10-11 November 2020

PaNOSC & ExPaNDS Annual Meeting

WP3 – Data Catalogues for Photon and Neutron Science

Alun Ashton
ExPaNDS – Paul Scherrer Institute

Tobias Richter
PaNOSC – European Spallation Source ERIC

November 2020

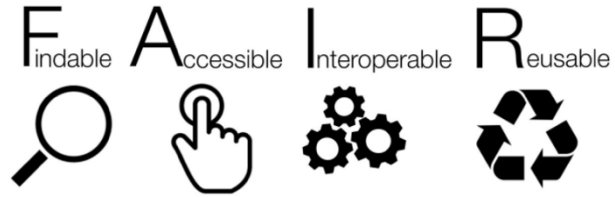
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

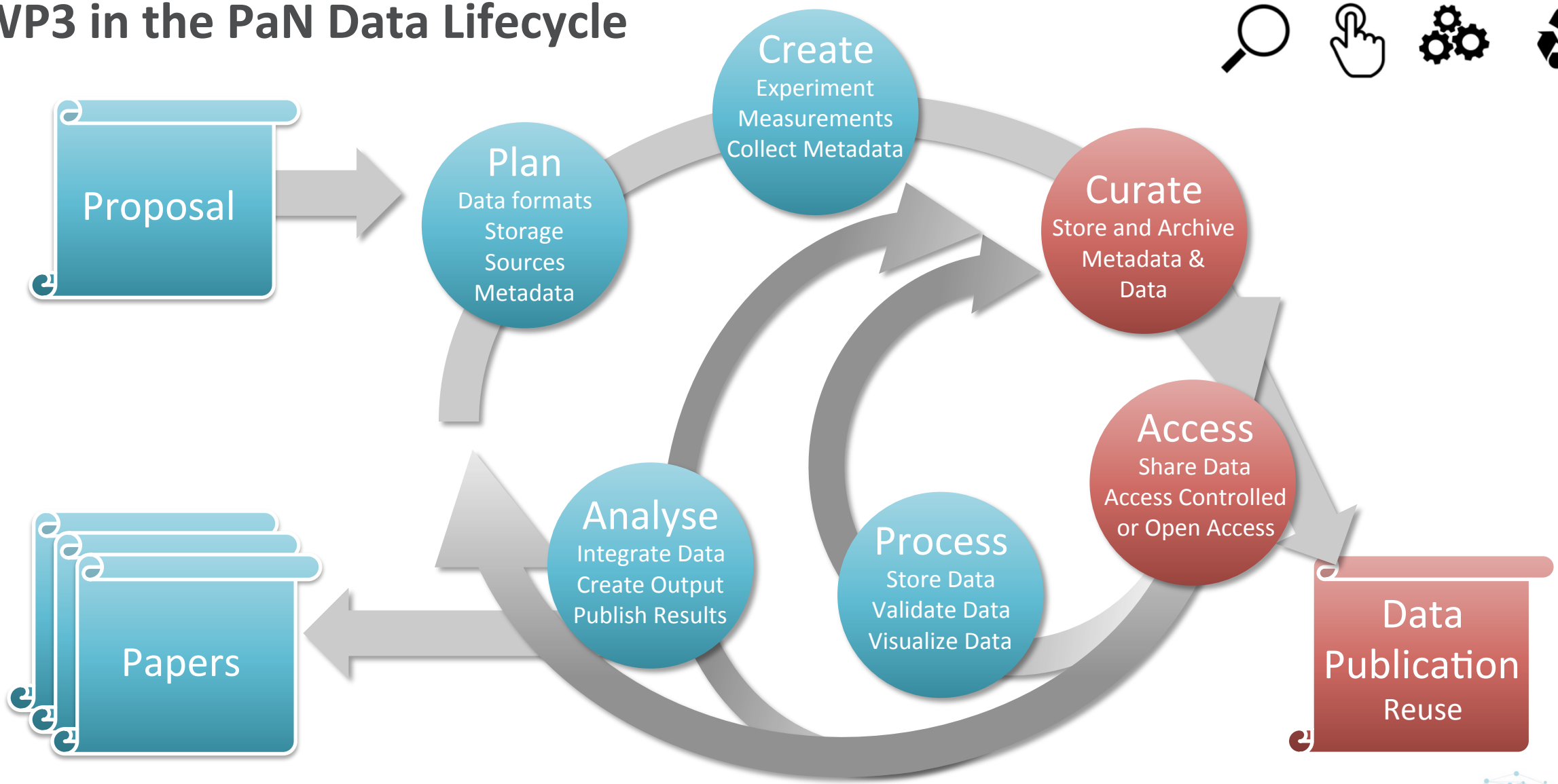
Overview

- Introduction
- Use cases
  - Photon and Neutron User – Federated PaN Search API
  - Interdisciplinary Researcher – Repository Metadata Harvesting
- Summary

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.

WP3 in the PaN Data Lifecycle

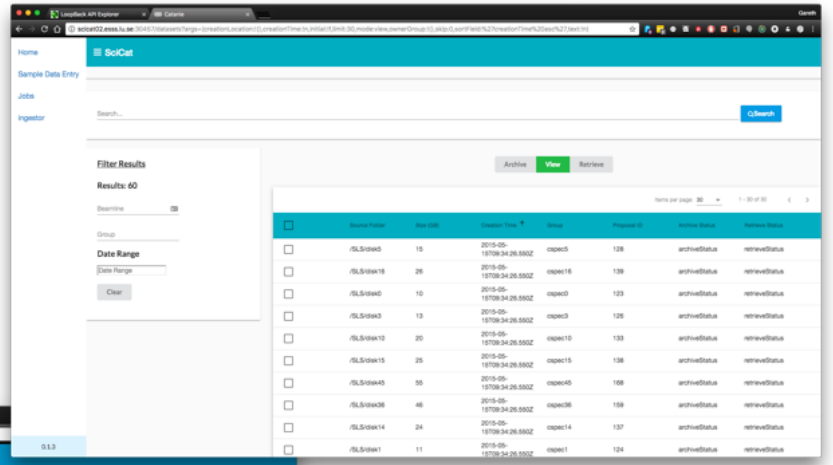


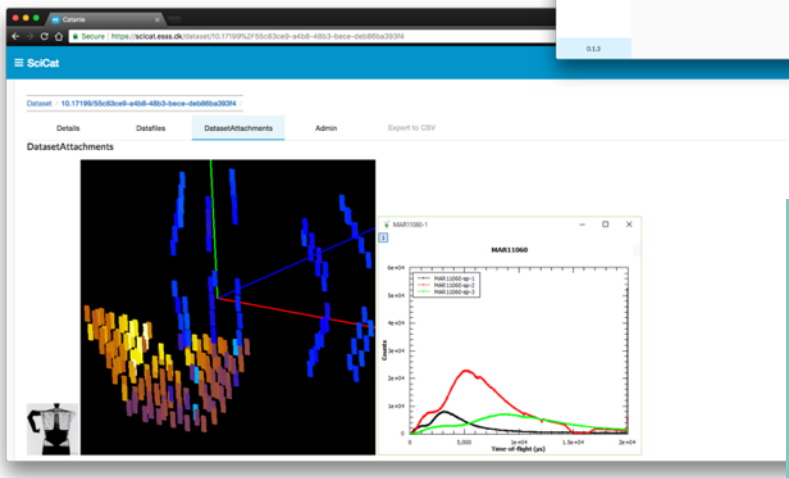


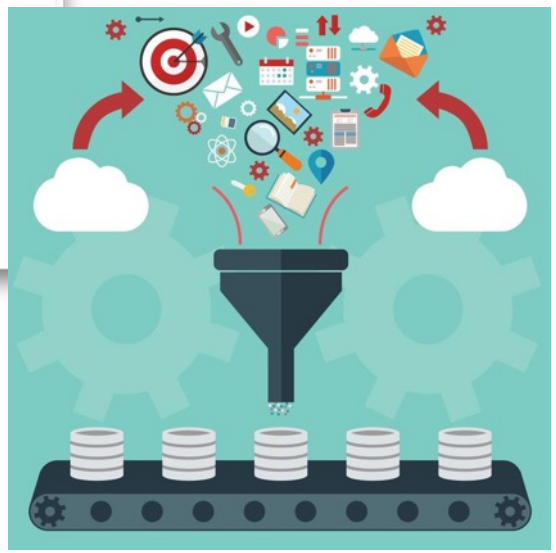
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Data Catalogues – Common Functions

- Catalogues are an entry point to other EOSC services, often provided by the facilities, like Download, Processing, Analysis or Visualisation Services.
- Store reference to raw data and the relevant metadata to identify interesting dataset to work with.
  - Information available often includes:
    - Proposal Information
    - Sample Information
    - Experimental Parameters
    - Previews
    - Provenance
    - Relationship to derived datasets
    - Relationship to people
- Catalogues also Facilitate Data Re-Use, Sharing and Publication via PIDs and DOIs.







ORCID
Connecting Research and Researchers

doi

panosc
photon and neutron open science cloud

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

Different starting conditions at different facilities

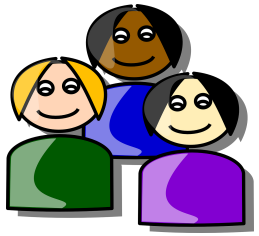
6 European PaNOSC facilities and 10 national ExPaNDS facilities

- Few are in construction and generate no user data yet
- Most have a (meta-)data catalogue
  - Implementations are often institute specific
  - Others collaborate on SciCat or ICAT
  - Some catalogues stores sufficient data required to locate a specific dataset
  - Many publish data after an embargo with a DOI
- Most write HDF5 data files
  - All aim to follow the NeXus standard for raw data
  - Follow common schema for data, metadata

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.

Use case One – A PaN facility user

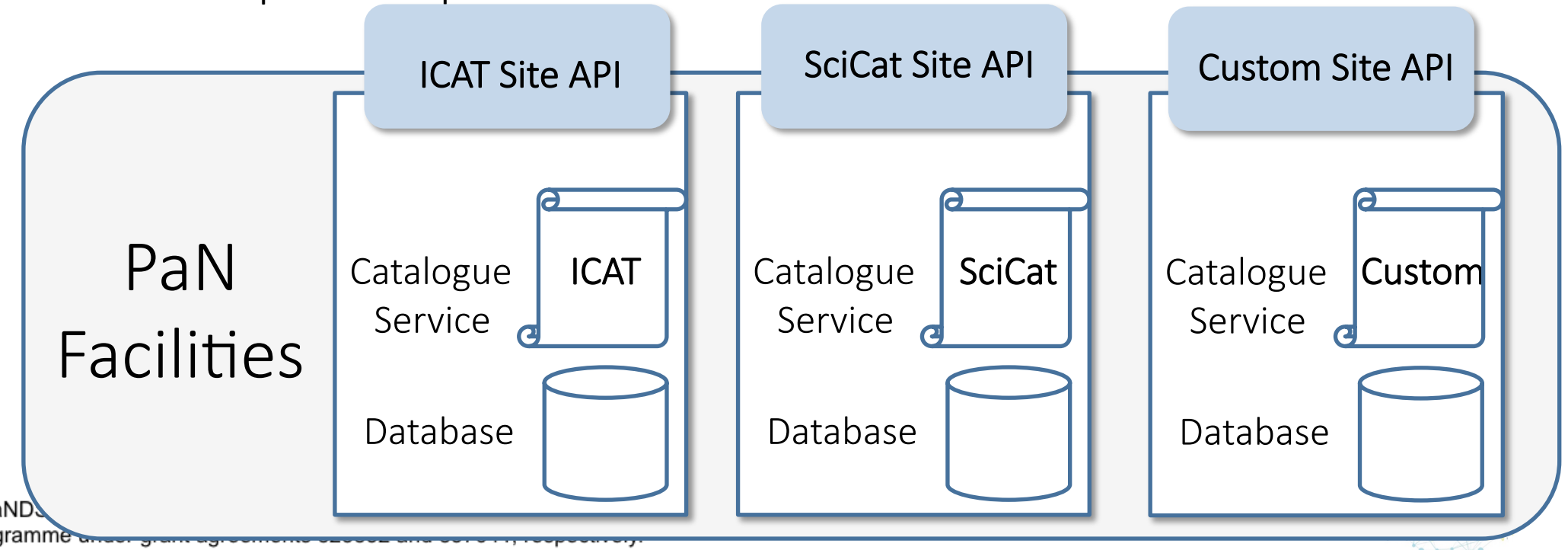
PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.



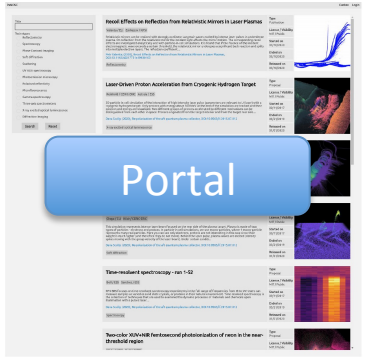
PaN Facility Users/Data Owners

- Securely search for their own data at a facility
- Securely search for their own data across facilities
- Searching for open data
  - Based on Experiment Configuration
  - Based on Sample and Sample Conditions

- Why?
  - Process/reprocess data (WP4)
  - Integrate Interoperable data (WP4)
  - Publish data/Link to publication (WP2)



PaNOSC and ExPaNDS
and innovation programme



Abstraction

Common REST API

Adapter Interface Layer

ICAT Site API

Common REST API

Adapter Interface Layer

SciCat Site API

Common REST API

Adapter Interface Layer

Custom Site API

PaN Facilities

Catalogue Service

ICAT

Database

Catalogue Service

SciCat

Database

Catalogue Service

Custom

Database

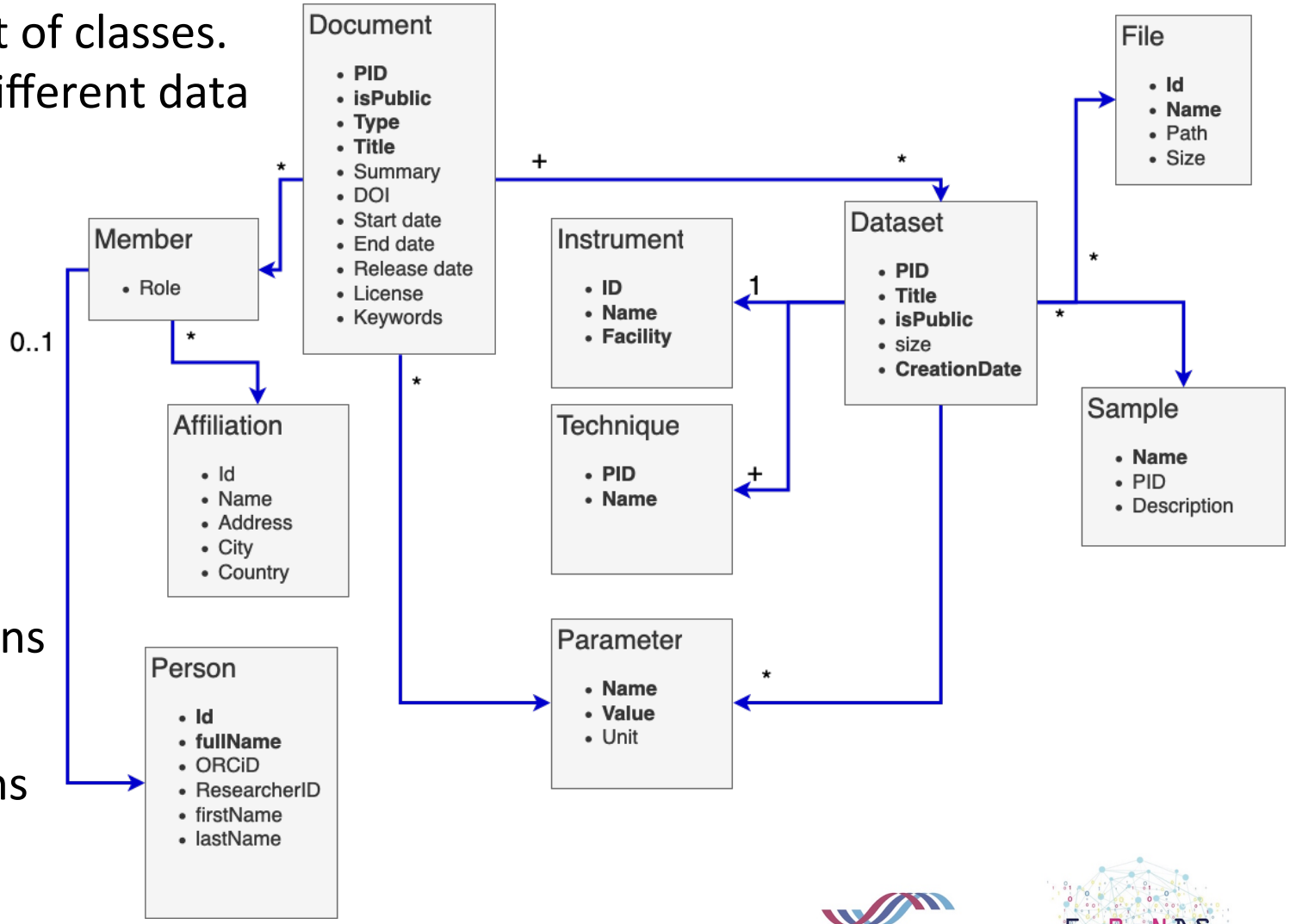
PaNOSC and ExPaNDS
and innovation programme

Common REST API (PaN API V1.0)

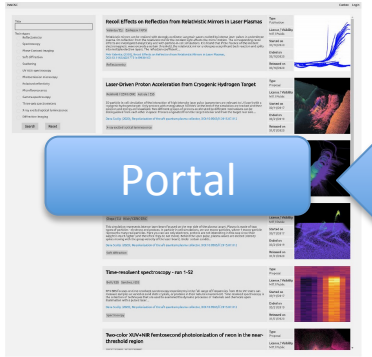
REST API operates on a fairly generic set of classes. That simplifies the implementation at different data catalogues.

Domain Specific Value of the API lies in curated dictionaries ontologies and local mappings for:

- Experimental Techniques
- Parameters for
  - Sample and Sample Conditions
  - Experiment Configuration
- Supported Units with Conversions
- Roles for People



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.



Portal

Federated Data Search Service

Common REST API

Common REST API

Common REST API

Abstraction

Adapter Interface Layer

Adapter Interface Layer

Adapter Interface Layer

ICAT Site API

SciCat Site API

Custom Site API

Federated Processing Service

PaN Facilities

Catalogue Service

ICAT

Database

Catalogue Service

SciCat

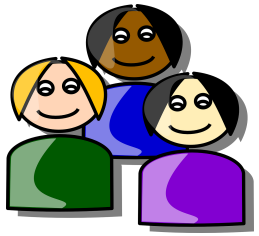
Database

Catalogue Service

Custom

Database

PaNOSC and ExPaNDS
and innovation programme



PaN Facility Users/Data Owners

- Securely search for their own data at a facility
- Securely search for their own data across facilities
- Searching for open data
  - Based on Experiment Configuration
  - Based on Sample and Sample Conditions

- Why?
  - Process/reprocess data (WP4)
  - Integrate Interoperable data (WP4)
  - Publish data/Link to publication (WP2)

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.

Supporting PaN Research

Technique Development

- Pulmonary disease studies in animal models to resolve the full pathway from gas intake to gas exchange.
- Mainly focuses on technical details of acquisition and post-processing steps.
- Published as an open access dataset (https://doi.org/10.16907/7eb141d3-11f1-47a6-9d0e-76f8832ed1b2) and may serve for future studies of lung structure and function in normal and pathological conditions at a new level of resolution and detail.
- ~2Tb of data published February 2020

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.

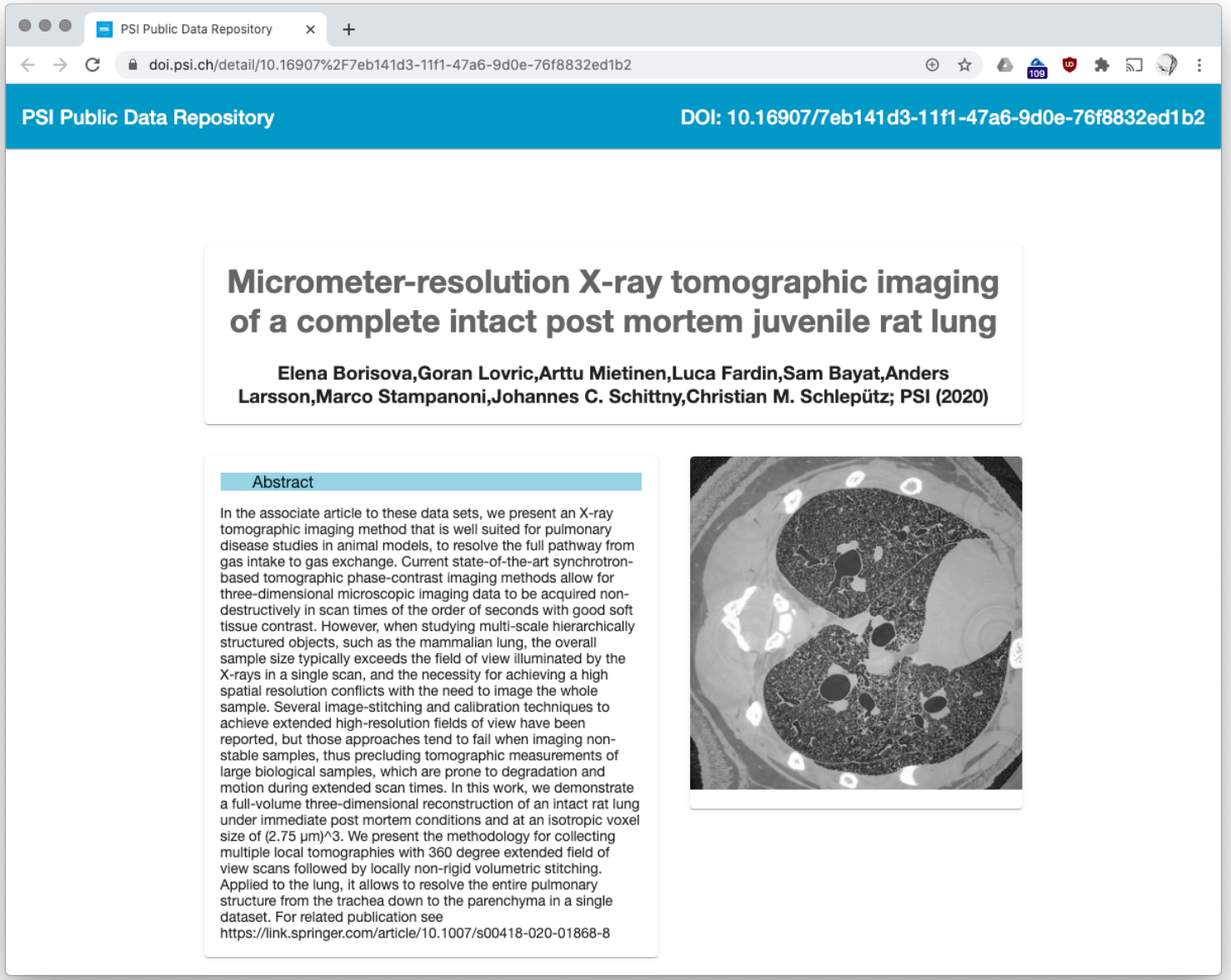
Use case Two – Interdisciplinary researchers

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.

Supporting non-PaN

How to reach the wider community

- Datasets from Photon and Neutron facilities can be useful in many scientific domains
- The domain specific search API requires some knowledge and investment from the potential user
- Generic (EOSC-hosted) data repositories fill the gap: They are available to everyone and carry information on datasets from a wide range of disciplines



imaging rat lung

panosc photon and neutron open science cloud

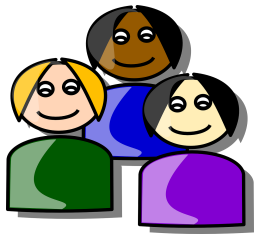
ExPaNDS European Open Science Cloud Photon and Neutron Data Services

Harvesting by Data Repositories

How to get PaN data out there

- Data is pushed to existing services for immediate and wide reach
- Facilities operate public endpoints providing access via known protocols and metadata schemas. PaNOSC chose OAI-PMH, due to broad support.
- Any foreign repository is free to harvest (copy) the entire public metadata collection.
- Meta-data schema is limited to what is common to most science areas (Dublin Core or equivalent). Extensions are possible, especially where there is a potential cross discipline agreement. Example: B2find carries “instrument”.
- This allows a generic user to type in their search terms into some widely known and adopted EOSC service.
  The results will be generated from the fully public metadata in the local database on the service. The original datasets still reside at the facility.

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.



Interdisciplinary Data Repositories

Abstraction

OAI-PMH	OAI-PMH	OAI-PMH
Adapter Interface Layer	Adapter Interface Layer	Adapter Interface Layer
ICAT API	SciCat API	Custom Site API

PaN Facilities

Catalogue Service	ICAT	SciCat	Custom Catalogue
Database			

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.

Common Metadata Expression

PaNOSC and ExPaNDS facilities work together to find common vocabularies for metadata, based on community standards like the NeXus file format most facilities use for raw data.

Finding common ground with other communities makes all exposed metadata more accessible.

We have ongoing discussions around:

- Roles for persons involved with the data
- Experimental or Measurement Technique
- Sample & Parameters
- Instrument & Parameters

Photo by Science in HD on Unsplash

Photo by Jossuha Théophile on Unsplash

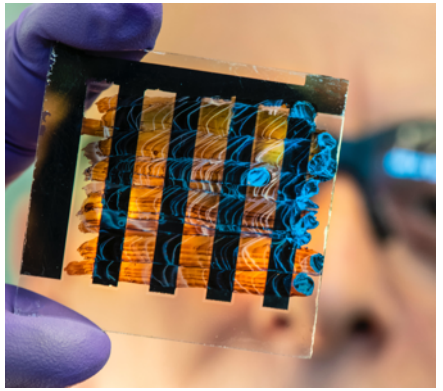


Photo by Science in HD on Unsplash

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.

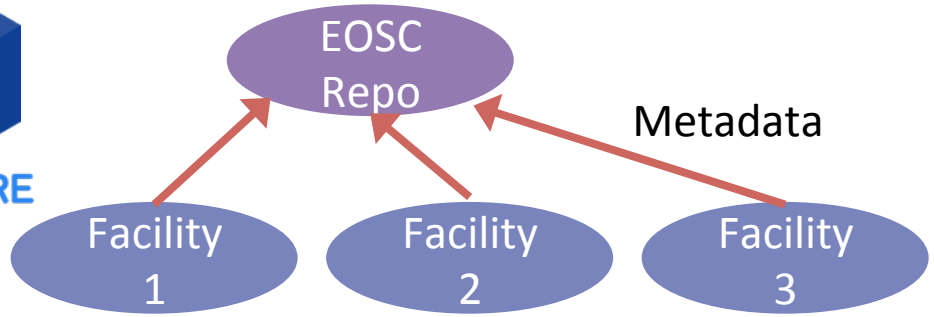
Summary

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.

Two Routes for Opening Datasets

	Harvesting
Exposure Method	Push Metadata resides in third party EOSC repository
Time to Market	Quick (existing solution)
Data under embargo	Cannot be exposed
Richness of Metadata	**Common Schema (Dublin Core initially)**
Target User Group	Citizen Scientists, Interdisciplinary Science Community

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
SciCat
B2FIND
OpenAIRE



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.

panosc
photon and neutron open science cloud

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

Prerequisites

Choosing a Catalogue

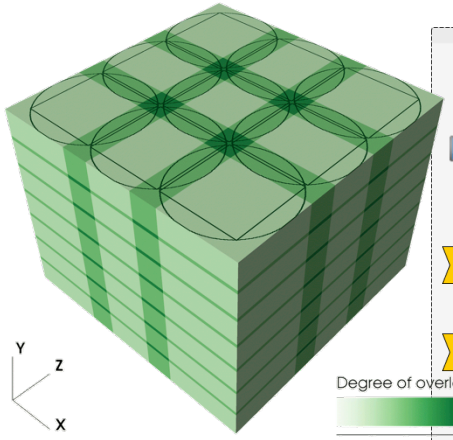
Minimum Product

Facility Integration

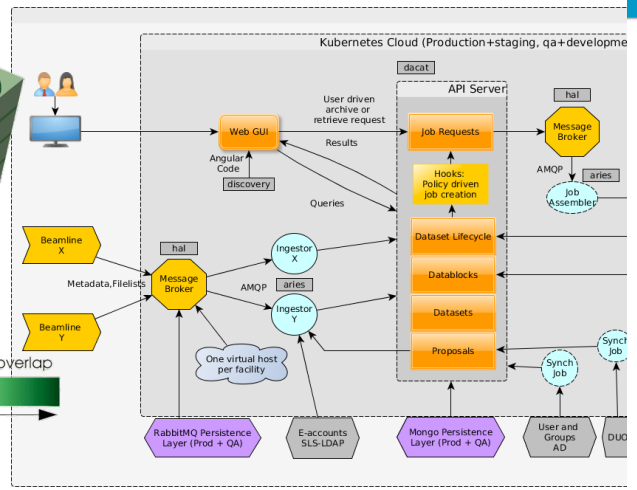
PaN Implementations

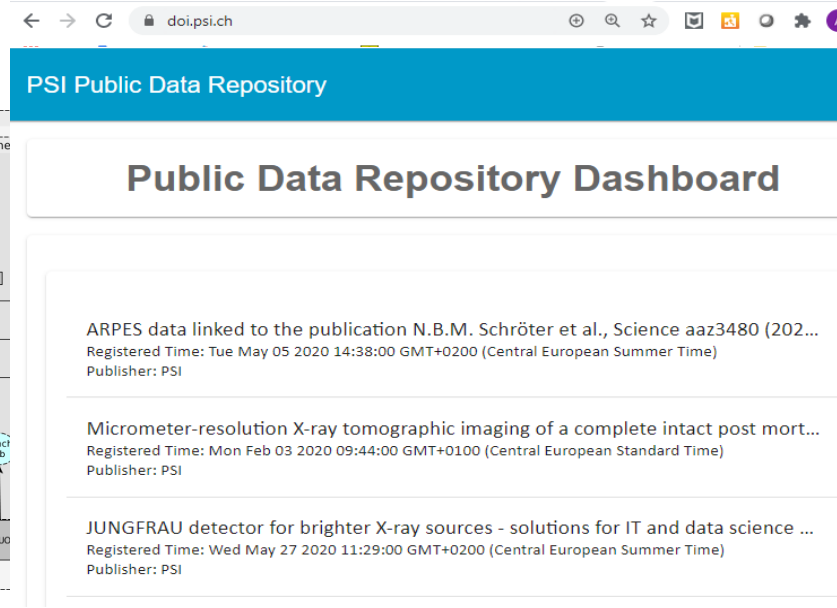
EOSC Integration

PaN Facilities



PBs of data a year





PaN Communities

Findable Accessible Interoperable Reusable

B2FIND

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
http://doi.org/10.17616/R31NJMKD
SciCat

OpenAIRE

panosc
photon and neutron open science cloud

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

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.

Summary

Major activities completed in the past year:

- OAI-PMH implementations and harvesting ongoing
  - with help from B2FIND and OpenAIRE
- PaN search API v1.0 defined and implementation ongoing
- Report on status, gap analysis and roadmap towards harmonised and federated metadata catalogues for EU national Photon and Neutron RIs

Upcoming deliverables:

- PaN Ontology v1.0 to be released in Spring 2021
- Federated search demonstrator to be released in Spring 2021

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.

PaNOSC & ExPaNDS Annual Meeting

Thank you

alun.ashton@psi.ch

tobias.richter@ess.eu

wp3-expands.eu@desy.de (joint address for both 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.