Outcomes of the PaN EOSC projects for users and user offices

European User Offices Meeting (EUOM) 14 June 2022

Presenters: Jean-François Perrin (ESRF) + Patrick Fuhrmann (DESY)
on behalf of PaNOSC + ExPaNDS partners

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Quick reminder : The Projects Cheat Sheet

	PaNOSC	ExPaNDS
EU Call	HORIZON 2020 INFRA-EOSC-04	HORIZON 2020 INFRA-EOSC-5B
Description	Cluster of ESFRI PaN Sources	EOSC PaN Data Services
Partners	ESRF, ILL, ESS, EU-XFEL, CERIC-ERIC, ELI-DC, EGI	DESY, ALBA, DLS, ELETTRA, EGI, HZB, HZDDR, Max IV, PSI, Soleil, UKRI
Observers	GEANT EU-DAT National RI's	
Linked 3rd Party	DESY STFC CESNET	
Start – End (Duration)	2018-12-01 – 2022-11-30 [4 Years]	2019-09-01 – 2023-02-28 [3 ½ Years]
Coordinators	A. Götz, G. Bodera	P. Fuhrmann, S. Servan
Budget	12 M Euros	6 M Euros
Home Page	PaNOSC.EU	ExPaNDS.EU
Twitter	@PaNOSC_eu #PaNOSC	@ExPaNDS_eu #ExPaNDS
GitHUB	github.com/panosc-eu	Github.com/expands-eu

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Evolution of PaN with respect to EU projects.

Policies	Common data policy		**FAIR** data policy	Data Management Plans
Analysis		Software Catalogue	Remote analysis	Jupyter
AAI	UmbrellaID		AARC Blueprint	eduTeams
Training		e-neutron		Training platform

2010 2015 2018 2021

DiTARI
And more
Proposal

Caterina's Talk

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

What did we achieve and what will we have achieved until the end of our projects?

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

10 Primary Outcomes of PaNOSC and ExPaNDS

1. **FAIR data policy** and **DMPs**
2. **FAIR assessment** and common **PID** framework
3. Standardised metadata (**Nexus/HDF5**, PaN ontologies)
4. **Federated search API** for PaN data catalogues
5. **Open Data portal** for searching + downloading data
6. Community **AAI UmbrellaId**
7. **JupyterLab notebooks** and **Nexus/HDF5** files visualisation
8. **Remote data analysis** with VISA + data analysis pipelines
9. **Simulation** software for simulating experimental data (SIMEX)
10. **PaN-learning** platform (pan-learning.org + pan-training.org)

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Data Policy Frameworks

Feb 2011 : https://www.panosc.eu/wp-content/uploads/2019/05/PaN-data-D2.1_PolicyFramework.pdf

Since then DPs have been published in our RIs (ILL, ISIS, ESRF, …). They are all based on the key principles of the PaN-Data Framework

But since the environment and comprehension has evolved.
The new texts feature: FAIR, processed/analyzed data, electronic logbook, DMPs, … and reflect on best practices for writing DP and implementing.

PaNOSC: https://doi.org/10.5281/zenodo.3738497

ExPaNDS: https://doi.org/10.5281/zenodo.5205825

Guidance Note, Key Policy Elements: https://zenodo.org/record/6090282

The process of updating the DP has started or will start in most of our RIs. User Offices are key stakeholders in this process, it is important that you get access to these documents.

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Data Management Plan (DMP)

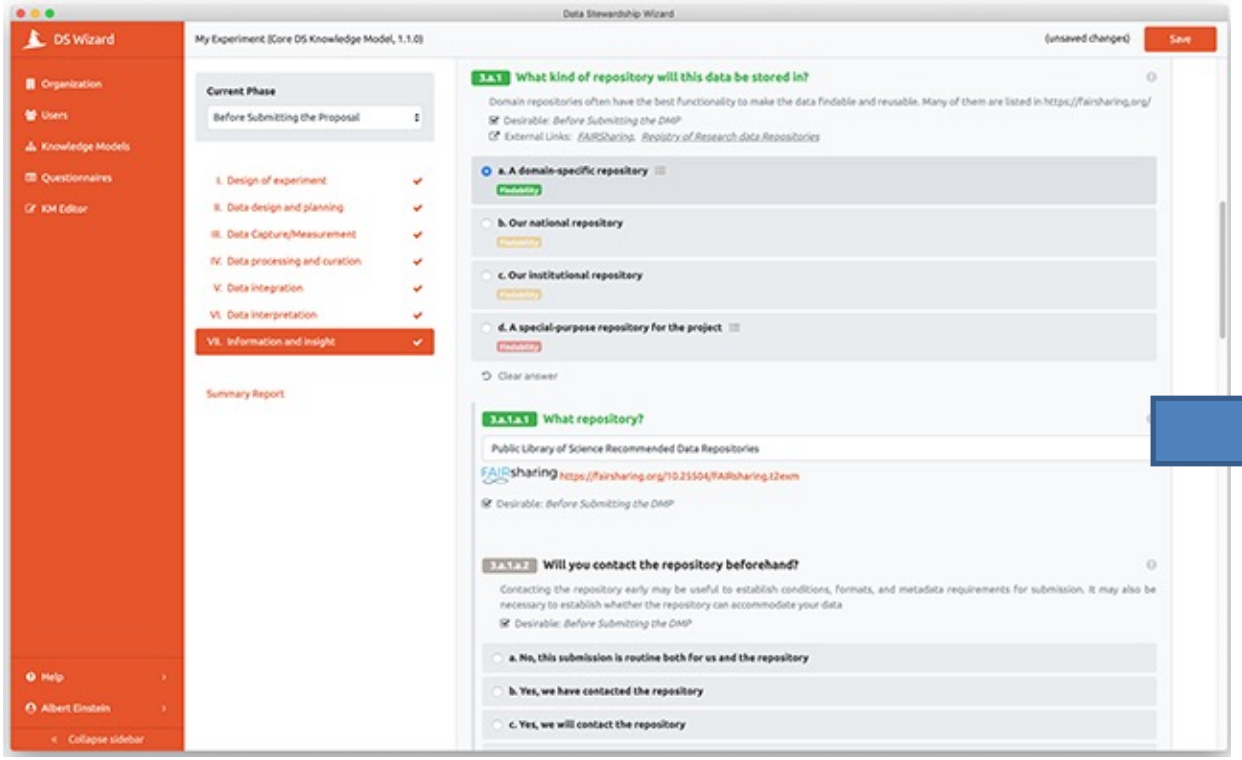
A DMP is document that should help the scientists to properly curate and manage their data
DMPs are more and more frequently requested by funders (H2020, ANR, DFG, …) .

They could also be of interest for RI to raise the awareness of our users regarding data management questions and to collect information on the potential re-use of the data.
With DMP users should know what to expect for the RI and what are there responsibilities.

We (ESRF, ESS, …) are opening new DMP services for users to:
- Help them fill their duty in regards of the funders
- Improve their understanding of our DP and implementation

User offices may receive request from users asking help for filling their funders DMP, a DMP service will definitely help to clarify what users should expect from RI and what do we expect from them.

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



- A DMP service is a web form
- Prefilled based on RI's Data Policy and practice
- This is a live document, enriched by information from the experiment
- Users have just to fill in a limited set of information (was the experiment successful?, Who could be interested by the data, …)
- They can generate the document for their funder.

3a. How will data and metadata be stored and backed up during the research?

✖ *This question has not been answered yet!*

✖ *This question has not been answered yet!*

Backups

All data are archived and ingested into the metadata catalogue automatically right immediately they are acquired.

Beyond 2 months the data is migrated to tape storage and remains accessible via the ESRF data portal. Two copies are stored on separate tape systems. Writing data back from tape to disk is fully automatic when access to the data is requested via the data portal (data.esrf.fr). Writing data back from tape to disk can also be done manually by the system administrators. This may be requested via a JIRA helpdesk ticket.

Backup Responsible: The IT infrastructure management is in the remit of the ESRF UNIX team. This team also makes sure that disk storage systems and tape storage are upgraded and maintained as required.

3b. How will data security and protection of sensitive data be taken care during the research

People who can access the data during the project

Access to the data is restricted to the experimental team for the duration of the embargo period (3 years). During this period the principal investigator can add other users as collaborators (via the data portal). Collaborators have access to the data and the electronic logbook as any other member of the experiment. After the embargo period data are made accessible to everyone i.e. open data. The PI can shorten the embargo period on request.

Identity is managed by the ESRF Single Sign On (SSO) platform. The implementation of Access Control Lists (ACLs) on the ESRF file systems allows fine-grained access control to directories and files depending on the user's identity. Access to

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Data transfer services for users

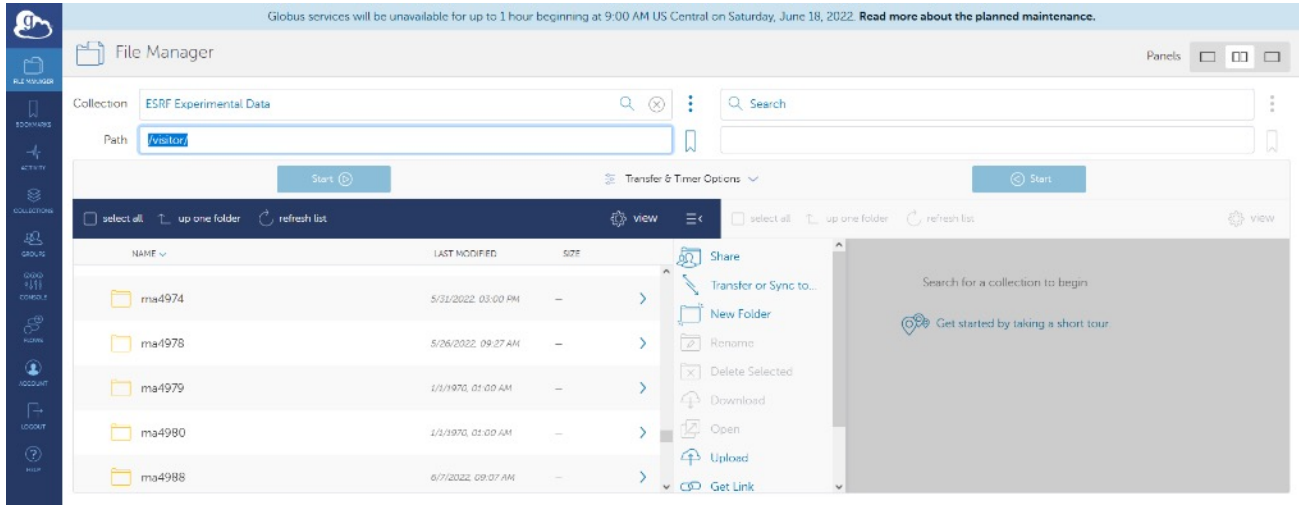
Solution for users to transfer data (especially large datasets) to their home lab/univ

We all usually propose data transfer solutions like:

- Data portal download
- SSH Based (RSYNC, SFTP, SCP)

What about very large data? data larger than 1TB, transfers spanning many days:

- Globus Connect is the solution that was identified to address these needs

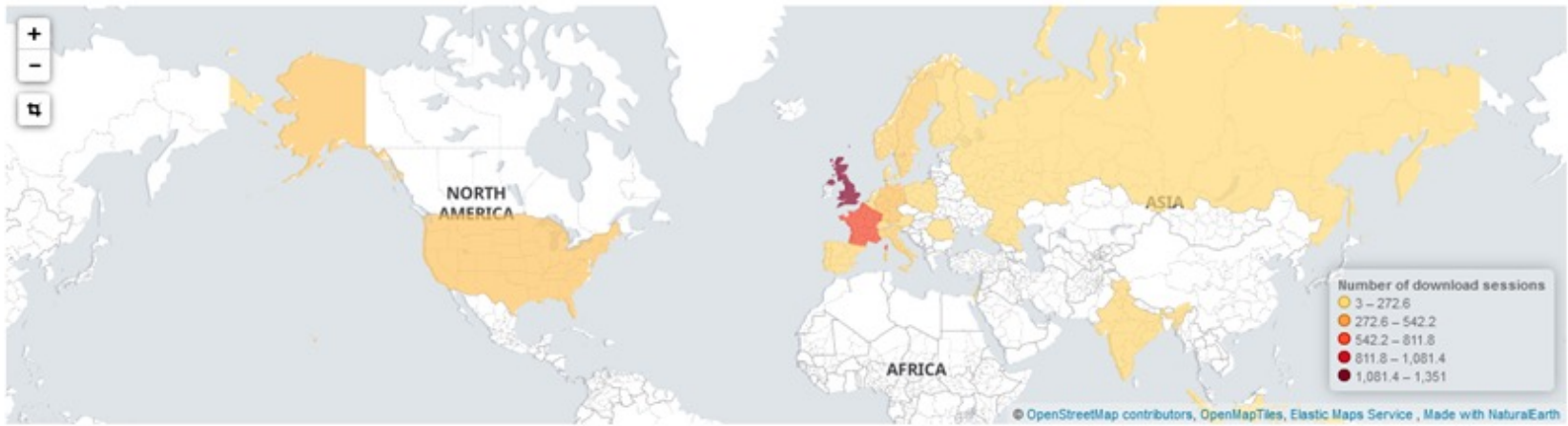


Users can transfer very large data sets reliably (+100TB) if they have the capacity to host these data.
Graphical and user friendly solution. Fits a large number of use cases
Already implemented by DESY, diamond, ESRF, EU-XFEL, PSI, Max IV, SOLEIL,

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Globus/Usage

Globus/Regions



Globus/Top users

username.keyword: Descending	Sum of nbytes
[redacted]	40.187TB
[redacted]	23.645TB
[redacted]	21.049TB
[redacted]	20.73TB
[redacted]	20.036TB
[redacted]	19.273TB

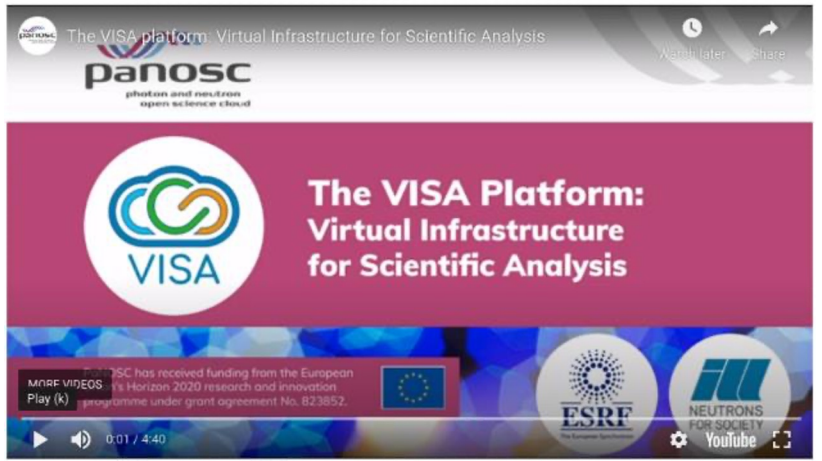
Globus/Top file

file.keyword: Descending
/data/visitor/mx2333/id30b/20220227/RAW_DATA/baduv/baduv-101_x11_cs/ref-baduv-101_x11_cs_1
/data/visitor/mx2333/id30b/20220227/RAW_DATA/baduv/baduv-101_x11_cs/ref-baduv-101_x11_cs_1
/data/visitor/im62/id16b/20201210/WO2C1_ht_120ms_25nm_bis_/WO2C1_ht_120ms_25nm_bis_v1_0
/data/visitor/ma4857/id19/RR2_3/stitched/RR2_3_z0016bits/z00_0010.tif
/data/visitor/ma4857/id19/RR2_3/stitched/RR2_3_z0016bits/z00_0003.tif
/data/visitor/ma4857/id19/RR2_3/stitched/RR2_3_z0016bits/z00_0001.tif

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Users need remote data analysis

- Why it's important
  - next generation of data analysis in globalised research
  - **remote access** to facilities
  - **Mitigate the impact of large datasets** (time to scientific articles, difficulties processing data, …)
- What we do
  - develop and deploy **VISA** platform
  - make **Jupyter** notebooks available at all sites
- Provides remote data analysis services in a web browser with access to
  - Experimental data
  - Scientific software
  - Compute resources
  - Support (IT and Scientific)



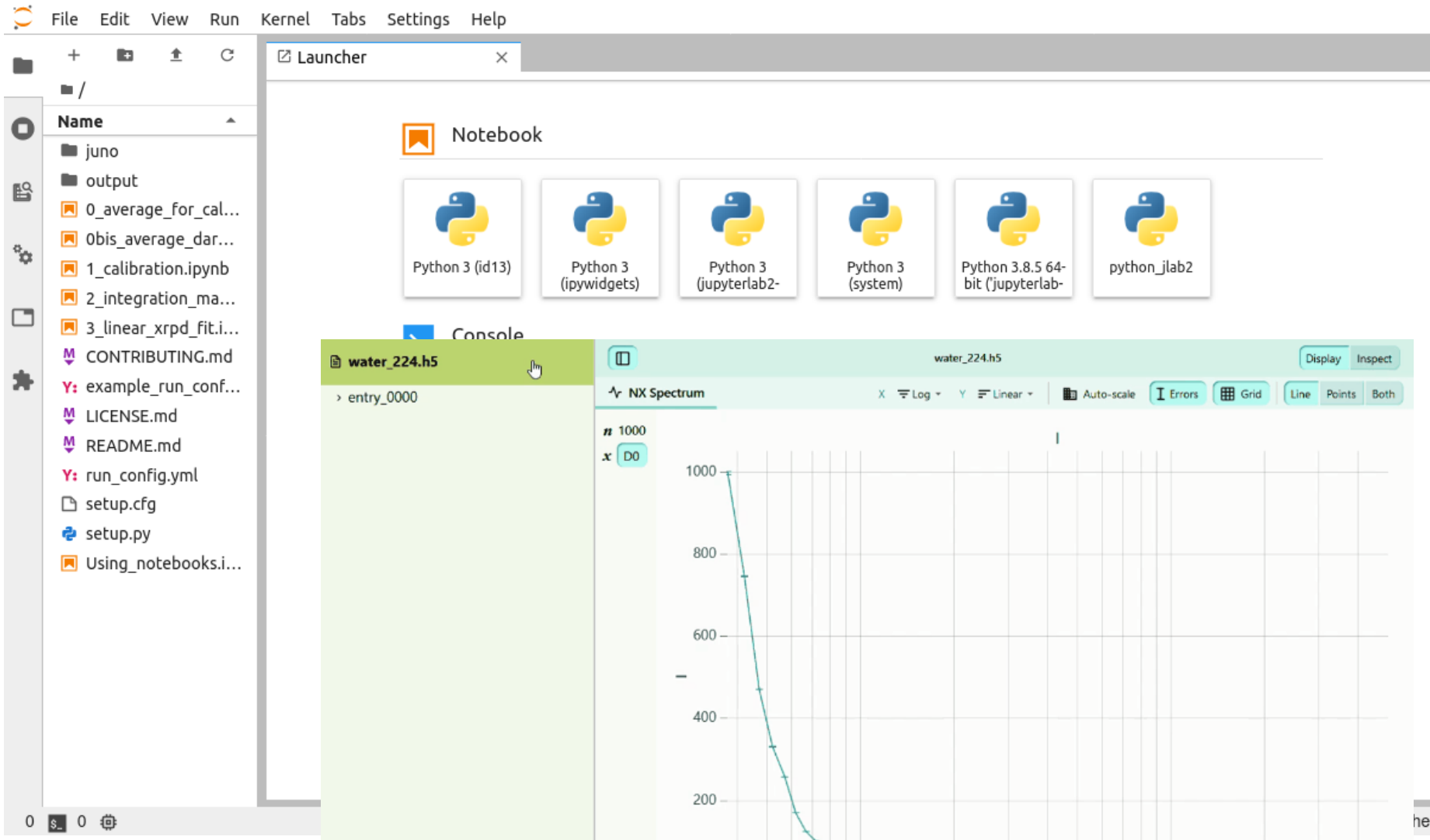
https://bit.ly/VISA-video

panosc photon and neutron open science cloud

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

Remote analysis + visualization with Jupyter notebooks

JupyterLab has been widely adopted as remote analysis tool



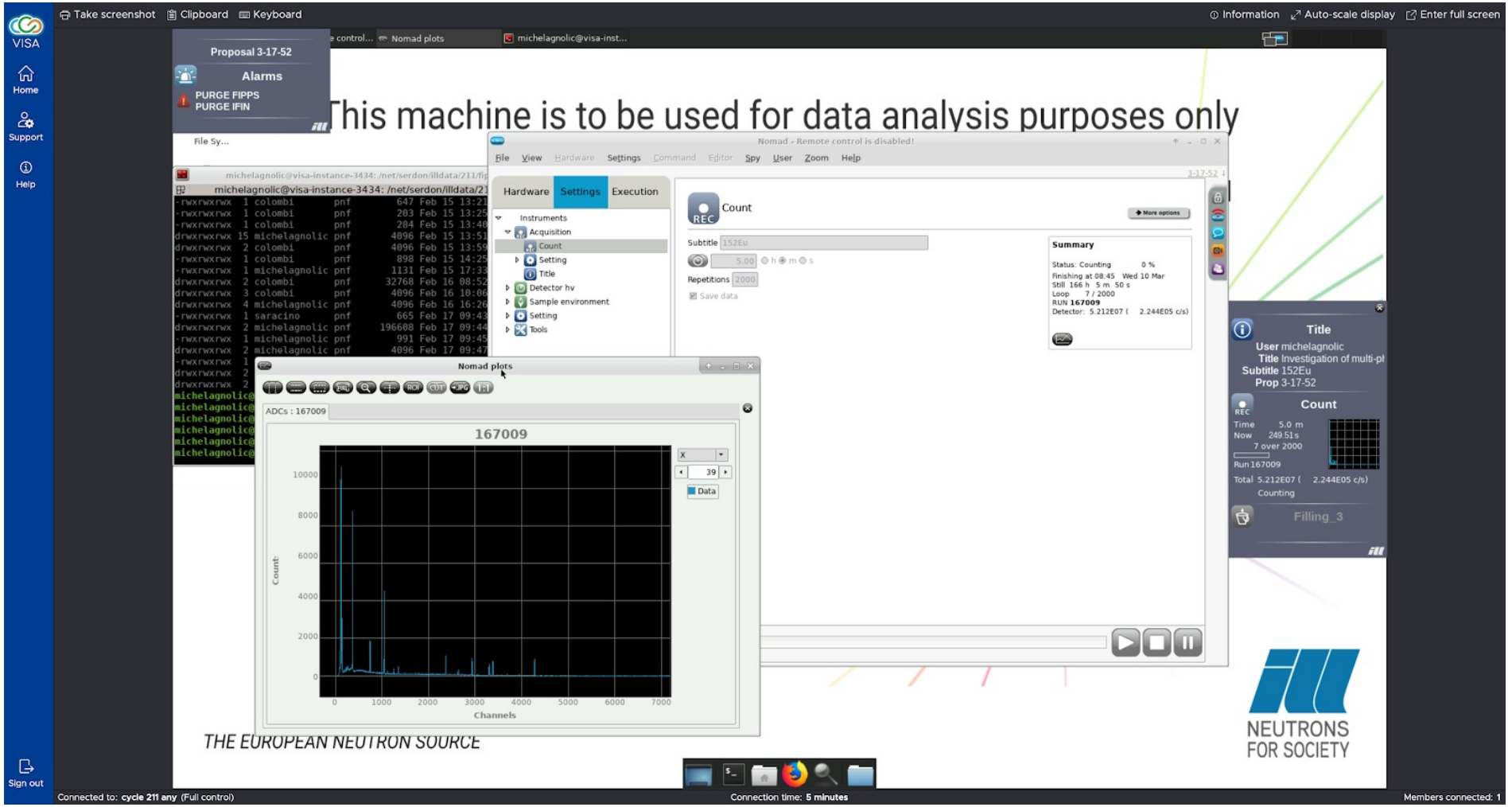
PaNOSC and ExPaNDS projects have received funding from the Europe
and innovation programme under grant agreements 823852 and 857641,

VISA

Users get access to a desktop, like if they were on RI site.

They can exchange with other scientist and receive support through screen sharing

Fully automated workflow, VMs are created by the users and destroyed after N days, data are preserved.



Remote data analyses services are being deployed in the different RI
They should facilitate the processing of the data for the users and help to mitigate the impact of the growing data sets.

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Authentication and Identity Management (AAI)

- Originally
  - Scientist needed a different identity at each facilities to access their services.
    - Lots of passwords to remember and to loose
    - Difficult with cross facility services.

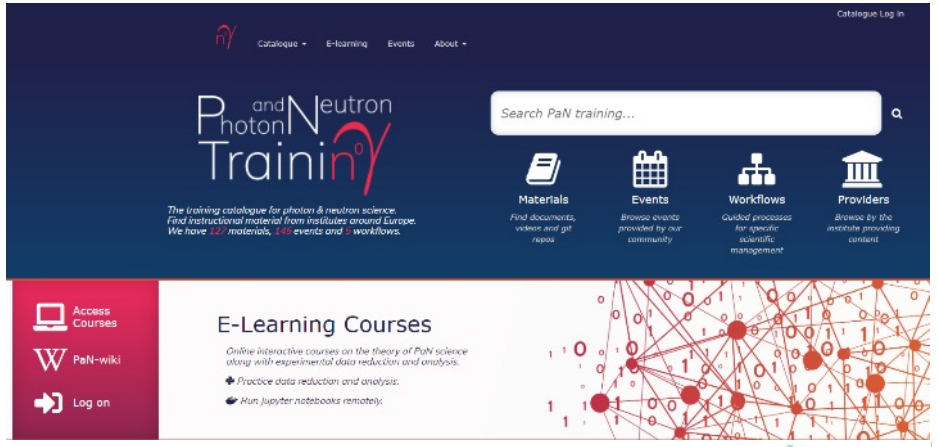
- UmbrellaID
  - Scientist only needs one identity with UmbrellaID (on top of home identity)
  - 'Catch all' identity providers are no longer state of the art and not trusted.
- Very soon (Fall 2022)
  - Scientist only needs the one identity from his/her home facility!
  - Single Sign On: for cross facility services and beyond (EOSC AAI federation)
  - Security: We are part of the European wide CERT system.
  - Extended metadata (email, affiliation, ORCID, …) will be available when the users login.

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

training and learning platform

- The lack of a central platform for PaN Teaching and Learning has been identified by the facilities
- We evaluated available, state of the art technologies for teaching and data collection platforms
- We introduced a PaN training platform
  - **to create**/store courses and
  - **to collect** existing material
- We are **re-using** successful projects developed by
  - **Elixir** (TeSS)
  - **SINE2020** e-neutrons and
  - **AAI UmbrellaID/eduTeams**

https://pan-training.eu/



PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Training and Learning Platform

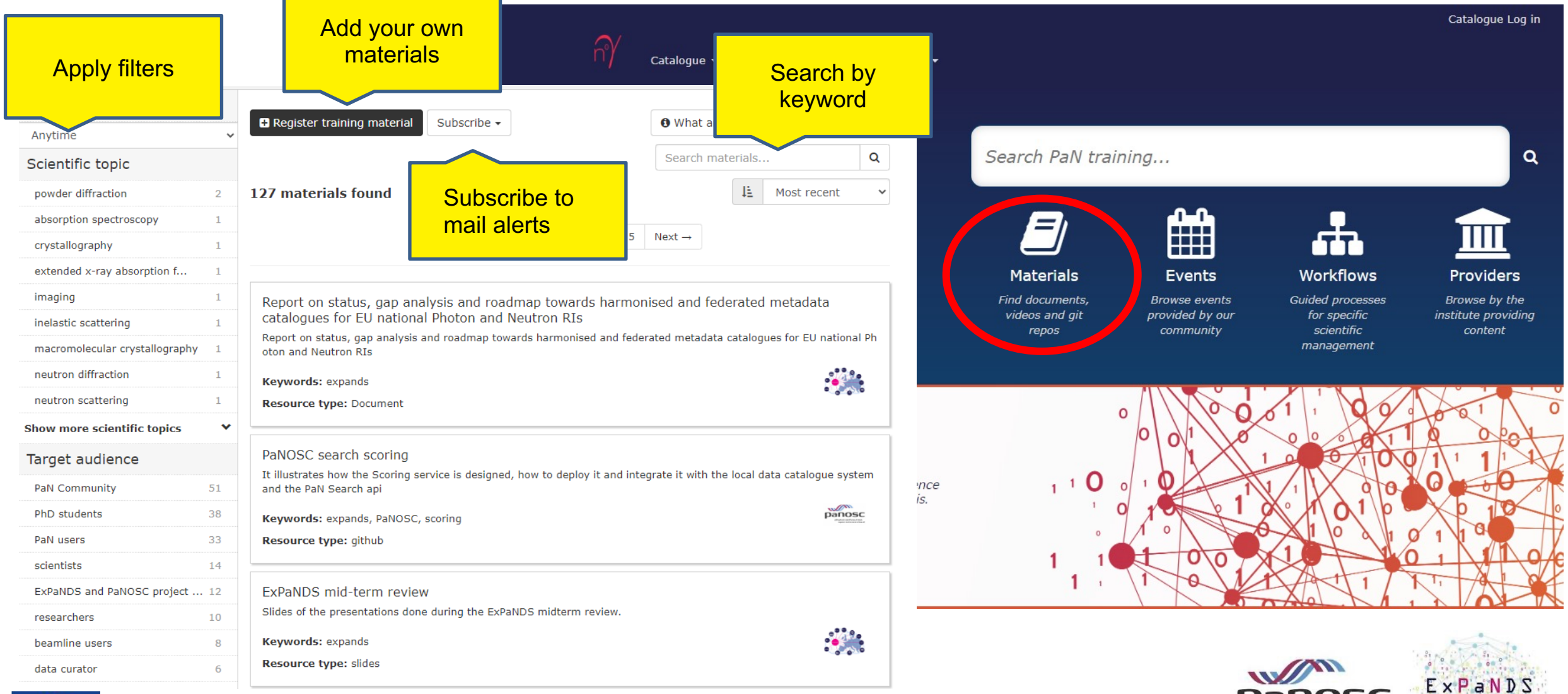
https://pan-training.eu/

Apply filters

Add your own materials

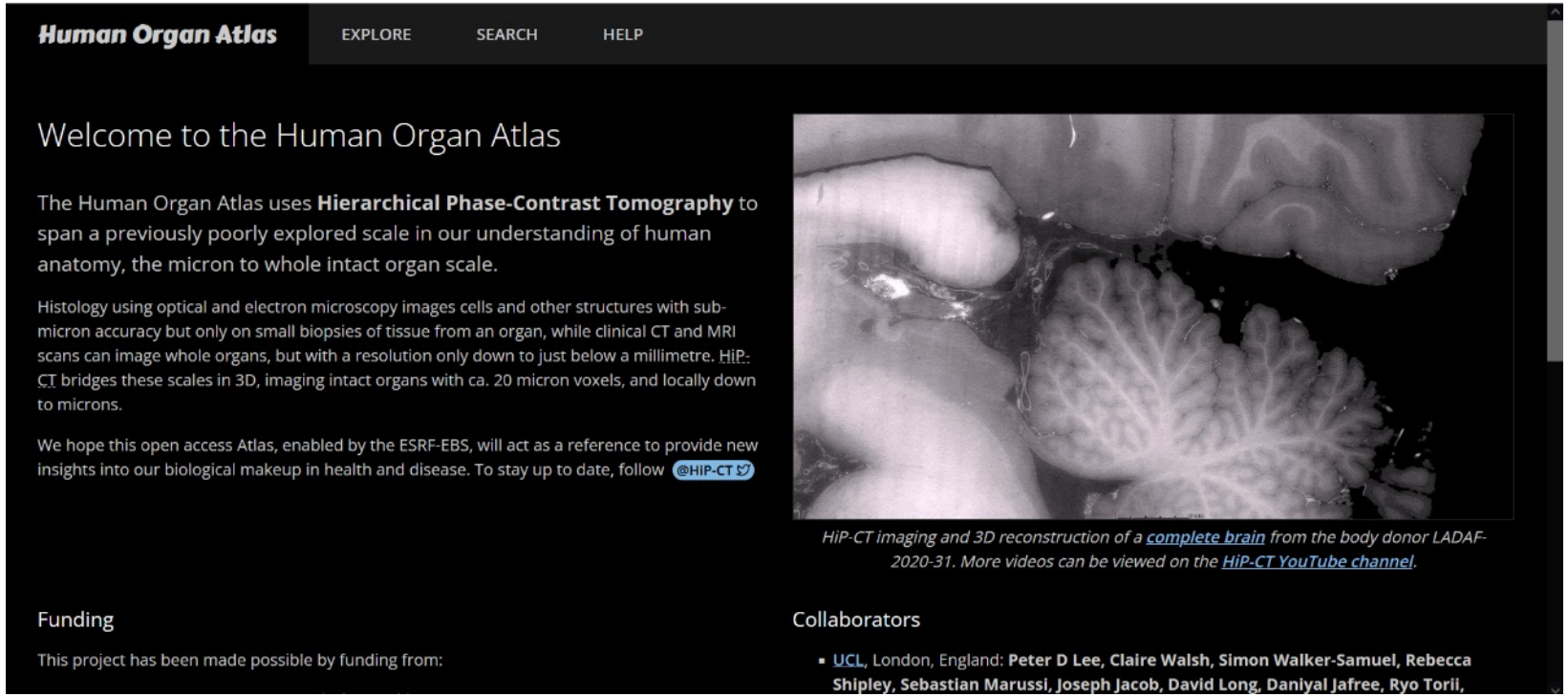
Search by keyword

Subscribe to mail alerts



PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Open data brings a new type of users



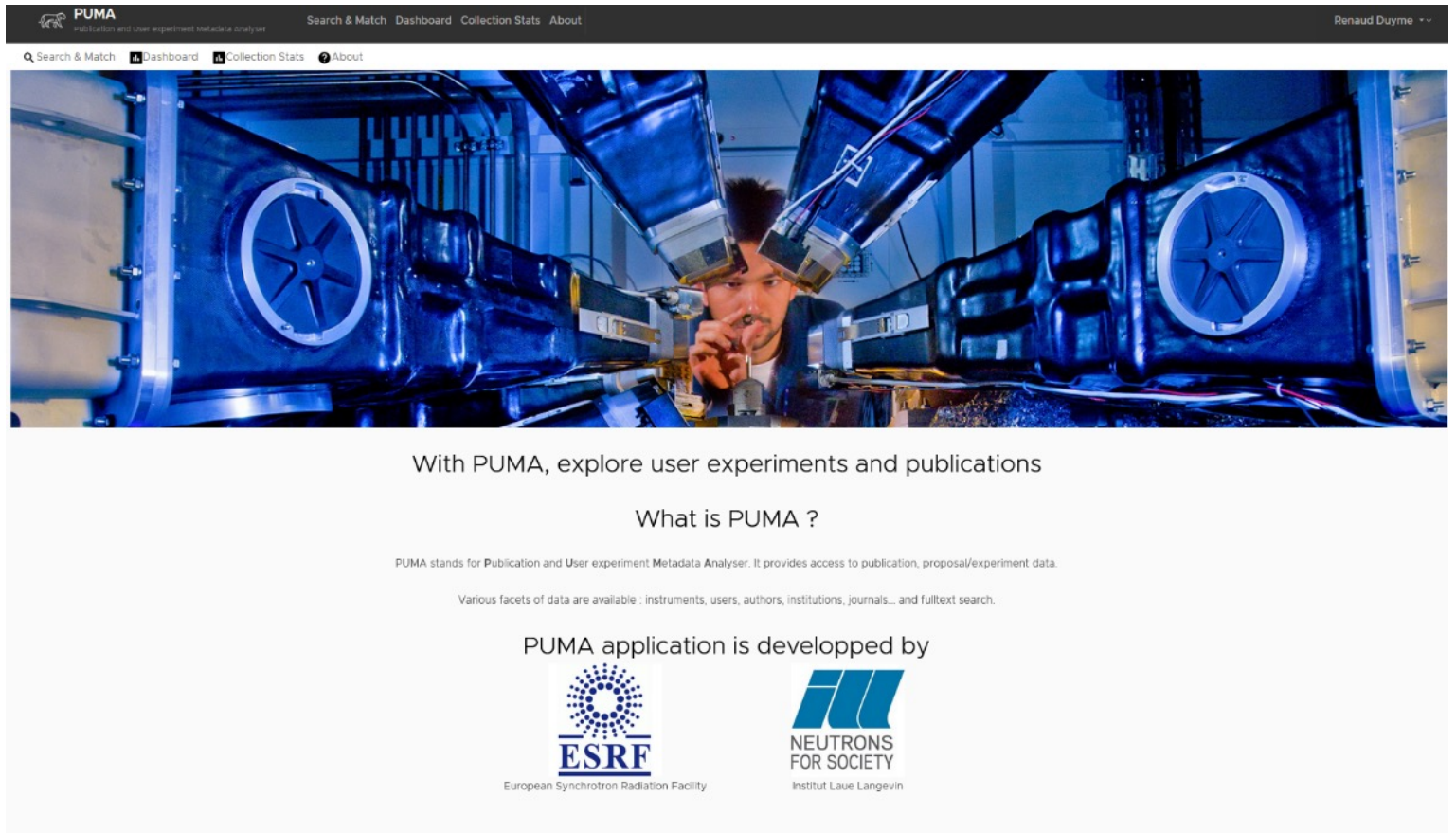
https://human-organ-atlas.esrf.eu/

How do we support these new users that do no submit scientific proposal?
What is the role of user offices?
How do we count these publications?

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

PUMA

- PUMA V1 : Built for the ILL through H2020 funded project FILL2030 –
  - PUMA BI – Business Intelligence (**Pu**blication **Ma**tcher)
- PUMA V2 : ESRF Funded project STREAMLINE : sept 2019 => sept 2022. Migration and enhancement.
  - **P**ublication and **U**ser experiment **M**etadata **A**nalyser



STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

Challenge

Need for reporting on **activity and scientific impact of research facilities**.

- Current stats are usually relying on institutional repositories tools : **Publication statistics** (citations, journals) using publications list coming from library repositories. **Proposals stats** using data from User Office databases.
- **Instrument scientists** also have to report on **their instrument**: publications, proposals, science topics covered, and techniques used.
- **Facility management board, business/industrial offices** regularly request specific reports based on countries, user institutions (educational/gov/corporate...)

These analysis are **time-consuming** and not always possible to achieve.

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

Existing solutions

Required solution : **CRIS (Current Research Information System).**

A CRIS is a database or information system to store, manage and exchange contextual metadata for the research activity funded by a research funder or conducted at a research-performing organisation. Ex: Clarivates InCites (TM), Elsevier SciVal (TM).

These allow **publication** analysis but are however **not tailored to hold specific entities** used by research facilities : **instruments, proposals**, datasets.

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

PUMA is a CRIS for research facilities

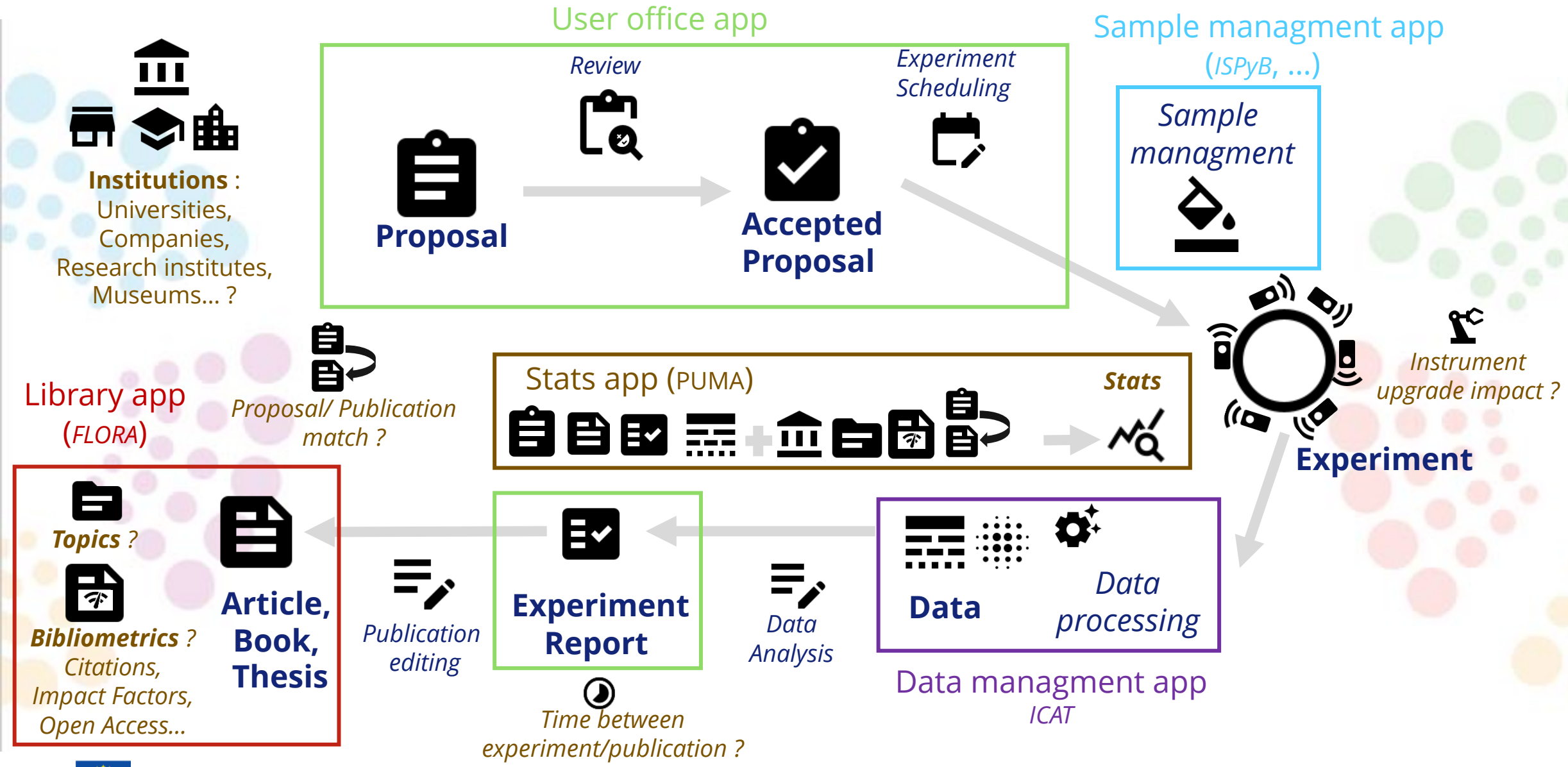
Managers, director of research, instrument staff need an efficient way to get overview of science produced by their instruments. PUMA aggregates meta data of proposals/experiments and publications documents : instruments, authors, institutions, abstract, topics, journals, citations.

PUMA provides :

- dashboards for facility instruments (proposal and publications metrics, citation/journal metrics, institution maps...)
- advanced search feature within all indexed documents.

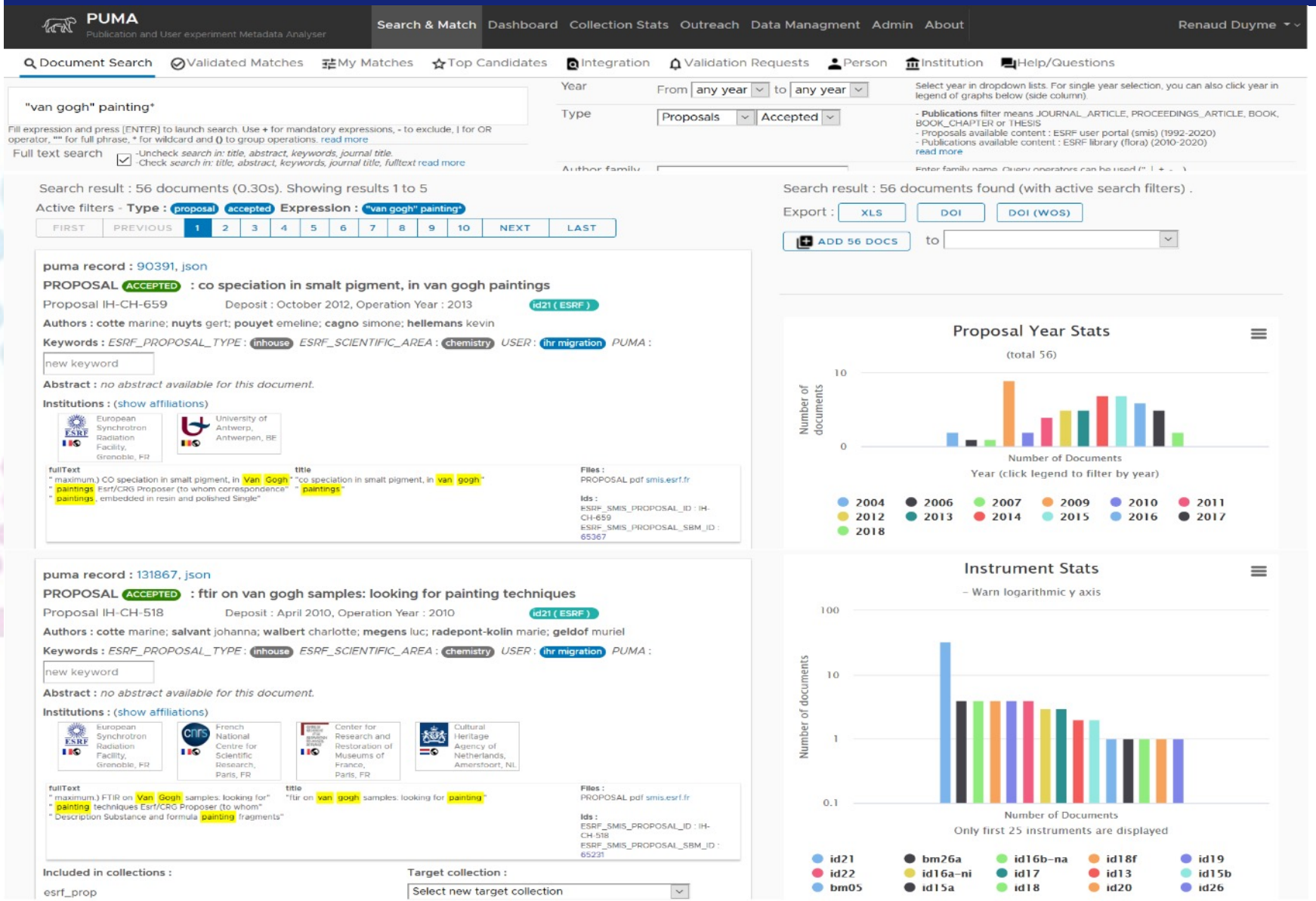
STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

PUMA in research facility architecture



STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

PUMA Search Module



- Search proposal or publication
- Full text search within pdf

Left Column (result):

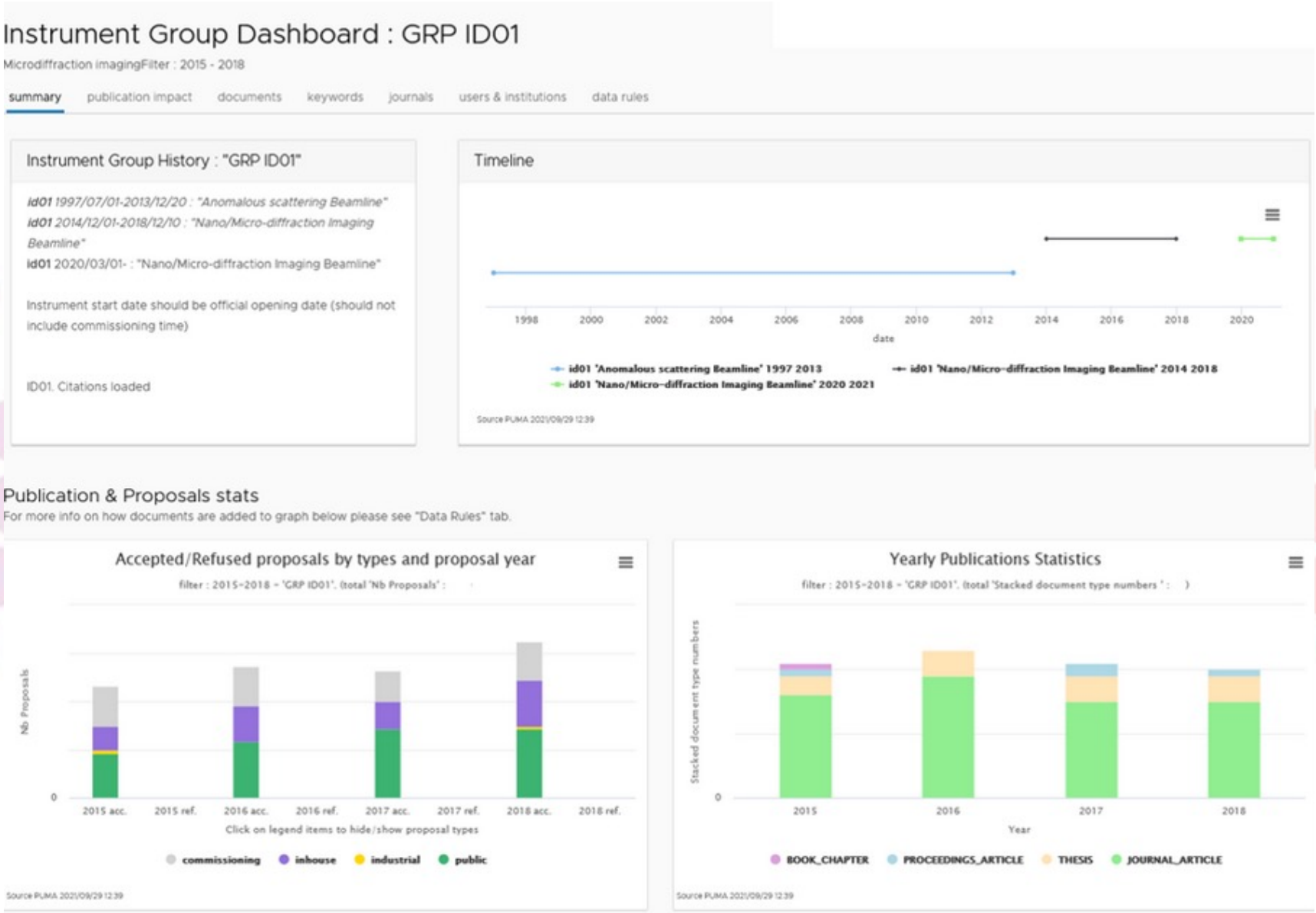
- Record detail, highlight search match, keywords, instrument, institution, pdf link…

Right Column (aggregated graphs):

- Document by year
- Instruments used
- Main institutions
- Main authors
- Topics/Keywords…

ınder grant agreement No. 870313.

Instrument dashboard – History, proposals and publications by type



STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

Instrument dashboard – shift usage

User countries - Experiments

Number or experimentation shifts used by country

filter : 2017–2022 – 'GRP ID01'.
An experimentation shift is associated to 'Principal Investigator' ('main proposer') institution's country.
warn : co-proposer and other experiment participants are not used to avoid associating a single experiment to multiple countries

	total:1611.50	count
	654	
FR	438	
DE	160.50	
GB	69	
CH	69	
SE	54	
AT	41	
RU	29	
NL	24	
IL	18	

1 - 10 of 16 rows

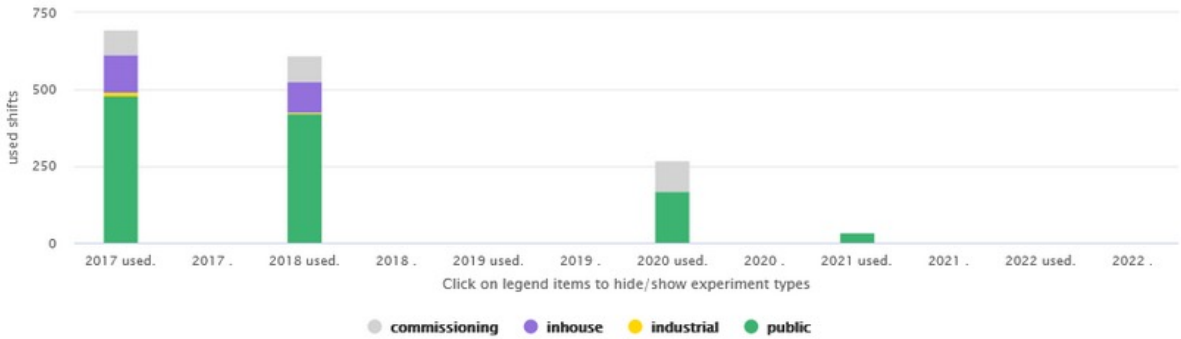
1 / 2

EU ZOOM

NO_INST** : User affiliation not resolved to an official institution (or at least country/city) by PUMA. Reason : affiliation resolver not ran on record or no result found for affiliation. See : Affiliation to institution match stats

Experimentation used shifts by type and year

filter : 2017–2022 – 'GRP ID01'.
Cancelled experiments and associated shifts are removed.
(total 'used shifts' : 1611.50)



Source PUMA 2022/02/03 14:27

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

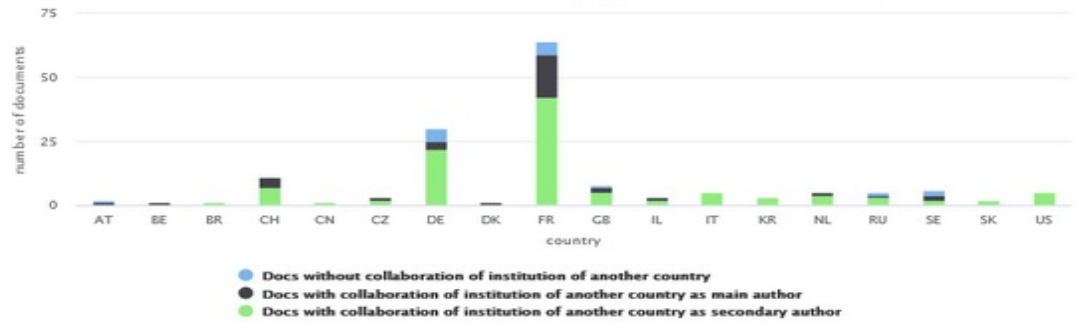
Collection report - Country collaboration

Country collaboration in collection

A document in collection can have authors coming from multiple countries. Below are displayed country collaboration figures

Collaboration rate of each country 2018–2021 esrf_grp_id01_proposal

All countries with collaboration are included in this graph (total 'number of documents' : 156)



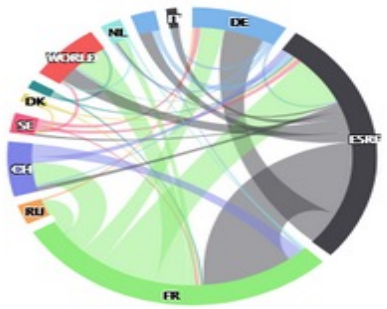
Source PUMA 2022/02/07 11:46

Country	Collab (Main Auth.)	Collab (Secondary Auth.)
AT	1	0
BE	1	0
BR	0	1
CH	4	7
CN	0	1
CZ	1	2
DE	3	22
DK	1	0
FR	17	42
GB	2	5
IL	1	2
IT	0	5
KR	0	3
NL	1	4
RU	1	3
SE	2	2
SK	0	2
US	0	5

1 - 18 of 18 rows

Country collaboration stats 2018–2021 esrf_grp_id01_proposal

Only ESRF members countries are included in this graph.



Source PUMA 2022/02/07 11:46

From	To	Nb docs
DE	ESRF	1
FR	RU	7
ESRF	DE	15
FR	CH	10
FR	DE	9
FR	ESRF	22
SE	DE	2
SE	FR	1
SE	ESRF	2
DK	FR	1
DK	ESRF	1
DK	SE	1
BE	FR	1
BE	ESRF	1
BE	WORLD	1

1 - 15 of 36 rows

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.

ILL and ESRF are willing to set up a collaboration to extend the use of PUMA, are you interested?

If yes do not hesitate to contact us.

Thanks you.

STREAMLINE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870313.