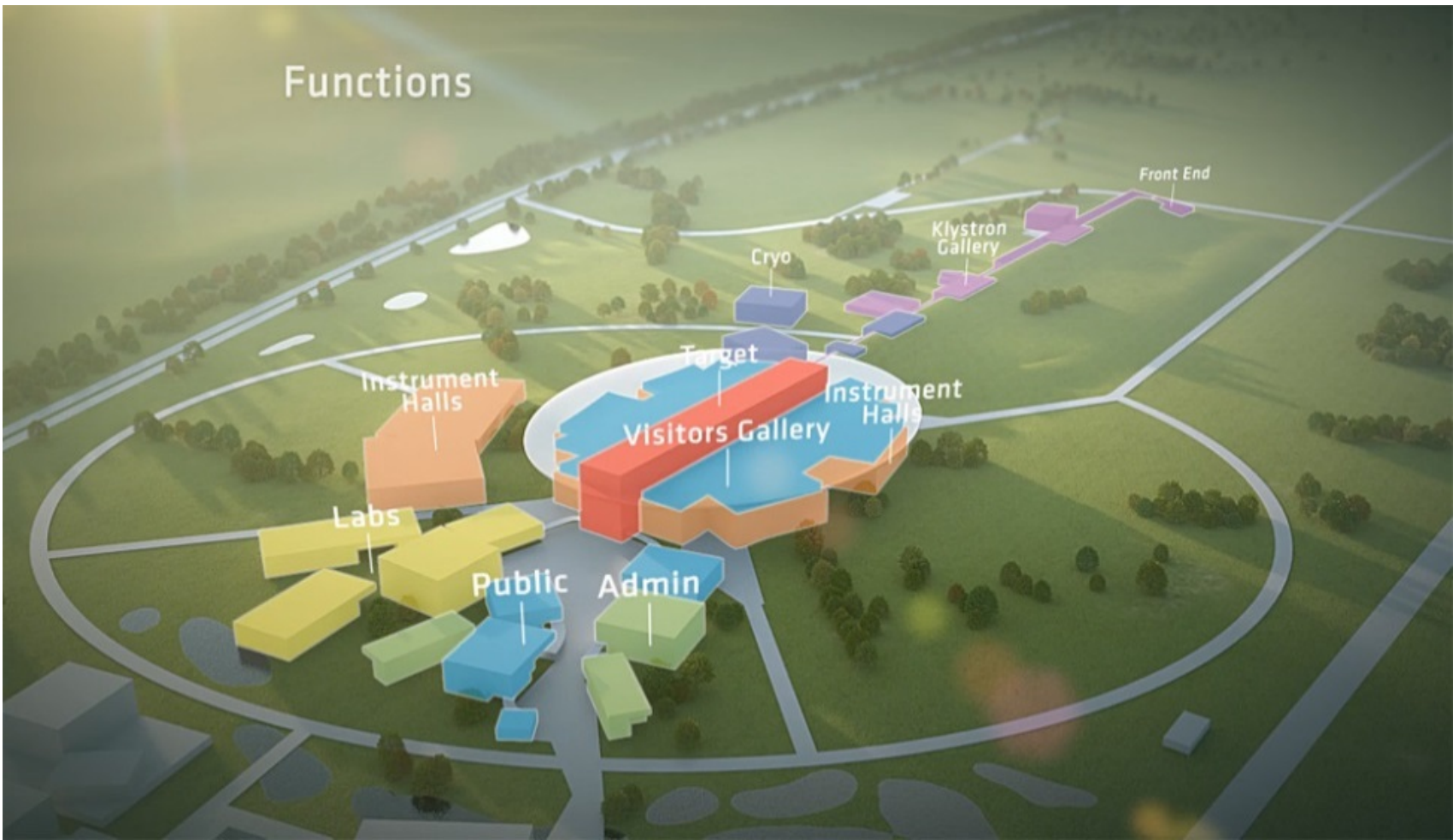


Where Will ESS Be Built?

- ESS is located in southern Sweden adjacent to MAX-IV (A 4th generation light source)
- To provide a world-class material research center for Europe

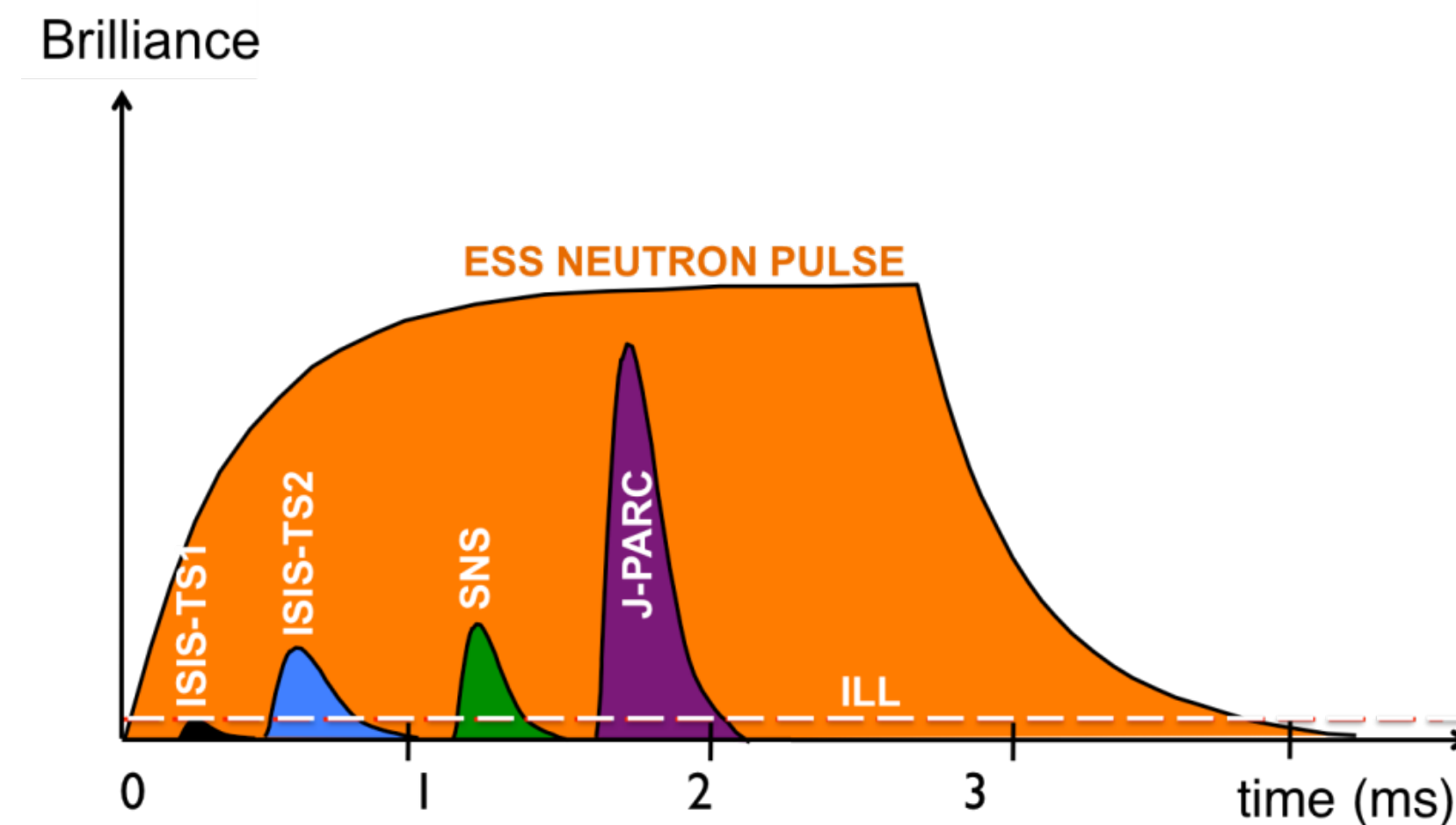
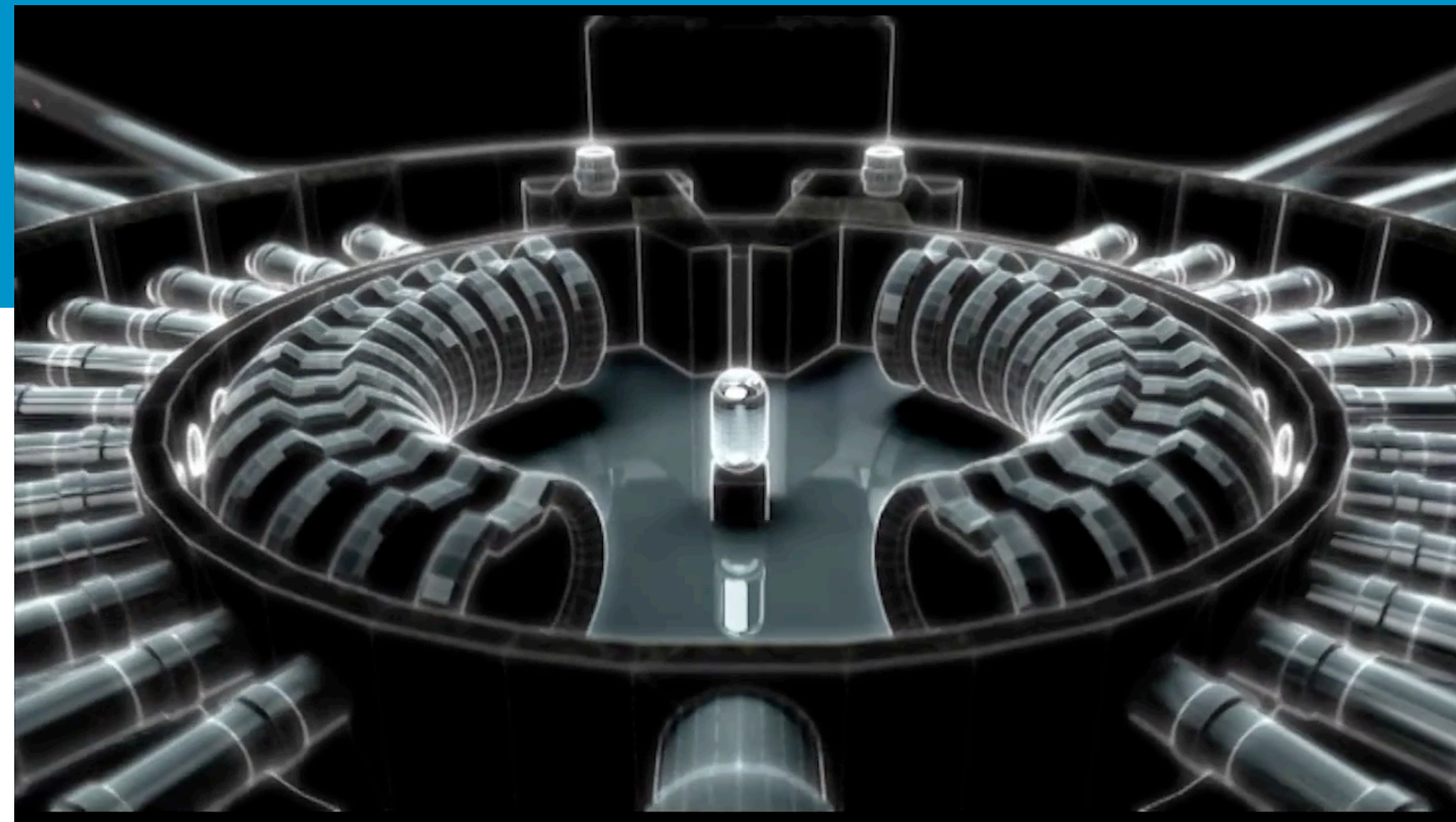


What will ESS look like?

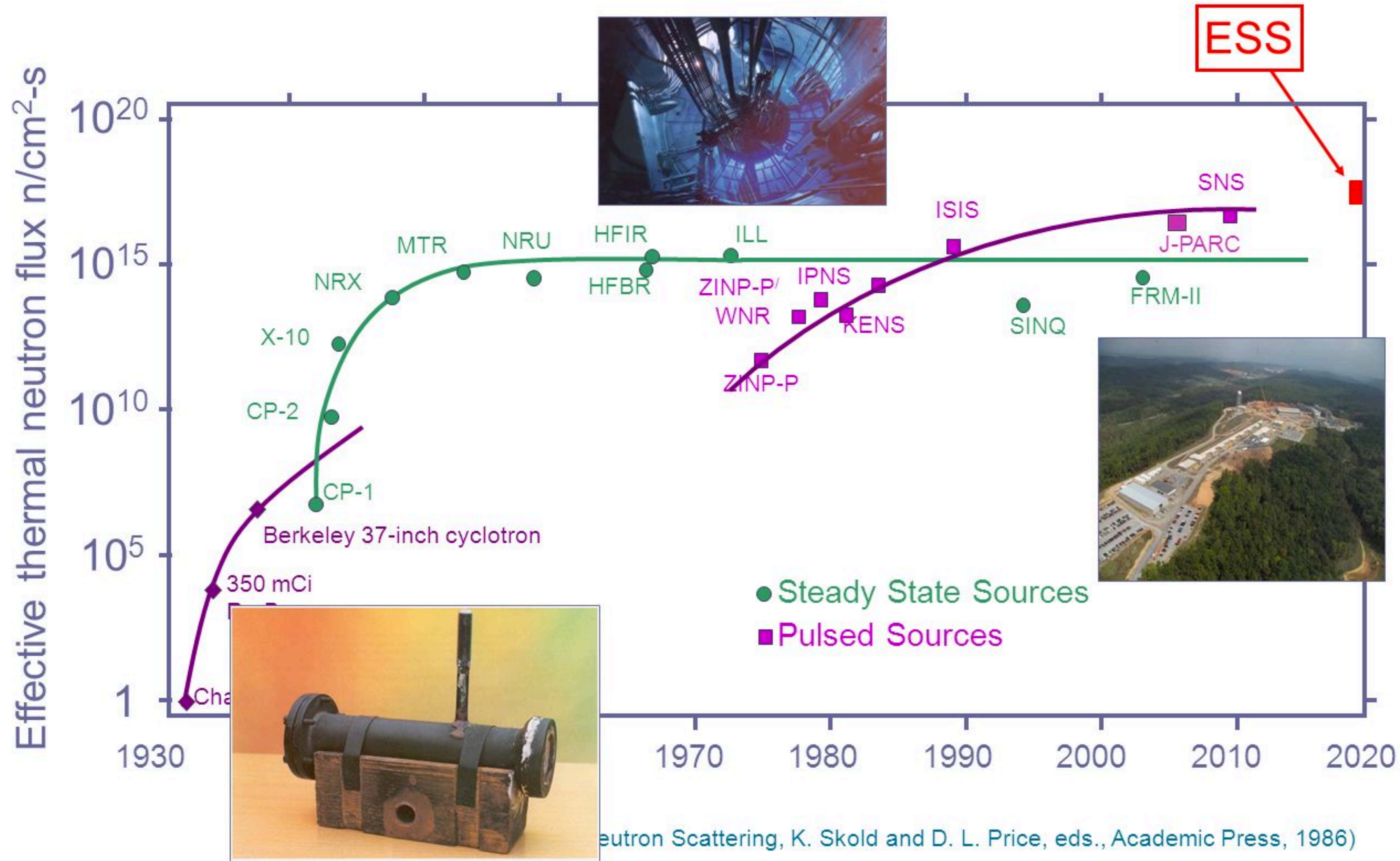


What is Different About ESS?

- The average proton beam power will be 5 MW
 - Average neutron flux is proportional to average beam power
 - 5 MW is five times greater than SNS beam power
- The total proton energy per pulse will be 360 kJ
 - Beam brightness (neutrons per pulse) is proportional to total proton energy per pulse
 - 360 kJ is over 20 times greater than SNS total proton energy per pulse



Evolution of neutron sources



Neutron Science



Energy Environment and climate Medicine and health Electronics and IT Manufacturing and industry Natural world Heritage science

Hydrogen-fuelled society Sub-zero survival Disease resistant crops Tackling chemical waste in the pharmaceutical industry

Tracking cholesterol Super superconductors Infection sensors Stress relief in the air

Flexible plastic solar cells Enhanced oil recovery



How many bytes per neutron?

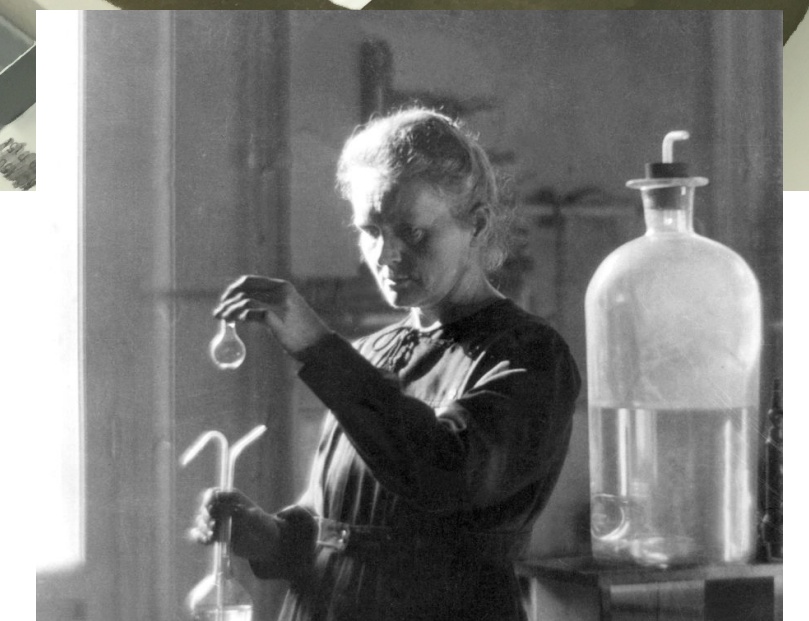
- 100 bits = 1 neutron
- We expect to have about 34 PB in first year of user operations
- 20 million PIDs
- Currently we have about 0.5 PB
- Need to be able to identify all this data with its correct metadata



Scientific metadata

“... is often notoriously incomplete. Additional quantities and assumptions necessary to interpret the data may initially only be recorded on scraps of paper, hard-coded into analysis software or only exist in the experimenter's head”

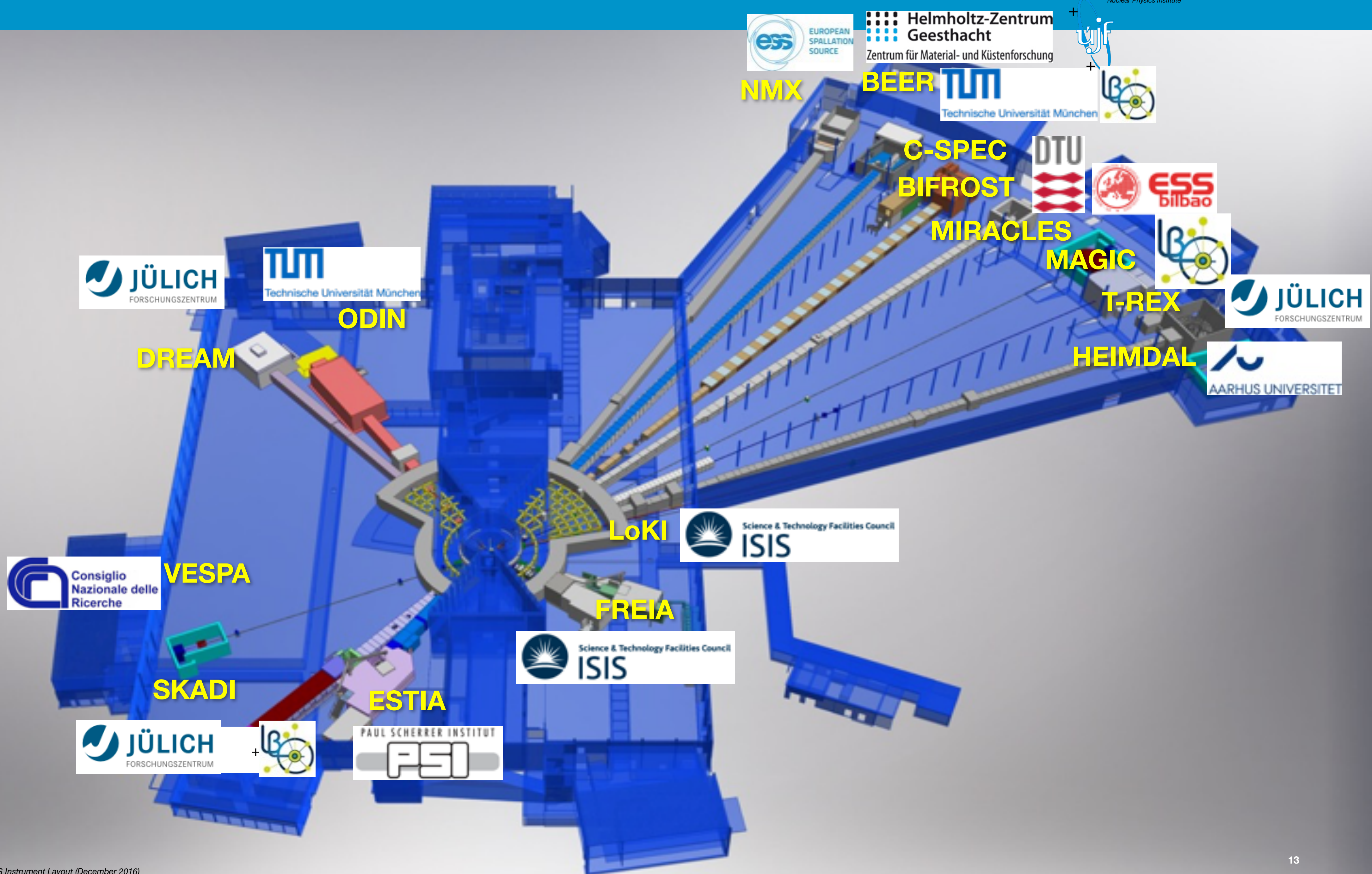
Clive Davenhall - Digital Curation Centre



How to prevent metadata from going AWOL?

- Each instrument has its own type of data
- Own methods of data reduction/analysis
- We use the same format across instruments
HDF5 file format with NeXus metaformat
- Metadata will be generated automatically and added to our data catalogue







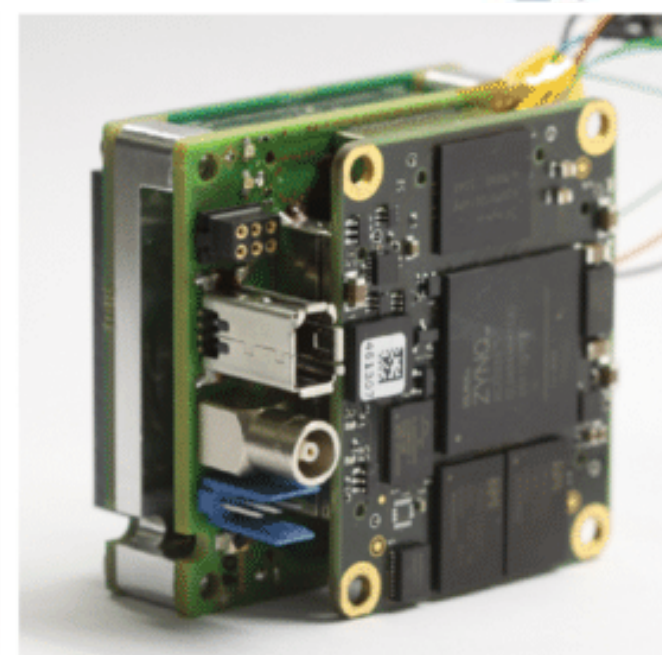
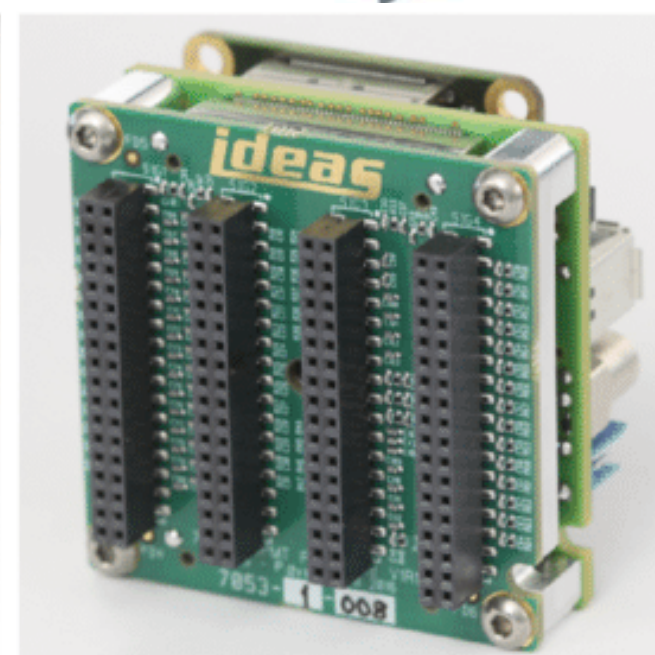
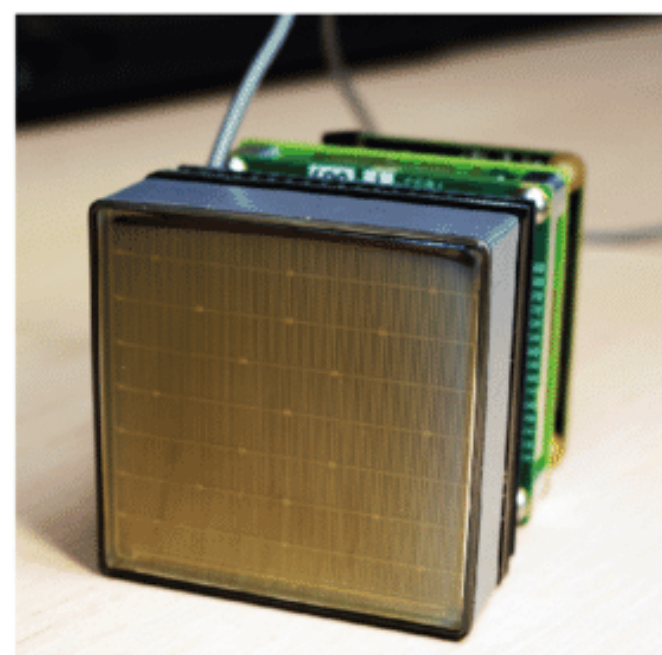
Matthew Buckley
@physicsmatt



I didn't realize how much of data science was just getting data in a format where you could do science.

20.02 · 17/01/2019 · [Twitter Web Client](#)

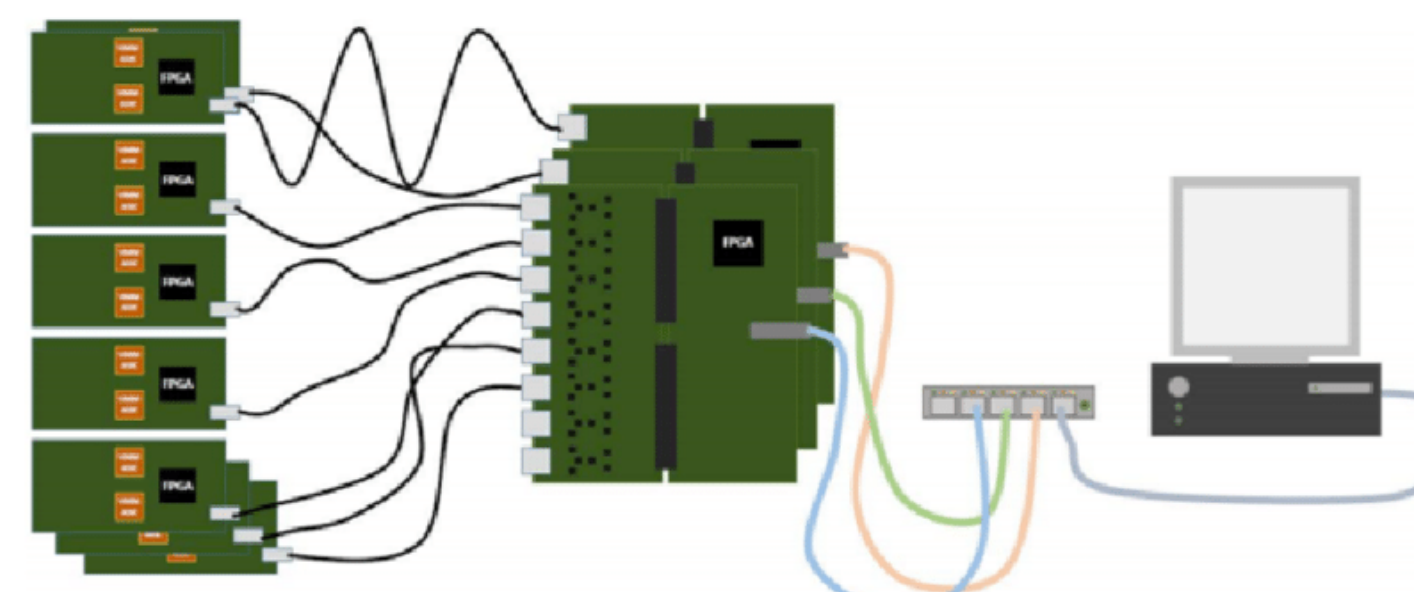
127 Retweets 707 Likes



Front End Hybrids

SRS

DAQ Toolbox

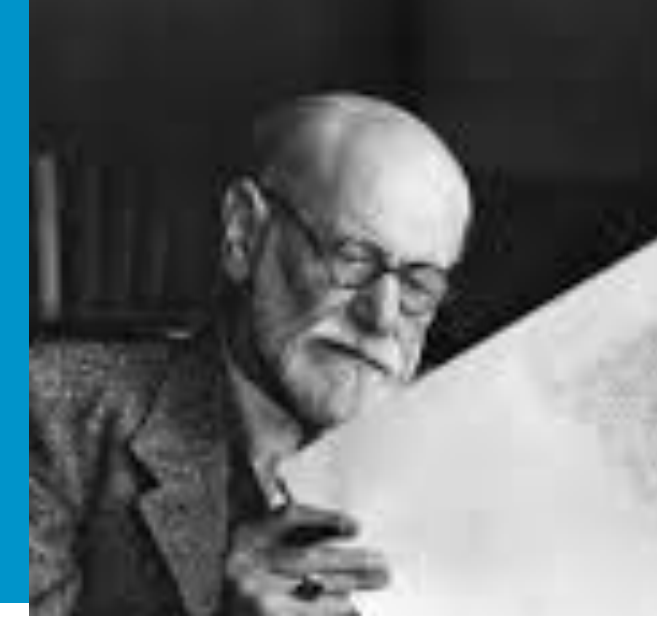


We need to identify

- Users - ORCID
- Published Datasets - DataCite DOI
- Unpublished data - Handle.net
- Proposals - ?
- Experiments - ?
- Instruments - ?
- Samples - ?



What do scientists want?



- Improved ability to do science with data
- A way of storing and retrieving data
- Upgrade from USB drive in desk drawer
- More control over metadata
- Ability to add their own custom tags
- We invited science users to a data curation workshop and asked them for input
- They responded with emails, screenshots, excel spreadsheets, photos ...



Users handle metadata in different ways

- Handwritten logs
- Excel spreadsheets
- Excel spreadsheets printed out and pasted into handwritten logs



	1:2:4 dChCl:hGlycerol:H2O/D2O + 20%hCl2hTAB_SANS	2015-10-10T19:56:53	00:15:08	10.0046	Edler,Arnold,Jackson,Fernandez,Heenan	
32381	HCHH Urea 50C 4h SANS	2015-10-10T20:13:08	00:15:08	10.0062	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32382	Actually HCHH_50C_SANS	2015-10-10T20:57:39	00:17:29	11.5630	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32383	1:2:4 dChCl:hGlycerol:H2O/D2Oc_SANS	2015-10-10T21:27:19	00:30:31	20.1903	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32384	1:2 h-ChCl:h-Urea + 0.1% d-SDS_SANS	2015-10-10T21:58:16	00:07:46	5.1232	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32385	1:2:4 dChCl:hGlycerol:H2O/D2Oa_SANS	2015-10-10T22:06:51	00:30:15	20.0051	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32386	1:2:4 dChCl:hGlycerol:H2O/D2Oc_SANS	2015-10-10T22:37:32	00:30:16	20.0110	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32387	1:2 h-ChCl:h-Urea + 0.1% d-SDS_SANS	2015-10-10T23:08:15	01:00:21	40.0024	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32388	1:2 h-ChCl:h-Urea + 0.2% d-SDS_SANS	2015-10-11T00:09:03	01:00:22	40.0062	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32389	1:2 h-ChCl:h-Urea + 0.5% d-SDS_SANS	2015-10-11T01:09:52	01:00:21	40.0013	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32390	1:2 h-ChCl:h-Urea + 1% d-SDS_SANS	2015-10-11T02:10:42	01:00:26	40.0045	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32391	1:2 h-ChCl:d-Urea + 0.1% h-SDS_SANS	2015-10-11T03:11:34	01:00:23	40.0079	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32392	1:2 h-ChCl:d-Urea + 0.2% h-SDS_SANS	2015-10-11T04:12:27	01:00:23	40.0081	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32393	1:2 h-ChCl:d-Urea + 0.5% h-SDS_SANS	2015-10-11T05:13:18	01:00:23	40.0048	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32394	1:2 h-ChCl:d-Urea + 1% h-SDS_SANS	2015-10-11T06:14:09	01:00:22	40.0101	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32395	1:2 h-ChCl:d-Urea + 0.1% d-SDS_SANS	2015-10-11T07:15:00	01:00:28	40.0037	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32396	1:2 h-ChCl:d-Urea + 0.2% d-SDS_SANS	2015-10-11T08:15:55	01:05:49	40.0067	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32397	1:2 h-ChCl:d-Urea + 0.5% d-SDS_SANS	2015-10-11T09:22:13	01:00:27	40.0094	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32398	1:2 h-ChCl:d-Urea + 1% d-SDS_SANS	2015-10-11T10:23:09	01:00:27	40.0081	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32399	H2O_SANS	2015-10-11T11:24:04	00:30:12	20.0079	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32400	D2O_SANS	2015-10-11T11:54:44	00:30:14	20.0078	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI
32401	H2O_TRANS	2015-10-11T12:25:32	00:12:04	8.0014	Edler,Arnold,Jackson,Fernandez,Heenan	UNSPI

23-26/10/15

FIGARO EXPERIMENT glycerol/ChCl DES + SDS, G2TAB
DPRC, DMPC.

9-13-612

local contact: Richard Campbell.

Adrian Sanchez-Fernandez, Karen Edler, Tom Arnold.

trial of troughs Delrin - D₂O ok
Macor - D₂O curvature??

Delrin - hDES (ChCl:glycerol 1:2)
=> issues with beam hitting window?
add paper spacers 0.67 mm thick.
(7 sheets of paper)

Using trough sample changer "wrong way around" (PS@3022mm)

Direct Beam 1 $\theta = 0.623^\circ$
#548799 S2H = 040 S3H = 020
CHOP = 7% FOM = 30
S2W = 44 S3W = 32
ATW = 5.0
30 min @ 1329 c/s

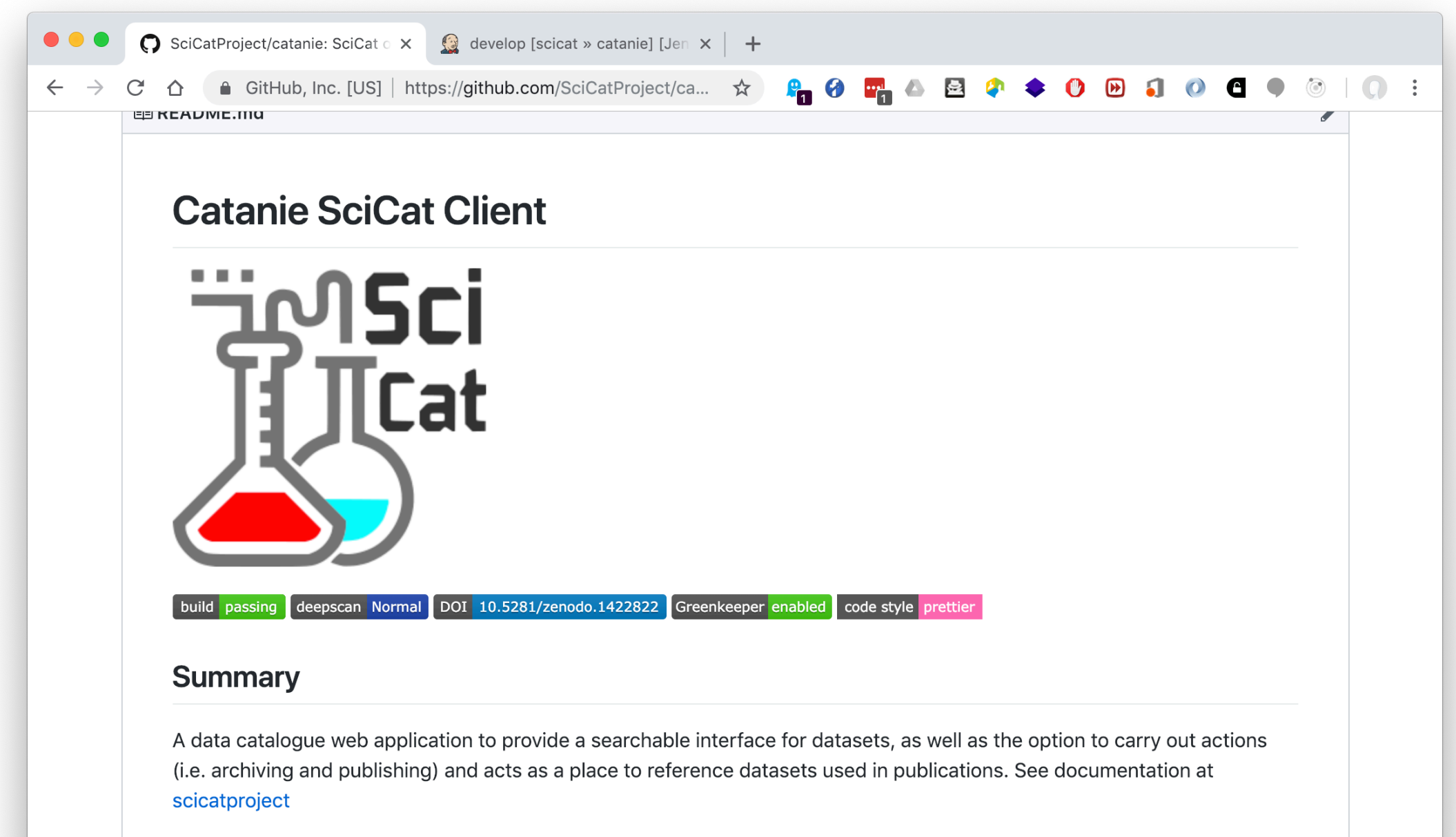
Direct Beam 2 $\theta = 3.79^\circ$
#548800 S2H = 48 S3H = 1.6
CHOP = 7% FOM = 30
S2W = 44 S3W = 32
ATW = 0.40
20 min @ 7629 c/s

D₂O in Delrin trough PZ A1 6min #548802 (3211 c/s)
~4.5ml A2 45min #548803 (3401 c/s)

D₂O in Macor trough P3 A1 6min #548806 (3194 c/s)
(2.5 ml) A2 11min #548807 (2200 c/s)

SciCat - a Scientists' Data Catalogue

- Manage scientific metadata for users
- Access to data
- PIs/users can publish data and create DOIs
- Open source, available on github.com/scicatproject



Why not use existing tools?

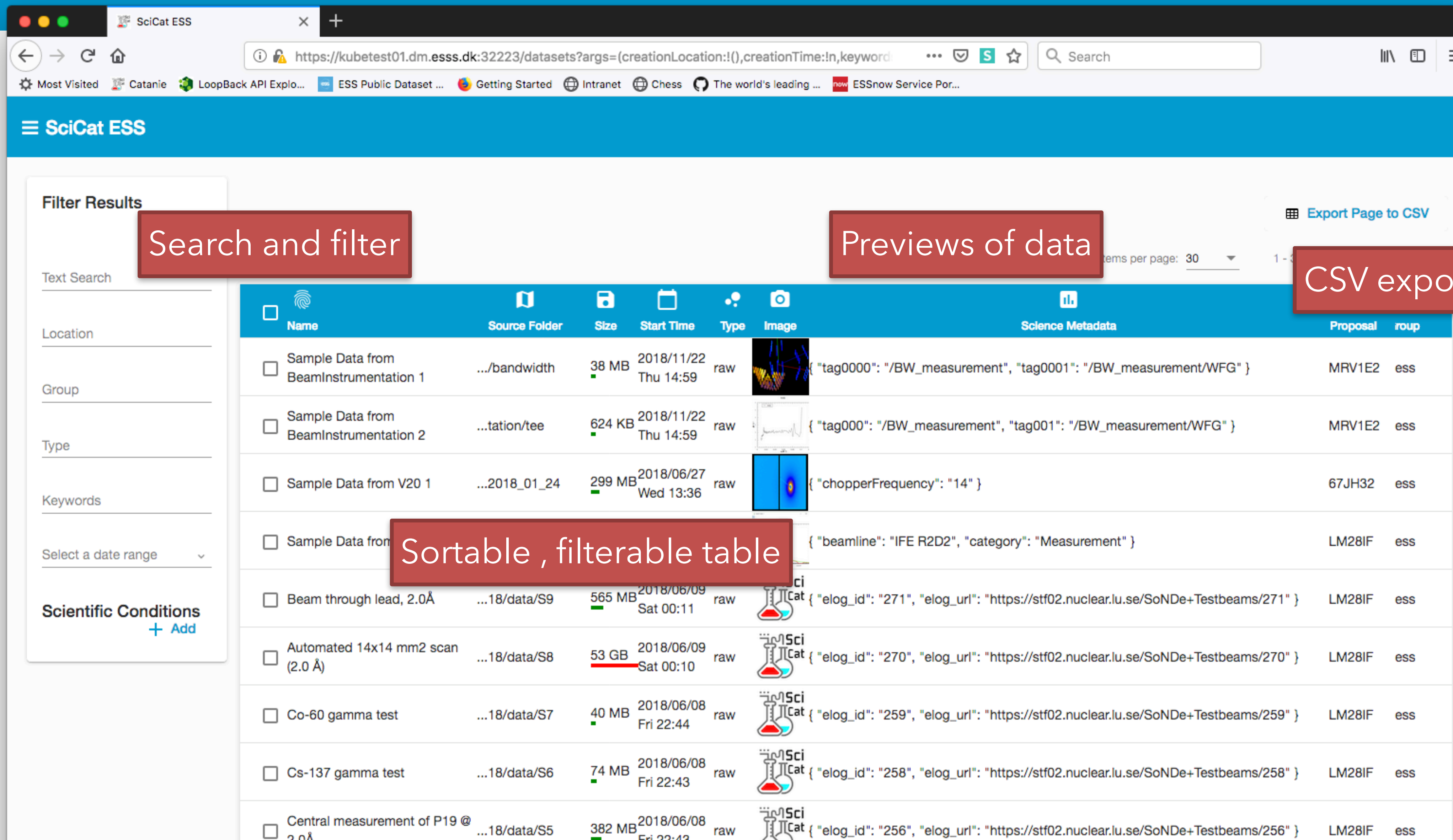
- Scientists don't necessarily know what they are looking for in a given experiment
- Between proposal writing, acceptance and lab time goals can change a lot
- "we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know" - Donald Rumsfeld
- We need to be able to capture this in metadata
- Structured metadata needs to have structure defined in advance
- metadata needs to be *unstructured*





**Basic research is what I am doing
when I don't know what I am doing
Werner von Braun.**

Data Catalogue - SciCat



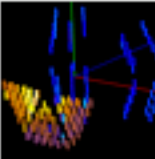
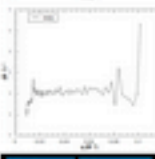
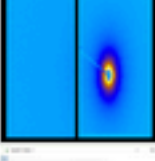





The screenshot displays the SciCat ESS web interface. At the top, there is a navigation bar with the SciCat ESS logo and a search bar. Below the navigation bar, there is a filter sidebar on the left with sections for Text Search, Location, Group, Type, Keywords, and Scientific Conditions. The main content area features a table of data entries. The table has columns for Name, Source Folder, Size, Start Time, Type, Image, Science Metadata, Proposal, and Group. Several red callout boxes are overlaid on the image: 'Search and filter' points to the filter sidebar, 'Previews of data' points to the image column, 'CSV export' points to the 'Export Page to CSV' button, and 'Sortable, filterable table' points to the table header.

Search and filter

Previews of data

CSV export

Sortable, filterable table

<input type="checkbox"/>	Name	Source Folder	Size	Start Time	Type	Image	Science Metadata	Proposal	Group
<input type="checkbox"/>	Sample Data from BeamInstrumentation 1	.../bandwidth	38 MB	2018/11/22 Thu 14:59	raw		{ "tag0000": "/BW_measurement", "tag0001": "/BW_measurement/WFG" }	MRV1E2	ess
<input type="checkbox"/>	Sample Data from BeamInstrumentation 2	...tation/tee	624 KB	2018/11/22 Thu 14:59	raw		{ "tag000": "/BW_measurement", "tag001": "/BW_measurement/WFG" }	MRV1E2	ess
<input type="checkbox"/>	Sample Data from V20 1	...2018_01_24	299 MB	2018/06/27 Wed 13:36	raw		{ "chopperFrequency": "14" }	67JH32	ess
<input type="checkbox"/>	Sample Data from						{ "beamline": "IFE R2D2", "category": "Measurement" }	LM28IF	ess
<input type="checkbox"/>	Beam through lead, 2.0Å	...18/data/S9	565 MB	2018/06/09 Sat 00:11	raw		{ "elog_id": "271", "elog_url": "https://stf02.nuclear.lu.se/SoNDe+Testbeams/271" }	LM28IF	ess
<input type="checkbox"/>	Automated 14x14 mm2 scan (2.0 Å)	...18/data/S8	53 GB	2018/06/09 Sat 00:10	raw		{ "elog_id": "270", "elog_url": "https://stf02.nuclear.lu.se/SoNDe+Testbeams/270" }	LM28IF	ess
<input type="checkbox"/>	Co-60 gamma test	...18/data/S7	40 MB	2018/06/08 Fri 22:44	raw		{ "elog_id": "259", "elog_url": "https://stf02.nuclear.lu.se/SoNDe+Testbeams/259" }	LM28IF	ess
<input type="checkbox"/>	Cs-137 gamma test	...18/data/S6	74 MB	2018/06/08 Fri 22:43	raw		{ "elog_id": "258", "elog_url": "https://stf02.nuclear.lu.se/SoNDe+Testbeams/258" }	LM28IF	ess
<input type="checkbox"/>	Central measurement of P19 @ 2.0Å	...18/data/S5	382 MB	2018/06/08 Fri 22:43	raw		{ "elog_id": "256", "elog_url": "https://stf02.nuclear.lu.se/SoNDe+Testbeams/256" }	LM28IF	ess

SciCat ESS test

localhost:4200/datasets/20.500.12269%2FBRIGHTNESS%2FSONDE0011

SciCat ESS test

Datasets / 20.500.12269 / BRIGHTNESS / SONDE0011 /

Details Datafiles Attachments Admin Export to CSV

About the data

Name Beam through lead, 2.0Å

Description 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. github.com/ess-dmsc/ess_file_formats/wiki/SONDE

Owner Ramsey Al Jebali

Keywords ["SoNDe", "neutron", "detector"]

PID 20.500.12269/BRIGHTNESS/SONDE0011

Source Folder /users/detector/experiments/sonde/IFE_june_2018/data/S9

Structural information

Type raw

Version 2.8.1

Proposal LM28IF

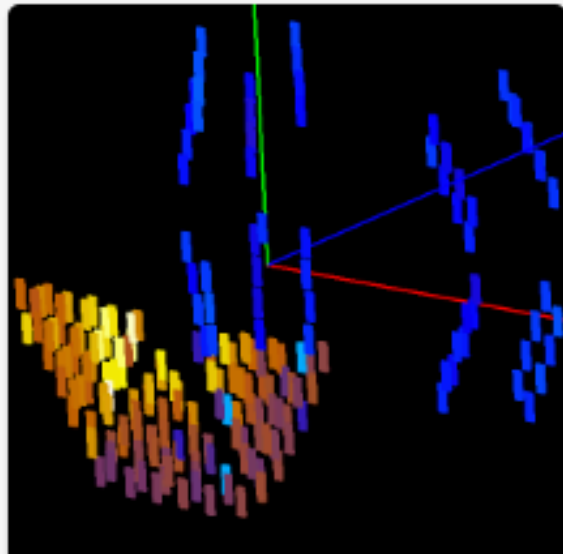
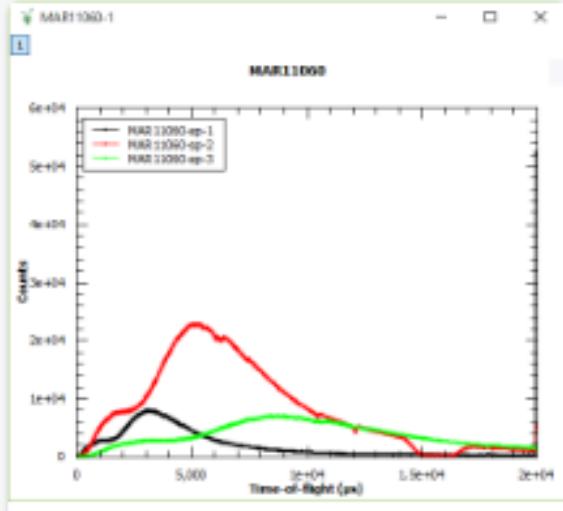
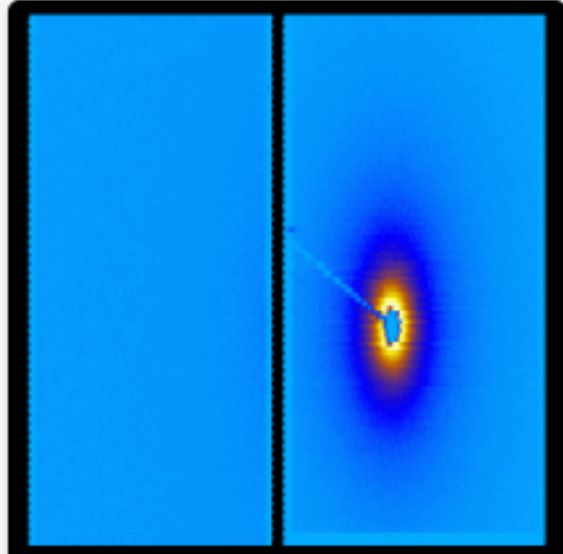
Sample SAMPLE001

Size 565 MB

Administrative information

Creation Time 2018/06/09 00:11

Principal Investigator Ramsey Al Jebali

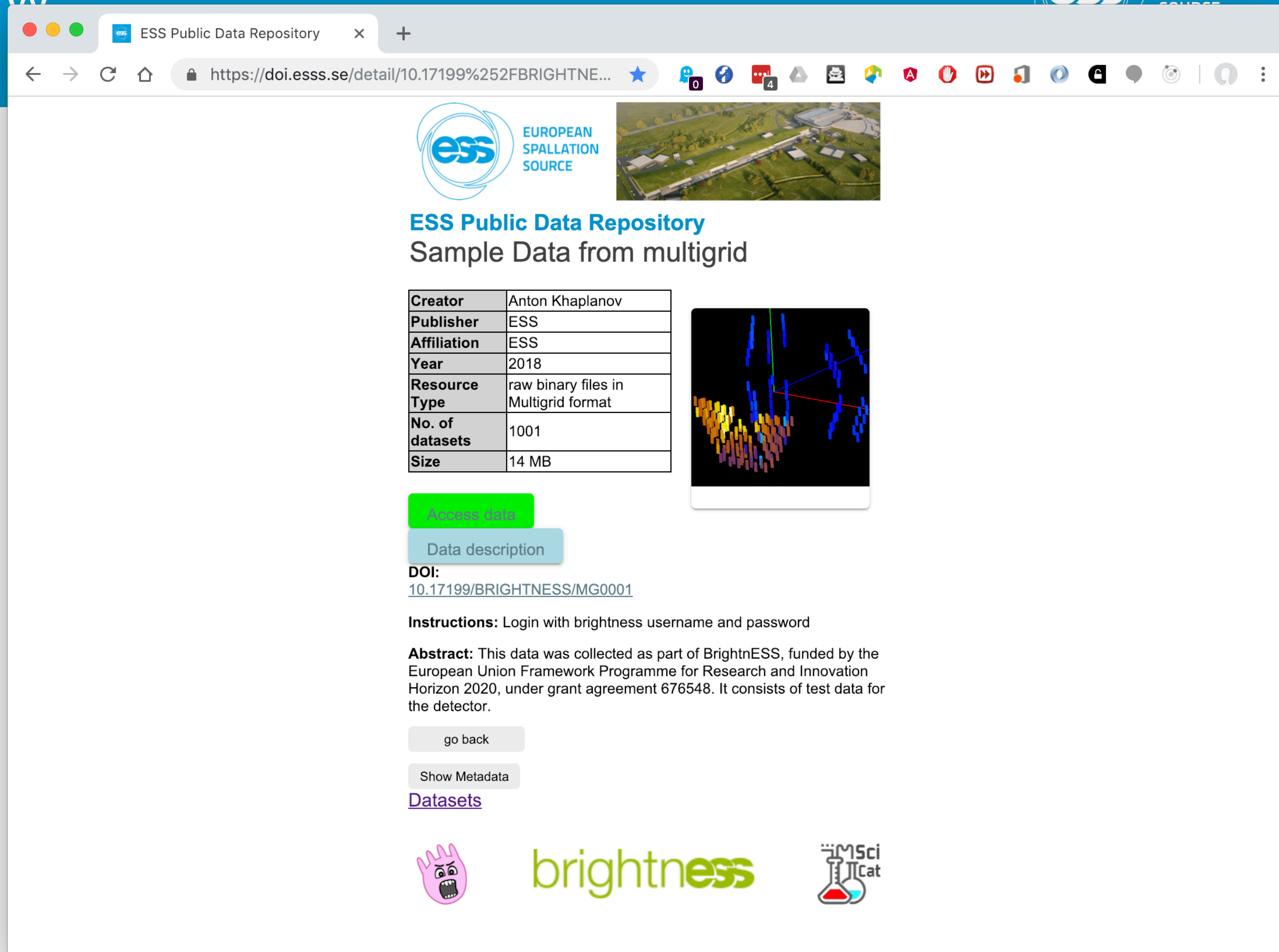
- Scientists have asked to improve the user experience
- “Make it more like Google!”
- “Make it fast!”
- We have hired a UX consultant to help with user needs

Greater access to data and metadata with PIDs




- Scientist clicks on a link
- Can access their data instantly via download
- Can preview the data
- Can link to proposal
- Link to sample

Publishing workflow




ESS Public Data Repository

https://doi.esss.se/detail/10.17199%252FBRIGHTNE...



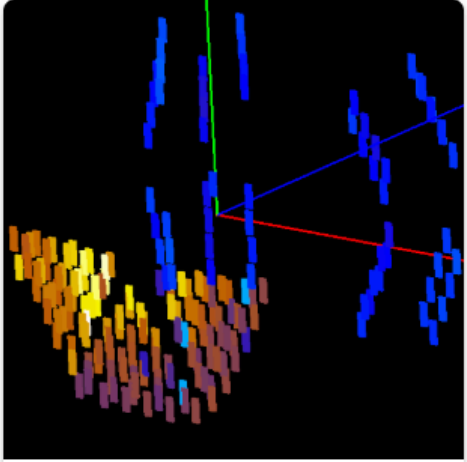
EUROPEAN
SPALLATION
SOURCE



ESS Public Data Repository

Sample Data from multigrid

Creator	Anton Khaplanov
Publisher	ESS
Affiliation	ESS
Year	2018
Resource Type	raw binary files in Multigrid format
No. of datasets	1001
Size	14 MB



[Access data](#)

[Data description](#)

DOI:
[10.17199/BRIGHTNESS/MG0001](https://doi.esss.se/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](#)

[Show Metadata](#)

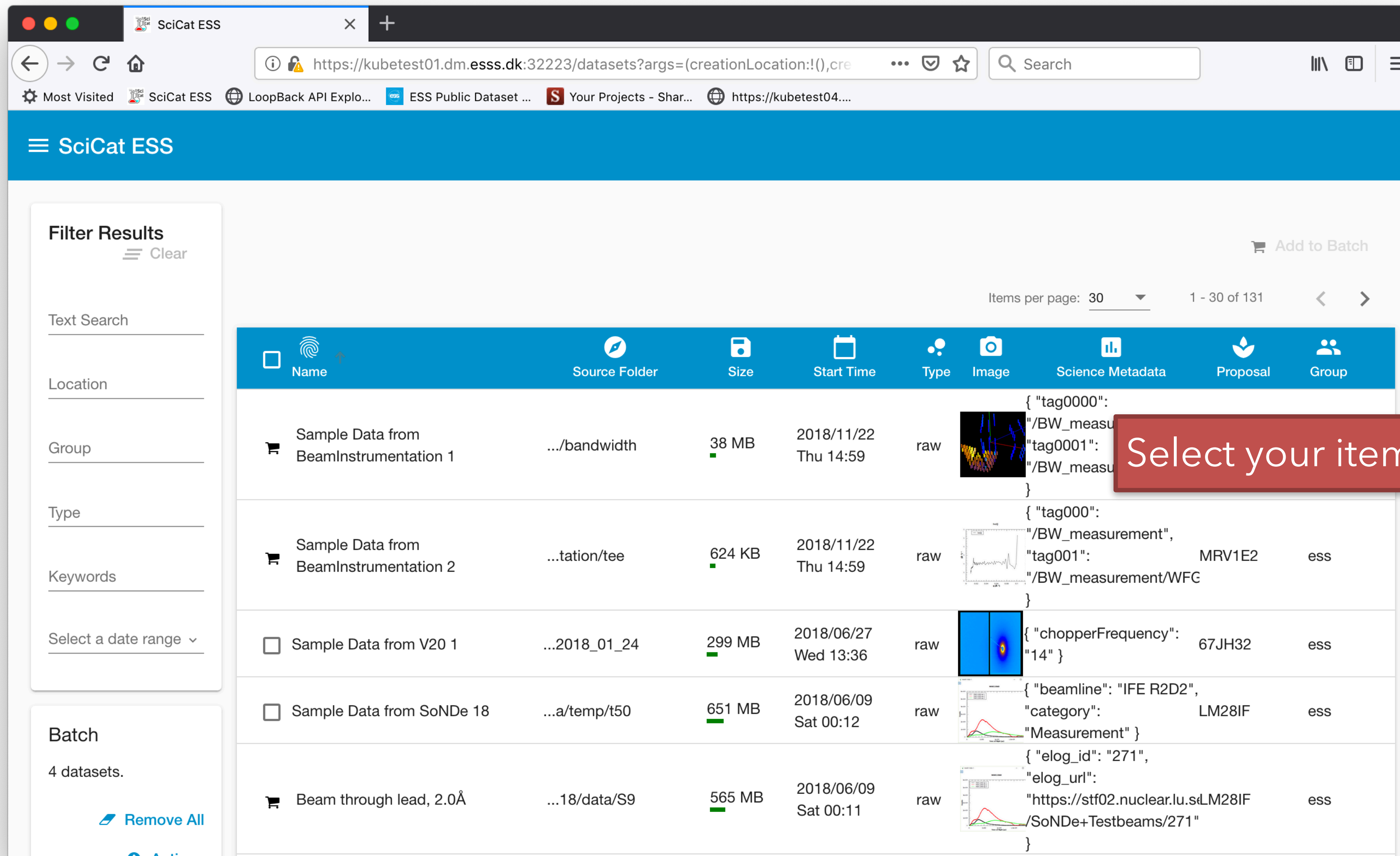
[Datasets](#)



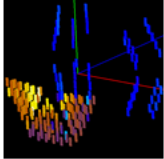
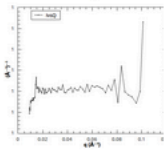
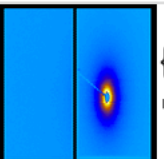
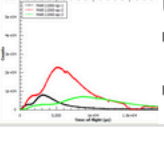
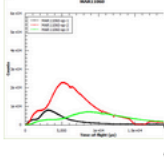
brightness



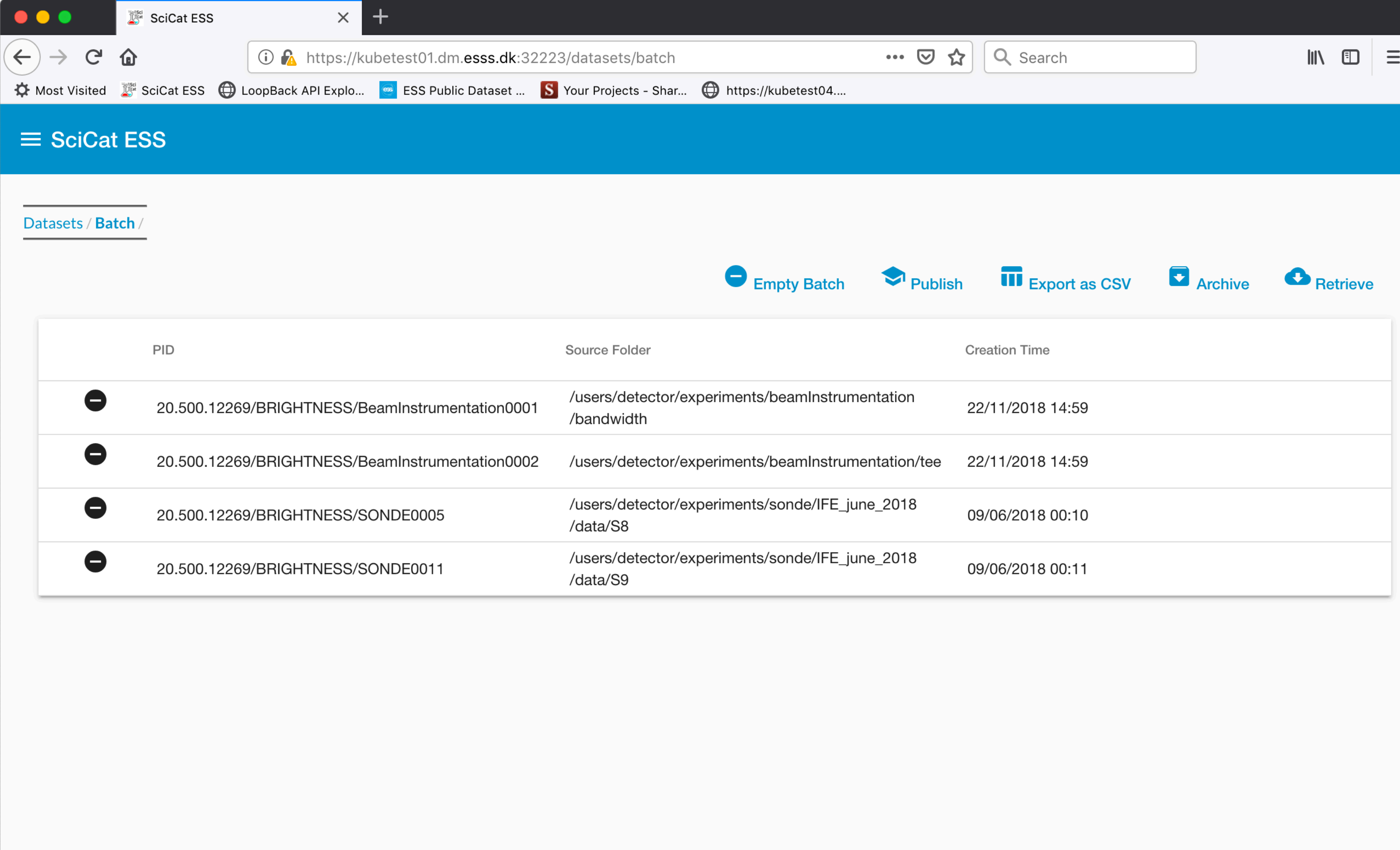
DOI shopping cart



The screenshot shows the SciCat ESS web interface. The browser address bar displays the URL: `https://kubetest01.dm.esss.dk:32223/datasets?args=(creationLocation:!(),cre`. The page title is "SciCat ESS". On the left, there is a "Filter Results" sidebar with options for Text Search, Location, Group, Type, and Keywords. The main content area displays a table of datasets with columns: Name, Source Folder, Size, Start Time, Type, Image, Science Metadata, Proposal, and Group. The table contains five rows of dataset information. A red callout box with white text is overlaid on the right side of the table, stating "Select your items and add to shopping cart".

Name	Source Folder	Size	Start Time	Type	Image	Science Metadata	Proposal	Group
<input checked="" type="checkbox"/> Sample Data from BeamInstrumentation 1	.../bandwidth	38 MB	2018/11/22 Thu 14:59	raw		<pre>{ "tag0000": "/BW_measur", "tag0001": "/BW_measur" }</pre>		
<input checked="" type="checkbox"/> Sample Data from BeamInstrumentation 2	...tation/tee	624 KB	2018/11/22 Thu 14:59	raw		<pre>{ "tag000": "/BW_measurement", "tag001": "MRV1E2", "/BW_measurement/WFC" }</pre>	MRV1E2	ess
<input type="checkbox"/> Sample Data from V20 1	...2018_01_24	299 MB	2018/06/27 Wed 13:36	raw		<pre>{ "chopperFrequency": "14" }</pre>	67JH32	ess
<input type="checkbox"/> Sample Data from SoNDe 18	...a/temp/t50	651 MB	2018/06/09 Sat 00:12	raw		<pre>{ "beamline": "IFE R2D2", "category": "Measurement" }</pre>	LM28IF	ess
<input checked="" type="checkbox"/> Beam through lead, 2.0Å	...18/data/S9	565 MB	2018/06/09 Sat 00:11	raw		<pre>{ "elog_id": "271", "elog_url": "https://stf02.nuclear.lu.se/LM28IF/SoNDe+Testbeams/271" }</pre>		ess

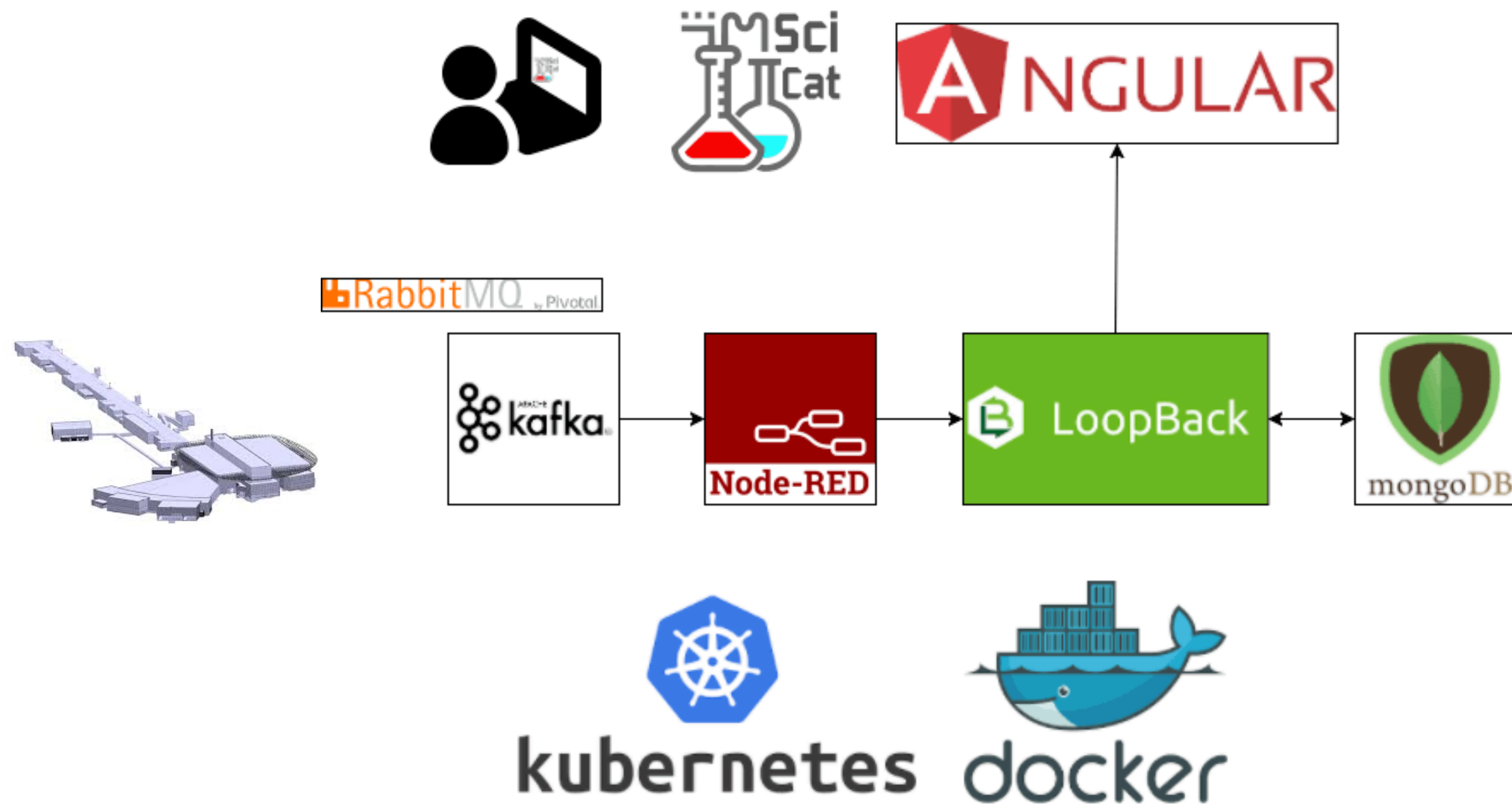
Publish your data at click of button



The screenshot shows a web browser window with the SciCat ESS interface. The browser address bar shows the URL `https://kubetest01.dm.esss.dk:32223/datasets/batch`. The page title is "SciCat ESS". Below the header, there is a navigation bar with the text "SciCat ESS". The main content area shows a breadcrumb "Datasets / Batch /" and a toolbar with buttons for "Empty Batch", "Publish", "Export as CSV", "Archive", and "Retrieve". Below the toolbar is a table with the following data:

PID	Source Folder	Creation Time
20.500.12269/BRIGHTNESS/BeamInstrumentation0001	/users/detector/experiments/beamInstrumentation /bandwidth	22/11/2018 14:59
20.500.12269/BRIGHTNESS/BeamInstrumentation0002	/users/detector/experiments/beamInstrumentation/tee	22/11/2018 14:59
20.500.12269/BRIGHTNESS/SONDE0005	/users/detector/experiments/sonde/IFE_june_2018 /data/S8	09/06/2018 00:10
20.500.12269/BRIGHTNESS/SONDE0011	/users/detector/experiments/sonde/IFE_june_2018 /data/S9	09/06/2018 00:11

SciCat Architecture

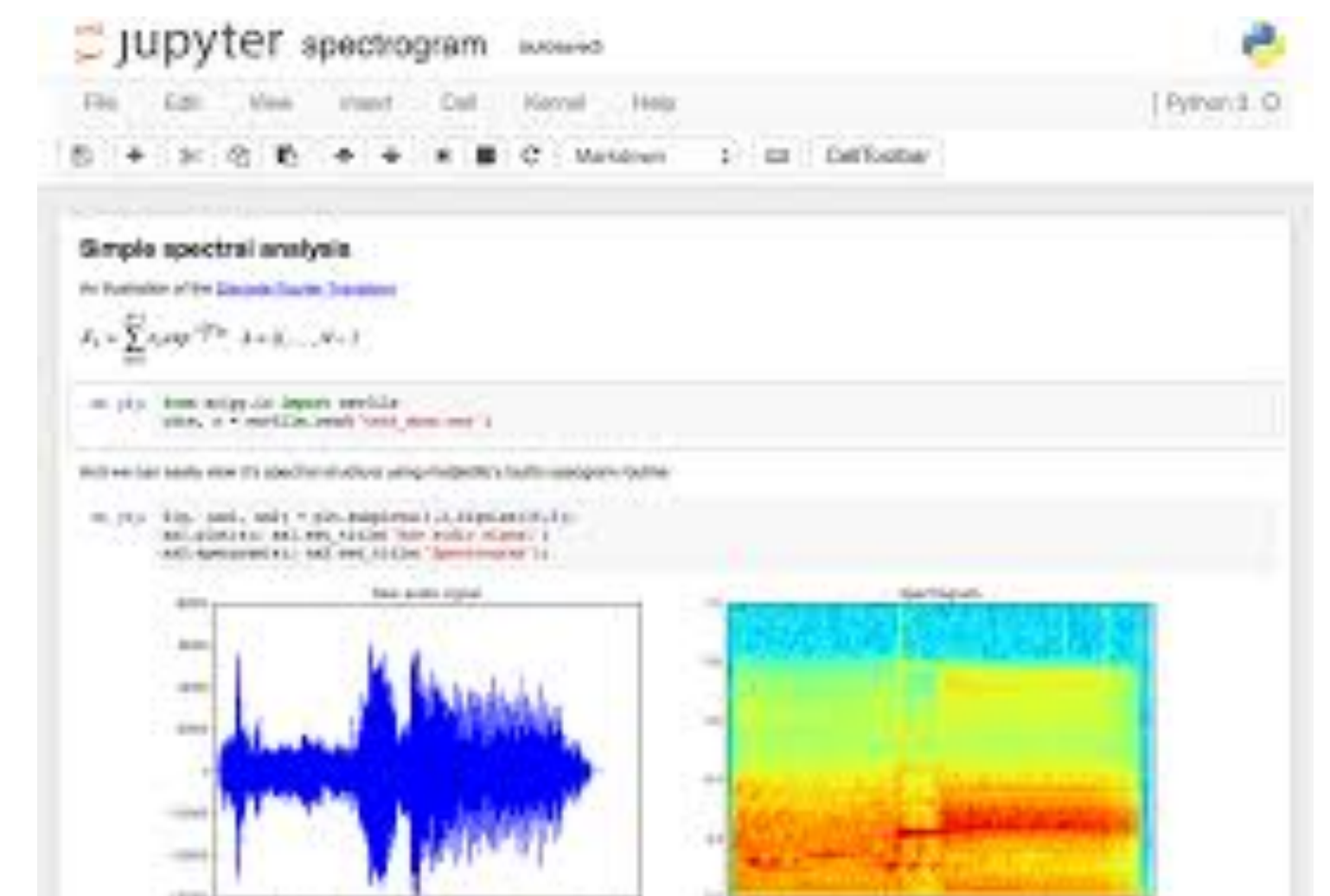


Lessons learned: Importing legacy data

- Legacy data is very hard to curate
- Metadata is not available in a lot of cases
- Hard to tell if important/not important
- By importing our legacy data, we can test our data pipeline
- 250,000 files, varying in size from 1kB to 100 GB
- Different formats can be hard to handle
- Moving everyone to use the same format is also difficult ...

Future plans

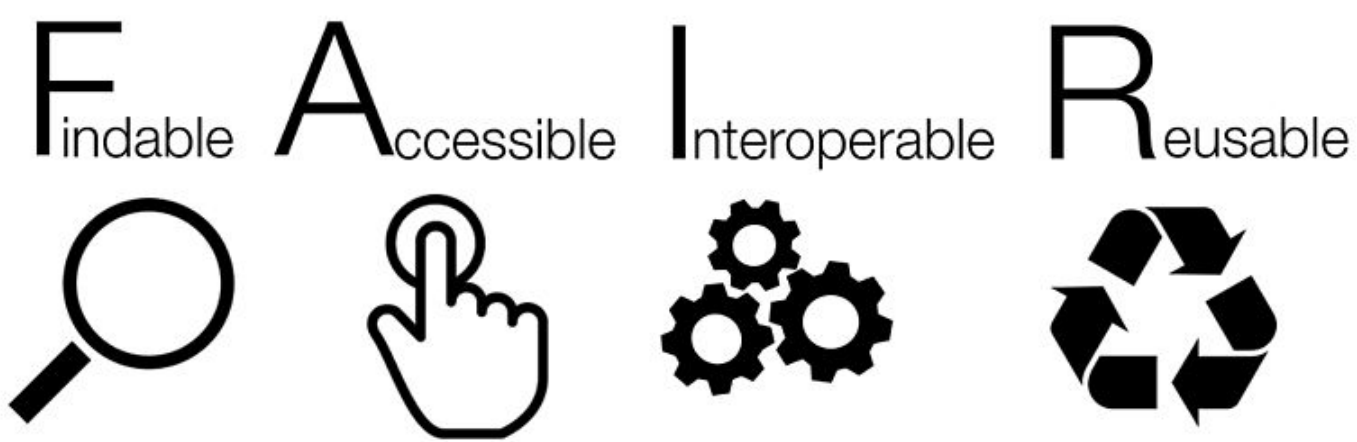
- Scientists want to analyse large datasets using cloud systems
- We are members of PANOSC - Photon And Neutron Open Science Cloud
- Goal is to fulfill analysis needs for science data
- Link with EOSC hub
- Users will be able to analyse data with Jupyter notebooks



- ESS will have lots of new data soon
- PIDs are important part of data infrastructure
- Our new data catalogue, SciCat will be able to provide improved access to science data for users

Thank you





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