PaNOSC
Photon and Neutron Open Science Cloud
H2020-INFRAEOSC-04-2018
Grant Agreement Number: 823852

Deliverable: D3.4 Implementation Report from Facilities

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Project Deliverable Information Sheet

Project Reference No.	823852
Project acronym:	PaNOSC
Project full name:	Photon and Neutron Open Science Cloud
H2020 Call:	INFRAEOSC-04-2018
Project Coordinator	Andy Götz (andy.gotz@esrf.fr)
Coordinating Organization:	ESRF
Project Website:	www.panosc.eu
Deliverable No:	D3.4
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	2022-07-31
Actual Delivery Date:	2022-07-XX
EC project Officer:	Flavius Alexandru Pana

Document Control Sheet

Document	Title: Implementation Report from Facilities
	Version: 1
	Available at: https://github.com/panosc-eu/panosc
	Files:
Authorship	Written by: Tobias Richter (ESS), Fredrik Bolmsten (ESS), Massimiliano Novelli (ESS), Luis Maia (XFEL), Axel Bocciarelli (ESRF), Alex de Maria (ESRF), Balázs Bagó (ELI-ALPS), Emiliano Coghetto (CERIC-ERIC)
	Reviewed by: Teodor Ivănoaica (ELI ERIC)
	Approved: Andy Götz (ESRF)

List of Participants

Participant No.	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institut Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure Delivery Consortium (ELI-DC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Executive Summary

This last deliverable of the Data Catalogue Work Package (WP3) in PaNOSC summarises the per-partner progress towards delivering integrated services as part of the European Open Science Cloud. It is the first deliverable that prominently features output from task 3.4 of the grant agreement with focus on the integration of data production facilities with the data catalogues. The task held a best practices workshop that was reported on in a previous milestone. This deliverable concludes the per-partner integration efforts to make cataloguing part of their data workflows.

Extending on the previous deliverables, the partners took the opportunity to iterate the progress to make local API implementations and data repository mappings compliant with the common search API that was the subject of previous work. At this point, 3 (ESRF, ESS, CERIC) out of the 6 PaNOSC partners have been able to deploy a search API that complies with the agreed minimum functionality. The other 3 have implemented the search API and are working on completing the scoring before the end of the project. Much effort is still devoted to achieve basic compliance of the API and to run the scoring (ranking) of results. Consequently, few partners are engaged in the curation of data, focusing on making locally catalogued datasets compliant with the common mapping for parameters, techniques, roles, etc.

In addition, this report starts off by covering the progress partners have made towards making their public data harvestable by EOSC services and other third parties, exposing data using established APIs, schemas and services. At the time of this writing, some partners have managed to deploy a working endpoint that is correctly and continuously harvested. Others were able to deploy the endpoints, but no public data could be made available.

There are 4 months left in the project, so we expect most partners to progress further and complete their implementations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Introduction

This document is the final deliverable of Work Package 3 in PaNOSC. WP3 subtask 3.4 to integrate data from heterogeneous data sources is a particular focus of this report. It is important especially for partners that operate spatially distributed facilities, like ELI and CERIC ERIC. These sites want to have a one stop shop for their data catalogue users, despite producing data in several sites. This deliverable covers the work each partner has implemented to this end.

In addition, this report follows on from the development of the common search API (D3.1), the development of a demonstrator implementation (D3.2) and providing the search as a service (D3.3) in that it also features the per-partner implementation effort towards the common search facility. This effort will enable users to submit a single search query to retrieve matching results from all connected sites in a common result list. Mapping between the PaNOSC search data model and legacy parameters and keywords in the local databases can require significant effort. Also, the mapping between measurement units is non-trivial and requires testing. Encoding the technique in a way that is compliant with the PaNET ontology, developed in collaboration with ExPaNDS, is a data curation task that was not previously done at most facilities.

Lastly the submission of public datasets to third-party EOSC cross-discipline repositories was an early goal of the WP. This uses existing protocols (OAI-PMH) to interface to third party services (OpenAIRE and B2Find). The text will detail and quantify the respective achievements.

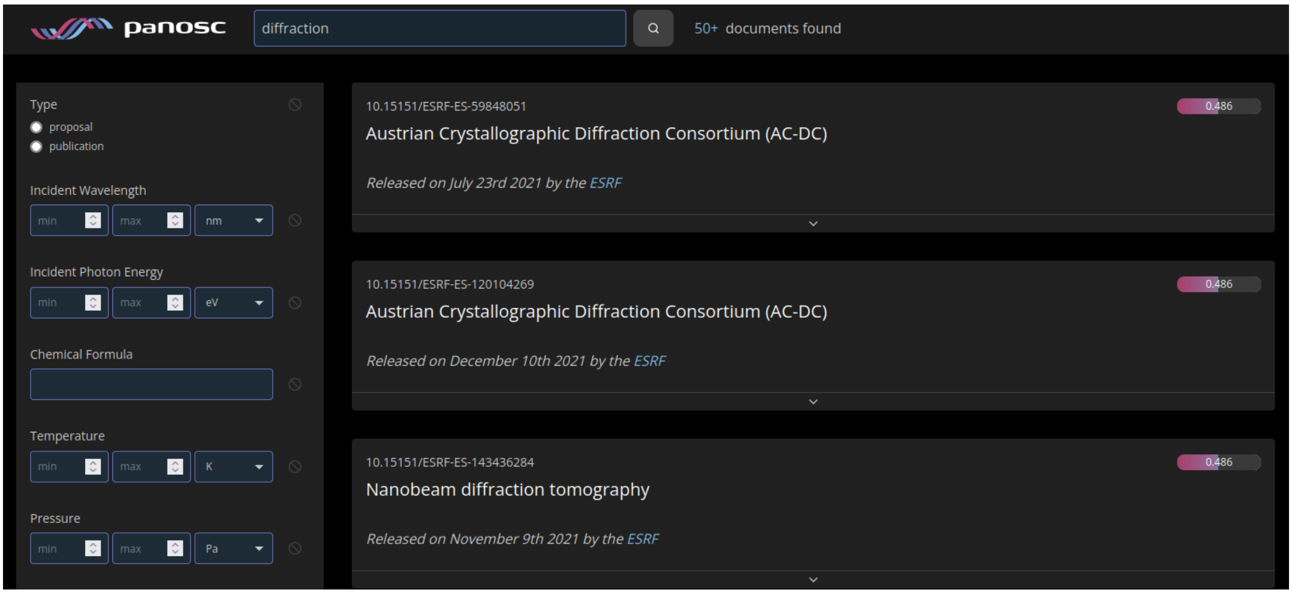
Before covering the cumulative results per partner, the following two small sections will give an update on the relevant common improvements made to the federated service and what the work package considers as a minimum baseline functionality.

Demonstrator Frontend Update

The data portal frontend, accessible at https://data.panosc.eu, has been reviewed and continuously improved since the last search API deliverable. Two face to face meetings were held for that purpose in 2022: one at the ESRF in April and a workshop session at the PaNOSC face to face meeting in June in Prague. The “search engine”-like experience has been improved, with the user landing on a front page which includes a simple search bar. Once the user has entered a list of terms to search for, a results page will be presented that allows the user to refine the terms and offers further filtering options, based on the PaNOSC agreed parameters, techniques, etc. The results from the different facilities are merged and sorted based on the relevancy score, described in detail in deliverable D3.3. In the current state, the data portal frontend makes it easy to submit queries to the PaNOSC federated search to evaluate the quality of the results returned. This process is also helping us to improve scoring and to define a set of queries that can function as testing baseline and run automatically by each facility or at the federated level.

Figure 1 shows the results for the query “diffraction”. Each tile shows the DOI, the first line of the title and the date when the result was released as a publicly accessible dataset. The score has been highlighted with a coloured bar which allows for quick evaluation of the relevancy of the result.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852



10.15151/ESRF-ES-59848051 0.486
Austrian Crystallographic Diffraction Consortium (AC-DC)
Released on July 23rd 2021 by the ESRF

10.15151/ESRF-ES-120104269 0.486
Austrian Crystallographic Diffraction Consortium (AC-DC)
Released on December 10th 2021 by the ESRF

10.15151/ESRF-ES-143436284 0.486
Nanobeam diffraction tomography
Released on November 9th 2021 by the ESRF

Figure 1: Query results for the term "diffraction"

Users can expand the result tiles to read the full title and also the list of people involved with the acquisition, creation and curation of the specific dataset. A click on the corresponding link takes the user to the DOI landing page set up by the facility hosting the dataset. On the landing page, the user has full access to all content and actions the facility offers. Common examples are displaying further metadata information or plots as content and download or processing options for actions.

Minimum Compliance Level for Inclusion into Federated API

One early outcome of the roll-out of the data portal frontend was that the success of the search portal depends highly on the relevancy of the results. When showcasing the first releases of the federated search, it became obvious that users are easily frustrated when presented with search results that, upon inspection, do not contain the query terms submitted or otherwise do not match the filtering parameters. In addition, the set of results returned was filled to the number of the results requested with random unrelated items. Results with incorrect, invalid or zero score are confusing, degrade the user experience and make it hard to evaluate differences between good and bad results.
For that reason the PaNOSC partners decided on some minimum criteria to have a facility included in the federated search. In addition there are goals that are meant to be achieved, but optional for now.

Partners have held a number of workshops and sessions for the last year to support the federated search API, local search API and scoring system to different facilities, including compliance review reports. At the last PaNOSC face to face meeting held in Prague, partners decided on the following minimum compliance level criteria:

- *Public access:* The API can allow authentication, but must return public results without it.
- *Ranking and relevance*: The results returned need to be ranked with relevance scores that are greater than 0 and lower than 1 and are calculated according to the reference scoring algorithm. Any filtering by parameters etc needs to be honoured. If in doubt results are to be discarded.

- *Correctness:* The API needs to comply with the expected reference implementation and return well formed results matching the data model or flag error conditions appropriately.

The ranking and relevance criteria requires that results should not be returned if a facility may not be able to decide if a dataset matches filters and terms entered by the user. This is relevant during the period when dataset curation and mapping of local parameter space to PaNOSC search terms is still ongoing. For example, if a user is interested in only data measured in a particular photon energy range, say 1 - 5 keV, and the facility does not have that information in a way that it can be automatically mapped and converted (for example from a wavelength in the metadata), the result needs to be suppressed even if it matches other search terms. Obviously, the ambition should be to include and map as many of the agreed parameters as possible in the long run. This effort will require some effort and will benefit from the testing done through the data portal frontend to highlight omissions.

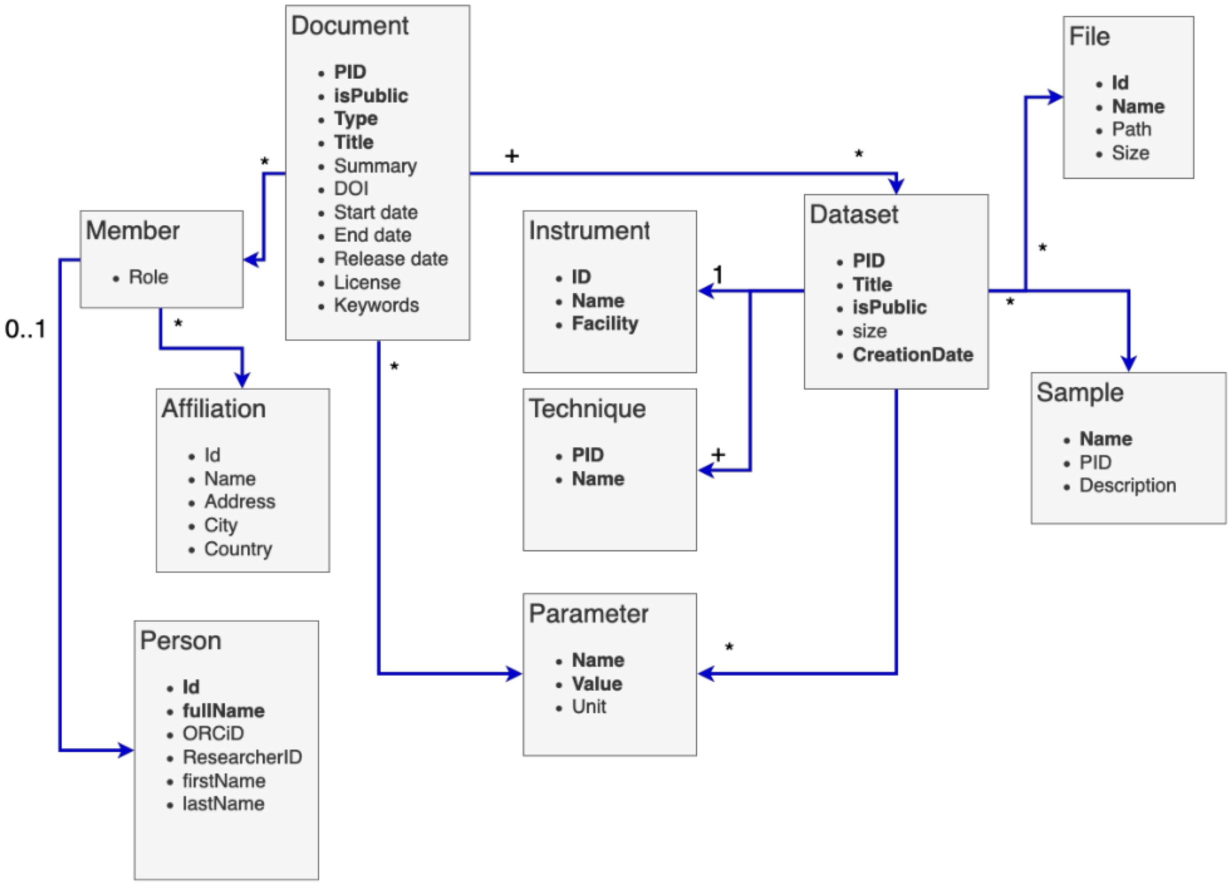


Figure 2: PaNOSC data model

Regarding the correctness of the results, we expect that each facility endpoint returns information conforming with the PaNOSC data model defined in deliverable D3.1, visible in figure 2.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

The data portal is configured to retrieve documents and therefore interacts only with the document endpoint of the search API. We expect that the results returned by each facility are compliant with the following data structure:

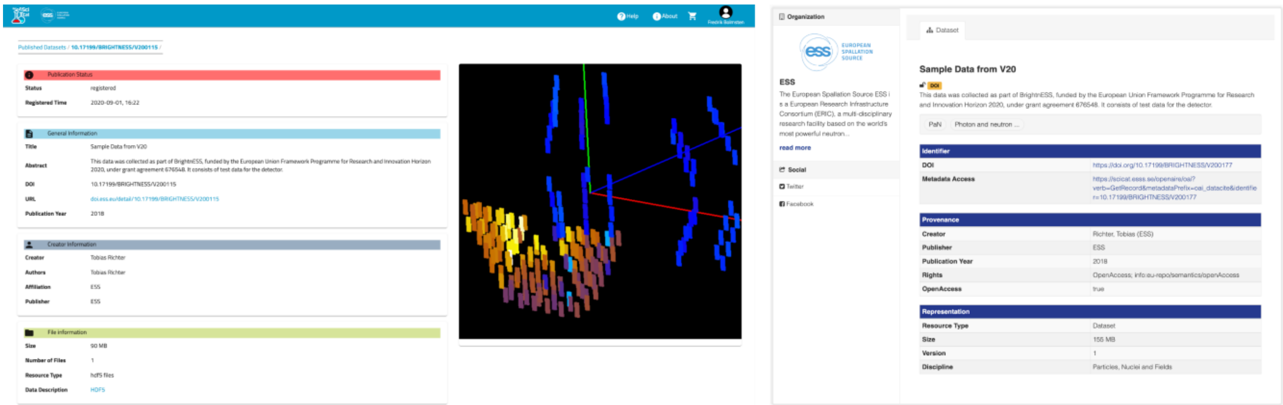
```
{
 "pid" : "assigned local pid",
 "isPublic": true,
 "type": "publication",
 "title": "document title",
 "summary": "document summary",
 "doi": "registered doi",
 "score": relevancy scoring as integer between 0 and 1 not included,
 "releaseDate": "document release date, format 2020-01-01T00:00:00.000Z",
 "datasets": [
  {
   "pid": "dataset assigned local pid",
   "title": "dataset title",
   "isPublic": true,
   "size": total size of the dataset in bytes,
   "creationDate": "dataset creation date, format 2019-12-11T12:48:03.000Z",
   "parameters": [
    {
     "name": "parameter name",
     "value": "parameter value",
     "unit": ""
    },
      …
   ],
   "techniques": [
    {
     "pid" : "technique official pid as in PaNET techniques ontology”,
     "name" : "technique official name as in PaNET techniques ontology”
    }
   ],
   "Members" : [
    {
     "Person" : {
      "fullName" : "person full name"
     }
    }
   ]
  }
```

If fields parameters, techniques and members are not available, the search api should return an empty array.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Status of OAI-PMH Endpoints for Harvesting

ESS



Published Datasets / 10.17199/BRIGHTNESS/V200115 /

Publication Status

Status	registered
Registered Time	2020-09-01, 16:22

General Information

Title	Sample Data from V20
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.
DOI	10.17199/BRIGHTNESS/V200115
URL	doi.ess.eu/detail/10.17199/BRIGHTNESS/V200115
Publication Year	2018

Creator Information

Creator	Tobias Richter
Authors	Tobias Richter
Affiliation	ESS
Publisher	ESS

File information

Size	90 MB
Number of Files	1
Resource Type	hdf5 files
Data Description	HDF5

Organization

EUROPEAN SPALLATION SOURCE

ESS

The European Spallation Source ESS is a European Research Infrastructure Consortium (ERIC), a multi-disciplinary research facility based on the world's most powerful neutron...

read more

Social

Twitter

Facebook

Dataset

Sample Data from V20

DOI

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.

PaN   Photon and neutron ...

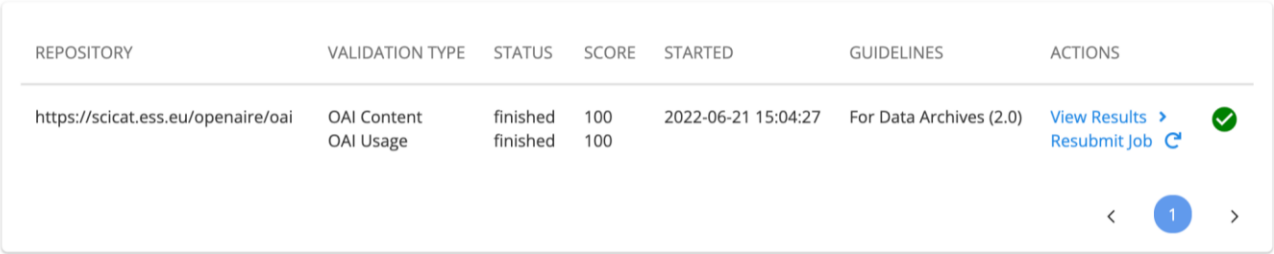
Identifier	
DOI	https://doi.org/10.17199/BRIGHTNESS/V200177
Metadata Access	https://scicat.esss.se/openaire/oai?verb=GetRecord&metadataPrefix=oai_datacite&identifier=10.17199/BRIGHTNESS/V200177

Provenance	
Creator	Richter, Tobias (ESS)
Publisher	ESS
Publication Year	2018
Rights	OpenAccess; info:eu-repo/semantics/openAccess
OpenAccess	true

Representation	
Resource Type	Dataset
Size	155 MB
Version	1
Discipline	Particles, Nuclei and Fields

Figure 3: A published dataset in the ESS metadata catalogue(left) and B2BFind(right)

ESS is currently being constructed and has no operational beamlines, data has however been produced at other facilities in conjunction with testing of detector prototypes, development of the control software and through similar activities. This data has been stored in our metadata catalogue SciCat (https://scicat.ess.eu/) and a total of 454 published datasets are being made available to OpenAire and B2Find through the OAI-PMH interface.



REPOSITORY	VALIDATION TYPE	STATUS	SCORE	STARTED	GUIDELINES	ACTIONS	
https://scicat.ess.eu/openaire/oai	OAI Content OAI Usage	finished finished	100 100	2022-06-21 15:04:27	For Data Archives (2.0)	View Results › Resubmit Job ↻	✓

Figure 4: Successful validation of OAI-PMH endpoint in OpenAire

There is a difference in the level of information being ingested and shown by OpenAire and B2FIND as OpenAire is only using the minimum set of metadata associated with Dublin Core, while after discussions with the B2Find community we were able to improve the metadata mapping to expose more domain specific information such as the size of the dataset. Using the following links will lead to the organisational pages for ESS at the different OAI-PMH providers:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

- B2FIND
  http://b2find.eudat.eu/organization/about/ess

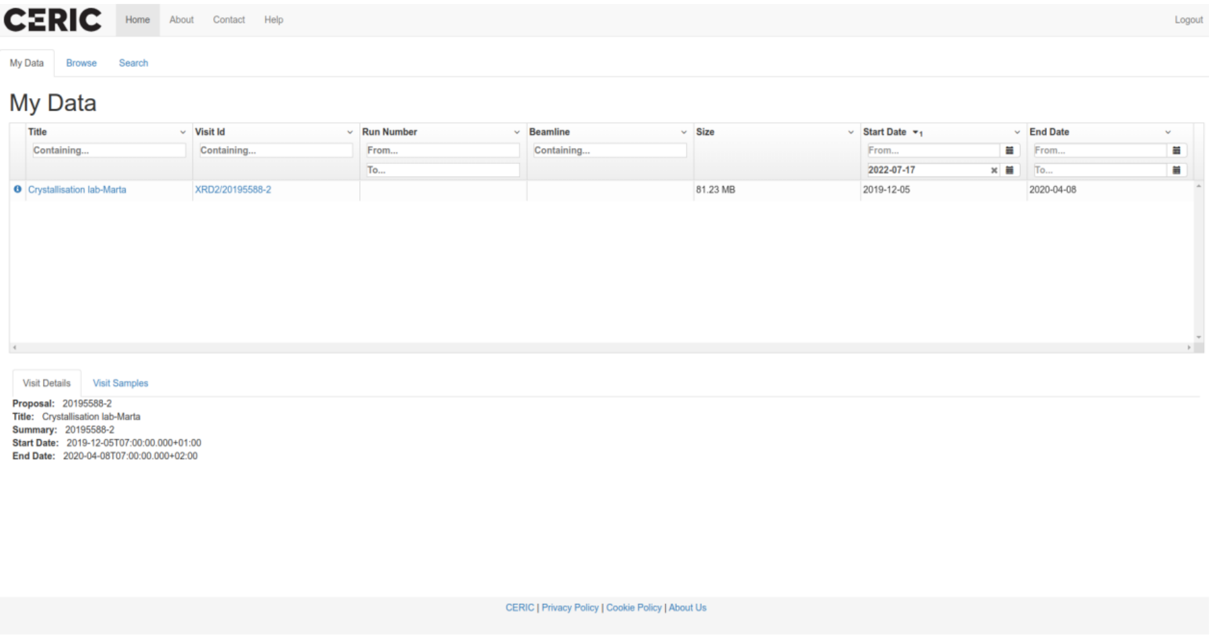
- OpenAIRE
  https://explore.openaire.eu/search/dataprovider?datasourceId=re3data_____::211db6e7f48f3d310552765226b9154e

Current metadata that is being mapped for datasets:

- Title
- Abstract
- Issued
- Publication year
- Creator
- Affiliation
- Publisher
- Version
- Subject
- Size (New)

CERIC-ERIC

ICAT (https://icatproject.org/) is the metadata catalogue in use at CERIC-ERIC. CERIC provides an OAI-PMH endpoint using the ICAT OAI-PMH component https://github.com/icatproject/icat.oaipmh that enables harvesting of metadata of the available open data. The endpoint is available at https://data.ceric-eric.eu/oaipmh/request?verb=Identify.



My Data

Title	Visit Id	Run Number	Beamline	Size	Start Date	End Date
Containing...	Containing...	From... To...	Containing...		From... 2022-07-17	From... To...
Crystallisation lab-Marta	XRD2/20195588-2			81.23 MB	2019-12-05	2020-04-08

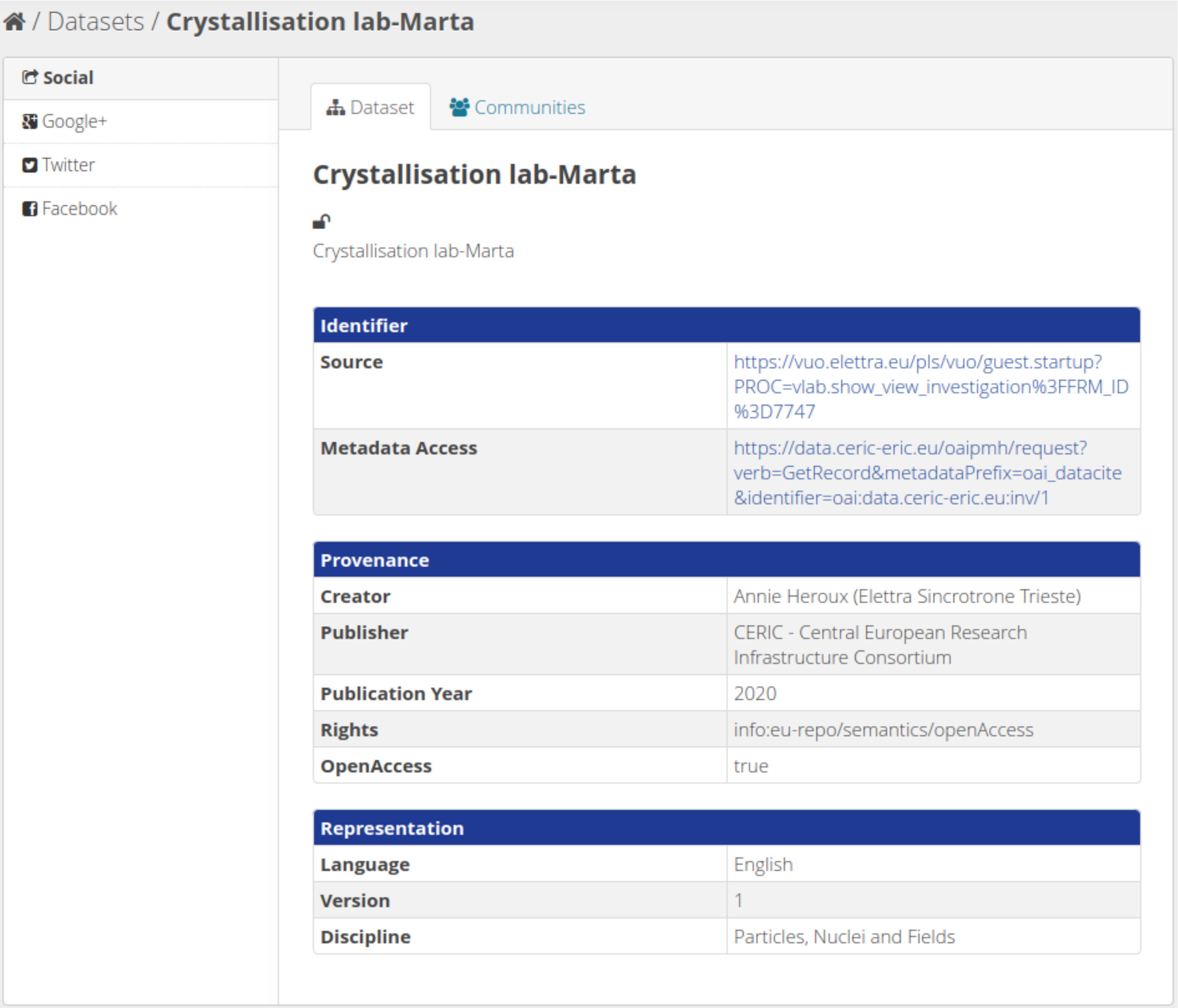
Proposal: 20195588-2
Title: Crystallisation lab-Marta
Summary: 20195588-2
Start Date: 2019-12-05T07:00:00.000+01:00
End Date: 2020-04-08T07:00:00.000+02:00

Figure 5: A published dataset in the CERIC metadata catalogue

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

The number of open datasets is very small. CERIC exposes metadata of a total of 20 datasets from 7 investigations, because the scientific data policy, which defines an embargo period of at least 3 years, was introduced in the second half of the PaNOSC project period.

CERIC offers continuous access to some of its instruments in a short time, with a fast evaluation procedure. Successful experiments are scheduled within a month from proposal submission, after evaluation by the facility. This access type is called Fast-Track Access. Three different types of Fast-Track Access are currently offered for feasibility studies (to test the feasibility of experiments or measurements for a maximum 48 hours per instrument), for commissioning (to perform measurements with the newest instruments and contribute to their commissioning) and for COVID-19 related research.
Fast-Track Accesses have a shorter embargo period so it is expected that the number of open-access datasets will increase significantly due to data resulting from Fast-Track Accesses.



Crystallisation lab-Marta

Crystallisation lab-Marta

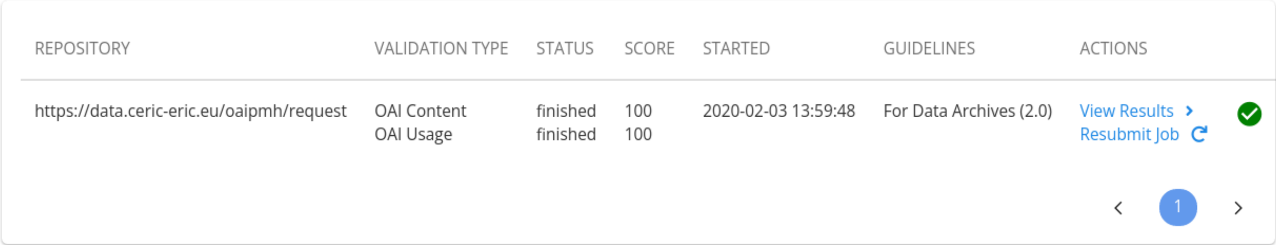
Identifier	
Source	https://vuo.elettra.eu/pls/vuo/guest.startup?PROC=vlab.show_view_investigation%3FFRM_ID%3D7747
Metadata Access	https://data.ceric-eric.eu/oaipmh/request?verb=GetRecord&metadataPrefix=oai_datacite&identifier=oai:data.ceric-eric.eu:inv/1

Provenance	
Creator	Annie Heroux (Elettra Sincrotrone Trieste)
Publisher	CERIC - Central European Research Infrastructure Consortium
Publication Year	2020
Rights	info:eu-repo/semantics/openAccess
OpenAccess	true

Representation	
Language	English
Version	1
Discipline	Particles, Nuclei and Fields

Figure 6: A published dataset on a B2F test repository

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852



REPOSITORY	VALIDATION TYPE	STATUS	SCORE	STARTED	GUIDELINES	ACTIONS
https://data.ceric-eric.eu/oaipmh/request	OAI Content OAI Usage	finished finished	100 100	2020-02-03 13:59:48	For Data Archives (2.0)	View Results › Resubmit Job ✓

‹ 1 ›

Figure 7: Successful validation of OAI-PMH endpoint in OpenAire

CERIC-ERIC is registered on Re3Data, B2FIND and OpenAire and has been harvested and validated by B2FIND. At the time of writing this deliverable, CERIC-ERIC datasets are available on a B2FIND test repository only. Recently B2Find has been requested to expose CERIC-ERIC datasets on B2FIND and OpenAire public repositories.

When CERIC-ERIC datasets will be exposed to B2FIND and OpenAire public repositories they will be available at

- B2FIND
  http://b2find.eudat.eu/organization/about/ceric

- OpenAIRE
  https://explore.openaire.eu/search/dataprovider?datasourceId=re3data_____::e79f66fcea0a894d66ecff3da5e52311

ELI ERIC

The Extreme Light Infrastructure (ELI), with facilities located in the Czech Republic, Hungary and Romania. The ELI facilities, built as individual construction projects, are now starting to transition towards operating as an integrated organisation.

The ELI ERIC have a central service that provides OAI-PMH for harvesting metadata. This service is available on the following link: https://data.eli-laser.eu/oai2d.

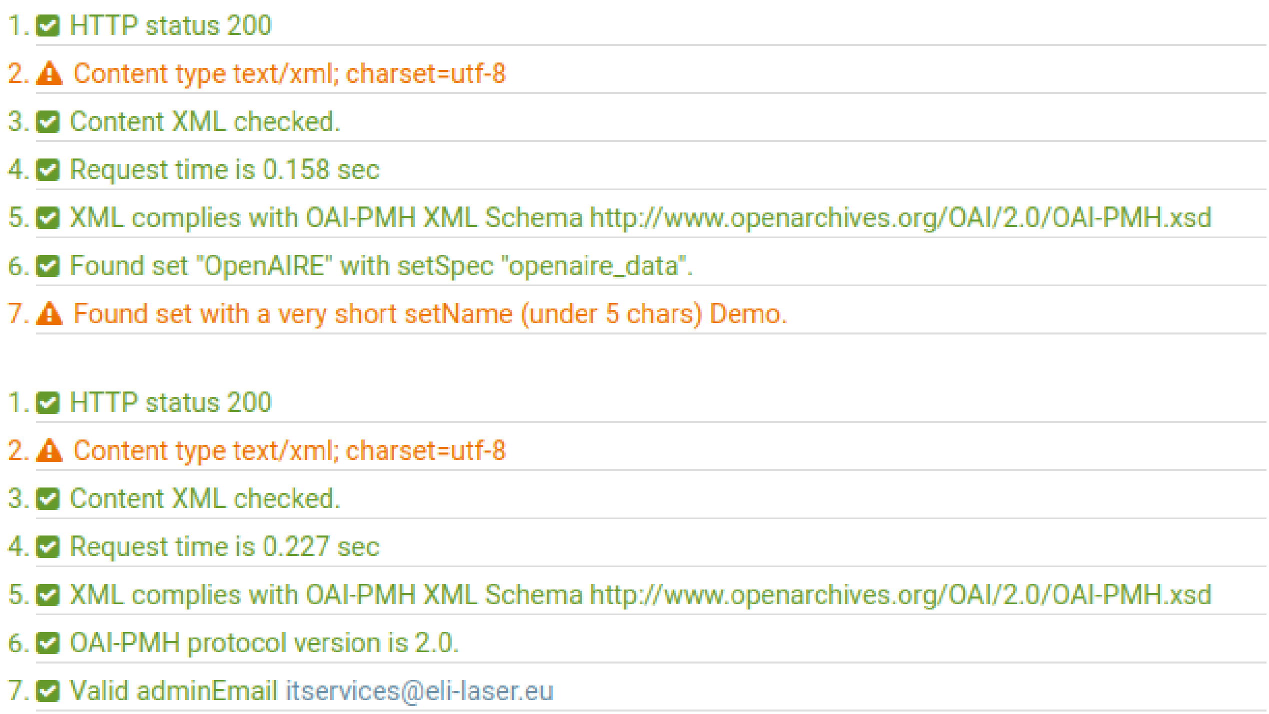
The OAI-PMH endpoint has two different sets, one “demo” set for the demonstration data and an “openaire_data” set for harvesting by the OpenAIRE service.

The OAI-PMH endpoint supports the following relevant metadata formats:

- OAI Dublin Core (http://www.openarchives.org/OAI/2.0/oai_dc.xsd)
- OAI DataCite (http://schema.datacite.org/oai/oai-1.1/)

The registration of ELI OAI-PMH endpoint to the **OpenAIRE** and **B2FIND** metadata harvesting services is started and expected to be finalised in the next couple of weeks. (The ELI ERIC Data Repository is registered with Re3data https://www.re3data.org/repository/r3d100013889 and the registration process has started with both OpenAIRE and B2FIND)
However, at this point, ELI facilities do not have any publicly available data, therefore the endpoint does not have real metadata for harvesting.

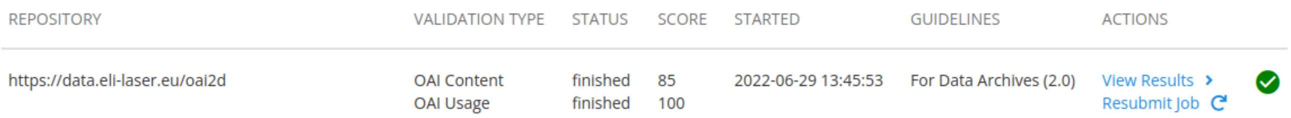
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852



1. ✓ HTTP status 200
2. ⚠ Content type text/xml; charset=utf-8
3. ✓ Content XML checked.
4. ✓ Request time is 0.158 sec
5. ✓ XML complies with OAI-PMH XML Schema http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
6. ✓ Found set "OpenAIRE" with setSpec "openaire_data".
7. ⚠ Found set with a very short setName (under 5 chars) Demo.

1. ✓ HTTP status 200
2. ⚠ Content type text/xml; charset=utf-8
3. ✓ Content XML checked.
4. ✓ Request time is 0.227 sec
5. ✓ XML complies with OAI-PMH XML Schema http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
6. ✓ OAI-PMH protocol version is 2.0.
7. ✓ Valid adminEmail itservices@eli-laser.eu

Figure 9: The validation results from "validator.oaipmh.com"



REPOSITORY	VALIDATION TYPE	STATUS	SCORE	STARTED	GUIDELINES	ACTIONS	
https://data.eli-laser.eu/oai2d	OAI Content OAI Usage	finished finished	85 100	2022-06-29 13:45:53	For Data Archives (2.0)	View Results › Resubmit Job	✓

Figure 8: The validation result from OpenAIRE

XFEL

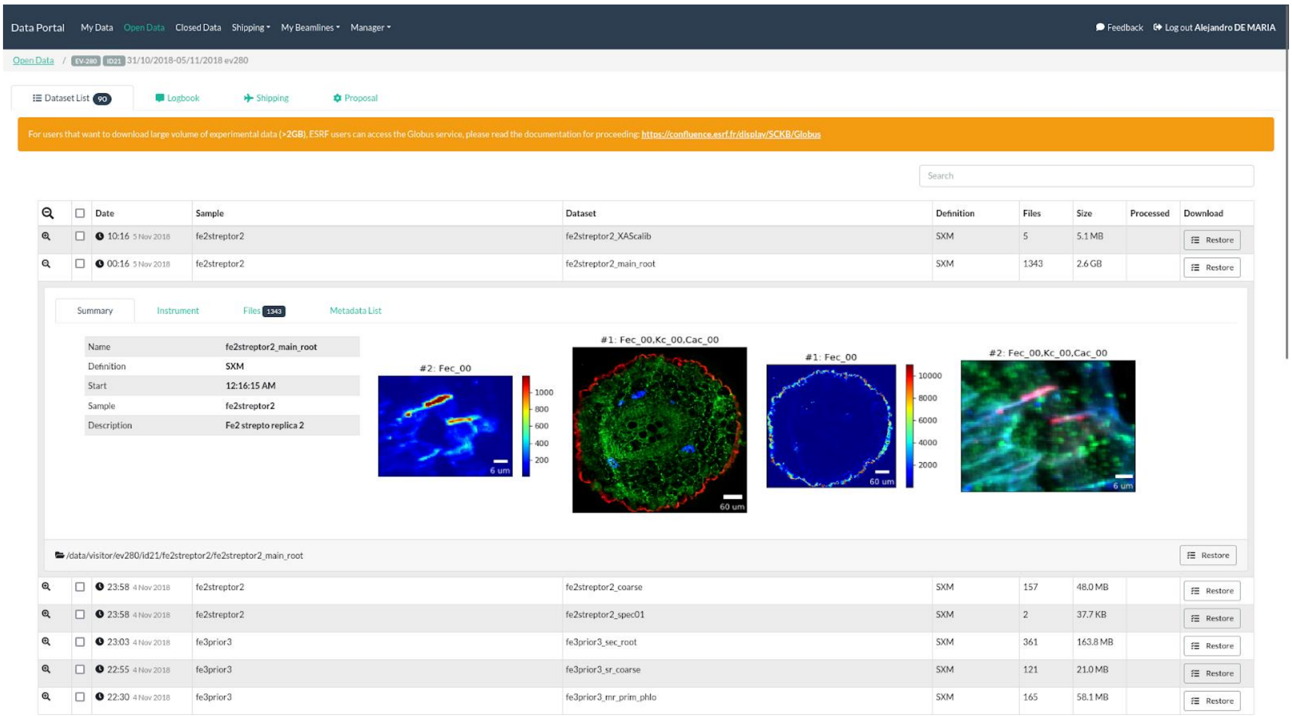
European XFEL, despite not being a distributed facility, provides its users with near real-time remote access to the data collected on their proposal. Additionally, European XFEL provides open access to example data taken in the facility's main 6 instruments. XFEL data is primarily stored on storage close to the instrument and migrated to the central repository.

The orchestration of the migration, documentation and exposure of all the files is done by the XFEL metadata catalogue – myMdC.

As described in the previous sections, myMdC is also used to provide the community with an OAI-PMH endpoint for the integration with EOSC and the Search API endpoint for the supply of the PaNOSC Federated Search service.

ESRF

ICAT (https://icatproject.org) is the metadata catalogue used at the ESRF since 2014 and currently the data and metadata from around 35 beamlines is automatically captured and catalogued in real time.
Any registered person can access to the public experiments through the data portal (https://data.esrf.fr)



Data Portal My Data Open Data Closed Data Shipping ▾ My Beamlines ▾ Manager ▾ Feedback Log out **Alejandro DE MARIA**

Open Data / EV-280 ID21 31/10/2018-05/11/2018 ev280

☰ Dataset List 90 | Logbook | Shipping | Proposal

For users that want to download large volume of experimental data (>2GB), ESRF users can access the Globus service, please read the documentation for proceeding: https://confluence.esrf.fr/display/SCKB/Globus

Search

	□	Date	Sample	Dataset	Definition	Files	Size	Processed	Download
	□	10:16 5 Nov 2018	fe2streptor2	fe2streptor2_XAScalib	SXM	5	5.1 MB		Restore
	□	00:16 5 Nov 2018	fe2streptor2	fe2streptor2_main_root	SXM	1343	2.6 GB		Restore

Summary | Instrument | Files 1343 | Metadata List

Name	fe2streptor2_main_root
Definition	SXM
Start	12:16:15 AM
Sample	fe2streptor2
Description	Fe2 strepto replica 2

/data/visitor/ev280/id21/fe2streptor2/fe2streptor2_main_root Restore

	□	Date	Sample	Dataset	Definition	Files	Size	Processed	Download
	□	23:58 4 Nov 2018	fe2streptor2	fe2streptor2_coarse	SXM	157	48.0 MB		Restore
	□	23:58 4 Nov 2018	fe2streptor2	fe2streptor2_spec01	SXM	2	37.7 KB		Restore
	□	23:03 4 Nov 2018	fe3prior3	fe3prior3_sec_root	SXM	361	163.8 MB		Restore
	□	22:55 4 Nov 2018	fe3prior3	fe3prior3_sr_coarse	SXM	121	21.0 MB		Restore
	□	22:30 4 Nov 2018	fe3prior3	fe3prior3_mr_prim_phlo	SXM	165	58.1 MB		Restore

Figure 10: Example of a public dataset published in the data portal (https://data.esrf.fr/investigation/135816585/datasets)

ICAT's community has developed an OAI-PMH component (https://github.com/icatproject/icat.oaipmh) that has been deployed at the ESRF. It enables harvesting of metadata of around 2500 investigations (https://icatplus.esrf.fr/oaipmh).

OAI-PMH is used by third-party software as B2Find(http://b2find.eudat.eu). B2Find is an interdisciplinary discovery portal for research output that allows free term search. It allows to discover ESRF data easily.

The ESRF B2FIND repository can be found at:

- http://b2find.eudat.eu/organization/about/esrf

ILL

ILL use DataCite as the registration agency of DOIs for the data and the data catalogue has been registered with Re3Data since 2016 (https://www.re3data.org/repository/r3d100012072). The ILL provides an OAI-PMH endpoint that enables the harvesting of metadata of the available open data (https://data.ill.fr/openaire/oai). Currently, metadata concerning more than 2500 proposals since 2013 is available through this endpoint. The endpoint has been registered with

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

OpenAire since April 2021.

Search API Endpoints and Data Curation

The list of facilities providing data to a running instance of the federated API is provided in the browser at the root URL where the API is running. The current deployment at ESS can be reached at the URL:

https://federated.scicat.ess.eu/

This URL provides information on the configuration of the instance running. At the time of writing this deliverable, the following JSON object is returned:

```
{
        "Uptime_seconds":345642.127604473,
        "Uptime":"96:00:42",
        "Api_version":"v2.7",
        "Docker_image_version":"v2.7",
        "hosting_facility":"ESS",
        "Environment":"production",
        "Data_providers":[
                "https://searchapi.maxiv.lu.se/api",
                "https://icatplus.esrf.fr/api",
                "https://search.panosc.ess.eu/api",
                "https://dacat.psi.ch/panosc-api"],
        "Provider_timeout_ms":10000,
        "Default_limit":100
}
```

As we can see from the information provided, the ESS instance of the federated search API pulls data directly from search APIs running on ESS, PSI, ESRF, and MaxIV.

The up-to-date list of active data providers can be found in the following configuration file in the Github repository:
https://github.com/panosc-eu/panosc-federated-search-service/blob/master/.env

This allows anyone to checkout the latest commit on the master branch and run a fully functional instance of the PaNOSC federated search connected to all active data sources locally on their machine.

Facilities who want their results to be included in the PaNOSC Federated Search, must submit a pull request in the official repository that includes the file mentioned above updated with the URL of their implementation PaNOSC Search API. The pull request creation will initiate a review of their endpoint compliance. If they meet the minimum compliance level previously defined, they will be integrated and become active facilities in providing results.

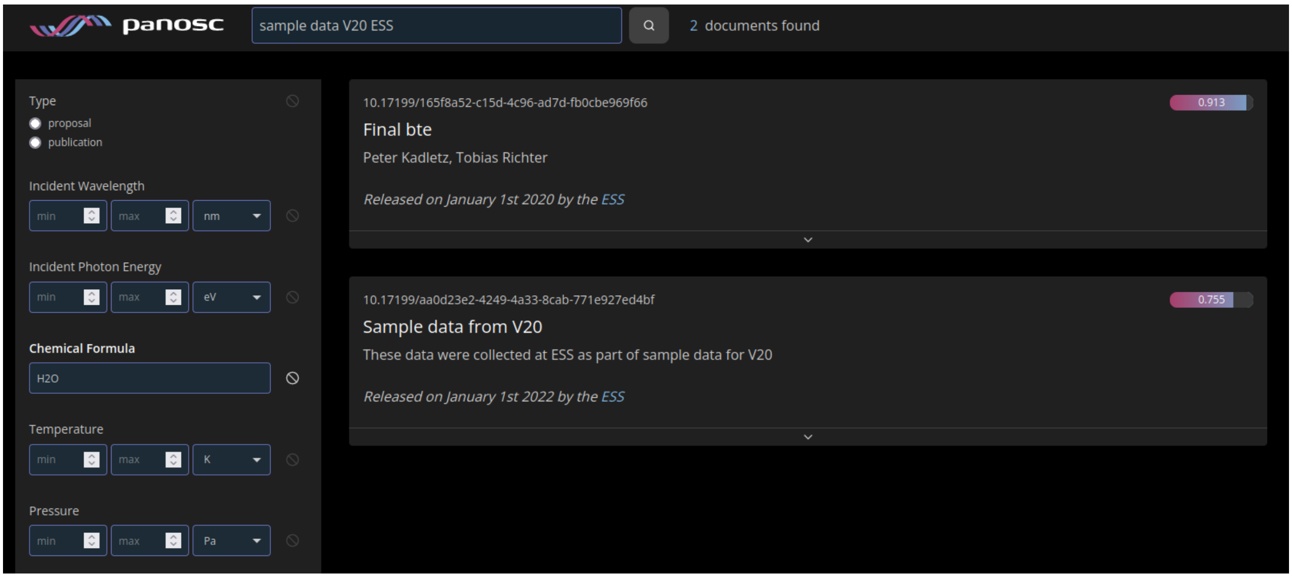
ESS

The ESS focus has been on leading the work to deploy the first version of federated search and its testing. Currently the federated API is deployed on ESS infrastructure and is accessible at the URL https://federated.scicat.ess.eu

The reference and SciCat implementations of the PaNOSC search API have been maintained and kept in sync. Since the last deliverable, the techniques service, developed by PSI under the ExPANDS project, has been integrated. The local instance of the PaNOSC search API instance, accessible at the URL https://scicat.ess.eu/panosc-api, offers full integration with the scoring and techniques services.

Data Curation

The ESS facility has around three thousand entries between datasets and documents that are publicly available. The PaNOSC community previously agreed to guarantee access and mapping for the following parameters: wavelength or photon energy, sample chemical formula, sample temperature and sample pressure. In the current SciCat implementation of the PaNOSC search api (https://github.com/SciCatProject/panosc-search-api) these parameters (including their units) are passed transparently to the SciCat query endpoints. This leverages the built-in search functionalities present in SciCat. It means however that the data has to be curated accordingly, which is currently in progress. A number of datasets have been labelled with the proper techniques and the scientific metadata has been curated to contain the correct entries for PaNOSC. For example a search by Chemical Formula results in matches, as shown in figure 11. When the user insert the conditions shown in figure 11 The data portal will construct the following query and submit it to the federate search:



sample data V20 ESS

2 documents found

Chemical Formula: H2O

10.17199/165f8a52-c15d-4c96-ad7d-fb0cbe969f66 — 0.913

Final bte

Peter Kadletz, Tobias Richter

Released on January 1st 2020 by the ESS

10.17199/aa0d23e2-4249-4a33-8cab-771e927ed4bf — 0.755

Sample data from V20

These data were collected at ESS as part of sample data for V20

Released on January 1st 2022 by the ESS

Figure 11: Portal search by chemical formula

```
{
 "include": [
  {
   "relation": "datasets",
   "scope": {
   "include": [
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

```
    {
     "relation": "parameters",
     "scope": {
      "where": {
       "and": [
        {
         "name": "sample_chemical_formula"
        },
        {
         "value": {
          "like": "H2O"
         }
        }
       ]
      }
     }
    }
   ]
  }
 }
],
"query": "sample data V20 ESS",
"limit": 50
}
```

When this query is received by the PaNOSC search API SciCat implementation, the condition on the parameters is converted into a condition on a scientific metadata key named *sample_chemical_formula* with value “H2O”. No mapping is applied or performed in the PaNOSC search API. If the sample chemical formula has been named differently on a specific dataset, ESS has two options: rename the specific scientific metadata entry or create a new one with the correct name and the same value.

The curation effort also includes techniques. They will be stored in our metadata catalogue and users will be able to query them directly through the PaNOSC search API. ESS has also deployed the technique service developed under the ExPaNDS project. This service is seamlessly integrated in the search API and completely transparent to the user. It provides easy access to the PaNET ontology and has the task to convert technique names to technique PIDs and expand the portion of the filter related to techniques and capture the hierarchical structure of the technique ontology. The filter expansion allows users to find datasets and documents which are tagged with more narrowly defined techniques even if a more generic technique is searched.

Step-by-step walkthrough of the technique expansion process: A user submits a query for documents containing datasets marked with the technique “neutron probe“ through the data portal. The data portal converts such a request with an additional condition on techniques as follows:

```
{
 “relation” : “techniques”
 “where” : {
  "pid": "http://purl.org/pan-science/PaNET/PaNET00101"
 }
}
```

Once this condition reaches the PaNOSC search API (SciCat and reference implementation), it is submitted to the techniques service which converts it to the following condition:

```
{
 “relation” : “techniques”
 “where” :
 {
  "pid": {
    "inq": [
      "http://purl.org/pan-science/PaNET/PaNET00101",
      "http://purl.org/pan-science/PaNET/PaNET01018",
      "http://purl.org/pan-science/PaNET/PaNET01019",
      "http://purl.org/pan-science/PaNET/PaNET01119",
      "http://purl.org/pan-science/PaNET/PaNET01120",
      "http://purl.org/pan-science/PaNET/PaNET01126",
      "http://purl.org/pan-science/PaNET/PaNET01127",
      "http://purl.org/pan-science/PaNET/PaNET01016",
      "http://purl.org/pan-science/PaNET/PaNET01235",
      "http://purl.org/pan-science/PaNET/PaNET01217",
      "http://purl.org/pan-science/PaNET/PaNET01100",
      "http://purl.org/pan-science/PaNET/PaNET01234",
      "http://purl.org/pan-science/PaNET/PaNET01237",
      "http://purl.org/pan-science/PaNET/PaNET01236",
      "http://purl.org/pan-science/PaNET/PaNET01240",
      "http://purl.org/pan-science/PaNET/PaNET01242",
      "http://purl.org/pan-science/PaNET/PaNET01239",
      "http://purl.org/pan-science/PaNET/PaNET01243",
      "http://purl.org/pan-science/PaNET/PaNET01017",
      "http://purl.org/pan-science/PaNET/PaNET01246",
      "http://purl.org/pan-science/PaNET/PaNET01245",
      "http://purl.org/pan-science/PaNET/PaNET01247",
      "http://purl.org/pan-science/PaNET/PaNET01248",
      "http://purl.org/pan-science/PaNET/PaNET01249",
      "http://purl.org/pan-science/PaNET/PaNET01250",
      "http://purl.org/pan-science/PaNET/PaNET01189",
      "http://purl.org/pan-science/PaNET/PaNET01276",
      "http://purl.org/pan-science/PaNET/PaNET01277",
      "http://purl.org/pan-science/PaNET/PaNET01278",
      "http://purl.org/pan-science/PaNET/PaNET01298",
      "http://purl.org/pan-science/PaNET/PaNET01299"
    ]
  }
}
```

The same conversion will happen if the user manually submits the name of the technique:

```
{
    “Relation” : “techniques”
    “Where” : {
     "name": "neutron probe"
    }
   }
```

As explained above, this conversion is completely transparent to the user. The PaNOSC search API extracts the technique condition from the query if present, submits a request to the techniques service and integrates the results back in the query, which is then submitted to the metadata catalogue system.

CERIC-ERIC

ICAT is the metadata catalogue in use at CERIC-ERIC. At the time of writing this deliverable, CERIC has just upgraded the ICAT metadata catalogue to the latest version (v5.0.0) which introduces among other things the concept of Technique and Affiliation. Given that ICAT is the metadata catalogue in use, it has been decided to deploy the implementation of the PaNOSC search API developed by the ICAT community (https://github.com/ral-facilities/datagateway-api). To be extended soon with the ranking and the scoring service. The base URL of the search API endpoints is available at https://data.ceric-eric.eu/search-api.
At the time of this deliverable, CERIC has submitted a PR request to have its search API endpoint included in the federated search and it is currently working through the details to meet the inclusion criteria mentioned in the introduction of this document.

Furthermore, CERIC has just deployed the latest release of its metadata ingestion system which will be discussed in more detail in the “Integration of Data Source” section.

Data Curation

CERIC has hundreds of investigations saved in the storage system that can be accessed through the user office platform by authorised users (PI, participants etc.), but very few of these investigations are open access for the reasons mentioned above. That’s why we are kindly asking scientists to upload to the storage system old data and fast-track related data as well in order to make them open access and respect the FAIR principles.

CERIC supports the PaNET techniques, in fact each dataset is associated with one or more PaNET techniques during the file metadata harvesting process and this relation is recorded into ICAT during the metadata ingestion process.

All the PaNOSC roles have been mapped to the user office roles, but currently only the principal investigator role has been implemented for the Search API and OAI-PMH interfaces.

ELI

The ELI facilities have a common service to provide the search API endpoint. The currently deployed data catalogue is an Invenio RDM instance and a search API connector was developed to support the search API endpoint for ELI.

The connector uses the search engine component of Invenio RDM to serve the desired functionality of the search API. The applied solution does not require any modification of the code base of the framework, it only relies on the proper configuration and the usage of existing

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

features (e.g.: The connection between a publication and a dataset is made available by using the "references" and "is referenced by" metadata information based on the DataCite scheme).

The scoring is being implemented by using the PaNOSC search scoring service. The scoring service is populated by using the metadata available on the PaNOSC search API. This solution is in principle data catalogue service independent, so it can be applied to any facility that already has support for the PaNOSC search API. The following link is an example Jupyter notebook about populating the scoring system, which uses the reference implementation of the search API with the test json based data catalogue:

https://github.com/panosc-eu/panosc-search-scoring/blob/master/notebooks/PSS-ReferenceSearchAPI-Integration.ipynb

The search API service of ELI is available on the following link:

https://panosc-search.eli-laser.eu

ELI has submitted a PR request to have its search API endpoint included in the federated search and it is currently testing the details to meet the baseline inclusion criteria mentioned in the introduction.

Data Curation

ELI facilities do not have open data to supply the Federated Search API, however demo metadata is available through the Search API endpoint of ELI.

The metadata schema of the catalogue of ELI is based on the DataCite metadata schema (https://schema.datacite.org/). The ELI facilities do not have publicly available data, but ELI-ALPS has already made experiments with commissioning users. Based on the data from these experiments the planned metadata scheme was extended to support the required parameters where those parameters are applicable. The used metadata scheme was extended with the following roles to support the defined roles in the Search API:

- Principal Investigator (mapped to the "Project Leader" role in the DataCite scheme)
- Local Contact (mapped to the "Contact Person" role in the DataCite scheme)
- Experimenter (mapped to the "Researcher" role in the DataCite scheme)
- Collaborator (mapped to the "Other" role in the DataCite scheme)

The PaNET support for the metadata was implemented by using the subject field of the DataCite metadata. The subject field contains the human readable name and the PID of the related technique.

The parameters are currently supported by using special triplets in the subject field of the metadata. This solution is not ideal, but Invenio RDM does not yet support the alteration of the metadata scheme through configuration. However, this feature is on the roadmap of Invenio RDM.

XFEL

European XFEL (https://www.xfel.eu/), has 6 scientific instruments, which are performing experiments since late 2017. Their metadata are stored in MyMdC (https://www.xfel.eu/), the metadata catalogue used within European XFEL.

The data and metadata information are made available to the experiment team immediately

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

after data is taken, MyMdC used to store the metadata and provide the necessary access to the data files. In accordance with the European XFEL Data Policy (https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html), data and metadata have a default embargo period of 3 years, during which access is restricted to the experimental team. After the embargo period data and metadata became public, being possible for external people to request access to it. Using MyMdC any registered user can query live data from all the proposals (s)he has access to.

European XFEL metadata provides RESTful APIs that allow their metadata to be queried, including the "Search API" defined within WP3. The current list of implemented methods can be found at https://in.xfel.eu/metadata/api-docs/index.html (default available APIs) and https://in.xfel.eu/metadata/api-docs/index.html?urls.primaryName=PaNOSC%20API%20Docs (PaNOSC search APIs). This endpoint (https://in.xfel.eu/metadata/api) currently serves metadata related to all Open Data (and embargoed) experiments, provided that valid authentication is provided. This makes the implementation non-compliant, although XFEL is working with their data policy and security groups to lift the need for authentication. In addition, the scoring implementation is missing, which at the time of publishing of this document, is under development.

Data Curation

In order to fulfil the PaNOSC search API, XFEL did integrate on MyMdC the techniques defined on project https://expands-eu.github.io/ExPaNDS-experimental-techniques-ontology, however, it is the responsibility of the Instrument Experts and users to select them and correctly assign it to the taken data.

ESRF

There are more than 40 beamlines currently performing experiments at the ESRF. Their data and metadata are archived and stored in the ICAT metadata catalogue (https://icatproject.org/). Following the ESRF data policy (http://www.esrf.eu/datapolicy), the data is made available as soon as the data is produced and accessible via the data portal (https://data.esrf.fr). Data and metadata will be publicly available after the embargo period (3 years) during which access is restricted to the experimental team, represented by the PI. Each investigation has its own persistent identifier (DOI) that is obtained from Datacite (https://datacite.org/). The data of the Human Organ Atlas project (https://human-organ-atlas.esrf.fr/) has been made accessible from the search API (cf.Improvements and documentation updates).

The local catalogue (https://data.esrf.fr) authentication uses openID allowing users and staff to access their data during the embargo period. The search API uses anonymous access to access public metadata and data. It is also possible to create an account via the ESRF User portal (https://smis.esrf.fr/). Every user that participates in an experiment has a role. The main roles are: Principal investigator, local contact, scientist, participant, etc. The PI of an experiment can add a user as a collaborator. This allows sharing the data (https://www.youtube.com/watch?v=FDUFPpnllxE). More than 900 parameters have been identified as metadata (https://gitlab.esrf.fr/icat/hdf5-master-config/-/blob/master/hdf5_cfg.xml), and are captured and stored automatically during the data acquisition. One of these parameters is the technique which will be made mandatory in the future. The PaNOSC search API has been developed within the ICAT community by STFC as part of the new ICAT API called datagateway-api (https://github.com/ral-facilities/datagateway-api). The scoring has been integrated by the ESRF to comply with the PaNOSC federated search to share and display results with a score > 0.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ILL

The data catalogue of the ILL (https://data.ill.fr) provides access to both embargoed and open data. A search interface allows users to obtain metadata concerning proposals based on proposal ID, instrument, reactor cycle and also more open, full-text searches.

Concerning the Search API of WP3, a production version has been in operation since May 2021 (https://data.ill.fr/fairdata/api). This endpoint serves metadata related to all proposals that have Open Data (currently more than 2500).

Currently the deployed end point does not support result scoring. This is the main reason ILL data cannot be queried from the federated API or frontend. It is planned to deploy the scoring reference implementation before the end of the project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Report from Integration of Data Sources

Task 3.4 in the grant agreement deals with the integration of data production facilities with the data catalogues. This is especially important for heterogeneous and distributed facilities. This task held a best practices workshop that was reported on in a previous milestone. This deliverable concludes the per partner integration efforts to integrate cataloguing into their data workflows.

ESS

ESS is not part of task 3.4 and hence has no specific activity to report here.

CERIC-ERIC

Since CERIC-ERIC provides open access to 61 instruments of several European RIs data integration is one of the main tasks in order to have a complete data and metadata curation and management cycle.

CERIC decided to adopt a first-stage solution that consists of collecting data produced in the facilities into a single storage system. Elettra Sincrotrone Trieste is the statutory seat and provides the IT infrastructure to CERIC, therefore, for CERIC Services are using the storage system already integrated with the user office and more specific in the proposal management system.

Collecting all the data produced by all the CERIC partners is a big challenge due to the heterogeneity of their IT infrastructures, control systems and acquisition systems as well as their geographical location.
We started with a pilot project which makes users and scientists able to upload data collected during an experiment using a standalone tool that can be installed both in a Linux and in a Windows desktop. This tool interfaces with the central storage system and with the user office platform, and it allows uploading data associated with a specific proposal, both during the acquisition and after the experiment was performed.

Once the data of a whole proposal, produced by one or more instruments, is uploaded to the central storage, it can be possible to populate ICAT with the related metadata; for this purpose, CERIC developed a modular metadata dispatching system, based on FastAPI, which can extract metadata from some kind of source (currently from user office and data files) and can ingest it to some kind of destination (currently ICAT).
In order to extract metadata from files, we developed, and we are going to continue to develop, harvesters which are Python modules, instrument specific, which can extract metadata from single data files. This development needs to involve the scientists responsible for the related instrument, to help to parse data acquired and to choose metadata of interest to be ingested into the catalogue for future searches.

Regarding the metadata, in order to have some level of standardization, we decide to adopt a subset of this EOSC list https://eosc-edmi.github.io/properties plus some parameters coming from the proposal system as common metadata among all partners.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Summing up the main achievements:

- We managed to transfer from the Austrian partner to the central storage system some test data associated to an investigation which was present in the proposal system;
- The last version of the metadata dispatcher has been recently deployed with approximately ten harvesters;
- The latest version (5.0.0) of ICAT has been deployed, and is ready to store metadata of dataset produced by the instruments of all partners;
- Metadata coming from 20 datasets and 5 investigations has been ingested into ICAT and can be searchable from the Search-API

Future steps:

- Get more partners involved in transferring data;
- Develop more harvesters in order to be able to extract metadata from more and more data file formats.

ELI

ELI facilities are located in the Czech Republic and Hungary, together with a third facility in Romania is expected to join the ERIC at a later stage. Each of the ELI facilities has its own infrastructure both for research and IT purposes. Each facility deals with the integration of the different research instruments, which are sometimes black box systems with only a defined interface to support the exact research activity without having a proper interface or a description to assist the data ingestion.

The ELI facilities are now preparing their 1st ELI ERIC Call for users, which will validate the Data Policy Implementation activities[1] and evaluate the minimum level of compliance of the existing functional research data handling mechanisms.

Considering the FAIR mission statement made by ELI by adopting the FAIR Data Policy, a decision and effort have been made, using the context provided by the PaNOSC project but not limited to only PaNOSC, to have shared services to provide a unified look for ELI Users and for the partner scientific communities, via projects and collaborations like EOSC or PaNOSC. The usage of common services gives the freedom of uniquely defined data pipelines, which fits the needs of the different scientific instruments and ensures the quality of the data and the user-facing services, these common services help ELI Facilities fulfil their commitment to FAIR data handling.

As described in the previous sections ELI uses shared services to provide the OAI-PMH endpoint for the integration with EOSC and the Search API endpoint for the supply of the PaNOSC Federated Search service. Both of these previously mentioned services are built upon a shared metadata catalogue service as shown on Figure 12 about the architecture of the shared services.

[1] https://zenodo.org/record/6515903#.Yt8kt-xBwUo

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

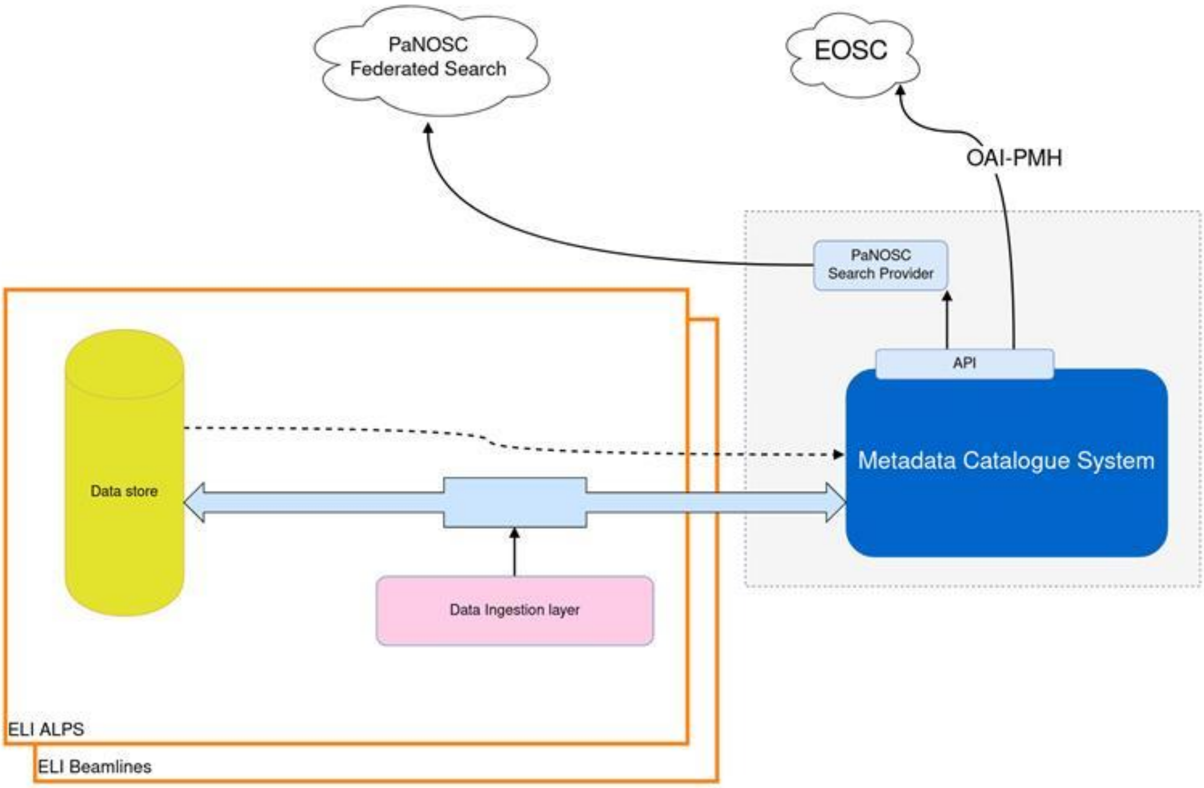


Figure 12: Architecture diagram of shared services at ELI

XFEL

European XFEL, despite not being a distributed facility, provides its users with near real-time remote access to the data collected on their proposal. Additionally, European XFEL provides open access to example data taken in the facility's main 6 instruments.

XFEL data is primarily stored on storage close to the instrument and migrated to the central repository. The orchestration of the migration, documentation and exposure of all the files is done by the XFEL metadata catalogue - myMdC.
As described in the previous sections, myMdC is also used to provide the community with an OAI-PMH endpoint for the integration with EOSC and the Search API endpoint for the supply of the PaNOSC Federated Search service.

ESRF

At the ESRF, the data and metadata are captured automatically in real time for most of the instruments (40+). The implementation started when the ESRF endorsed the data policy (https://www.esrf.fr/datapolicy) in 2015.
Given the variety and number of instruments there are currently multiple data acquisition software in place. The newest and most used is Bliss (https://www.esrf.fr/BLISS) but the legacy control system (SPEC) plus some bespoke control software systems have also been adapted to support the ESRF data policy i.e. they ingest data and metadata into ICAT.
The implementation in this heterogenous environment is possible thanks to the intensive use of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Tango devices (https://www.tango-controls.org). Each beamline/instrument has its own metadata Tango device which manages the proposal, sample, dataset and data acquisition folder and the control software pushes the metadata and data that will be then handled by a specific software (known as the ingester) in charge of the preservation of the data.

The main role of the ingester is the persistance of the data and metadata. This includes the storage of the data in the metadata catalogue ICAT, the minting of DOIs when required, and the storage of the data in tape.

The above describes the process for all raw data produced at the ESRF. Developments are on-going to support ingesting processed data by the end of 2022 for beamlines whose users need processed data rather than raw data.

ILL

No reported activity for task 3.4

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Summary and Outlook

Driven by Scientific excellence data produced by cutting-edge research facilities has been in continuous evolution for many years. This has been the case with Photon and Neutron (PaN) facilities, continuously developing standards, tools and services to support their users' experiments. In this continuous process, PaNOSC Work Package 3, Task 3.4 put together a multi-facility group of data experts to work together on defining and developing a fledgling service for a PaN Data Commons to support the development of common search capabilities for open data from PaN facilities.
The most important outcome of the T3.4 activities is having the scientists engaged in identifying and implementing missing scientific techniques from the NeXUS vocabularies and ontologies, which allows the data experts to develop meaningful data services for PaN and EOSC users' communities.

The work package has opened up data from a number of facilities to the science community at large and the researchers in the field of photon and neutron science specifically. Most of the public data held by PaNOSC partners is now accessible from third-party repositories fed by the OAI-PMH endpoints that have been implemented and deployed.

The search API has been defined, implemented, federated and shown to deliver results relevant to scientists working in the field. At the time of writing the number of local implementations that are part of the federation are lacking behind the ambitions of having all PaNOSC partners implementing the search API with scoring. The main holdup for partners has been the deployment and verification of the scoring. Most have made enough progress to be in the verification phase for inclusion at least and are in the process of completing the deployment of the scoring. All partners need to improve the quality of curation and mapping of parameters, techniques, etc over time. Once the service can be rolled out to real users we can expect a feedback loop to develop from real users of the first federated PaN Data Commons service.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852