PaNOSC

Photon and Neutron Open Science Cloud

H2020-INFRAEOSC-04-2018

Grant Agreement Number: 823852

Deliverable: D3.2 Demonstrator Implementation

(federation of search APIs)

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:	D3.2
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	2021-03-31
Actual Delivery Date:	2021-03-24
EC project Officer:	René Martins

Document Control Sheet

Document	Title: Demonstrator Implementation
	Version: 1
	Available at: https://github.com/panosc-eu/panosc
	Files:
Authorship	Written by: Lajos Schrettner (ELI) & Tobias Richter (ESS)
	Contributors: Fredrik Bolmsten (ESS), Balázs Bagó (ELI), Teodor Ivănoaica (ELI-DC), Jiří Majer (ELI), Stuart Caunt (ILL), Jamie Hall (ILL), William Turner (ILL), Luis Maia (EuXFEL), Sandor Brockhauser (EuXFEL), Henrik Johansson (ESS), Alejandro de Maria Antolinos (ESRF), Emiliano Coghetto (CERIC), Alessandro Olivo (CERIC), Silvia da Graca Ramos (ExPaNDS - DLS), Alun Ashton (ExPaNDS - PSI)
	Reviewed by: Andy Götz
	Approved: Jordi Bodera

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	The 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

Overall, the work package is advancing according to plan. This document marks the delivery of a federated search demonstrator for open data, a proof-of-concept centralised search service to access all PaNOSC sites at the same time. The API of federated search provides essentially the same service as the base search API reported in Deliverable 3.1. The prominent added feature is that search queries are propagated to a configurable number of underlying search providers, returned results are aggregated and provided to the client. The providers are expected to be the search service instances implemented by all PaNOSC and ExPaNDS (https://expands.eu) partner sites, but any further facilities can join this effort. There is an agreement with the cataloguing work package of the ExPaNDS project that the API can be used as a common target by both projects. The federated search service allows to find datasets and data publications from any number of configured sites based on relevant domain specific metadata and can be used by third parties to find data that has been released from any facility imposed embargo period, as well as by the original researchers.

The resulting service will be an important entry point for anyone to use the EOSC visualisation, processing or data transfer services that are being developed in other PaNOSC WPs. Specifically, there has been joint work on a custom graphical search user interface in cooperation with PaNOSC WP4 as part of the data portal development task

The main purpose of this demonstrator is to form a checkpoint for development. It helps identify issues in the design and plan of the development, by putting items to a real-world test. This deliverable summarizes the design decisions taken for the federated search demonstrator, describes its architecture, some implementation and deployment details. The demonstrator has been deployed and tested at partner facilities, but it is packaged and documented in sufficient detail so that deployment by interested parties is relatively straightforward. It will continue to be used for further testing and as a basis of the development of a production system.

Introduction

This document summarises the development of a federated search demonstrator created in the PaNOSC Work Package 3. It explains the development and many of the decisions made as well as open tasks for future development.

The ultimate goal of the domain specific search service developed in PaNOSC is to provide a unified way across facilities for scientists to find data using a variety of parameters. These parameters could include the facility, instrument or experiment technique used to collect it or its main experiment parameters including source (X-ray or neutron beam) characteristics, sample information or detector details, or the name of investigators related to it, that originate from the different institutes in PaNOSC.

Following up from the development of the common search API (D3.1) it is a natural next step for all PaNOSC to be targeted by a single search query, aggregating matching results from all sites in a common result list. This federated search demonstrator fulfils this requirement as a proof of concept implementation. It is useful to test and evaluate the compliance of the search implementations at partner catalogues, explore the commonality in metadata vocabularies and it can serve as a basis for a production ready service provided towards the EOSC (see Figure 1).

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

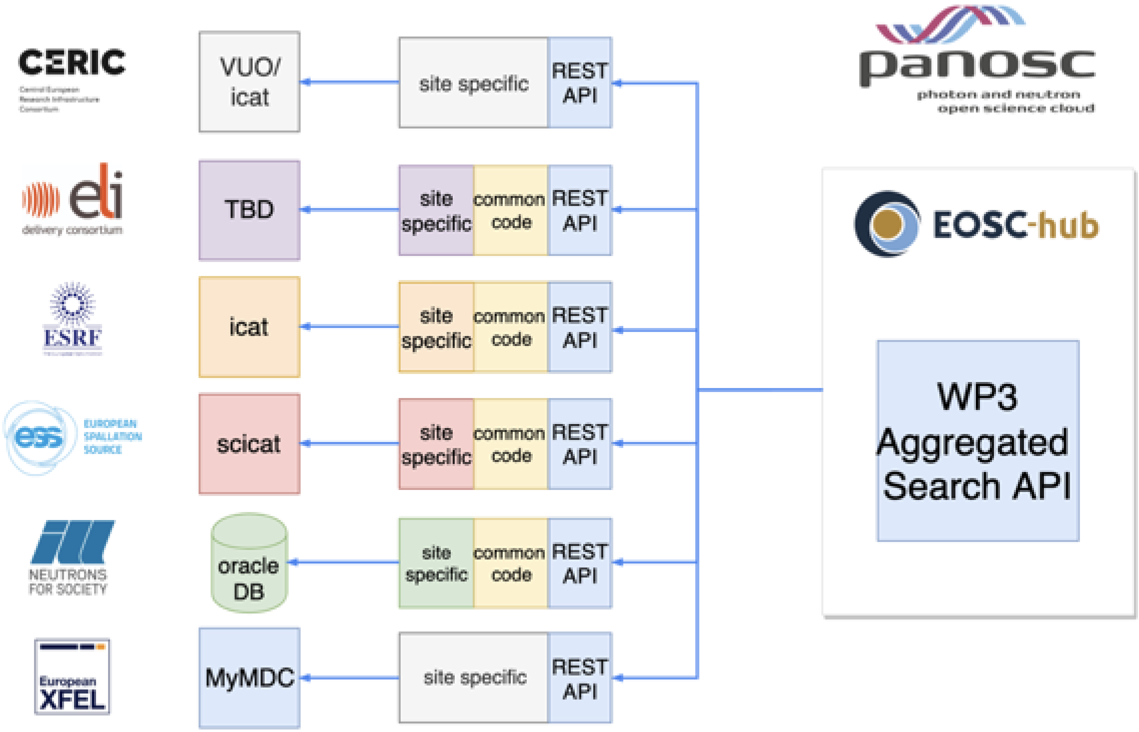


Figure 1: Overview of targeted deployment at partner sites and use of search API.

ExPaNDS, the INFRAEOSC-05b project, European Open Science Cloud (EOSC) Photon and Neutron Data Service, with an additional ten national Photon and Neutron Research Infrastructures (PaN RIs) is working closely with PaNOSC on common objectives, especially in this data catalogue work package. ExPaNDS partner facilities have committed themselves to roll out compliant implementations of the search API that can be federated in the same search provider. This enables users to carry out domain specific searches throughout most PaN RIs in Europe. It has to be pointed out that ExPaNDS facilities take an active role in the PaNOSC development activities with their expertise.

The description of the data structures used to compose a search query is a defining part of the API. Results of the search are returned in the same data model, shown in Figure 2, as it was released with deliverable D3.1. The federated search demonstrator uses the same semantics and also the identical data model. As a result, any client that can be used to query an individual data catalogue compliant with the common search API will by design also work with the federated, aggregated service.

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

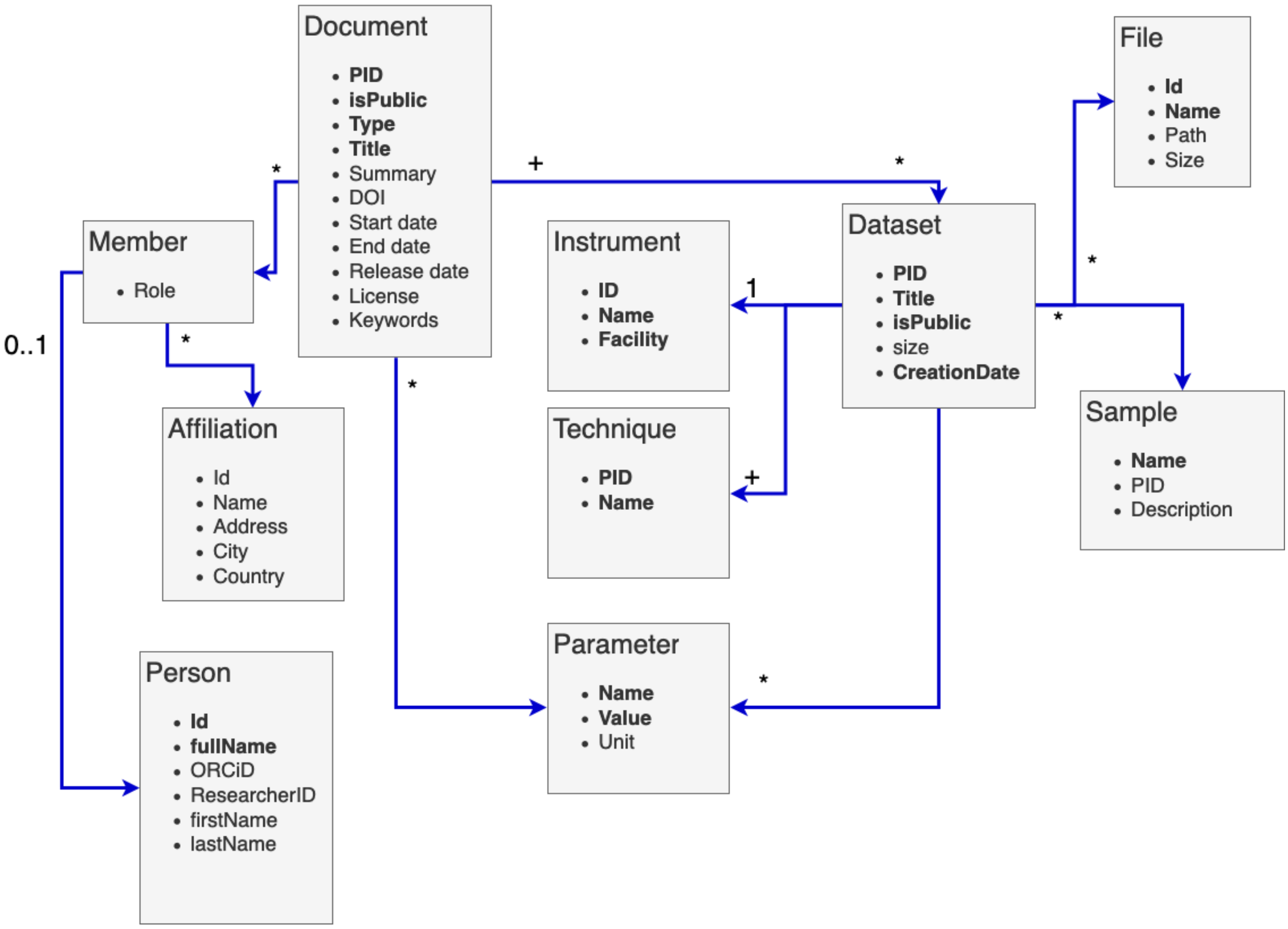


Figure 2: Search API data model

As background the documentation about the search API and the federated search service is open and available on the PaNOSC confluence platform https://confluence.panosc.eu. Confluence is a content collaboration tool used to help teams collaborate and share knowledge therefore it is well suited to act as a documentation repository for the common search API and more generally for the PaNOSC project. The search API calls are described in the OpenAPI format (https://www.openapis.org/) and a tool called Swagger UI is used to visualize and interact with the API's resources without having any of the implementation logic in place (https://confluence.panosc.eu/x/TwGm). The search API calls are based on a well-defined and common data model whose entity-relationship schema can be found at https://confluence.panosc.eu/x/IQD8. The federated search demonstrator service API is identical to the base search API, and thus is not documented separately.

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

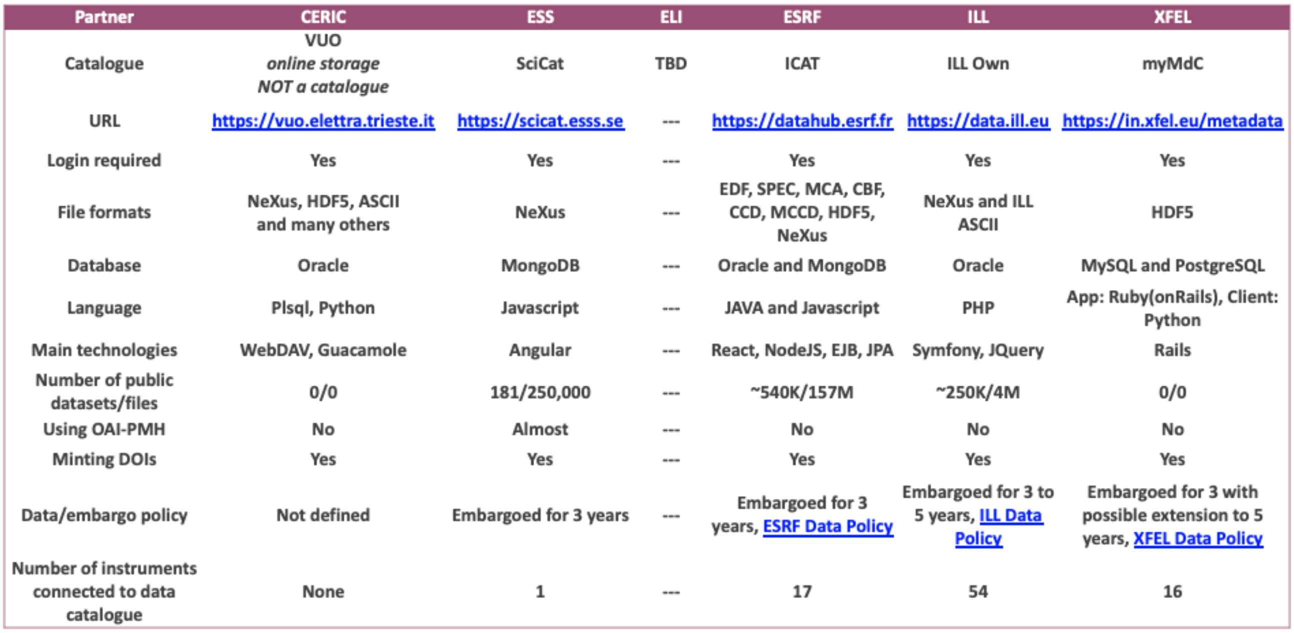
Progress in the WP since the last deliverable

After the previous deliverable, work progressed towards implementing a federated search service based on the API definition. Substantial effort has been put by the partners into setting up their search API instances, the so-called search providers in order to provide an environment for federation.
GitHub continues to be the main repository for both the code and the documentation and Confluence is being used more and more to draw up documentation, guides and functional diagrams, moreover, it stores also some per-site implementation information like the endpoints of the federated search API (provider).
The WP participants regularly meet on monthly video conferencing to review and discuss the ongoing WP activities. Since the very start of the ExPaNDS project the WP meetings have seen a growing involvement of ExPaNDS people and are now regularly held together since both projects can benefit one another and the photon and neutron community as a whole. When needed further meetings were scheduled to discuss and deepen single topics like the Ontologies and the Federated Search Demonstrator.
The spread of the Coronavirus pandemic with the consequent travel restrictions has made it impossible to have face to face meetings and workshops so, of necessity, the partners actually continued to interact exclusively via video conferences.
The Federated Search API provides an identical API to the standard Search API (defined in D3.1). Therefore, a client of the Federated Search API can query for results using the same syntax as the Search API and results are returned in the same format. This has the advantage that client applications (such as the Portal of WP4) can be easily configured to point to individual site Search APIs or the Federated Search API without any logical changes.
To assist in the development and testing of the Federated Search API it was decided that each facility would provide either a working endpoint with a minimal set of data or a simple JSON data file containing relevant metadata about documents and datasets. Each site would therefore have data present in the aggregated results of the demonstrator. The Search API code available on GitHub (https://github.com/panosc-eu/search-api) allows for the use of a simple JSON data file as a data source. Several partner sites have therefore successfully reused the same code for providing valid API endpoints. All sites with Open Data have therefore been incorporated in the Federated Search Demonstrator and have data available from the search results.

Progress at Partners

One of the main objectives of PaNOSC is to open up the large repositories of experimental data collected from measurements and experiments at photon and neutron European Research Infrastructures (RIs) to the scientific communities at large under the open data and FAIR data principles. Due to the large number and volume of the datasets, being able to identify what data is and is not relevant before processing or transferring them is a strong requirement. To this end, the PaNOSC facilities have deployed data catalogues that store metadata and support domain specific searches. Table 1 shows the status before this project was started, with a number of different catalogues, based on different technologies.

Table 1: Overview of Data Catalogues deployed at the PaNOSC facilities and some related information from the start of PaNOSC (Status 2019)



Partner	CERIC	ESS	ELI	ESRF	ILL	XFEL
Catalogue	VUO *online storage* *NOT a catalogue*	SciCat	TBD	ICAT	ILL Own	myMdC
URL	https://vuo.elettra.trieste.it	https://scicat.esss.se	---	https://datahub.esrf.fr	https://data.ill.eu	https://in.xfel.eu/metadata
Login required	Yes	Yes	---	Yes	Yes	Yes
File formats	NeXus, HDF5, ASCII and many others	NeXus	---	EDF, SPEC, MCA, CBF, CCD, MCCD, HDF5, NeXus	NeXus and ILL ASCII	HDF5
Database	Oracle	MongoDB	---	Oracle and MongoDB	Oracle	MySQL and PostgreSQL
Language	Plsql, Python	Javascript	---	JAVA and Javascript	PHP	App: Ruby(onRails), Client: Python
Main technologies	WebDAV, Guacamole	Angular	---	React, NodeJS, EJB, JPA	Symfony, JQuery	Rails
Number of public datasets/files	0/0	181/250,000	---	~540K/157M	~250K/4M	0/0
Using OAI-PMH	No	Almost	---	No	No	No
Minting DOIs	Yes	Yes	---	Yes	Yes	Yes
Data/embargo policy	Not defined	Embargoed for 3 years	---	Embargoed for 3 years, ESRF Data Policy	Embargoed for 3 to 5 years, ILL Data Policy	Embargoed for 3 with possible extension to 5 years, XFEL Data Policy
Number of instruments connected to data catalogue	None	1	---	17	54	16

In order to get relevant information about the datasets held to the user communities, the project aims to deploy two different mechanisms. One is to open a service for harvesting metadata on all open access data to the EOSC aggregating repositories, namely OpenAIRE and B2Find. This will make the data available to the wider EOSC community, together with the data from other domains.

Alternatively, the common search Application Programming Interface (API) will provide a uniform way via a web portal or computer program to interrogate all data catalogues with specific searches in order to find data of interest by an individual.

The partners keep track of their progress towards providing a search API endpoint at the confluence page https://confluence.panosc.eu/display/wp3/Search+API. Table 2 presents a snapshot of the current status at partner sites.

Table 2: Overview of catalogue URLs for the search API endpoints and the status of exposed data (Status March 2021)

Facility	**Search provider base URL**	**Dataset Exposure**
CERIC-ERIC	http://panosc-search.apps.okd2.ceric.fedcloud.eu/api	Placeholder information
ELI	no endpoint	
ESRF	https://icatplus.esrf.fr/api/	Live data
ESS	https://scitest.esss.lu.se/panosc-explorer	Live data with public datasets

ILL	http://data.ill.fr/fairdata/api	Snapshot from end of 2020
XFEL	https://in.xfel.eu/metadata/api-docs/index.html	Live data (password required)

CERIC-ERIC

In the beginning of 2021, the ICAT data catalogue was formally adopted. Populating it with metadata for selected open access investigations is in progress. ICAT services run on cloud resources (https://cericat.apps.okd2.ceric.fedcloud.eu). CERIC's users can create DOIs for their investigations, the provider used is DataCite with prefix 10.34967. It is foreseen that soon it will be possible to create DOIs for the Dataset as well.

Last year CERIC registered to re3data.org, OpenAIRE and B2FIND. At the very beginning a CERIC-ERIC specific OAI-PMH implementation was deployed to expose an endpoint (https://data.ceric-eric.eu/oai-pmh/oai2?verb=Identify). It still serves metadata related to a few test proposals. Since ICAT has been adopted work is ongoing to replace our OAI-PMH implementation with the one provided by the oai-pmh ICAT plugin (https://github.com/icatproject/icat.oaipmh).

The default implementation of the WP3 search API (https://github.com/panosc-eu/search-api) has been deployed (http://panosc-search.apps.okd2.ceric.fedcloud.eu/). This endpoint currently serves a snapshot of metadata related to 3 investigations with data in the public domain. In the near future the default work package Search API will be replaced by the ICAT search API plugin.

Given a significant increase in the number of data saved in the storage infrastructure in the last year it is foreseen during this year (2021) to increase the number of metadata collected in the ICAT catalogue. As a consequence, as soon as they will become open-access they will be available to OpenAire and B2Find through the ICAT OAI-PMH interface.

ELI

The Extreme Light Infrastructure (ELI) consists of complementary facilities located in the Czech Republic, Hungary and Romania. The ELI facilities, built as individual construction projects, are now coming together as an integrated organization, the ELI European Research Infrastructure Consortium (ELI ERIC), that will be in charge of their joint operations.

The ELI sites are in the process of building and commissioning their experiments and beamlines. That means no publishable experimental data is available at this point. Current work has concentrated on designing and building the necessary background infrastructure to capture, secure, store raw experiment data, together with its associated auxiliary data and metadata. To this end ELI is actively participating in multiple PaNOSC activities, working on the development and custom integration of different PaN tools and best-practices. As part of this development and integration process, ELI sites have deployed and tested the development versions of the search API of PaNOSC WP3, and they are now further continuing the work on developing a custom UI. Further, ELI took a leading part in defining and implementing the federated search demonstrator.

ESRF

There are more than 40 beamlines currently performing experiments at the ESRF. Their data and metadata are archived and stored in a metadata catalogue named ICAT (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/).
An OAI-PMH endpoint has been developed and provides access to the high level metadata of all the investigations done at the ESRF (https://icatplus.esrf.fr/oaipmh). Besides, a search API has been deployed inline with the PANOSC deliverable that exposes a subset of the public data (https://icatplus.esrf.fr/api/datasets). It is foreseen during this year (2021) to enrich the metadata provided by OAI-PMH, include all public data in the PANOSC search API as well as register the repository in OpenAire (https://www.openaire.eu/).

ESS

ESS is currently being constructed and currently 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 made available to OpenAire and B2Find through the OAI-PMH interface. Discussions are ongoing with B2Find to improve the metadata mapping to expose more domain specific information. Finally the search API (https://scitest.esss.lu.se/panosc-explorer) has been fully connected to the live SciCat instance and is making 2895 datasets explorable.

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. We 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 harvesting of metadata of the available open data (https://data.ill.fr/openaire/oai). Currently metadata concerning more than 1500 proposals since 2013 is available through this endpoint. The registration of this endpoint in OpenAire is currently in progress.

Concerning the Search API of WP3, a test implementation endpoint has been provided (https://data.ill.fr/fairdata/api). This endpoint currently serves metadata related to 250 proposals that have Open Data. The full implementation of the Search API (providing unauthenticated access to the metadata of all ILL open data) is currently in progress.

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

XFEL

European XFEL (https://www.xfel.eu/), has currently 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 is made available to the experiment team immediately 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.

European XFEL is currently providing, upon request, DOIs for their experiments (under DataCite prefix 10.22003). It is foreseen that soon DOIs will be generated automatically for all experiments on the last day of the experiment, making at that moment some of the experiment metadata publicly available.

European XFEL metadata provides RESTful APIs that allow their metadata to be queried, including a good percentage of 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. This endpoint currently serves metadata related to all Open Data (and embargoed) experiments, provided that valid authentication is provided. The full implementation of the Search API (providing unauthenticated access to the metadata and extra fields not yet implemented in MyMdC like Technics) is currently in progress.

Concerning the MyMdC OAI-PMH endpoint, it is currently under development. This endpoint will allow external providers to harvest only the metadata of the available open data experiments (https://in.xfel.eu/metadata).

Finally, for the purposes of the demonstrator, 6 real use cases (and respective real data) were prepared and exposed to the demonstrator (currently available at GitHub project).

Progress on controlled vocabularies (Ontologies)

The search API and its Data Model as described in the previous deliverable D3.1 and figure 2 provides a number of properties to search and filter on. These cover a large set of the anticipated searches. In the work package there is a task to develop controlled vocabularies for additional use cases that do not have defined properties in the Data Model. These are explicitly:

- Roles
- Techniques
- Parameters

The possible entries for those properties need to be narrowed down and aligned between the facilities, so

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

that a search for a given common parameter will give all possible results from all partners. In addition, the API also would allow a common curation of "sample" and "instrument". These are considered outside of the common scope, though and will be maintained independently by the facilities.

As part of the activities to develop and test the search demonstrator we have collected or gained access to a set of datasets per partner, with metadata that is representative for content that can be made accessible at the moment. A survey of the available information provided valuable input to a draft definition to choose common names for. As a first result the survey validated the choice of the API and Data Model by finding among the top common properties:

- sample name
- sample description
- dataset name
- investigation title
- start and end dates
- abstract

All of which map trivially into Data Model properties (the last three into the generic "Document" class).

Roles

For roles of a person associated with a dataset the survey found a relatively diverse set of names and concepts between data catalogues at partners. Some had no roles mapped. It was relatively simple to come up with a draft of a common scheme that everyone might be able to successfully map into, though. The resulting roles would be:

> **Principal Investigator** *(principal_investigator)* - the person that submitted the beamtime proposal and is the administrative responsible for the experiment and dataset(s). Often that role is limited to a single person, but there is no guarantee that has to be the case.
>
> **Local contact** *(local_contact)* - a person responsible for the local equipment at the photon or neutron facility, often the beamline or instrument scientists. Someone that might help with some aspects of data analysis or instrumental effects or aspects in the data. This can be more than one person.
>
> **Experimenter** *(experimenter)* - a person that was involved in performing the experiment. Often a person that was physically present. This class includes a special role of "main proposer" that some facilities differentiate and would be assigned to the person leading the group of experimenters attending the measurement.
>
> **Collaborator** *(collaborator)* - a person that is associated with the datasets in a different way. For example, a user that just mailed in a sample and got given access to the resulting data without ever performing the measurement might fall into this category.
>
> These four roles should form the starting point of a common mapping, which the demonstrator will help to evaluate and verify. The suggested spelling in brackets follows Python naming conventions (snake_case).

Parameters

Parameters in the API Data Model can be associated with Documents and Datasets. In the latter case they can describe properties of the Samples, or Instrument (or both). The survey of submitted parameters found a varying degree of metadata availability from the facilities. Some datasets had only two or three Parameters

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

associated with a dataset, others had dozens of key-value pairs. To come up with a common initial set to curate data on we established the following criteria:

- Parameters need to be presented in training data from at least one facility - this excludes future parameters from the discussion that do not provide immediate value
- Parameters need to be interpretable on their own, without secondary information - in the future combined searches may make this a noticeable limitation
- filtering on the Parameter would be a valuable search criterion for a person not involved in the original data collection - this for the moment excludes many diagnostics parameters, which can still be accessed, but will at the moment not receive a standard mapping

With those limitations in place we only extracted the following candidates for the initial round of mappings:

- wavelength or photon energy
- sample chemical formula
- sample temperature
- sample pressure

While this looks like a short list, one has to keep in mind that changing and verifying the mapping of these parameters consistently can require a large amount of manual or semi-automated work at the facilities when historic datasets need to be curated and changed. The discussion also already highlighted a number of issues that can be solved on these five examples now, that any future mapping will benefit from. One is the issue of unit conversion, that the API already supports. The supported conversions do not include the change from wavelength to photon energy, which will require special treatment. Especially if we also want to support a potential conversion to neutron energy. It was also highlighted that legacy data often would have unknown or inconsistent units applied, which prevents any automatic mapping. In addition, matching the values is non-trivial considering values are often stored as strings, get converted to floating point numbers with limited precision and then undergo arithmetic processing due to the potential unit conversions. So even an exact match needs to allow some fuzziness. In addition, for many experiments the metadata recorded does not refer to a single value, but a range. That is the case when the values represent scans of values, parametric measurements and the like. We may need to add special treatment for that into the API or the local implementations.

Techniques

ExPaNDS have been working closely with PaNOSC on the Experimental technique ontology, especially by defining the terms that should be available in the controlled vocabulary. For the experimental techniques, the idea was to use atomic classes to describe the techniques, here are the four main classes: experimental physical process, experimental probe, functional dependence and technique purpose (naming may change after consultation with all PaNOSC/ExPaNDS partners). Each class will be defined as a hierarchy/taxonomy. This is work in progress. Once available the added classification can be added to the data catalogues and the search can be augmented by the technique filter, without any modification to the API.

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

Overview of the Demonstrator Implementation

General Overview and Considerations

The REST framework LoopBack was chosen for the implementation of the search API and is also part of the solution stack for the federated search service. LoopBack is a mature Open Source project that is backed and maintained by IBM, this should give it a level of sustainability for the future. It is following the REST architectural style, it also has a number of features built in such as JSON filters, ordering, pagination and authentication. The subset of features from LoopBack that has been used has been documented in detail below, this enables other REST frameworks to be used for the local implementation (adaption to the facility data catalogue), as long as they adhere to the specified documentation. In addition to documenting the API a test harness has been developed that can be used to verify any implementation.

Results from individual API endpoints are returned with a score that is determined by each facility. The demonstrator uses this score in the aggregation process to sort the full set of results of all the sites. No specific logic is embedded in the demonstrator application to provide its own scoring mechanism.

The federated search demonstrator repository is available on GitHub at https://github.com/panosc-eu/search-api/tree/dev/federated_search_api. Anyone trying to use or implement it can raise issues and submit pull requests there.

Data Model

Classes that may be returned by API calls have an id property allowing reference to them in subsequent calls like GET /datasets/{id}. This id may be an internal identifier of the local metadata catalogue. It should be considered ephemeral and should not be retained by the client beyond the current session. Some amount of effort should be taken by the server to make this property globally unique to prevent clashes. The value should be restricted to the characters 0-9A-Za-z_.~-. Future versions of the API may pose more stringent restrictions on this property for federation purposes.

Some classes have a `pid` property. This is a persistent identifier that is supposed to not change over time and may be stored in the client for later referral. It also allows cross references to objects in remote repositories. The value should be a well-established persistent identifier such as a DOI, a Handle, an ORCID-iD, or a ROR. If such a PID is not available for the object, a locally assigned identifier in the metadata catalogue is acceptable, as long as it is guaranteed to be stable.

As each site will generate `pid`’s independently there is a risk for collusion e.g. two datasets getting the same `pid`. To ensure that this does not happen the convention was used that each site's `pid` contains a prefix that is used when responding to queries from the federated search. This convention is to be superseded by an algorithmic solution in later versions of the federated search service.

Units and Conversions

Measurement values and their associated units must be first class citizens in a data catalogue and a search API. The adequate choice of units for a particular quantify, experiment and sample is important to the user carrying out the experiment. Users would like to see the units displayed back to them how they were entered.

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

So storing everything in SI units is not acceptable. That does have implications on the search. Just matching values in the database engine would be easy, but there is no universally accepted choice of units. For these reasons conversions need to be performed on the fly.

The choice of which units to be supported in the search was created based on observations from current metadata catalogues and domain knowledge. When supplying units in a parameter query, the quantity will be converted for comparison with the value stored in the database. Before returning the results, the relevant quantity is converted to the unit supplied by the user in the query. For example, when querying a parameter using keV, the keV quantity will be converted and compared to the value stored in the database. Results will also be returned in the same unit that the user provided in the query, in this case keV. This enables the user to easily compare the results of the search with their filter, but also to sort results by that parameter. The functionality for this is provided by the math.js library (https://mathjs.org/), which also includes support for the chosen units and prefixes.

Support for searching ranges of a value is included, this will enable users to specify a specific range a parameter needs to reside between. This is essential as filtering metadata values after conversions with exact matches with only work in a few cases. How this is specified in the API is borrowed from loopback and illustrated in the example use cases below.

At the moment, unit conversion is handled at the providers as part of the original search API implementation. Further investigation is needed to determine whether some of the logic behind unit transformations is worthwhile to be moved to the federated search service.

Architecture

The general architecture of the federated search service is shown in Figure 3.

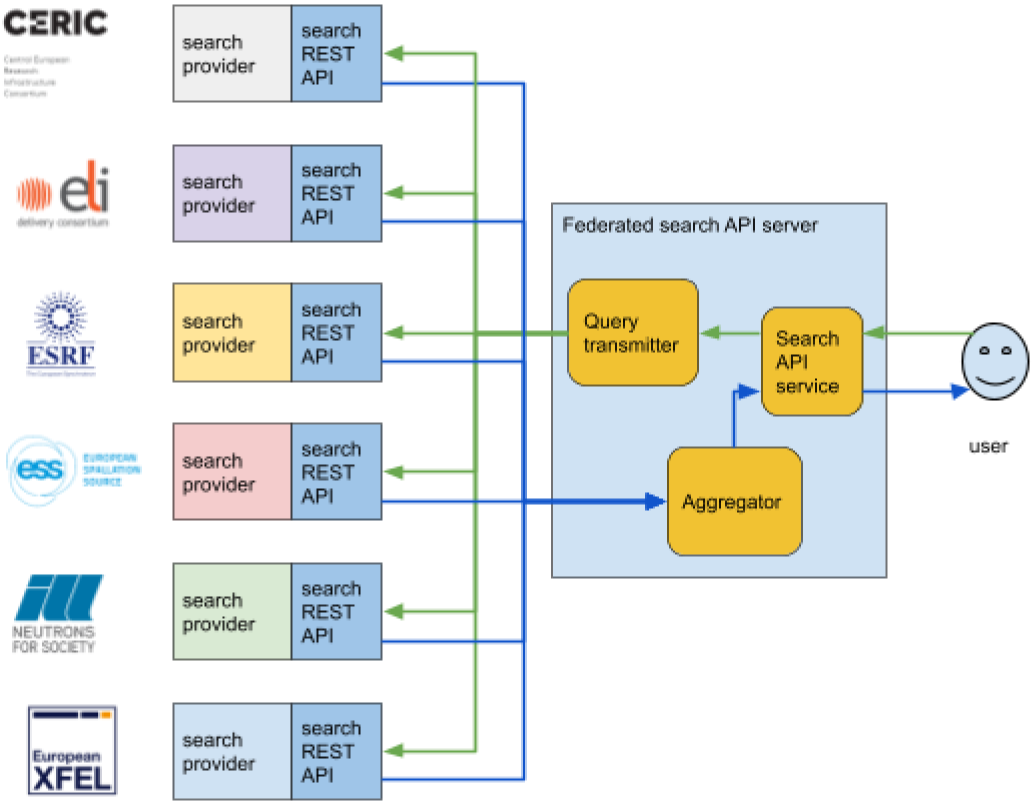


Figure 3: Federated search demonstrator architecture.

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

The architecture as illustrated here refines the one shown in Figure 1, most importantly the internals of the federated search API server. The server has a startup parameter for the list of providers to which the queries have to be forwarded, and from which results have to be collected. The list of providers cannot be changed during the operation, the server must be restarted to modify this list.

The federated search server has three main functional components:

Search API service: This service exposes the same API as the search providers. It accepts queries from the clients, then forwards these queries to the query transmitter component. Then it collects the results from the aggregator component through a callback mechanism and sends the results back to the clients in json format.
Source:https://github.com/panosc-eu/search-api/blob/dev/federated_search_api/search-api/server/server.js

Query transmitter: The transmitter component forwards the query to each of the configured providers in parallel. The original query as formulated by the client is forwarded to the providers, because the federated search and the search providers have identical interfaces.
Source:https://github.com/panosc-eu/search-api/blob/dev/federated_search_api/search-api/server/connectors/distributedConnector.js

Aggregator: The aggregator component receives and combines the intermediate results into the final result based on the type of the query and sends it to the search API service through a callback. Each record from a particular provider is also extended with a “provider” field to indicate the source of the record.
If the result of the input query is a list (concatenated from the result lists of the providers), then it sorts the result based on the score parameter in descending order. If the query requests for the count of the relevant items, then it returns the sum of the counts returned by the providers.
Source:https://github.com/panosc-eu/search-api/blob/dev/federated_search_api/search-api/server/aggregator.js

The federated search service has been implemented using the LoopBack 2.42.0 framework and NodeJS 10. The demonstrator in its present form contains only a limited number of sanity checks regarding possible network error data consistency conditions. A production ready version needs to be enhanced with more elaborate network handling and error handling in general. Also, the test data volumes for the demonstrator have been very low, it is expected that higher volumes require some architecture changes (e.g. caching) to provide acceptable real-time performance.

Example Queries

Like for the description of the end points of the search API, the demonstrator search API on GitHub has been complemented with a few datasets that illustrate search use cases we intended to address. Test cases on the demonstrator API ensure the correct results are returned. This test harness can also be run against other implementations for verification.
A list of example queries can be found on the project confluence page:

https://confluence.panosc.eu/display/wp3/Demonstrator+test+cases, some examples are given below as well.

Example 1: Search for documents with a title containing “X-ray”

This example shows a search for documents whose title contains the string “X-ray”. Two documents are returned that are from the same provider (facility). The score field is set to zero by all providers in the current implementation.

Endpoint: GET/Documents
Filter: `{"where":{"title":{"like":"X-ray"}}}`
Expected response:

```
[
  {
    "pid": "10.16907/7eb141d3-11f1-47a6-9d0e-76f8832ed1b2",
    "isPublic": true,
    "type": "publication",
    "title": "Micrometer-resolution X-ray tomographic imaging of a complete intact post
mortem juvenile rat lung",
    "score": 0,
    "provider": "http://psi-provider:3000/api"
  },
  {
    "pid": "10.16907/c01b594e-e42e-45f9-913e-5001fd283aee",
    "isPublic": true,
    "type": "publication",
    "title": "Synchrotron radiation X-ray tomographic microscopy datasets for
Atlantocarpus, Lambertiflora and Mugideiriflora from the Early Cretaceous of eastern
North America and Portugal",
    "score": 0,
    "provider": "http://psi-provider:3000/api"
  }
]
```

Example 2: Retrieve documents related to “AWA RF” source

This example has the same structure as the previous one, but uses a different substring to search for. This time one document is returned, and we assume the user is interested in the datasets that belong to this document, so the next example retrieves these.

Endpoint: GET/Documents
Filter: `{"where":{"title":{"like":"AWA RF"}}}`
Expected response:

```
[
  {
    "pid": "10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc",
    "isPublic": true,
    "type": "publication",
    "title": "Data of of a global sensitivity study of the PSI Injector 2, PSI Ring Cyclotron, IsoDAR and AWA RF gun",
    "score": 0,
    "provider": "http://psi-provider:3000/api"
  }
]
```

Example 3: Retrieve datasets that belong to a specific document

The query in this example aims to retrieve the datasets that belong to a document recently returned. We can use the pid field of the document to uniquely identify it, and although the query is forwarded to all providers, only the one storing the document returns any result, which is a dataset associated with the document.

Endpoint: GET/Datasets

Filter:

```
{ "where":{
      "documentId":"10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc"
      }
}
```

Expected response:

```
[
  {
    "pid": "20.500.11935/09dc7cdd-0c6a-473c-88a6-fc1cd4d29b4a",
    "title": "uq-amr/ring",
    "isPublic": true,
    "creationDate": "2019-12-14T09:32:26.000Z",
    "score": 0,
    "documentId": "10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc",
    "instrumentId": "20.500.11935/4821438d-2359-42da-8e4f-a9134c855c9d",
    "provider": "http://psi-provider:3000/api",
    "parameters": [],
    "techniques": []
  },
  {
    "pid": "20.500.11935/0ca89b9c-219a-460b-8a4b-6d10c68c1d3c",
    "title": "uq-amr/ring",
    "isPublic": true,
    "creationDate": "2019-11-26T11:34:12.000Z",
    "score": 0,
    "documentId": "10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc",
    "instrumentId": "20.500.11935/4821438d-2359-42da-8e4f-a9134c855c9d",
    "provider": "http://psi-provider:3000/api",
    "parameters": [],
    "techniques": []
  },
  {
    "pid": "20.500.11935/0d50ecd4-02cf-41d2-9a6d-8606e677890c",
    "title": "uq-amr/isodar",
    "isPublic": true,
    "creationDate": "2019-12-14T15:38:51.000Z",
    "score": 0,
    "documentId": "10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc",
    "instrumentId": "20.500.11935/4821438d-2359-42da-8e4f-a9134c855c9d",
    "provider": "http://psi-provider:3000/api",
    "parameters": [],
    "techniques": []
  },
  {
    "pid": "20.500.11935/15a650fe-9f20-430c-9b1d-a3ee2eb72087",
    "title": "uq-amr/AWA-gun",
    "isPublic": true,
    "creationDate": "2020-01-09T10:08:41.634Z",
    "score": 0,
    "documentId": "10.16907/fd7d6880-7a0f-4b52-942d-35e23b77d0dc",
    "instrumentId": "20.500.11935/1587bac7-402d-4410-9f62-30eef9f950fe",
    "provider": "http://psi-provider:3000/api",
    "parameters": [],
    "techniques": []
  }
]
```

Example 4: Get documents that have ions mentioned in their title

This example shows that search results coming from different providers are aggregated into a joint list.

Endpoint: GET/Documents

Filter: `{"where":{"title": {"like": " ion"}}}`

Expected response:

```
[
  {
    "pid": "e9d5bec3e066e01f5e6efd93e7a94d1208e66098",
    "isPublic": true,
    "type": "Proposal",
    "title": "Laser driven ion acceleration",
    "score": 0,
    "provider": "http://eli-1-provider:3000/api"
  },
  {
    "pid": "doi:10.5442/ND000001",
    "isPublic": true,
    "type": "publication",
    "title": "Neutron study of the topological flux model of hydrogen ions in water ice",
    "summary": "The familiarity of water ice means we often overlook its non-trivial
character illustrated, for example, by the many snowflake morphologies resulting from
disordered combinations of covalent and hydrogen bonds between hydrogen and oxygen atoms
in water ice’s most common phase (Ih) that keep the H_2 O molecular character. Using
neutron diffraction on the flat-cone diffractometer E2 at BER-II, Helmholtz-Zentrum
Berlin, we probe the atomic scale configuration in the Ih phase of water ice to test
theories that describe this “disordered” state as exhibiting a form of topological order
characterized by an emergent gauge field. We find excellent agreement between low-
temperature experiment and analytical theory, which even allows us to estimate the density
of defects charged under this emergent gauge field. The development of quantitative models
of water ice paves the way for further studies to develop a comprehensive atomic-scale
understanding of this most commonplace of solids.\nThe merged untransformed datasets from
the flat-cone diffractometer E2 at the neutron source BER II is given in the Nexus/HDF5
file format. The calculated reciprocal space and the simulation are stored as HDF5
files.",
    "releaseDate": "2018-03-07T00:00:00.000Z",
    "license": "CC0-1.0",
    "score": 0,
    "provider": "http://hzb-provider:3000/api"
  }
]
```

Example 5: Retrieve documents related to absorption based techniques

This example shows a more complex query with a filter of related objects. We can retrieve documents based on attributes of related datasets which meet the requirements of the filter. Also, the query matched documents at two different providers whose results are merged into a single list.

Endpoint: GET/Documents

```
Filter: {"include":[ {
             "relation":"datasets",
             "scope":{"include":[{
                   "relation":"techniques",
                   "scope":{"where":{"name": {"ilike": "Absorption"}}}
             }]}
       }]}
```

Expected response:

```
[
  {
    "pid": "10.5072/panosc-document2",
    "isPublic": true,
    "type": "proposal",
    "title": "PaNOSC Test Proposal",
    "score": 0,
    "provider": "http://ess-provider:3000/api",
    "datasets": [
      {
        "pid": "20.500.12269/panosc-dataset3",
        "title": "PaNOSC Test Dataset 3",
        "isPublic": true,
        "creationDate": "2020-05-05T15:01:02.341Z",
        "documentId": "10.5072/panosc-document2",
        "instrumentId": "20.500.12269/f0637030-9f89-4398-8f01-09211145efa1",
        "techniques": [
          {
            "pid": "20.500.12269/panosc-tech2",
            "name": "x-ray absorption"
          }
        ]
      },
      {
        "pid": "20.500.12269/panosc-dataset4",
        "title": "PaNOSC Test Dataset 4",
        "isPublic": true,
        "creationDate": "2020-05-05T15:01:02.341Z",
        "documentId": "10.5072/panosc-document2",
        "instrumentId": "20.500.12269/d3dd2880-637a-40b5-9815-990453817f0e",
        "techniques": [
          {
            "pid": "20.500.12269/panosc-tech2",
            "name": "x-ray absorption"
          }
        ]
      }
    ]
  },
  {
    "pid": "10.16907/c01b594e-e42e-45f9-913e-5001fd283aee",
    "isPublic": true,
    "type": "publication",
    "title": "Synchrotron radiation X-ray tomographic microscopy datasets for
Atlantocarpus, Lambertiflora and Mugideiriflora from the Early Cretaceous of eastern
North America and Portugal",
    "score": 0,
    "provider": "http://psi-provider:3000/api",
    "datasets": [
      {
        "pid": "20.500.11935/1873b589-58bb-40d5-a2d1-67e0b651a268",
        "title": "PP43780a",
        "isPublic": true,
        "creationDate": "2008-10-10T19:22:02.000Z",
        "documentId": "10.16907/c01b594e-e42e-45f9-913e-5001fd283aee",
        "techniques": [
          {
            "pid": "20.500.11935/bd65160d-38ba-45f7-ba1d-4200595e7067",
            "name": "Absorption contrast x-ray tomographic microscopy"
          }
        ]
      }
    ]
```

```
   }
]
```

Deployment Instructions

The demonstrator service can be deployed as docker containers.
Source code of provider and federated search is available on GitHub:

- provider:
  https://github.com/panosc-eu/search-api
- federated search:
  https://github.com/panosc-eu/search-api/tree/dev/federated_search_api/search-api

Both of them are buildable by the docker container tool.

Example dataset from all facilities can be served up locally for testing purposes from the following GitHub branches and files:

- CERIC: https://github.com/panosc-eu/search-api/blob/dev/ceric/data/db.json
- ELI: https://github.com/panosc-eu/search-api/blob/dev/ELI/data/db.json
- ESS: https://github.com/panosc-eu/search-api/blob/ESS/data/db.json
- ESRF: https://github.com/panosc-eu/search-api/blob/ESRF/data/db.json
- ILL: https://github.com/panosc-eu/search-api/blob/ILL/data/db.json
- XFEL: https://github.com/panosc-eu/search-api/blob/dev/xfel/data/db.json
- HZB: https://github.com/RKrahl/search-api/blob/hzb_data/data/db.json

Detailed Deployment Procedure

To build the provider container:

```
$ git clone https://github.com/panosc-eu/search-api.git
$ docker build --tag search-api-provider:1.1 .
```

To build the federated search container:

```
$ git clone --single-branch --branch dev/federated_search_api https://github.com/panosc-eu/search-api.git
$ cd search-api
$ docker build --tag federated-search-api:1.1 .
```

To start the federated search environment with test providers:

- start providers with the inbuilt data:

```
$ docker run -d --name provider_1 -p 3001:3000 search-api-provider:1.1
```

- start providers with the example JSON data:

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

```
$ wget -O /tmp/db.json https://raw.githubusercontent.com/panosc-eu/search-api/dev/ceric/data/db.json
$ docker run -d --name provider_2 -p 3002:3000 \
-v /tmp/db.json:/home/node/app/data/db.json search-api-provider:1.1
```

- start federated search API:

```
$ docker run -d --network host --name federated_search -p 3000:3000  \
-e PROVIDERS=”http://localhost:3001/api,http://localhost:3002/api” federated-search-api:1.1
```

To instead start the federated search environment with external providers:

```
$ docker run -d --name federated_search -p 3000:3000  \
-e PROVIDERS=”[API URL of provider 1],[API URL of provider 2]” federated-search-api:1.1
```

The federated search API will be available at http://localhost:3000/api URL or through the LoopBack UI: http://localhost:3000/explorer
Additionally, support for docker-compose has been added that enables easy setup for testing and deployment. Using docker-compose we generate an environment containing the federated search-api as well as two local providers with example data. To use this setup issue following commands:

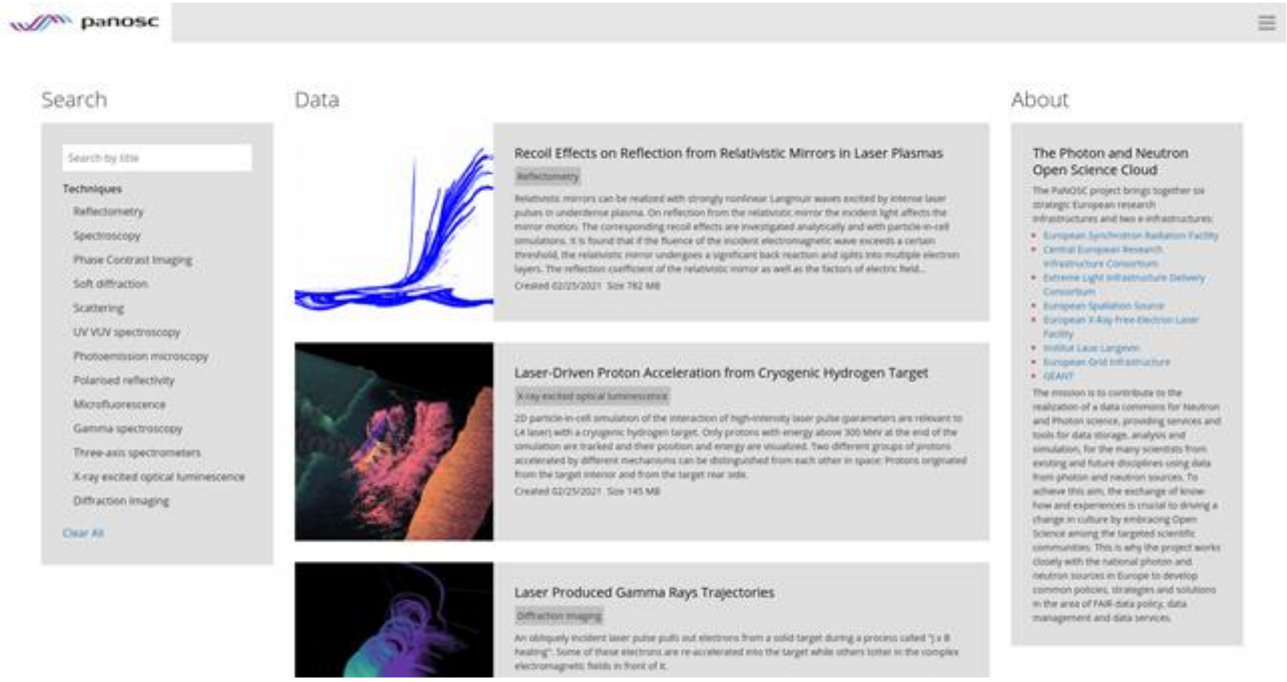
```
$ git clone -b dev/federated_search_api https://github.com/panosc-eu/search-api.git
$ docker-compose -f docker-compose-test.yaml up –build
```

The explorer of the federated search-API can be found on `localhost:3000` where it is easy to generate a dataset query generating a response from both the local providers.
The deployment instructions have been tested at ELI, PSI, CERIC and ESS. Tests have been made both with test providers and with external providers (for instance using CERIC external provider at http://panosc-search.apps.okd2.ceric.fedcloud.eu/api/). Apart from identifying some typos on test providers’ data that have since been fixed, the trials confirmed the interoperability of the implementations and established the federated demonstrator as a useful tool to achieve compliance.

Frontend and WP4

The frontend of WP4’s PaN Portal aims to integrate the Search API in its full capability. The search API is used as the main backend solution for the catalogue part of the Portal. As such it will populate the Portal with valuable data and enable advanced searching capabilities. After browsing Documents (e.g. publications, proposal, use cases) as well as related Datasets, the Portal is going to offer EOSC users with the possibility to analyse the data selected on compute environments integrated to the cloud.

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



Search

Search by title

Techniques

- Reflectometry
- Spectroscopy
- Phase Contrast Imaging
- Soft diffraction
- Scattering
- UV VUV spectroscopy
- Photoemission microscopy
- Polarised reflectivity
- Microfluorescence
- Gamma spectroscopy
- Three-axis spectrometers
- X-ray excited optical luminescence
- Diffraction imaging

Clear All

Data

Recoil Effects on Reflection from Relativistic Mirrors in Laser Plasmas

Reflectometry

Relativistic mirrors can be realized with strongly nonlinear Langmuir waves excited by intense laser pulses in underdense plasma. On reflection from the relativistic mirror the incident light affects the mirror motion. The corresponding recoil effects are investigated analytically and with particle-in-cell simulations. It is found that if the fluence of the incident electromagnetic wave exceeds a certain threshold, the relativistic mirror undergoes a significant back reaction and splits into multiple electron layers. The reflection coefficient of the relativistic mirror as well as the factors of electric field...

Created 02/25/2021 Size 762 MB

Laser-Driven Proton Acceleration from Cryogenic Hydrogen Target

X-ray excited optical luminescence

2D particle-in-cell simulation of the interaction of high-intensity laser pulse (parameters are relevant to L4 laser) with a cryogenic hydrogen target. Only protons with energy above 300 MeV at the end of the simulation are tracked and their position and energy are visualized. Two different groups of protons accelerated by different mechanisms can be distinguished from each other in space: Protons originated from the target interior and from the target rear side.

Created 02/25/2021 Size 145 MB

Laser Produced Gamma Rays Trajectories

Diffraction imaging

An obliquely incident laser pulse pulls out electrons from a solid target during a process called "j x B heating". Some of these electrons are re-accelerated into the target while others totter in the complex electromagnetic fields in front of it.

About

The Photon and Neutron Open Science Cloud

The PaNOSC project brings together six strategic European research infrastructures and two e-infrastructures:

- European Synchrotron Radiation Facility
- Central European Research Infrastructure Consortium
- Extreme Light Infrastructure Delivery Consortium
- European Spallation Source
- European X-Ray Free-Electron Laser Facility
- Institut Laue Langevin
- European Grid Infrastructure
- GÉANT

The mission is to contribute to the realization of a data commons for Neutron and Photon science, providing services and tools for data storage, analysis and simulation, for the many scientists from existing and future disciplines using data from photon and neutron sources. To achieve this aim, the exchange of know-how and experiences is crucial to driving a change in culture by embracing Open Science among the targeted scientific communities. This is why the project works closely with the national photon and neutron sources in Europe to develop common policies, strategies and solutions in the area of FAIR data policy, data management and data services.

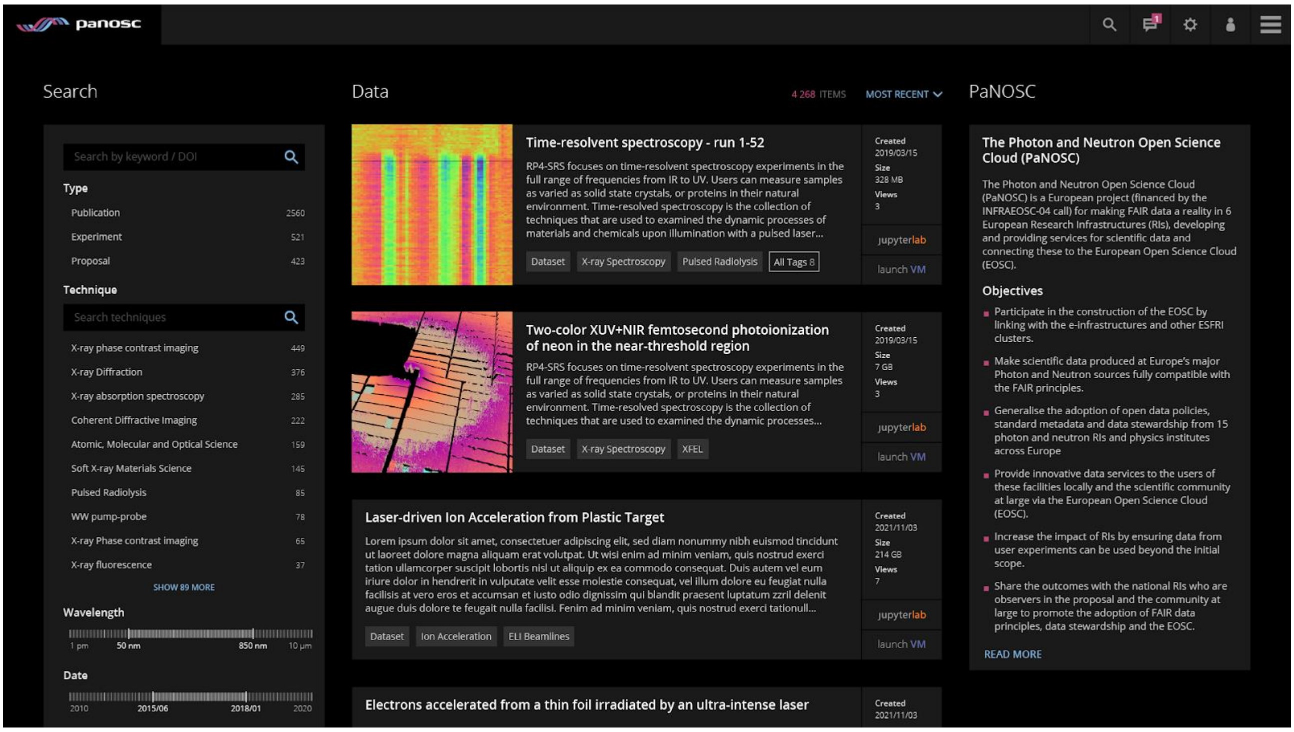
Figure 4: Screenshot of the federated search demonstrator web frontend.

Current integration of the Search API should be considered a proof of concept showcasing the data fetching capabilities and providing very basic title search and filtering based on keywords (Figure 4). The frontend can be deployed on docker using the Dockerfile provided in its git repository, it is recommended to use the development branch. The build command to use is:

```
$ docker build --tag frontend:0.1 . --build-arg="SEARCH=<SEARCH_API>/api"
```

As for the demonstrator above, a docker-compose setup has been created to automate the deployment procedure - starting the frontend on port 8080 and Search API with the demo database on port 8081. Run docker-compose up in the root of the repository. This setup is also being used on an internal demonstrator server at ELI Beamlines.

A more advanced search interface is currently under development and will be available as Portal development progresses. Standalone versions of the search interface as well as a standalone search query generator are also to be delivered as part of work on WP4's Portal frontend. Figure 5 shows a design mockup of the advanced search integration.



panosc

Search

Search by keyword / DOI

Type

Publication	2560
Experiment	521
Proposal	423

Technique

Search techniques

X-ray phase contrast imaging	449
X-ray Diffraction	376
X-ray absorption spectroscopy	285
Coherent Diffractive Imaging	222
Atomic, Molecular and Optical Science	159
Soft X-ray Materials Science	145
Pulsed Radiolysis	85
WW pump-probe	78
X-ray Phase contrast imaging	65
X-ray fluorescence	37

SHOW 89 MORE

Wavelength

1 pm 50 nm 850 nm 10 μm

Date

2010 2015/06 2018/01 2020

Data 4 268 ITEMS MOST RECENT

Time-resolvent spectroscopy - run 1-52

RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes of materials and chemicals upon illumination with a pulsed laser...

Dataset | X-ray Spectroscopy | Pulsed Radiolysis | All Tags 8

Created 2019/03/15 Size 328 MB Views 3

jupyterlab launch VM

Two-color XUV+NIR femtosecond photoionization of neon in the near-threshold region

RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes...

Dataset | X-ray Spectroscopy | XFEL

Created 2019/03/15 Size 7 GB Views 3

jupyterlab launch VM

Laser-driven Ion Acceleration from Plastic Target

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Fenim ad minim veniam, quis nostrud exerci tationull...

Dataset | Ion Acceleration | ELI Beamlines

Created 2021/11/03 Size 214 GB Views 7

jupyterlab launch VM

Electrons accelerated from a thin foil irradiated by an ultra-intense laser

Created 2021/11/03

PaNOSC

The Photon and Neutron Open Science Cloud (PaNOSC)

The Photon and Neutron Open Science Cloud (PaNOSC) is a European project (financed by the INFRAEOSC-04 call) for making FAIR data a reality in 6 European Research Infrastructures (RIs), developing and providing services for scientific data and connecting these to the European Open Science Cloud (EOSC).

Objectives

- Participate in the construction of the EOSC by linking with the e-infrastructures and other ESFRI clusters.
- Make scientific data produced at Europe's major Photon and Neutron sources fully compatible with the FAIR principles.
- Generalise the adoption of open data policies, standard metadata and data stewardship from 15 photon and neutron RIs and physics institutes across Europe
- Provide innovative data services to the users of these facilities locally and the scientific community at large via the European Open Science Cloud (EOSC).
- Increase the impact of RIs by ensuring data from user experiments can be used beyond the initial scope.
- Share the outcomes with the national RIs who are observers in the proposal and the community at large to promote the adoption of FAIR data principles, data stewardship and the EOSC.

READ MORE

Figure 5: Design mockup of the frontend with more advanced search UI in the left column

Lessons Learned in the Development Process

The following items came out of discussion and as a result of the development. They inform the direction of future work.

Changes to API needed: The project would benefit from a dynamic method for generating user interface elements, based on the data available in the data catalogues and the adoption of common vocabularies. An example of this would be the set of techniques that occur in the databases. It would be useful to present these as filter options to the user to assist formulating a valid and useful query. Other information is stored as instances of Parameters (see Figure 2). For the demonstrator an endpoint has been added to query the set of parameter values found in the databases. Using these queries, the resulting user interface is able to list available techniques to the user as seen in Figure 4 and 5. Unless a different solution will be found an updated API specification will be issued to meet these requirements.

Ontology feedback: the application of the above solution to query all data catalogues for potential techniques and parameters highlighted the lack of a defined common vocabulary, especially for the legacy datasets from before the start of PaNOSC that are no longer subject to embargo periods. This puts some focus on the work to harmonize the dictionaries of metadata terms between facilities and the needs to curate existing and future data to comply with these new boundary conditions.

Complications with the mapping: care must be taken that internal identifiers from different providers do not collide with each other. The objects in the data model are identified either through PID, DOI or internal IDs, the usage of locally unique identifiers works when using data that is coming from only from one site, but presents a problem when combining data from multiple sites as multiple objects may use the same identifiers. To provide globally unique identifiers, a transparent prefix-based scheme could be implemented. In this scheme unique names are assigned to the providers, these names then used as identifier prefixes where collisions may occur. The prefixes need to be handled by the federated search server, adding/removing them as the identifiers are communicated between the providers and the client.

Units handling: Units handling is delegated entirely to the providers and works as described above in this document. This scheme works but has the consequence that unit handling code is duplicated at each site and is more complex than it could be because each provider needs to be able to handle each kind of unit as its input. The situation could be improved by converting quantities to predefined (e.g. SI) units at the federated search server, then providers would need to deal with (convert to/from) these units. Further investigation is needed to determine the advantages and drawbacks of centralized vs. distributed unit handling schemes.

Authentication: For the purposes of WP4 and to benefit photon and neutron facility users in general it would be extremely important to also provide access to the dataset search via an authenticated interface to access data still under embargo. This would enable users to benefit from PaNOSC services to process and analyze their own data, one of the main use cases of facility computing infrastructure. For this demonstrator it was decided to postpone the implementation of authentication to a later stage. In principle the required components are available for the API. But a later integration to authentication enables the work package to take full advantage of the deliverables by WP6, which is in the process of rolling out a common authentication infrastructure that gives access to centrally federated user ids.

Next Steps

Following the release of the search API this delivery of the federated search demonstrator was the second important step establishing a PaN domain specific dataset search through PaNOSC. The main aim of the demonstrator was to allow a complete functional test of the API, the compliance of partner implementations and to provide a proof of concept of a federated production service. Already during the implementation phase of the demonstrator some issues with the API and around the planned services were discovered. This corroborates the usefulness of this iterative development approach. Obviously further more thorough testing with the federated search demonstrator will commence to ensure there are a number of compliant implementations of the common API in its most up to date version.

With the implementation work, as in the software development, progressed this far, a focus area that needs to progress to achieve success is the definition of useful controlled vocabularies to formulate the search. both cataloguing work packages (ExPaNDS and PaNOSC) make good progress on the ontologies, namely the list of parameters, their mappings with unit conversions, the hierarchy of experimental techniques as well as a list of roles for dataset members. Draft versions are now available for most of these. With the demonstrator in place the applicability of the dictionaries can be tested against all partner databases and the target use cases can be evaluated. This will require changes to the local data repositories, either by changing data records or introducing local mappings between local terms and common ones.

In addition, four main areas have been identified that should be visited in more detail to take the aggregator to a functional PaN search service, which will require some more work on the implementation. Either on the implementation of the aggregator or for the API.

- The search providers have a pagination feature which allows to divide the returned data of the queries. This feature is also a requirement for the federated search, but is lacking the implementation or test.
- A query in the federated search will potentially return a large number of datasets spanning multiple

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

sites, there is therefore a need for a score that can be used for sorting of the results. The API requires each dataset to return a score indicating how well the search has matched, which does allow a ranking by relevance of the results. However, the ranking in the demonstrator has not been fully tested (it also requires the pagination functionality) and there has not yet been an agreed upon way of calculating this score. Some information for potential options has been collected and the presence of the demonstrator will facilitate the required exploratory work.

- The current version of the federated search provider does not support dynamic search provider endpoint changes. In the future it would be useful to change the list of providers without having downtime. This would allow seamless integration of the new search providers.
- Integration of authentication, for which work package 6 is preparing important groundwork, would make the search service much more valuable for the facilities' users and hence would be of most value to the mission of the photon and neutron facilities. Once an authentication service is available in normal operations, there are no foreseen obstacles in adapting this to a specific web service like the search API.

Again, the purpose of the demonstrator is to provide a better view on these issues and to highlight addressing which ones has the most impact to successful rollout of a search service.

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