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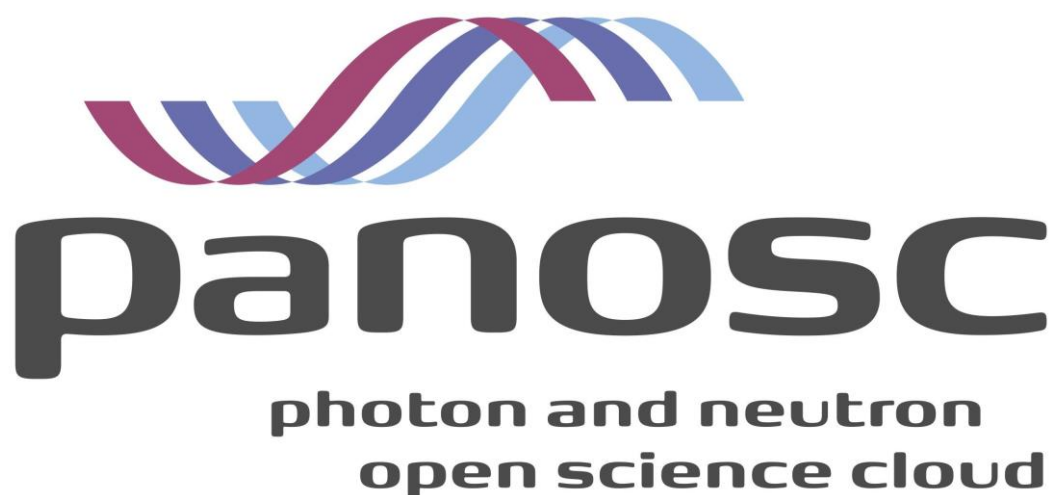# Photon and Neutron Open Science Cloud
H2020-INFRAEOSC-04-2018
Grant Agreement Number: 823852

Deliverable:
Publicly Accessible Demonstrator (4.4)

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:	D4.4
Deliverable Type:	Demonstrator
Dissemination Level	Open
Contractual Delivery Date:	30/11/2022
Actual Delivery Date:	22/12/2022
EC project Officer:	Flavius Alexandru Pana

Document Control Sheet

Document	Title: Publicly Accessible Demonstrator
	Version: 22 December 2022
	Available at: https://github.com/panosc - eu/panosc/tree/master/Submitted Deliverables
	Files: 1
Authorship	Written by: Fabio Dall'Antonia
	Contributors: Jiri Majer, Axel Bocciarelli, Loic Huder, Thomas Vincent, Emiliano Coghetto, Stuart Cuant, Teodor Ivanoaica, Michael Schuh
	Reviewed by: Andy Götz
	Approved: Andy Götz

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	ELI European Research Infrastructure Consortium (ELI-ERIC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

Table of Content

Introduction

The aspects of the FAIR principle, that is: Findability, Accessibility, Interoperability and Reusability (respectively Reproducibility) of scientific data, constitute not only a desirable, but meanwhile to a large extent even expected and required concept of good scientific practice. This holds for facility/user-exclusive (embargoed) data and open data alike.

The two major goals of the PaNOSC project are: first, making FAIR data a reality at the contributing research institutions (RIs), respectively facilities and, second, adding value to the European Open Science Cloud (EOSC) by providing means to expose and re-use open data on a federated and, to an extent, public level, including training material and services.

The PaNOSC work package four (WP4) has contributed to the aims of the project by means of software tool and software service development for scientific data analysis, that use and re-use raw scientific data with the aims of processing it and extracting meaning from the derived data, suited to produce publishable results.

Outcomes of WP4 include

- Infrastructure and services to utilise the Jupyter notebook ecosystem on distributed compute clusters
- Frameworks to inspect HDF5 format data (the de facto standard at all partner RIs) in the Jupyter environment, in local applications or embedded in web-services like data catalogues
- VISA: a cloud-based web platform for data analysis, offering both a remote desktop environment for GUI software and JupyterLab service in - optionally team-shared - virtual machine instances

The goal to align the outcomes horizontally has been successful as most of the products have been adopted, or are in the process of being adopted, by all of the partners. Moreover, the software developed is open-source, that is, source code of the respective software is freely accessible on GitHub and follows the best practices for software development. Deployed software services, as far as already used in production at the facilities, have been registered or are being registered on the EOSC marketplace.

The present demonstrator addresses a missing aspect, namely the connection of the aforementioned outcomes to open data. Demonstrator scenarios, as explained in the main part, start from a public and completely open (i.e. authentication-free) search portal. While such openness is desirable for encouraging re-usability, in reality both security and resource sustainability reasons justify the control over authentication to services and authorization of access in case of compute hardware use – as opposed to anonymous usage – even if considering a demonstrator use case. In the scope of this deliverable, open-access is understood in the following manner:

- URLs are public, i.e. not restricted to a local network and not requiring proxies or VPNs
- While authentication is required at certain stages, which may imply registration on first usage, there are no prerequisites to being registered (e. g. having been awarded beam-time at one of the facilities), meaning that there is no exclusivity of usage.

Demonstrator scenarios

All use cases demonstrated in this section have a common starting point: the PaNOSC open data search portal at https://data.panosc.eu. The public portal (no authentication required) leads to the landing page of the web-site, serving a web front-end with a text query field to initiate the search. At the backend level, the portal employs a federated search API developed in PaNOSC WP3 (https://github.com/SciCatProject/panosc-search-api), connecting to open data search endpoints at the RI data catalogues. Upon retrieval of query matches, a result page is shown. Each data catalogue endpoint is connected to a facility-side scoring service based on a common algorithm (developed in WP3), in order to provide a relevance metric associated with a returned result. Upon aggregation of results, the back-end sorts the results by their scores. The result page features the optional use of search filters, such as experiment technique following the PaNET ontology (https://expands-eu.github.io/ExPaNDS-experimental-techniques-ontology), and important physical parameters associated with the experiment (probe wavelength, sample temperature, etc.)
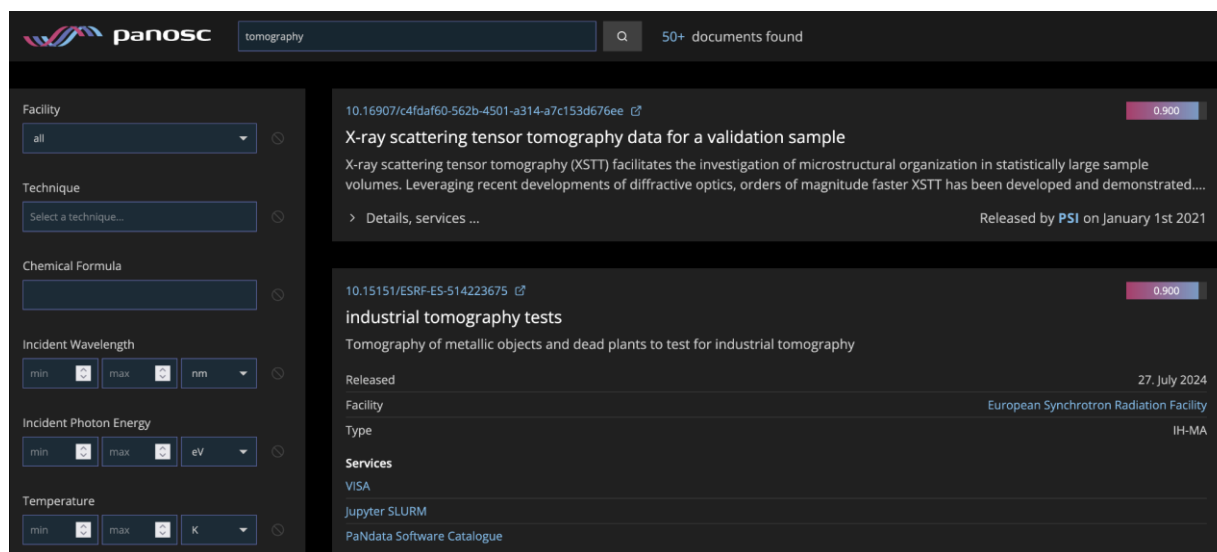

Fig. 1 Example of text query result representation in the PaNOSC search portal

Inspection and re-use of open data can start from the search portal following different routes. First, the result items on the portal redirect back to the DOI landing pages of the respective data catalogue, and some of the catalogue services have embedded tools to visualise and pre-analyse the raw data sets linked to by the DOI page. Second, search results can be expanded in the portal user interface and reveal URL links to existing services outside the data catalogues. Currently VISA is a demonstrator service accessible from the search portal in the case of European XFEL, ESRF and CERIC-ERIC.

Partner	Data catalogue (DOI target)	Result Granularity	Connected data analysis services	Service access level
ESRF	https://data.esrf.fr/doi /	Proposals or Experiments	https://visa.esrf.fr https://jupyter-slurm.esrf.fr/	Authentication
ILL	https://doi.ill.fr/	Proposals		
European XFEL	https://in.xfel.eu/metadata/doi/	Proposals	https://visa.xfel.eu	Authentication [1]
ESS	https://doi.ess.eu/detail			
ELI-ERIC				
CERIC-ERIC	https://vuo.elettra.eu/pls/vuo/open _access_data_portal.show_search	Proposals or Experiments	h5nuvola (from within catalogue)	Authentication [2]

[1] Helmholtz AAI, accepts also non-RI identity providers like GitHub, UmbrellaID
[2] Requires registration to Elettra's User Office (VUO), but no pre-conditions for registration

All search result entries provide a common link to the PaNdata Software Catalogue at https://software.pan-data.eu. This resource includes software services and tools developed for processing data from PaN facilities.

VISA has been designed to work with an authentication service such as Keycloak, that accepts OpenIdConnect tokens through an authentication and authorisation infrastructure (AAI). By accepting information from multiple remote identity providers, a PaN community (and beyond) level of openness is achieved. In the scope of PaNOSC (WP6), the UmbrellaID was envisaged as a common way to authenticate at respective facilities.

Open data access at ESRF

The options to follow from a given search result entry are, besides the DOI landing page of the experimental data at the ESRF catalogue, VISA and Jupyter-SLURM as data analysis services. Currently these two service links directly to the respective ESRF service home page without specifying data to be pre-selected.

The ESRF data catalogue exposes all experiment proposal entries with a DOI to the search API, whether these are still under embargo (public release data in the future) or have become public already. On the DOI landing page, embargoed datasets are flagged as such and access is available only to the experimental team. Open data are accessible without restriction - upon following the respective access button, a login as user or anonymous login are available. After the login process, which requires no special registration for open data, the actual data catalogue page with a listing of data sets is opened, ready for browsing.
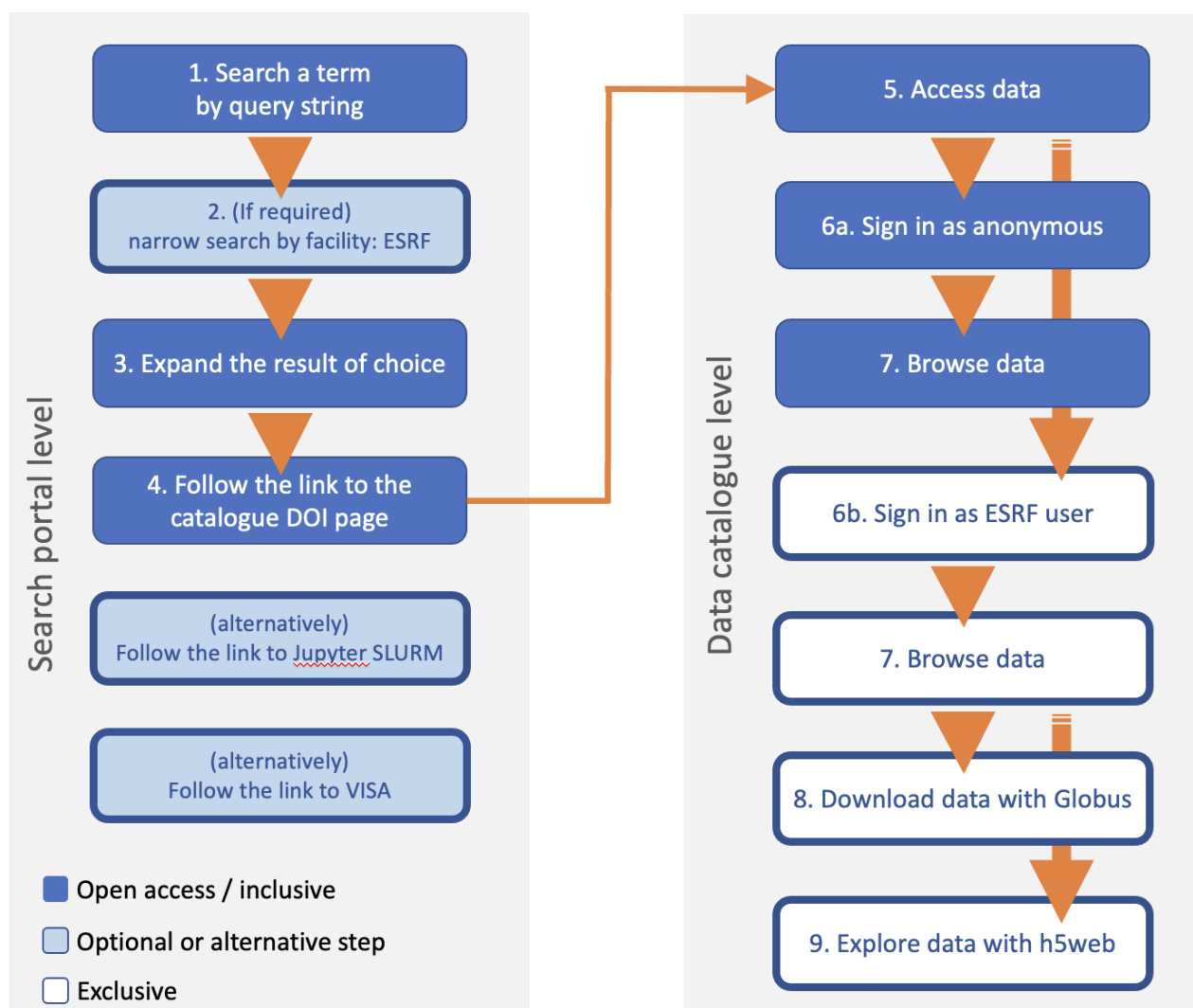
Figure 2. *Schematic workflow of open data accessibility at ESRF. The essential chain of unrestricted user action steps starting from the initial search for data of interest by keyword terms and ending with the inspection of data sets is highlighted by blue boxes*

Open data is not yet downloadable or visualisable by h5web at the moment because open proposals have been archived and they need to be restored from tape to disk-drive. Currently the path of the restored datasets is not automatically accessible by the download service or the h5web service. The path has to be specified manually. The data use options are added to the above work-flow to showcase the principal concept and desired behaviour in the future, whereas currently the relevant services work only with recent proposal data at regular disk storage paths, which applies mainly to data sets still under embargo. In practice, this makes the download and visualisation services exclusive to ESRF users, and more precisely: the subset of experiment team members. This will evolve in the future to provide access to viewing of open data for unauthenticated users too. More sophisticated viewers are being tested for large datasets e.g. https://human-organ-atlas.esrf.fr/reconstructions/

7

Figure 3. *Demonstration of the ESRF workflow by screenshots. In the depiction of the federated data search portal on top, steps 1 to 3 have already been performed: among the cross-facility results matching the "cancer cell" term, those from ESRF were filtered, and the entry of interest, ihls3112, has been expanded. Clicking on the panel close to the DOI mention on top (4) will open a new browser tab with the DOI landing page on the catalogue site. General access is available (5), since this is open data. On the following data portal login page one has to sign in anonymously (6) and reaches the dedicated data portal page for the proposal of interest (7).*

Open data access at CERIC-ERIC

Currently the only service available for a given search result entry is the DOI landing page of the experimental data at the CERIC-ERIC data catalogue.

The CERIC-ERIC data catalogue exposes to the search API only the experiment proposals that have become public. All the public experiment proposals are exposed with their corresponding DOIs and every DOI takes the user to the corresponding experiment proposal landing page in the CERIC-ERIC data catalogue.

Open data is available without restriction to registered users of the data catalogue. Registration to the data catalogue is available to everyone: there are no prerequisites to becoming a registered user.

An intermediate login page is shown when the user asks for access to the data on the corresponding DOI landing page. After the login process the actual data catalogue page opens with a listing of experiments associated with the selected experiment proposal ready for browsing, analysis and downloading.

A tool called h5nuvola (https://github.com/ElettraSciComp/h5nuvola) is available for data browsing and analysis. H5nuvola is a web-based HDF files viewer with additional functionalities: it implements file browsing, data visualisation services and selective exporting of data in the web browser. Users are required to authenticate to use h5nuvola.
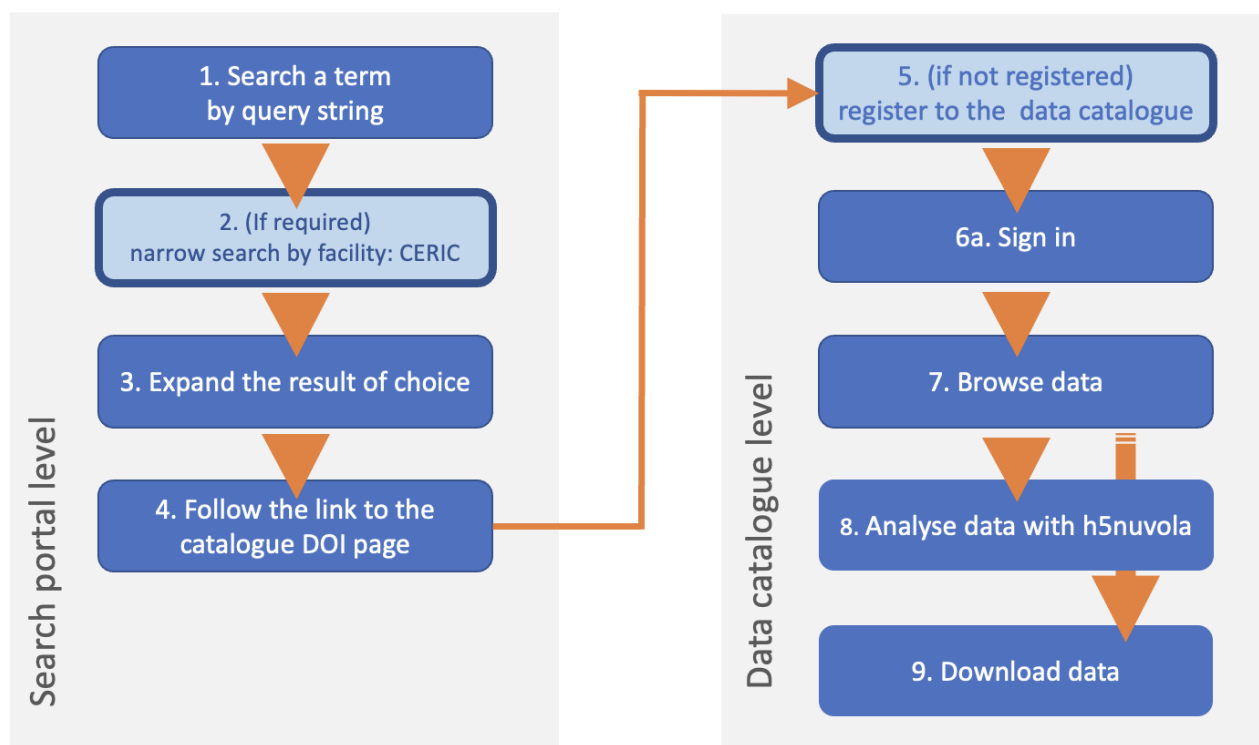
Figure 4: *Schematic workflow of open data accessibility at CERIC-ERIC, same color scheme as for ESRF.*
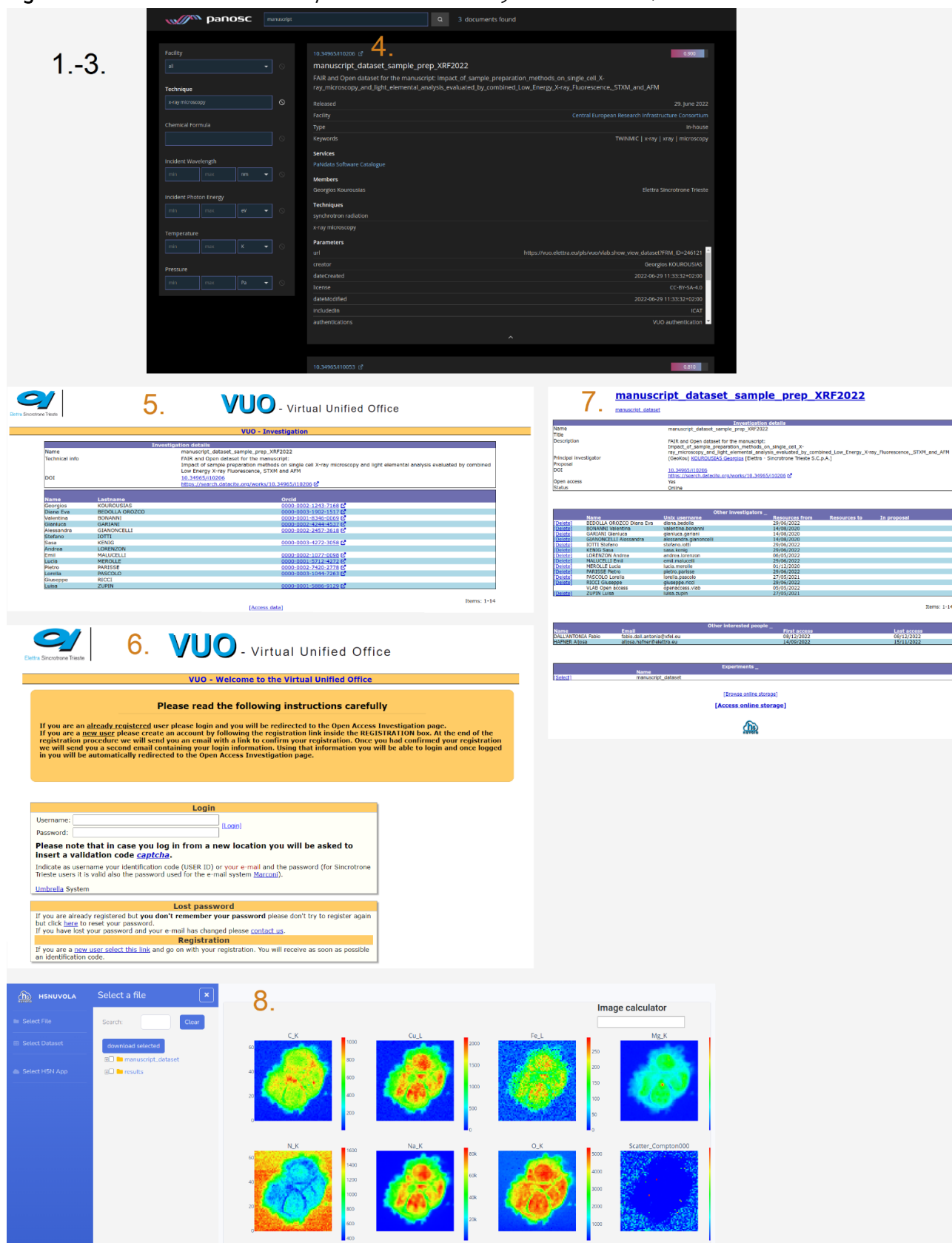

Figure 5: *Demonstration of the CERIC-ERIC workflow by screenshots.*

Figure 4 shows the corresponding demonstration of the CERIC-ERIC workflow by screenshots. In the depiction of the federated data search portal on top, steps 1 to 3 have already been performed: the results matching the "manuscript" term were filtered by the PaNET experiment technique "x-ray microscopy" and the entry of interest, "manuscript_dataset_sample_prep_XRF2022", has been expanded. Clicking on the panel close to the DOI mention on top (4) will open a new browser tab with the DOI landing page on the catalogue site. On the following data portal login page the user has to log in (6) and once logged in he reaches the dedicated data portal page for the proposal of interest (7). From here the user can browse, download or analyse the data using the tool H5nuvola (7, 8).

Open data access at European XFEL

Starting from the PaNOSC data search portal, currently two open proposals can be retrieved from European XFEL, as exposed to the search API by the metadata catalogue myMDC. These proposals represent collections of representative experiment data taken from various user proposals post embargo period, or commissioning proposals. These data sets cover a range of experimental techniques. The public DOI pages to which the search result entries link back mark the proposals as open, but do not yet provide the full functionality of the data catalogue interface for data set browsing or the Globus download service. One can get to this level only after signing in as EuXFEL user. Currently open data will be made available to any scientific guest user on request by e-mail. Coverage of the community re-use case with exploration and download on data catalogue level is being worked on; it will include a service for downloading via Globus.
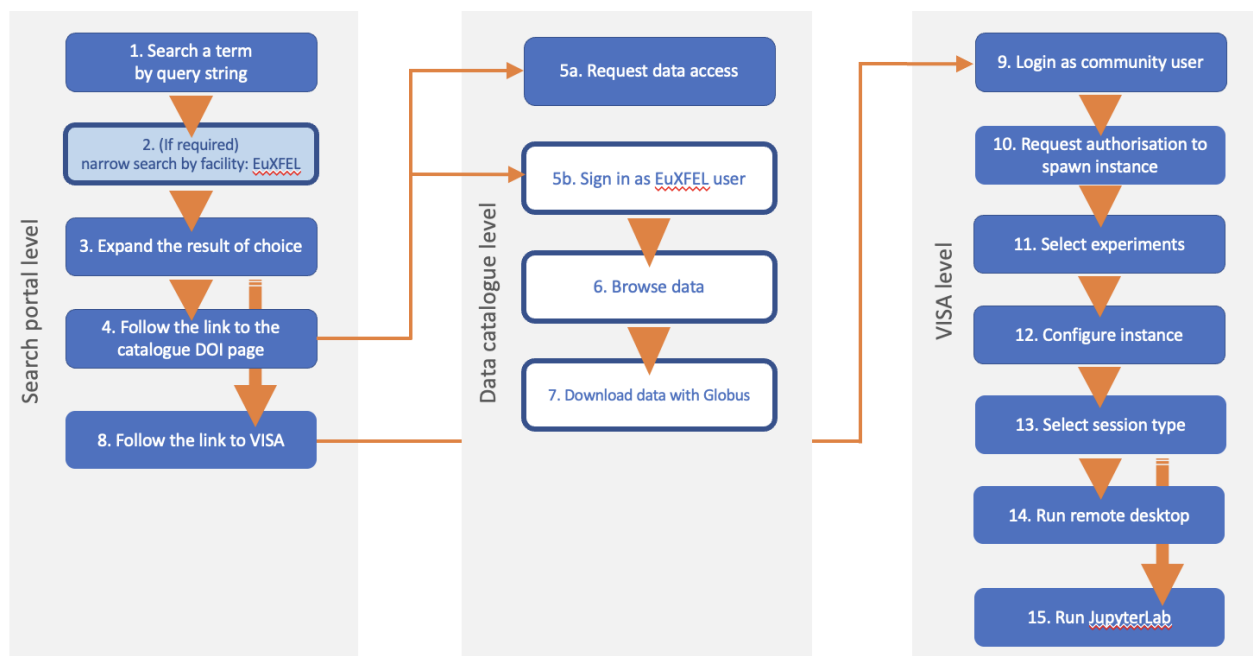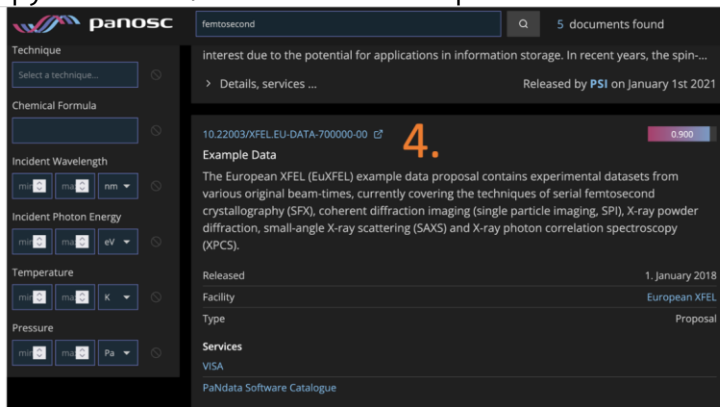

Figure 6. *Schematic workflow of open data accessibility at European XFEL, same color scheme as for ESRF.*

European XFEL results include service links to VISA. The deployed VISA service at https://visa.xfel.eu is accessible to all PaN community users after they have signed in to the Helmholtz AAI system with an authorised identity provider including GitHub and soon

11

UmbrellaId, and requested authorisation to build and use a VISA instance on the cloud cluster, hosted by DESY.
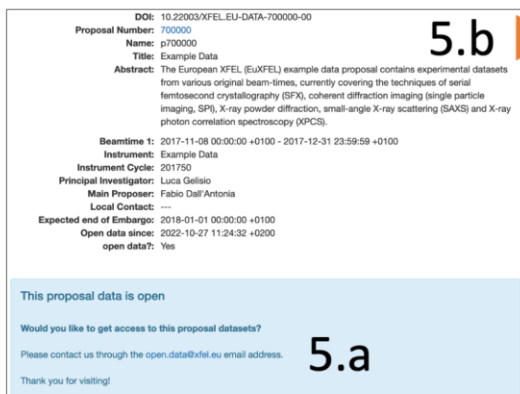
Since the portal link to VISA directs to a proposal-specific URL, the experiment data sets of the selected proposal are explorable in the VISA frontend (like they are in the data catalogue for XFEL users and will be, soon, for everyone). After selection of an experiment, the compute instance (virtual machine) can be configured and used in order to start a data analysis session based on JupyterLab and/or a remote desktop.
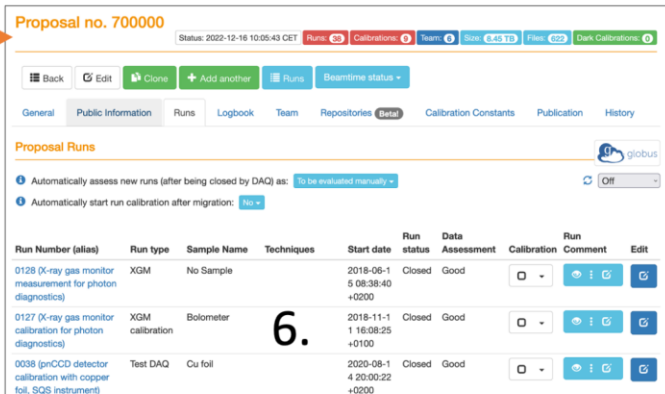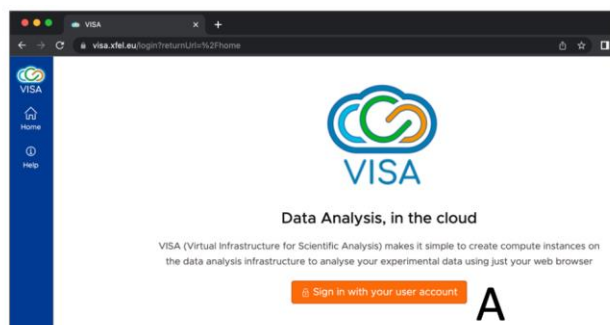

Figure 7. *Demonstration of the EuXFEL workflow, on data search portal and catalogue (myMDC) level, by screenshots. For guest users the workflow currently ends at step 5a.*

Figure 8. *Login to VISA (step 9 of the workflow) via authentication to the Helmholtz AAI.*
The login button (A) directs a user to the Keycloak authentication service, where the Helmholtz AAI is selected (B). Currently GitHub is an option for semi-anonymous identity provision (C). After external login to GiHub and confirmation by e-mail (D, E, not shown), the authentication information must be accepted ("Allow" button, F).
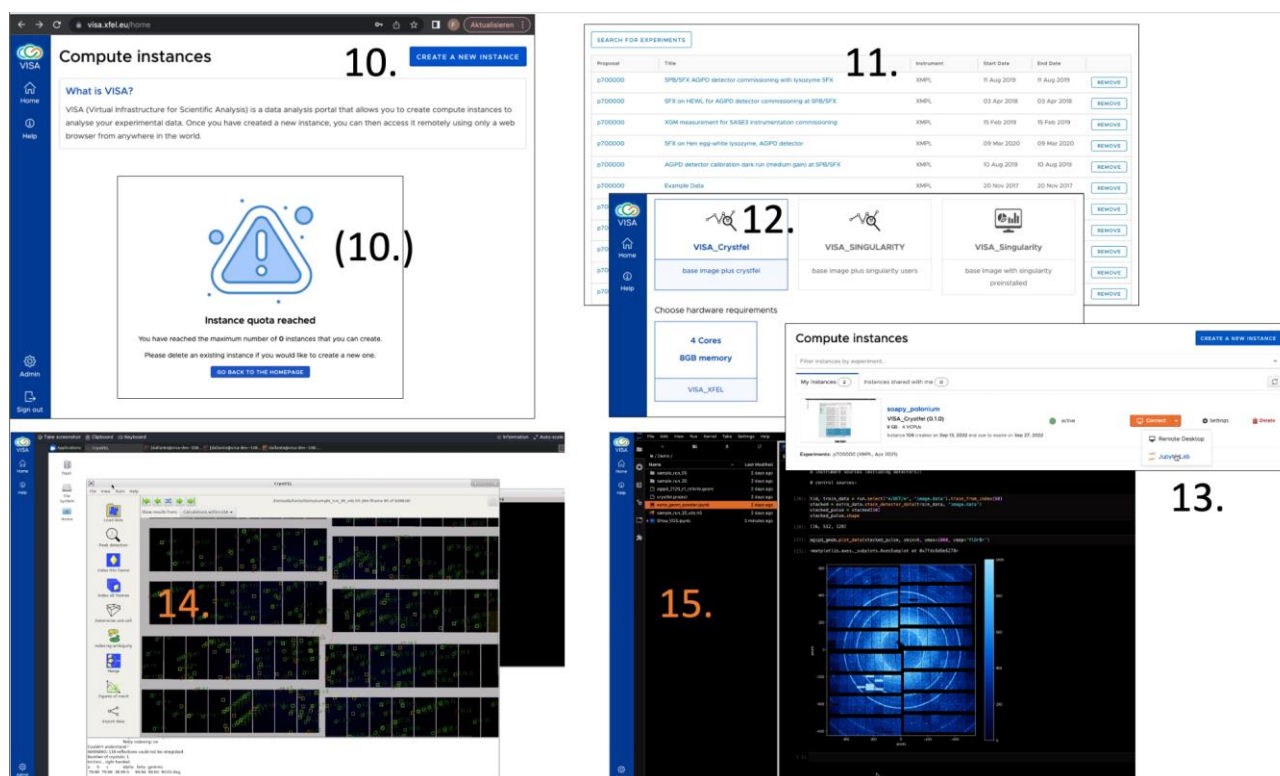
Figure 9. *Use of VISA as service for remote data analysis, steps 10 to 15 of the workflow. The instance creation request (10.) will show a message for guest users which requires contacting the system administrator for quota increase, which means authorisation to use the cloud-cluster infrastructure for the subsequent steps.*

Conclusion and Outlook

The next step for the demonstrator scenarios lies in the implementation of the full concept into the respective production services. Such activities fall into the future after the end of the PaNOSC project and must be catered for by regular RI IT activities or projects funded by future community grants. Ideally all of the services developed in the PaNOSC scope should allow for access to open data, and usage by PaN community members, not necessarily belonging to the registered facility user groups. This poses challenges with respect to limitations of resources like hardware infrastructure and security aspects like prevention of misuse.

Given the current budgets facilities cannot easily sustain open access to compute resources at a scale that would be expected for community-wide open data re-use. Demonstrator setups as well as the catering of FAIR services to the own users of facilities can however spark the interest and even demand of such services on a broader basis, which will in turn create user numbers that convince funding agencies to support investments in more hardware and manpower to maintain the technology run on such hardware.

Finally, for a truly federated approach of data storage, exposition and compute resources, one could imagine centralised services hosted by a dedicated organisation which receives direct funding from commissions like EOSC. Of course, a flawless interplay between such a data hub and the data-producing facilities requires optimised standards and workflows. On the route to establish such, the PaNOSC project has delivered valuable insights and outcomes that can be built upon.