Milestone Report

EuPRAXIA PP
Grant Number 101079773
MILESTONE: WP7-MLS7.1

EuPRAXIA-PP

16/04/2024

DocID		***Ver.***	***Status***
EuPRAXIA PP_WP7_MLS7.1_16.04.24		1.0	Final
Project	***Milestone***	***Grant N.***	
EuPRAXIA PP	WP7-MLS7.1	101079773	
Lead beneficiary:		***Due date:***	
IST		Feb 2024	

Benchmark within existing facilities on e-needs and data policy

Abstract

The following report presents a benchmark within existing facilities on e-needs and data policy, as per milestone M7.1 of the EuPRAXIA-PP project. The work-package group surveyed the landscape of research facilities that operate comparable sources and/or perform similar activities to the future EuPRAXIA sites. For the most part, the data policies in place were found to address open-data and FAIR principles issues and the PaNOSC data policy framework to establish an excellent starting point for the future EuPRAXIA data policy. Regarding e-needs, the tremendous data rates and volumes involved in the operation of these facilities require some of the most demanding data management infrastructures available, with instrument data rates of up to 10 GB/s, and yearly data volumes over 10 PB. These present a significant challenge not only for data acquisition but also for analysis and long-term storage and curation. While some commonalities between the multiple facilities can be found, there are very diverse solutions implemented, particularly on how to implement the open-data policies.

	Name	Institute	Date
Authored by	R. Fonseca S. Pioli	IST INFN	29/03/2024
Approved by WP Coordinator	R. Fonseca	IST	02/04/2024
Reviewed by Project Office	Falone - A. Ghigo -M. Ferrario - P. Campana - C. Pelliccione	INFN	12/04/2024
Approved by Project Coordinator	P. Campana	INFN	16/04/2024

The EuPRAXIA Preparatory Phase (EuPRAXIA-PP) has received funding from the European Union´s Horizon Europe research and innovation programme under grant agreement No. 101079773.

The information contained in this document reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

Index

List of Revisions **4**

List of Annexes **4**

1 Executive Summary **5**

2 Benchmark within existing facilities on e-needs and data policy **8**

3 Data policy benchmark **9**

3.1 PaNOSC FAIR Research Data Policy Framework 10

3.1.1 Key points *11*

3.2 Data Policies of Existing Facilities 14

3.2.1 European Synchrotron Radiation Facility *15*

3.2.2 Central European Research Infrastructure Consortium *16*

3.2.3 Elettra-Sincrotrone Trieste *17*

3.2.4 European Spallation Source *17*

3.2.5 Institut Max Von Laue - Paul Langevin *18*

3.2.6 SOLEIL Synchrotron *19*

3.2.7 European XFEL *20*

3.2.8 ISIS Neutron and Muon Source *21*

3.2.9 Diamond Light Source *22*

3.2.10 Paul Scherrer Institute *23*

3.2.11 Extreme Light Infrastructure *24*

3.2.12 Advanced Photon Source (US) *26*

3.2.13 Linac Coherent Light Source *26*

3.3 Data management plans 27

3.4 Overview 28

3.5 Conclusions 30

4 E-needs benchmark **31**

4.1 Data volumes and rates 32

4.1.1 Data reduction 33

4.2 Storage systems 34

4.3 Data Monitoring and Analysis 36

4.3.1 e-logbooks 37

4.4 Connection to HPC/HTC resources 38

4.5 Integrated systems control and data acquisition 38

4.6 Conclusions 39

5 References........ 41

List of Revisions

rev.	date	modification
0.1	29/03/2024	First Release
0.2	02/04/2024	Approved by WP coordinator
0.3	12/04/2024	Reviewed by Project office
0.4	16/04/2024	Reviewed by Project Coordinator
1.0	16/04/2024	Final

List of Annexes

rev.	date	document title	note
0.1	29/03/2024	List of Acronyms	
0.1	29/03/2024	Creative Commons Licenses	

1 Executive Summary

The following report presents a benchmark within existing facilities on e-needs and data policy, as per milestone M7.1 of the EuPRAXIA-PP project. It focuses on operating facilities offering the same type of sources and/or performing the same type of activities as EuPRAXIA is expected to offer in the operation phase, which is expected to start in 2028. A total of 13 facilities were considered for this benchmark, as well as the outcomes of the PaNOSC project. Although the two items under consideration, namely e-needs and data policy, are closely related, the former has a more technical nature focusing on the actual data management infrastructure, while the latter establishes the overall policy procedures regarding this data, in particular in what regards open-data and FAIR policies. As such, we decided to address these two benchmarks separately, while highlighting the connections between the two whenever relevant.

Regarding the ***data policy*** *benchmark our report finds that:*

- *The PaNOSC data policy framework covers in detail the data policy requirements for EuPRAXIA and similar facilities, in particular concerning open data and FAIR principles:*
  - *It makes data openly available after an embargo period, using an "appropriate license";*
  - *It establishes the facility as the curator of the data;*
  - *It proposes that the facility publishes the metadata in an online catalogue that is searchable and that will link the metadata and the raw data; data is to be preserved for a minimum of 10 years;*
  - *This is to be done in "well-defined" formats, with the facility making the tools required for reading the data publicly available;*
  - *It makes users responsible for ensuring that the data generated from the experiments includes accurate contextual information (metadata) such that it fulfils FAIR principles.*
- *The data policies of the facilities benchmarked generally follow these principles. The key points found were:*
  - *Data policy acceptance is a condition of facility access;*
  - *The main difference from the framework relates to the data retention period, with many sites committing only to 5 years, although striving for longer archival on a "best-effort" basis;*
  - *Different sites treat different data types (raw data, metadata, processed data, etc.) differently. As a rule, only raw data and associated metadata are preserved;*

- *Licenses used for raw data vary from several Creative Commons licenses: CC-BY, CC-BY-NC, and CC-BY-SA. One case distinguishes the case of "high-level metadata" which is to be distributed under a CC0 license (See the annex on Creative Commons Licenses for a brief discussion);*
- *Access to data is generally restricted to registered users, in some cases with data downloads being logged and the original PI having access to the information on who downloaded the data. Registration is generally open, but there are some legal caveats;*
- *Some facilities identify NEXUS/HDF5 as the preferred data format, others use proprietary (but open) formats;*
- *The data policies benchmarked do not clarify what data quality procedures are to be implemented and generally limit themselves to specifying that it is the responsibility of the users to ensure that the data generated from the experiments includes accurate contextual information;*
- *Only one of the surveyed documents sets a calendar for reviewing the policy;*
- *Only one of the surveyed documents refers to the "European Code of Conduct for Research Integrity ALLEA";*
- *The majority of the data policy documents reviewed do not refer to the user data management plans;*
- *The 2 facilities considered in the US do not publish data policy documents in a similar way to their European counterparts. The focus is mostly on data retention and the facilities do not implement (that we could find) open data / FAIR policies.*

As for the ***e-needs*** *benchmark, it was found that:*

- *The current data rate per instrument is in the range of 1 - 10 GB/s, leading to data rates on the order of 10 - 20 GB/s for a complete experimental setup, in bursts of varying duration, typically on the order of 30 minutes;*
- *Considering typical operation values the total data volume per year will be on the order of 5 - 10 PB. Note that the PaNOSC recommendation is to store the experimental data for 10 years, which would lead to total storage requirements on the order of 50 - 100 PB;*
- *Data reduction techniques can be implemented to lower these requirements by as much as an order of magnitude; however great care must be taken not to jeopardize the scientific content;*

- *The storage system to be implemented will need to cope with both the very large data rates and data volumes. The solutions implemented usually rely on a layered approach of storage systems of increasing capacity/decreasing performance;*
- *Tape storage is still the main choice for long-term data archiving, should EuPRAXIA decide to implement this on-site. However, we are considering moving the data to an external data repository, yet to be identified;*
- *The data management systems implement two categories of data processing, one for (next to) real-time data processing, and one for offline data processing. The former should provide feedback on the order of seconds on data that is currently being acquired, while the latter is to be applied to saved data;*
- *Electronic logbooks are implemented in several facilities, as a way of documenting experimental protocol and notes, and in connection with the data policies in each facility;*
- *Data processing requires considerable computational power. Most facilities have an on-site computing cluster for offline data processing, complementing specific hardware deployed for real-time data processing. Estimated requirements are on the order of 1 - 10 PFLOPS;*
- *To lower this requirement EuPRAXIA sites may collaborate with external supercomputing facilities. However, this has significant network necesities and needs extensive workflow automation work to be of practical use;*
- *Most systems use some form of integrated systems control and data acquisition. EPICS and Tango controls are the two most popular choices.*

The work package will continue to monitor the developments in the community regarding these topics and update the document with any new relevant information that may become available. As other components of the EuPRAXIA project mature this document may also be extended to focus on e-needs that are then identified.

2 Benchmark within existing facilities on e-needs and data policy

This document was prepared by Work-package 7 on E-Needs and Data Policy, as per milestone M7.1 of the EuPRAXIA-PP project. It reports on the benchmark activities focusing on the state-of-the-art "e-needs" and data policy of existing facilities offering the same type of sources and/or performing the same type of activities as EuPRAXIA is expected to offer in the operation phase, which is expected to start in 2028. The benchmark also targets community efforts in the field focus on data and leverages the previous work done by this work package on the preparation of the initial version of the Data Management Plan (DMP) for the operation phase.

The report is organized into 2 main sections, the first focusing on data policy, and the second on e-needs. The data policy for the EuPRAXIA operation phase will need to reflect its open-data nature, as well as its adoption of FAIR principles, and our benchmark attempts to focus on the data policy choices that were made by the community contributing to these issues. We investigate a community effort on this topic, namely the work done in the context of the PaNOSC project, and use it to frame the analysis of existing facilities. Our choice of benchmark facilities concentrated when possible on facilities also implementing open-data and FAIR policies, but it also included facilities in the United States, where the adoption of these principles is not yet at the same level as their European counterparts. We additionally attempted to include distributed facilities as this will also be the case for EuPRAXIA. Regarding the e-needs, the choice was made to target mostly on the facilities chosen for the data policy benchmark. In this case, however, we chose not to organize the benchmark by facility, but rather by topic, given that published information regarding the detailed e-needs of these facilities is somewhat scarce and scattered, which also reflects the very large diversity and very rapid evolution of data-related technology. This structure will be of greater contribution to laying the groundwork for the required data management infrastructure of EuPRAXIA sites.

Both topics are under active and rapid development in the community, and the work package will need to continue to monitor the evolution of them throughout the EuPRAXIA-PP project and revise this report accordingly, should the need arise. Additionally, as other components of the EuPRAXIA project mature, raising specific technical and policy needs not yet covered by this document, the working group will further extend this benchmark to cover them, in preparation for the final deliverable of the work package, "D7.3 - Final report e-needs and Data Policy".

3 Data policy benchmark

Open Science is defined by the “UNESCO Recommendation on Open Science” [1] as “making science more accessible, inclusive and equitable for the benefit of all”. The European Commission (EC) has adopted an Open Science policy for all EC-funded projects, and EuPRAXIA has implemented these principles since the beginning. To this end, the European Commission (EC) has funded the European Open Science Cloud Association (EOSCA) to promote and facilitate Open Science, and in particular, Open Data in scientific research. This is made clear by the European Data Strategy [2] which recognizes that EOSC has “the basis for a science, research, and innovation data space”. This is further supported on the report by the European Commission Expert Group on FAIR Data [3] which outlines guiding principles to make data Findable, Accessible, Interoperable, and Reusable (FAIR) [4], in addition to the directive 2019/1024 of the European Parliament and of the Council, dated 20 June 2019, both focusing on open data and the re-use of public sector information [5], highlighting the key importance in the European landscape.

The data policy for many facilities similar to EuPRAXIA in Europe, in particular photon sources, already reflects these principles. In Europe, these have long collaborated in what regards data strategy, particularly under the framework of LEAPS, the “League of European Accelerator-based Photon Sources”. This consortium, with the primary goal of ensuring and promoting the quality and impact of the research done at these facilities, published recently a paper [6] outlining the overall strategy for LEAPS in this matter. This document highlights several objectives that LEAPS facilities management should endorse and adopt to deal with the very large amounts of data being produced, while ensuring that the best use is made of this data. In this sense, the adhering to the principles of Open Science and in particular the adoption of FAIR principles for data is strongly recommended.

The LEAP communities (and the sister communities focusing on neutron sources, LENS) have contributed to the implementation of the EOSC open science platforms through two flagship projects, ExPANDS [7] and PaNOSC [8]. The European Open Science Cloud (EOSC) Photon and Neutron Data Service (ExPaNDS) project is a multi-national collaboration of national Photon and Neutrons research infrastructures and the European Grid Infrastructure (EGI), to deliver standardized, interoperable, and integrated data sources and data analysis services for Photon and Neutron facilities. The Photon and Neutron Open Science Cloud (PaNOSC) project is the science cluster representing Photon and Neutron European Research Infrastructures and connecting these to the European Open Science Cloud (EOSC). Together, these projects have accelerated the implementation of Open Science practices in the community. The outcomes of these projects have contributed to many recommendations related to data, in particular with data policy recommendations for these facilities.

The PaNOSC project published the “PaNOSC FAIR Research Data Policy framework” [9] in 2020, as deliverable D2.1 . The document is based on previous work done in the context of the FP7 PaN-data project [10], the “Principles and guidelines for access to research data from public funding” [11], and on the above-mentioned report by the European Commission Expert Group on FAIR Data. It proposes a common data policy framework for the photon and neutron community, which reinforces the role of FAIR principles in this context. It also presents relevant implementation notes, regarding many key issues such as the curation period, data licenses to adopt, or persistent identifiers.

This framework has seen widespread adoption by the community and serves as the basis for most of the published data policies of facilities similar to EuPRAXIA in Europe. As such, we chose this framework as a baseline to which the specific data policies of the multiple sites we analysed are compared to, and we begin by a detailed analysis of this framework.

3.1 PaNOSC FAIR Research Data Policy Framework

The PaNOSC FAIR Research Data Policy Framework [9] describes a common framework to be adopted for the management of scientific data by photon and neutron facilities under open access and FAIR policies. In particular, the framework aims at making previously measured data available for further analysis without the need to replicate the experiment, promoting data user and cross-disciplinary research, while improving scientific integrity and reproducibility of experiments by making raw data open to scrutiny by other researchers. The framework includes an analysis of its compliance to FAIR principles, done according to the Research Data Alliance FAIR Data maturity model [12].

The acceptance of the facility data policy is a condition for access; no beam time will be awarded without it. Additionally, the data policy framework refers to data management plans (DMP), defining all aspects of data management. The framework does not make it clear if supplying data management plans at the time of beam time application is mandatory. However, it does state that users are required to follow any recommendation provided by the facility regarding data management in general and DMPs in particular.

Overall, the policy framework makes a distinction between public research, that is research done through some form of open peer-review access, and research conducted through purchased (commercial) beam-time. The former is expected to be published in peer-reviewed scientific journals and openly accessible, while the results from the latter are under no obligation to do so.

Most of the responsibilities defined in the framework fall onto the facility. The key point is that the facility will apply "all reasonable efforts" to ensure accurate storing and curation of data, as well as access to it. Data is to be stored by the facility, which will act as a custodian for the data, for a minimum duration of 10 years, while metadata is expected to be stored forever. Users are responsible for ensuring that the data generated from the experiments includes accurate contextual information (metadata) such that it fulfils FAIR principles, and are encouraged to add additional data that will further augment the use of the experimental data.

The policy framework allows for an embargo period, starting from the end of the experimental session before data is made openly available. During this period access to the data is restricted to the experimental team. The facility and its staff may also access the data during this period for user support and to improve facility processes and performance.

Data will be curated in "well-defined" formats, and the means for reading the data will be made publicly available by the facility. While the policy does not define which format should be used, it does recommend the use of the NeXus [13] / HDF5 [14] format which has been adopted by a majority of photon and neutron sources. The facility will generate persistent identifiers for the data so that it can be referenced in any publication. Metadata will be made available by the facility in an online catalogue that will link the metadata, data, and reports. The framework also makes clear that the facility cannot be made liable in case data cannot be accessed or gets lost.

The proposed policy does not discuss in detail data quality assurance procedures. For the two main quality dimensions considered in EuPRAXIA data, namely fitness for use (data reliability and useability) and fitness for purpose (data includes information that allows end users to self-assess its fitness to their requirements), the policy only assigns the responsibility of the former to the facility, as the custodian of the data, and of the later to experimental teams, as it is up to them to ensure metadata quality.

3.1.1 Key points

3.1.1.1 Data

The framework defines two main types of data that are collected in the context of the experimental activity:

1. Data being generated by the experimental devices at the facility, referred to as **raw data**;
2. Additional information linked to the raw data that provides contextual details, referred to as **metadata**.

The framework also considers data that is collected/generated outside the context of the actual experimental procedure, namely:

3. Data providing additional contextual information regarding the experiment and datasets, such as information about samples used, or data processing scripts, referred to as **auxiliary data**.
4. Data that is obtained by processing raw data from the experiments, usually automatically at the facility, referred to as **processed data**.

Additionally, the framework defines **results** as data and other outcomes arising from the analysis of raw data, not including publications.

As a rule, raw data and associated metadata from public research will be made open access after the embargo. This data will be curated in “well-defined” formats, and the facility will make available the required tools for reading it when needed. Raw data should be stored for 10 years, and metadata should be stored permanently.

An “appropriate license” for open data will be used. Although not strictly specified in the policy, the framework recommends the adoption of one of the licenses from the Creative Commons family (e.g. CC-BY, CC-BY-NC, or CC0) [15]. Note that the CC-BY licenses oblige the users of the data to cite provenance appropriately, meaning that the data will be open access but not in the public domain, as would be the case of the CC0 license (See the annex on Creative Commons licenses for a brief discussion).

Processed data and auxiliary data, when available, will also follow the same rules, except that the facility only guarantees readability for facility-generated processed data. The facility will provide a means for the users to upload results and link them to the relevant raw data, but it will not be responsible for curating them (e.g. providing the required tools to read them).

Each dataset will be assigned a unique identifier that should be referenced when publishing results based on open-access data, usually in the form of a digital object identifier (DOI) [16].

3.1.1.2 Actors and responsabilities

The framework considers the following actors in terms of data policy, assigning them specific responsabilities:

1. The facility where the experiment is conducted, and its staff;
2. The experimental team, which is composed of a principal investigator (PI) and additional users;
3. Researchers using open data, who were not involved in the original experiment.

The facility assumes the role of custodian of the data, being responsible for storing and curating the data. This will be done on a "best efforts" basis, and the facility will not be liable in case of loss of data, temporary unavailability, or the inexistence of tools for reading it. The facility also ensures GDPR [17] compliance.

The experimental team is responsible for ensuring that raw and processed data are collected with accurate metadata to fulfil FAIR principles to a minimum defined by the facility, as well as to add auxiliary data. The PI can create/distribute copies of the raw data and request extensions/reductions to the embargo period.

The researchers using open data, while under no formal obligation to do so, are encouraged to contact the PI of the original experiment to inform them. The results from this research, as well as the methods used to obtain them, should also be made openly accessible whenever possible.

3.1.1.3 Good practices

The framework also identifies a set of additional good practices for both the facility and the experimental team, focusing on enabling FAIR principles to the fullest. In particular, the framework urges the experimental teams to make the metadata as complete as possible, as it will enhance the possibilities for data reuse (easier to search, retrieve, and interpret). The support this, the facility should further provide means for the capture of metadata items that are not automatically collected by instruments.

To improve reproducibility, the experimental teams should also provide a complete log of the protocol used, entered in the electronic logbook, if the facility provides one. All publications using facility data should follow best practices in citing any software tools used or developed for data analysis. Furthermore, authors are strongly encouraged to make all the analysis procedures and software tools available, so that others may reproduce the analysis from the raw data. These files would be deposited as auxiliary data associated with the original data.

3.2 Data Policies of Existing Facilities

Given EuPRAXIA's goal of serving users in ultra-fast science, (e.g. on high-resolution medical imaging, deeply penetrating positron annihilation spectroscopy for materials, and offering transformative capabilities for research on biomolecules, viruses, and microscopic processes) our choices for benchmarking data policies fell upon other facilities offering radiation sources for these types of applications, ranging from x-ray and high-energy electrons to high-power laser sources. Additionally, we chose to also include neutron sources, as they are also used in some of the application fields, and share a common set of data management practices and challenges, as discussed in the previous sections.

Our choice focused when possible on facilities implementing Open Data and FAIR policies, as these principles are at the core of the EuPRAXIA project, and present specific challenges that the data policies should address. For the most part, these principles are observed, at different levels, by European facilities. The same is not true, however, for other international facilities, that we nevertheless chose to include in this benchmark, for a more complete study. Additionally, we also attempted to include distributed facilities, i.e., organizations featuring more than one experimental facility, as this will also be the case for EuPRAXIA.

After careful consideration, we chose 13 different facilities according to the above criteria:

1. European Synchrotron Radiation Facility
2. Central European Research Infrastructure Consortium
3. Elettra-Sincrotrone Trieste
4. European Spallation Source
5. Institut Max Von Laue – Paul Langevin
6. SOLEIL Synchrotron
7. European X-ray free electron laser
8. ISIS Neutron and Muon Source
9. Diamond Light Source
10. Paul Scherrer Institute

11. Extreme Light Infrastructure
12. Advanced Photon Source (US)
13. Linac Coherent Light Source (US)

Should the need arise, further facilities may be considered before the end of the current project and included in the final report of this workgroup. We will also continue to monitor any relevant changes in data policy for these sites and include them in the final report.

3.2.1 European Synchrotron Radiation Facility

The European Synchrotron Radiation Facility is located in Grenoble, France. This research institute is supported by 22 countries. It attracts around 8,000 scientists each year who conduct over 2,000 experiments and produce approximately 1,800 scientific publications. The ESRF is one of the world's brightest synchrotrons, providing scientists with advanced tools to study materials and living matter. It has several beamlines used for various research purposes, such as Multiple-wavelength Anomalous Diffraction, high-resolution macromolecular crystallography, x-ray imaging and tomography techniques, x-ray micro diffraction, and studying materials under extreme conditions (pressure/temperature).

3.2.1.1 ESRF Data Policy

ESRF was one of the earliest sites to publish a data policy, in 2015, focusing on data ownership, curation, and archiving, as well as its open access. This data policy pre-dates the PaNOSC data policy framework and is based largely on the earlier recommendations of the PaN-data Europe Strategic Working Group [10]. However, given that the PaNOSC data policy framework is also based on these recommendations, there is a significant overlap with it.

The ESRF data policy differs in a few points:

1. It states that all data and metadata are subject to the data protection legislation of France, while PaNOSC refers to the EU GDPR [17], and makes GDPR compliance the responsibility of the facility;
2. Long-term storage is set at a minimum of 5 years, although ESRF will strive to store the data for 10 years;

3. The license adopted for the ESRF data archive is CC-BY [15];
4. Only data with metadata generated by ESRF software will be archived;
5. ESRF will not store or curate processed data;
6. Access to the online catalogue of the ESRF will be restricted to registered users, i.e., the data is open to access but subject to previous registration;
7. It does not make any reference to user data management plans.

The ESRF Data Policy has been updated very recently (December 2023) [18], with the main change introduced being the inclusion of the possibility of more extensive data curation, including processed data, triage and the use of lossy compression, or limiting the amount of data stored. The updated version also includes an explicit reference to FAIR principles.

3.2.2 Central European Research Infrastructure Consortium

The Central European Research Infrastructure Consortium, CERIC–ERIC, is a multidisciplinary Research Infrastructure for basic and applied research in all fields of Materials, Biomaterials, and Nanotechnology. It has been built by integrating leading national research facilities based in 8 countries (Austria, Croatia, Czech Republic, Hungary, Italy, Poland, Romania, and Slovenia) into a unique European entity supporting the production of basic knowledge and technology transfer, while promoting the mobility of researchers in an international multicultural scientific environment. As a whole, CERIC-ERIC provides several types of sources, from photons to neutrons and ions, as well as UV laser pulses, together with a multitude of instruments.

3.2.2.1 CERIC-ERIC Data Policy

The CERIC-ERIC data policy was published in 2021 [19] and follows the PaNOSC data policy framework very closely. It implements an open-access data policy and FAIR principles and makes clear in the document that the format chosen by CERIC-ERIC for the raw data is NEXUS/HDF5. Being a distributed infrastructure, this data policy adds to the PaNOSC framework a "field of application", specifying that the policy applies to all CERIC-ERIC Instruments and Partners facilities, highlighting that a common approach to the management of scientific data will greatly contribute to facilitating the work of scientists using more than one facility and to the overall transparency of the scientific process.

This data policy does not identify which license will be used by the data.

3.2.3 Elettra-Sincrotrone Trieste

The Elettra Sincrotrone Trieste is an international research centre located in Basovizza on the outskirts of Trieste, Italy. It is a multidisciplinary research centre of excellence that is open to the international research community. Elettra specializes in generating high-quality synchrotron and free-electron laser radiation. Elettra operates 28 beamlines that enable the most important x-ray-based techniques in the areas of spectroscopy, spectro-microscopy, diffraction, scattering, and lithography.

3.2.3.1 ELETTRA Data Policy

The latest ELETTRA Data Policy, published in 2022 [20], also follows the PaNOSC data policy framework very closely. It implements a scientific data policy following FAIR principles as recommended by the EU and the Research Data Alliance, and updates previous versions, in particular making the acceptance of the policy by the users of the facility a necessary condition for beamtime to be awarded. The document does not expand significantly from the PaNOSC data policy framework but makes clear that the data license to be used is the Creative Commons CC-BY-SA-4.0 license.

Additionally, the ELETTRA Data policy:

- Identifies the NEXUS/HDF5 format as a possible format;
- Assumes that ELETTRA will strive to curate the data for 10 years, with possible restrictions;
- Includes the FAIR Data Maturity Model analysis from the PaNOSC framework.

3.2.4 European Spallation Source

The European Spallation Source (ESS) is a research institute located in Lund, Sweden. The ESS is a collaboration between 13 European member countries, which are partners in its construction and operation. The facility is currently under construction and aims to provide advanced research capabilities in various scientific fields. The ESS is designed to generate neutron beams for scientific experiments and will contribute to advancements in materials science, life sciences, energy, and environmental research. The facility data management is supported by the ESS Data Management and Software Centre, which is situated in Copenhagen, Denmark.

3.2.4.1 ESS Data Policy

The ESS Data Policy, published in 2017 [21], was the result of an investigation on existing policies at that time from ILL, ESRF, ISIS, and Diamond Light Source, also analysed in this document. It applies to non-propriety use of beam-time only and is very similar to the policies available at the time. It also bears many similarities to the PaNOSC framework.

The key points are:

- Acceptance of policy is a condition of facility access;
- ESS will keep two copies of the raw data and metadata in physically different locations for at least 5 years following data collection;
- Raw and metadata will be moved to long-term curation for a minimum 10 years following a 3-year embargo period;
- ESS will act as a custodian of the data;
- The data policy does not consider processed data or auxiliary data;
- All raw data will be given a citeable identifier (in the form of a DOI);
- Post-embargo access to data is open and in line with ERIC statutes;
- The document does not make any reference to user data management plans.

3.2.5 Institut Max Von Laue – Paul Langevin

The Institut Max Von Laue – Paul Langevin (ILL), commonly known as Laue-Langevin, is a research institute located in Grenoble, France. The ILL is an internationally financed scientific facility and is one of the world centres for research using neutrons, providing one of the most intense neutron sources in the world. Every year, ILL attracts about 1400 researchers from over 40 countries, conducting about 1000 experiments selected by a scientific review committee. Research focuses primarily on fundamental science in a variety of fields: condensed matter physics, chemistry, biology, nuclear physics and materials science, etc.

3.2.5.1 ILL Data Policy

The ILL Data policy, published in 2018 [22], follows quite closely the recommendations of the PaN-data Europe Strategic Working Group [10] and, like the ESRF data policy, also has significant overlap with the PaNOSC framework.

The main differences are as follows:

- The license for the scientific data was chosen to be the CC-BY license. The data policy makes clear that, according to this license, users of such data are obliged to cite ILL and the corresponding experiments and scientific teams as the source of the data. As such, while the data is open, it does not reside in the public domain;
- Metadata that is automatically captured by instruments and auxiliary data (which is referred to in the policy document as "Proposal-related metadata") will also be curated. Other electronic metadata will not be curated in general;
- Access to open data is restricted to registered users of the "ILL Data Portal". Downloading of open data will be logged and the information made available to the PI, including the identity of the downloader;
- The policy further recommends that preprints, postprints, and reprints of publications related to ILL experiments should be submitted to the library staff;
- The policy does not make any reference to user data management plans.

3.2.6 SOLEIL Synchrotron

The SOLEIL Synchrotron is the French national synchrotron facility and is located near Paris, France. It is operated by the French National Centre for Scientific Research (CNRS) and the French Alternative Energies and Atomic Energy Commission (CEA). SOLEIL functions as an electromagnetic radiation source that spans a broad spectrum of energies, from infrared to x-rays. Additionally, it operates as a state-of-the-art research laboratory, specializing in advanced experimental methods for analysing matter at the atomic level. It also serves as a collaborative platform accessible to scientific and industrial communities.

3.2.6.1 SOLEIL Data policy

The SOLEIL Data policy was published in 2018 [23] and a summary was later published as a technical report in 2019 [24]. As with other older data policies, is also based on the model developed in the context of the European FP7 PaN-data Europe project. The model was adapted to SOLEIL considering the input from science data experts from CNRS and feedback from SOLEIL technical (IT) and scientific staff.

The data policy presents a slightly different structure from the reference PaNOSC framework but in essence, covers the same topics. The key points are as follows:

- Acceptance of the policy is a condition for beamtime allocation;
- The data policy requires that proposals include a data management plan;
- SOLEIL will be the custodian of raw data and associated metadata. Data will be stored for 5 years with SOLEIL striving for 10 years. As a rule, processed data, reduced data, and associated metadata will not be curated;
- Metadata is to be collected automatically by SOLEIL instruments, with users being asked to fill in auxiliary data (which in this policy also falls in the metadata category). This metadata will be available in a catalogue, which can be accessed online to browse and download data and metadata;
- Data generated at the SOLEIL beamlines will have a unique persistent identifier (DOI) for citing the results;
- After a 3-year to 5-year embargo period, data will be released under a CC-BY license with open access to anyone registered on the SOLEIL dedicated portal;
- While user data management plans are mentioned in the policy in reference to external institutions, they are not required by the policy.

3.2.7 European XFEL

The European X-ray free electron laser (European XFEL) is an x-ray research laser facility. It is an international project located in Schenefeld, Germany, with 12 participating countries; 9 shareholders (Denmark, France, Germany, Hungary, Poland, Russia, Slovakia, Sweden, and Switzerland) and 3 other partners (Italy, Spain, and the United Kingdom). The European XFEL is based on an underground 3.4 km long linear accelerator and produces the x-ray laser light from accelerated electrons using undulators. Science performed at the European XFEL spans many disciplines, from molecular biology to probing the characteristics of extreme states of matter and observing the behaviour of electrons within complex molecules.

3.2.7.1 European XFEL Data Policy

The “Scientific Data Policy of the European X-Ray Free-Electron Laser Facility GmbH” (Version 2) was approved in 2023 [25] and will enter into effect in 2025. Like other data policies analysed, this data policy follows the recommendations from the PaN-data Europe working group and the updated PaNOSC data policy framework. The structure of the data policy is very similar to the latter, but makes some important additions:

- It defines auxiliary data as "data that provides contextual information regarding the experiment.", and metadata as a subset of this;
- It adds "reduced data" to the types of data considered, meaning data was specifically identified from the raw data;
- Adds a reference to "quality of data services", meaning a strategy for storing and accessing scientific data according to their categories and usage purpose, detailed in a separate document;
- It distinguishes the types of licenses used depending on the data type. CC-BY will be used for all data (as many other facilities) except for "high-level metadata", where CC0 will be used instead. The choice of the CC0 for the latter is meant to enable automated harvesting of machine-accessible metadata by third parties;
- The policy states that processed data and the associated metadata hereto will not be curated for the long term; however, reduced data and the associated metadata will be curated up to a given volume. Additional auxiliary data will also be curated, again up to a given volume;
- The document goes into much detail regarding the warranty and liability regarding scientific data, essentially declaring the facility not to be liable in case of any issues.

3.2.8 ISIS Neutron and Muon Source

The ISIS Neutron and Muon Source is located at the Rutherford Appleton Laboratory in the United Kingdom. It produces beams of neutrons and muons that allow scientists to study materials at the atomic level. Each year the facility is used by up to 3000 scientists from over 30 countries, running 1200 different experiments. It supports a national and international community of scientists who use neutrons and muons for research in physics, chemistry, materials science, geology, engineering, and biology.

3.2.8.1 ISIS Data policy

The latest ISIS data management policy was published in April 2023 [26], describing how ISIS handles curation, access, ownership, usage, and storage of data collected at the facility. ISIS is committed to the principles of FAIR data, and, unsurprisingly, the overall data policy also bears significant similarities with the PaNOSC data policy framework and the ESS data policy. The key additions are as follows:

- Adds "facility generated reduced data", meaning raw data that has been automatically processed by the facility, to the types of data considered;
- Beyond a 3-year embargo period, raw data and associated metadata will be accessible for a minimum of 10 years under a CC-BY license by users registered at the ISIS computing infrastructure;
- All publications related to experiments carried out at ISIS are obliged to acknowledge ISIS and cite data DOIs of the experiments;
- The document sets a yearly calendar for reviewing the data policy, to "keep it as relevant and accurate as possible";
- The document does not reference user data management plans.

3.2.9 Diamond Light Source

The Diamond Light Source (DLS) is a research facility located at the Harwell Science and Innovation Campus in Oxfordshire, United Kingdom. It is the UK's national synchrotron light source science facility. The Diamond Light Source operates with an energy of 3 GeV and currently has 32 beamlines focusing on crystallography, microfocus spectroscopy, small angle x-ray scattering and diffraction, and hard x-ray probing, among other scientific and technical applications.

3.2.9.1 DLS Data Policy

The Diamond Experimental Data Policy, published in 2018 [27], is one of the few data policies reviewed that does not follow either the PaNOSC or the PaN-Data recommendations, and that does not reference FAIR principles, although focusing on making experimental data available through open access. It is a very concise policy document and provides an interesting alternative to the more "mainstream" approach followed by other facilities. The key points from the policy are as follows:

- Ensure transparency on the way experimental data produced at DLS is managed;
- The document focuses on "experimental data", defined as user-generated data and all associated metadata;
- Data will be stored at the facility for a minimum of 30 days following creation. Afterward, a single archive copy is created for long-term access;
- Following a 3-year embargo period, data will be made available under a CC-BY license;
- Users are required to acknowledge the source of data on any publications and publish on an open-access basis.

The document does not refer to any online catalogue, obligations in terms of metadata production, or mention of data identifiers being generated at the facility. The data policy does not reference any user data management plan.

3.2.10 Paul Scherrer Institute

The Paul Scherrer Institute (PSI) is a multi-disciplinary research institute for natural and engineering sciences located in Würenlingen, Switzerland. Every year, more than 2500 scientists from Switzerland and around the world come to PSI to use their unique facilities to carry out experiments. PSI operates large scientific research facilities, such as the Swiss Light Source SLS, the free-electron x-ray laser SwissFEL, the SINQ neutron source, the SμS muon source, and the Swiss research infrastructure for particle physics CHRISP, which offer out-of-the-ordinary insights into the processes taking place in the interior of different substances and materials.

3.2.10.1 PSI Data Policy

The PSI Data Policy applies to the external users of the PSI large-scale facilities, namely SLS, SwissFEL, SINQ, SμS, and CHRISP. The latest version was published in 2022 [28], which is an evolution from the original document published in 2016, and focuses on the ownership, curation, and access to research data and metadata produced at PSI or its facilities. The original document predates the PaNOSC recommendations, but the policy does strive to address open-data and FAIR principles, making it clear that the data will be published using a CC-BY-SA license.

The key points from the policy are as follows:

- The principal investigator is, by default, the owner of the data;
- PSI encourages users to include research data management (RDM) in the planning of their activity, and to define data management plans;
- Data activities must be per the "PSI instruction on Research Integrity";
- The policy separates between raw data and research data, the latter being defined as "the evidence that underpins the answer to the research question". It is the responsibility of the PI to make the distinction between the two, as these may depend on the specific research topic;
- It is the responsibility of the PI to ensure that all relevant research data and metadata are recorded and stored in an appropriate repository; PSI reserves the right to apply data reduction on all research data;
- The use of "Electronic Lab Journals" is recommended;
- Aims at research data retention for 10 years.

Finally, it should be noted that this was the only data policy surveyed to explicitly discuss the costs associated with data curation and storage. The former is to be assumed by PSI, while the latter is to be covered by the “data steward” which is to be defined when beam time is awarded.

3.2.11 Extreme Light Infrastructure

The Extreme Light Infrastructure (ELI ERIC) is one of the world’s largest and most advanced high-power laser infrastructures. The organization consists of three complementary facilities: ELI Beamlines, in Dolní Břežany, Czech Republic, ELI Alps, in Szeged, Hungary, and ELI Nuclear Physics (ELI NP), in Măgurele, Romania, and features many collaborations with universities and research labs across the world. ELI-ERIC operates several high-power, high-repetition-rate laser systems that enable the research of physical, chemical, materials, and medical sciences.

3.2.11.1 ELI-ERIC Data Policy

The ELI-ERIC data policy, published in 2022 [29], focuses on the data produced from experiments conducted at any of the ELI facilities. It is another example of a data policy document that does not follow the PaNOSC framework policy. It does, however, adhere to FAIR principles, and considers these to be of critical importance. The key points from the policy are as follows:

- ELI ERIC aims to preserve and manage data according to FAIR principles;
- A data management plan is required from users applying to do experiments;
- All data will receive a unique and persistent identifier, that can be used to reference the original data;
- A rich metadata format will be used and associated with the data, providing detailed provenance information and meeting domain-relevant community standards;
- Metadata will be registered/indexed into an online catalogue, making the data and metadata easily searchable and discoverable by other users;
- ELI ERIC will be the custodian of the data and will have a global scientific data management plan covering all data lifecycle.

The policy also states a set of objectives that ELI ERIC aims to fulfil, in particular:

- To develop tools for FAIR-by-design metadata collection and storage;
- To Preserve data for a minimum of 10 years;
- To Promote data usage by other scientists following an embargo period.

Finally, the document states that ELI-ERIC shall adopt a "Code of Conduct for Research Integrity and Ethics" based on “The European Code of Conduct for Research Integrity ALLEA (All European Academies)” [30].

3.2.12 Advanced Photon Source (US)

The Advanced Photon Source (APS) is a storage-ring-based high-energy x-ray light source facility located at Argonne National Laboratory in DuPage County, Illinois, United States of America. The APS attracts more than 5500 researchers each year to study a wide range of materials and phenomena using high-energy x-ray beams. It plays a crucial role in various fields, including materials science, chemistry, biology, and physics.

3.2.12.1 APS Data policies

The Advance Photon Source does not make public any data policy or publish any guidelines referring to open data and FAIR principles. To our knowledge, the only published data policies refers to Data Management and Retrieval Practices [31] and dates from 2019. The key points are:

- The facility cannot guarantee indefinite data archival;
- It is the responsibility of the users to meet the data management requirements of their home institutions and/or funding agencies;
- Long-term data retention and management is the responsibility of the users;
- Data ownership is set by the user agreement between the user and the facility.

3.2.13 Linac Coherent Light Source

The Linac Coherent Light Source (LCLS) was the first hard x-ray free electron laser and is located in San Mateo County, California, in the US. It is based on the SLAC linac, that provides high-current, low-emittance 5 – 15 GeV electron bunches at a 120 Hz repetition rate, and long undulator to bunch the electrons, leading to self-amplification of the emitted x-ray radiation, constituting the x-ray FEL. The SLAC linac also powers the FACET-II facility, that is used to develop advanced acceleration and coherent radiation techniques with high-energy electron and positron beams, and in particular, beam-plasma acceleration.

3.2.13.1 LCLS Data policies

Much like the APS facility, the LCLS does not make public any data policy or publish any guidelines referring to open data and FAIR principles. To our knowledge, the only published data policy refers to Data retention policy [32], updated in 2022. The key points are the same as for the APS facility:

- The facility cannot guarantee indefinite data archival;
- It is the responsibility of the users to meet the data management requirements of their home institutions and/or funding agencies;
- Long-term data retention and management is the responsibility of the users;
- Data ownership is set by the user agreement between the user and the facility.

LCLS does however provide an infrastructure for storing experiment data and metadata for a period of ten years. The facility will not be responsible for curating the data or set any kind of online catalogue for accessing it.

3.3 Data management plans

As mentioned above, the PaNOSC policy framework, as well as many of the benchmark sites, foresee the need for data management plans (DMP) to be prepared by each applicant team. This document will define the (experimental) data lifecycle, and ensure the implementation of any data policies defined by the facility, providing a better overview of data management in the facility during their project.

One of the tasks of the PaNOSC project regarding FAIR data was specifically to develop a template for the user DMPs. As a matter of fact, the PaNOSC and ExPaNDS projects developed a DMP template together [33], and the projects are currently working on implementing it for the research infrastructures involved [34].

The proposed DMPs should be aligned with the complete project workflow, starting with the proposal, which implies (for the user) to consider the data management before the experiment. The template proposes the so-called "active DMPs", that is, DMPs that are continuously updated, in particular with information from the facility systems. Also, since for many facilities funder specific DMPs are required, the work focused on establishing a common knowledge model (KM), that can then later be translated to facility-specific questionnaires. The KM should be lightweight and, where possible, generated automatically from the experiment proposal.

The paper has specific recommendations on the KM and on the tool to be used to help in developing the DMPs. The choice fell on the Data Stewardship Wizard [35], which can be customized to allow the specific KM required, as well as to include additional questions for each specific facility. It also allows for the majority of questions in the DMP to be answered by the facility, greatly simplifying the work of the users. This DMP can also be exported in a version that can be submitted to funders.

It should be noted however that the majority of the facilities in the PaNOSC project do not see the DMP as a mandatory step in the project workflow, and some of the facilities that do consider the DMP mandatory use prefilled DMPs and do not involve the users in the preparation of the DMPs. However, this may change in the future as funding agencies, both supporting the facilities and the users, are starting to demand DMPs for support.

3.4 Overview

The data policy to be implemented for future EuPRAXIA-generated data will implement open data policies following FAIR principles, following the UNESCO and EU recommendations, and the open science policy adopted for all EC-funded projects. Several other facilities in Europe, in particular, photon and neutron sources, have long been collaborating regarding data strategy, and in particular on the implementation of open science practices. The Photon and Neutron Open Science Cloud (PaNOSC) project is the science cluster representing Photon and Neutron European Research Infrastructures and connecting these to the European Open Science Cloud. Within this project, a common data policy framework for these facilities was proposed.

The PaNOSC data policy framework covers in detail the data policy requirements for EuPRAXIA and similar facilities, in particular concerning open data and FAIR principles:

- It makes data openly available after an embargo period, using an "appropriate license";
- It establishes the facility as the curator of the data;
- It proposes that the facility publishes the metadata in an online catalogue that is searchable and that will link the metadata and the raw data; data is to be preserved for a minimum of 10 years;

- This is to be done in "well-defined" formats, with the facility making the tools required for reading the data publicly available;
- It makes users responsible for ensuring that the data generated from the experiments includes accurate contextual information (metadata) such that it fulfils FAIR principles.

This framework has seen widespread adoption by the community, being in many cases adopted verbatim, and in some other cases with small modifications.

Our benchmark considered a total of 13 facilities covering sources ranging from x-ray and high-energy electrons to high-power laser and neutron sources. We focused, when possible, on facilities implementing open-data and FAIR policies, to include facilities from outside Europe, and to include distributed facilities. The data policies of these facilities were then compared against the PaNOSC data policy framework. The key points found were:

- Data policy acceptance is a condition of facility access;
- The main difference from the framework relates to the data retention period, with many sites committing only to 5 years, although striving for longer archival on a "best-effort" basis;
- Different sites treat different data types (raw data, metadata, processed data, etc.) differently. As a rule, only raw data and associated metadata are preserved;
- Licenses used for raw data vary from several Creative Commons licenses: CC-BY, CC-BY-NC, and CC-BY-SA. One case distinguishes the case of "high-level metadata" which is to be distributed under a CC0 license (See the annex on Creative Commons Licenses for a brief discussion);
- Access to data is generally restricted to registered users, in some cases with data downloads being logged and the original PI having access to the information on who downloaded the data. Registration is generally open, but there are some legal caveats;
- Some facilities identify NeXus/HDF5 as the preferred data format, others use proprietary (but open) formats;
- The data policies benchmarked do not clarify what data quality procedures are to be implemented and generally limit themselves to specifying that it is the responsibility of the users to ensure that the data generated from the experiments includes accurate contextual information;
- Only one of the surveyed documents sets a calendar for reviewing the policy;
- Only one of the surveyed documents refers to “The European Code of Conduct for Research Integrity ALLEA”;

- The majority of the data policy documents reviewed do not refer to the user data management plans;
- The 2 facilities considered in the US do not publish data policy documents in a way that is similar to that of their European counterparts. The focus is mostly on data retention and the facilities do not implement (that we could find) open data / FAIR policies.

Regarding the user data management plans, while the majority of the facilities do not consider these to be mandatory, the community has already done extensive work in attempting to establish a common framework for the DMPs, and there are proposals on how to implement these using automated data as much as possible.

3.5 Conclusions

The future development of a data policy for EuPRAXIA will leverage extensive work already carried out in the community, in particular by the PaNOSC project. The PaNOSC data policy framework, which had widespread adoption in the community, establishes an excellent starting point for a EuPRAXIA document, and this report concludes that it should be used as the basis for any of our future work.

The benchmark performed of the data policies of similar facilities supports this conclusion: for the most part all topics discussed were covered by the PaNOSC data policy framework. There were, however, some additional items that were identified that should be considered for a future EuPRAXIA data policy, namely the inclusion of a revision calendar, and the inclusion of a code of conduct for users regarding data. Following the example from some of the facilities, EuPRAXIA should also consider making user data management plans mandatory, ideally also following the recommendations from the PaNOSC project.

It should also be noted that the data policy will have important implications on other aspects of the operational phase of the project, in particular technological choices regarding data management that will further define the development and evolution of the data systems in the EuPRAXIA sites and their connection to external resources. Any final decision on the data policy will need to consider all these aspects.

The working group will continue to monitor the evolution of this topic throughout the project, and revise this report accordingly, should the need arise.

4 E-needs benchmark

The "Supporting the Transformative Impact of Research Infrastructures on European Research" report, published by the European Commission [36], was prepared by a High-Level Expert Group to assess the "Progress of ESFRI and Other World Class Research Infrastructures Towards Implementation and Long-Term Sustainability". This report identified the importance of better exploiting the data generated by the RI to achieve these goals, and to this end ESFRI has put pressure on the RIs to express the e-needs, meaning the quality and volume of interfaces with the external data networks, archives, HPC (high-performance computing) and HTC (high-throughput computing) resources needed for the full exploitation of the research data. This vision is further reflected by the 2021 update of the Strategy Report on Research Infrastructures in Europe, ESFRI Roadmap 2021 public guide [37], which establishes a minimal key requirement at the design stage to define the e-needs of the project, specifically the "vision on e-infrastructure requirements, including access policy and security measures ready", and to outline "interfaces with communication networks, distributed calculation or HPC/HTC".

These recommendations are of particular importance in the field of application of EuPRAXIA, where the amount of experimental data produced is increasing at unprecedented rates [6], [38]. This leads to significant challenges not only in the data acquisition and storage procedures but also in terms of data analysis and curation, in what is currently referred to as the "data deluge". Furthermore, considering the requirements imposed by open data policies and FAIR principles, this becomes a critical aspect of a project like EuPRAXIA. Fortunately, a significant amount of work has already been done in the community focusing on these aspects, and our project may greatly benefit from it. The main goal of this benchmark is therefore to assess the overall practices followed in what regards e-needs of existing facilities in the same field of applications.

We will focus mostly on the facilities chosen for the benchmark on data policy, for the same reasons presented earlier. However, contrary to the benchmark on data policy done in the previous sections, we chose not to organize this survey by facility, but rather by topic. The main rationale behind this choice relates to the scarcity of reference publications on this subject, in opposition to what happened to the data policies. This reflects the tremendous diversity and constant evolution of the technological aspects, as well as the need for continuous development and upgrade of data systems to keep up with the growing needs of experiments in these fields, as well as the challenges posed by open-data and FAIR policies. We therefore chose to organize this benchmark in alignment with the e-needs definitions presented above, focusing on the following topics, and concentrating on specific facilities for each of them:

- Data volumes and rates/reduction
- Storage systems
- Data monitoring and analysis
- Connections to HPC/HTC resources
- Integrated systems control and data acquisition

4.1 Data volumes and rates

The global data volumes and acquisition rates used on modern sources depend heavily on facility parameters, such as repetition rate, the number of beamlines/instruments, and their specific details, and all of these vary significantly within the surveyed facilities. Given the uniqueness of the EuPRAXIA facilities, and the uncertainties still surrounding implementation details, it is not possible at this time to directly extrapolate from existing facilities to the EuPRAXIA e-needs. Alternatively, we present an overview of repetition rates and instruments used, from the data point of view, so that we may establish a baseline from which to calculate EuPRAXIA needs as these decisions come to fruition.

One key element influencing the expected data rates and volumes is the repetition rate of the source. For the facilities that we analysed, this value can span over 5 orders of magnitude, ranging from ~ 10 Hz scale systems, such as the European Spallation Source [39] to MHz ranges such as the European XFEL [40]. The Linac Coherent Light Source (LCLS) in the U.S.A. has a base design of 120 Hz and has been recently upgraded [41] to operate up to the MHz range. Considering high-power laser sources, the ELI-ERIC facilities host a range of lasers, ranging from 10 Hz systems operating in the PW range [42] to ultrashort laser pulses operating at 100 kHz [43]. These sources can also present a complex time structure, of which the European XFEL, referenced above, is probably the most common example. This system delivers bursts of x-ray pulses, with an intra-train repetition rate of 4.5 MHz. Each burst features up to 2700 pulses, and the system will deliver 10 bursts per second. Ideally, the instruments and data acquisition systems should be able to capture data at this repetition rate, if not continuously, at least during relevant time scales.

The instruments used in these facilities also vary significantly, not only from facility to facility but also in the context of the same experiment, with data rates per instrument usually falling in the 1 - 10 GB/s ranges. We can consider as an example the x-ray detectors used in many of these facilities. The x-ray detectors used at the LCLS [44] range from 140k pixels to 2.3M pixels and are designed to operate at 120 Hz, leading to data rates of up to 0.5 GB/s. Another example is the JUNGFRAU family of x-ray detectors [45], designed at the Paul Scherrer Institut (PSI). While having been originally designed for use in the SwissFEL, these detectors have found widespread use worldwide (e.g. [46]) and are also capable of high-resolution/high-repetition rate acquisition [47], with data rates in the range of 1 - 4 GB/s. For acquiring x-ray data at kilohertz sampling frequencies, the GigaFROST system [48] is usually cited as an example. This system, also developed at PSI for tomography applications, is capable of data acquisition up to a maximum frame rate of 33.9 kHz, leading to data rates of up to 7.7 GB/s. If we now focus on experiments run at high-power laser systems [49], we see that the requirements are within the same order of magnitude. A typical example is that of a monochrome 16-bit 10 M pixel camera, operating in the near-infrared. At 50 frames per second, peak data rates are expected to be on the order of 1 GB/s.

When considering these aspects combined, the facilities must design data management systems capable of ingesting data rates in the order of 10 - 20 GB/s, and data volumes going up to 15 PB per year. For example, in the US, the data management system for LCLS [50] was designed to support up to 10 GB/s per instrument (beamline) and accumulated 11 PB from 2009 to 2017. The data systems being designed for the LCLS-II upgrade [51], [52], which increases the repetition rate of the system up to 1 MHz, would need to handle global data rates of up to 200 GB/s without data reduction (see below). Similarly, the data management system of the Advanced Photon Source (APS) [53], [54] also needs to deal with approximately 10 PB/year of raw experimental data from over 100 unique instruments. In Europe, the European Synchrotron Radiation Facility deploys data services [55] that can handle experimental data at a rate of up to 20 GB/s. ESRF is currently producing on the order of 10 PB/year [56]. Finally, the European XFEL is among the largest data producers of the facilities that we benchmarked. The data management infrastructure at the European XFEL [57] is designed to accommodate up to 15 GB/s from a single instrument during an experiment and can accept up to 2 PB/day. The data volume produced from the European XFEL has recently reached a critical milestone, having produced over 100 PB of raw experimental data since operations began in 2017.

4.1.1 Data reduction

Up until recently the complete set of experimental data from these facilities was preserved and no systematic data reduction procedures were in place. However, the unprecedented rates at which current light sources are currently producing data are putting significant strain on data management systems and, in particular, storage infrastructures, which are expensive and limited in lifetime, and have significant environmental impact. Several facilities are now considering data reduction procedures, with the dual aim of reducing the data rates at acquisition and the overall data volumes produced.

It should be noted that these procedures present an important risk, particularly in the context of open-data policy because data reduction must not eliminate yet-to-be-identified relevant data and/or introduce observational biases. Therefore, careful consideration regarding both the technical and policy issues concerning data reduction, ensuring that all relevant reduction procedures are well documented and justified in data management plans, and that quality procedures are taken to prevent the loss of scientific content.

As mentioned earlier, LCLS-II data systems [52] are considering the use of data reduction by implementing a data reduction pipeline to reduce the data rates by an order of magnitude from a peak rate of ~ 200 GB/s to 20 GB/s. This pipeline would consider several complementary techniques, such as lossless compression, feature extraction (considering only "relevant" regions), multi-event reduction (correlation over a sequence of events), and veto (software trigger, useful for experiments with a low hit rate). The latter is widely used in high-energy physics experiments. The pipeline also ensures that a fraction of non-reduced events is stored to promptly validate the correctness of the data reduction algorithms. This pipeline is seen as a "necessary component of the LCLS facility strategy to manage data storage costs and complexity."

A comparable approach is also being pursued at the European XFEL [58]. As mentioned above, raw data from the European XFEL has recently passed the 100 PB milestone, and reducing data is considered not to be avoidable anymore. The data reduction procedures, i.e. applying selection and transformation techniques to experimental data, are implemented at several points in the data lifecycle, ranging from data acquisition systems output to retroactively reducing cold data, and use different algorithms focusing either on the physics problem or specific detector configurations. Using these techniques, it was possible to avoid storage of about one-third of the expected volume of processed data in 2023 (about 7 PB). They also estimate that the same techniques can be applied retroactively to free up to 17 PB of storage space. The facility is aware that data reduction is intrinsically associated with the risk of compromising scientific throughput and, to mitigate these risks will produce extensive quality and validation metrics, as well as to address user concerns about data loss.

4.2 Storage systems

As seen in the previous section, the instruments used in these facilities can produce data at a rate in the 10 - 20 GB/s range for the duration of the data acquisition runs. The duration of these runs also varies significantly between facilities but, for each experiment, is limited in time. For example, at LCLS data collection run is usually performed in 5 - 30 minutes [59], [60] leading to approximately 200 GB of raw data. To be able to ingest these bursts of data at very high rates and simultaneously deal with the very large volumes that are produced, most facilities usually use a tiered approach with smaller capacity, higher bandwidth storage placed close to the detector systems, followed by progressively larger / slower performance storage levels. In practice, these systems usually consist of 3 levels:

1. A very high-performance (speed) storage connected directly to the instrument acquisition system;
2. An intermediate performance system used for data processing during the experimental campaign, usually installed on-site. Some facilities also provide additional systems for data analysis after the experimental campaign has finished;
3. A cold storage system, mainly for data archival purposes. These can be on or off-site and can implement redundancies.

The data system for LCLS [50] implements this hierarchy. Data from the detectors is fed into the so-called “data cache layer” using multiple 10 Gb/s ethernet connections, which is then transferred to the “fast feedback layer” that is capable of storing ~ 0.5 PB of data [60] while awaiting transfer to the next storage level. Data is then transferred using a 10 GB/s ethernet connection to the so-called “offline layer” which uses high-bandwidth cluster storage with 5 PB capacity at the time. Finally, data will be transferred to the HPSS tape system with over 20 PB of storage available. LCLS also allows data transfer to a high-end computing facility off-site for further data processing and storage (see below). Other facilities in the US follow similar approaches. For example, the storage infrastructure of the Advanced Photon Source facility [54] is backed by a storage system scalable to 15 PB of storage, that can support up to 30 GB/s of aggregate file I/O performance, which is integrated with a tape backup library for long term storage managed off-site by the Argonne Leadership Computing Facility (ALCF).

The European XFEL system [57] organizes the storage system into 4 layers: a first later (online) that can handle up to 15 GB/s from a single instrument during a data-taking run; a second layer (high-performance storage) that provides the necessary capabilities for data processing during and after experiments, with an aggregate read performance of > 175 GB/s; a third layer (mass storage) providing mid-term data access for detailed analysis; and finally a fourth layer (tape archive) providing long-term storage for over 10 years, in accordance to the facility data policy and open-data/FAIR principles. The system can accept and process up to 2 PB/day.

Another European example is the ESRF central storage system [55], [56] which features a 100 TB based storage system for fast data intake, and two 4.2 PB systems (additional storage of 2 × 5 PB was being procured). Long-term storage is ensured by a tape library that can scale up to 170 PB.

4.3 Data Monitoring and Analysis

The data volumes and rates present a significant challenge for end users, who must be able to extract scientific insight and results from the collected data, which is usually achieved by exhaustive processing of data that has been recorded to disk. However, the ability to make informed decisions in response to near real-time results is also of critical importance, allowing for the active steering of experiments, like tuning the source or the detectors, playing a key role in the success of the experiment. To this end, data processing at the surveyed facilities is generally into the following categories, which are closely associated with the storage levels analysed in the previous section:

1. Real-time or online data processing, which happens during data-taking runs, and is expected to provide fast feedback to users with latencies on the order of seconds. Until recently, this was mostly limited to monitoring detector images, but with recent advances in computing power and algorithms, it is becoming increasingly common to provide some processing at this level, such as tomographic inversion (e.g. [61]). This processing is usually associated with the fastest storage layer.
2. Offline data processing, taking place after data-taking, typically with turnaround times on the order of minutes or hours. This data processing is usually performed using on-site computing clusters with direct access to the second layer of storage (e.g. [50], [62]), or at the home institution of the experimental team after data is transferred there. Recently the possibility of streaming this data to high-end computing centres (HEC) for faster analysis is being explored (e.g. [59]).

One example of this type of data monitoring and analysis is that of the data systems at the European XFEL [62]. In these systems, online data processing is performed in the so-called “online computing clusters”, which are dedicated computing nodes for each of the European XFEL beamlines, located physically close to their end stations, and allowing online image correction and preview, with preview latencies of a few hundred milliseconds. Offline data processing is done at the Maxwell computer cluster at DESY and includes both automatic processing steps and additional user-triggered steps. The distribution of resources on the Maxwell cluster prioritizes ongoing experiments, allowing fast allocation of computing nodes for this purpose, and typical turnaround times of below one minute.

Another example is that of the LCLS [50]. The data systems at this facility focus primarily on displaying and analysing critical real-time information and are implemented using the so-called “analysis monitoring interface” (AMI) for graphical online monitoring, and the software framework psana [63] for data analysis. The AMI runs on dedicated monitoring nodes which are exclusive for each instrument and allows users to display and monitor information on-thefly. The facility deploys a computing infrastructure for data analysis [52], aiming to deliver near real-time analysis of data bursts (< 10 minutes). Support for LCLS-II data rates is estimated to require a computing capacity on the order of 10 PFLOPs.

It should also be noted that the implementation of open-data and FAIR policies by these facilities is enabling a third level of data processing, by allowing researchers to mine published public data in previously unknown ways or using methods developed after the experimental campaign. The Protein Data Bank [64] is frequently used as an example of such data processing through its use as training data for artificial intelligence models leading to the prediction of protein structures using a novel machine learning approach. To this end, many facilities, such as the ESRF [55] or the European XFEL [62] have included in their data processing pipelines the necessary resources to capture relevant metadata, mostly through automatic processes, but also allowing for user input. This metadata, together with raw data, is then published in online catalogues with open access.

4.3.1 e-logbooks

One particular item that has received significant attention in recent years is the electronic logbook, or e-logbook, which is also referred to in the data policy documents of many facilities, as described in the first section of the present report. The main goal of these documents, like traditional laboratory logbooks, is to document the experimental protocol and notes, as well as decisions, observations, and/or conclusions made during the experimental campaign. This is a key feature in the implementation of the data policy, greatly contributing to the validation and reproducibility of the results. In the context of data analysis, an e-logbook is also supposed to easily link to experimental data and metadata, allowing users to easily include relevant experimental observations.

Many of the facilities that were surveyed mention the implementation of e-logbooks and their integration into the data management system, such as the European Spallation Source [65], the ESRF [55], and the European XFEL [62]. Open-source solutions, such as Elog [66] developed by PSI, the ESRF e-logbooks [67], and Jupyter notebook-based solutions [68] have gained some popularity, but the community does not seem to have developed a standard solution with wide adoption. The high-power laser community has also been deploying e-logbooks [49], such as the tool deployed at the Helmholtz-Zentrum Dresden-Rossendorf (HZDR) [69], but no common solution has been adopted either.

4.4 Connection to HPC/HTC resources

As noted in the previous section, the computational requirements involved in performing data analysis of the large data volumes being produced in current light sources are in the multi PFLOP range [52]. While some facilities are in the process of deploying high-performance computing (HPC) / high-throughput computing (HTC) resources locally, such as the ESRF data center [56] (ESRF IT Strategy 2020) or the Maxwell cluster at DESY [57], it has long been envisioned that establishing a connection with remote computing facilities and in particular those falling in the high-end computing (HEC) category, would play a pivotal a role in allowing users to perform Big Data-intensive analytics and visualization on complex experiments [38]. This concept is being actively pursued in the US by combining high data rate experiments and online HPC processing, in what is referred to as the "Superfacility" model [70].

Some recent work done at LCLS [59] demonstrates the prospective of using a remote supercomputer system, in this case, the Cori and Perlmutter systems at the National Energy Research Scientific Computing Center (NERSC), for live interactive analysis of large data sets. A data workflow is implemented to automatically transfer the data from LCLS to NERSC servers once the measurement is concluded. This relied on a high-bandwidth network connecting the facilities (ESNET), achieving peak transfer rates of ~ 2.6 GB/s, which allowed the transfer of the data from an acquisition run (~ 200 GB) to be done in close to 3 minutes. Automating the process (both data moving and analysis job submission) allowed data to be analysed rapidly after its collection. This work has further been scaled to the exascale level of performance [60], with peak data transfer reaching 15 TB/day, to great success.

4.5 Integrated systems control and data acquisition

Given the extreme requirements of modern radiation sources, both in terms of high data rates and volumes and the use of complex and custom-made hardware, combined with the need to achieve greater efficiency and effectiveness in translating experimental data into knowledge and, to this end, the need of close integration of data analysis and rapid feedback, most facilities have adopted integrated systems control and data acquisition systems. While not usually considered in the general "e-needs" requirements, but rather as a part of overall systems design, this system is at the core of the data acquisition process, and its integration with all data management process is of key importance in the data lifecycle and its efficient use. These considerations are not new, and the community has long been developing solutions to address these needs.

Overall, the facilities that were surveyed used either the EPICS [71], originally developed at Argonne National Laboratory, or Tango Controls [72], originally developed at the ESRF distributed control systems. One notable exception is that of the European XFEL which uses a custom solution developed in-house named Karabo [73]. In all of the benchmarked facilities, these systems are tightly integrated into the operation of the system and the experimental diagnostics, which, for the most standard solutions, already provide the necessary software tools for connecting with the systems. It should also be noted that these systems are also responsible for automatically collecting metadata, while allowing for additional user input, as well as additional facility and operational parameters, and storing it centrally, typically in a metadata database system such as ICAT [74], or in custom designed database.

4.6 Conclusions

The specific e-needs of the EuPRAXIA project will depend heavily on technical decisions yet to be made, such as the repetition rate of the source or the number of beamlines and specific instruments that will be implemented, as well as the decisions on the data policy to be used, as discussed in the first part of this document. This benchmark allowed us to identify some reference values and design decisions that, after scaling to the appropriate dimensions of our project, will allow us to project the overall e-needs of our project.

The key findings of this benchmark were:

- The current data rate per instrument is in the range of 1 - 10 GB/s, leading to data rates on the order of 10 - 20 GB/s for a complete experimental setup, in bursts of varying duration, typically on the order of 30 minutes;
- Considering typical operation values the total data volume per year will be on the order of 5 - 10 PB. Note the PaNOSC recommendation is to store the experimental data for 10 years, which would lead to total storage requirements on the order of 50 - 100 PB;
- Data reduction techniques can be implemented to lower these requirements by as much as an order of magnitude; however great care must be taken not to jeopardize the scientific content;
- The storage system to be implemented will need to cope with both the very large data rates and data volumes. The solutions implemented usually rely on a layered approach of storage systems of increasing capacity/decreasing performance;
- Tape storage is still the main choice for long-term data archiving, should EuPRAXIA decide to implement this on-site. However, we are considering moving the data to an external data repository, yet to be identified;

- The data management systems implement two categories of data processing, one for (next to) real-time data processing, and one for offline data processing. The former should provide feedback on the order of seconds on data that is currently being acquired, while the latter is to be applied to saved data;
- Electronic logbooks are implemented in several facilities, as a way of documenting experimental protocol and notes, and in connection with the data policies in each facility;
- Data processing requires considerable computational power. Most facilities have an on-site computing cluster for offline data processing, complementing specific hardware deployed for real-time data processing. Estimated requirements are on the order of 1 - 10 PFLOPS;
- To lower this requirement EuPRAXIA sites may collaborate with external supercomputing facilities. However, this has significant network requirements and requires extensive workflow automation work to be of practical use;
- Most systems use some form of integrated systems control and data acquisition. EPICS and Tango controls are the two most popular choices.

As the EuPRAXIA project matures, and particularly when more technical design decisions concerning the future EuPRAXIA sites are made, raising this way specific e-needs not covered by this document, the working group will further extend this benchmark to lay the groundwork for the required data management infrastructure.

5 References

[1] UNESCO, ‘UNESCO Recommendation on Open Science’, UNESCO, 2021. doi: 10.54677/MNMH8546.

[2] *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*. 2022. [Online]. Available: http://data.europa.eu/eli/reg/2022/868/oj

[3] European Commission. Directorate General for Research and Innovation., *Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data.* LU: Publications Office, 2018. Accessed: Mar. 28, 2024. [Online]. Available: https://data.europa.eu/doi/10.2777/1524

[4] M. D. Wilkinson *et al.*, ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Sci Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

[5] *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information*. 2019. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2019/1024/oj

[6] A. Götz *et al.*, ‘LEAPS data strategy’, *Eur. Phys. J. Plus*, vol. 138, no. 7, p. 617, Jul. 2023, doi: 10.1140/epjp/s13360-023-04189-6.

[7] ‘ExPaNDS - European Open Science Cloud Photon and Neutron Data Service’. [Online]. Available: https://expands.eu/

[8] ‘PaNOSC - The Photon and Neutron Open Science Cloud’. [Online]. Available: https://www.panosc.eu

[9] A. Gotz *et al.*, ‘PaNOSC FAIR Research Data Policy framework’, May 2020, doi: 10.5281/ZENODO.3862701.

[10] R. Dimper, ‘PaN-data Europe, Deliverable D2.1 - Common policy framework on scientific data’, Dec. 2010. [Online]. Available: http://pan-data.eu/sites/pan-data.eu/files/PaN-data-D2-1.pdf

[11] OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD, 2007. doi: 10.1787/9789264034020-en-fr.

[12] Research Data Alliance FAIR Data Maturity Model Working Group, ‘FAIR Data Maturity Model: specification and guidelines - draft’, 2020, doi: 10.15497/RDA00045.

[13] ‘NeXus Data Format’. [Online]. Available: https://nexusformat.org

[14] ‘HDF5 Library & File Format’. HDF Group. [Online]. Available: https://www.hdfgroup.org/solutions/hdf5/

[15] P. Braak, H. De Jonge, G. Trentacosti, I. Verhagen, and S. Woutersen-Windhouwer, ‘Guide to Creative Commons for Scholarly Publications and Educational Resources’, [object Object], Oct. 2020. doi: 10.5281/ZENODO.4741966.

[16] The DOI Foundation, ‘Digital Object Identifier’. [Online]. Available: https://www.doi.org

[17] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[18] V. Favre-Nicolin, A. Götz, M. Krisch, and G. Martinez-Criado, ‘ESRF Data Policy 2024’. [object Object], 2024. doi: 10.15151/ESRF-DC-1534175008.

[19] ‘CERIC-ERIC Data Policy’, Central European Research Infrastructure Consortium, CERIC-ERIC, Feb. 2021. [Online]. Available: https://www.ceric-eric.eu/wp-content/uploads/2021/02/CERIC-Scientfic-Data-Policy.pdf

[20] ‘ELETTRA Scientific Data Policy’, Aug. 2022. [Online]. Available: https://vuo.elettra.eu/vuo/cgi-bin/download-tm4.py?frm_user_id=8707&frm_iddocumenttype=14&frm_iddocument=625063&frm_hash=2450356c44429054734c8bdfd9cbc309fa74512f

[21] ‘ESS Data Policy’, European Spallation Source, ESS-0081403, May 2017. [Online]. Available: https://indico.esss.lu.se/event/809/attachments/6187/8717/Data_policy_20170505.pdf

[22] ‘ILL Data Policy’, Institut Laue-Langevin, Mar. 2018. [Online]. Available: https://www.ill.eu/fileadmin/user_upload/ILL/3_Users/User_Guide/After_your_experiment/Data_management/ILL_data_management_policy_March_2018.pdf

[23] B. Gagey, Ed., ‘SOLEIL Data Management Policy’. Feb. 10, 2018. [Online]. Available: https://www.synchrotron-soleil.fr/en/file/15971/download?token=2iVYagN6

[24] M. Ounsy and B. Gagey, ‘SOLEIL Data Management Policy’, *Synchrotron Radiation News*, vol. 32, no. 3, pp. 23–24, May 2019, doi: 10.1080/08940886.2019.1608122.

[25] European XFEL, 'Scientific Data Policy of European X-Ray Free-Electron Laser Facility GmbH', [object Object], 2025. doi: 10.22003/XFEL.EU-TR-2025-001.

[26] W. Taylor *et al.*, 'ISIS Neutron and Muon Source Data Policy', ISIS Neutron and Muon Source, Apr. 2023. doi: 10.5281/ZENODO.7794219.

[27] 'DLS Experimental Data Policy', Diamond Light Source, Dec. 2018. [Online]. Available: https://www.diamond.ac.uk/Home/Legal-and-Compliance/Policies/Experimental-Data-Management-Pol.html

[28] A. Amato, A. Ashton, O. Bunk, M. Erat, and S. Janssen, 'PSI Data Policy', Paul Scherrer Institute, Apr. 2022. [Online]. Available: https://www.psi.ch/en/media/74588/download

[29] A. Weeks, G. Szabó, R. Hvězda, F. Gliksohn, and T. Ivănoaica, 'ELI ERIC Data Policy', May 2022, doi: 10.5281/ZENODO.6515903.

[30] ALLEA - All European Academies, *The European Code of Conduct for Research Integrity*. DE: ALLEA - All European Academies, 2023. Accessed: Mar. 28, 2024. [Online]. Available: https://doi.org/10.26356/ECoC

[31] 'Advanced Photon Source Data Management and Retrieval Practices', Argonne National Laboratory, Feb. 2019. [Online]. Available: https://www.aps.anl.gov/Users-Information/Help-Reference/Data-Management-Retrieval-Practices

[32] A. Perazzo and W. Kroeger, 'LCLS Data Retention Policy', Linac Coherent Light Source, Aug. 2022. [Online]. Available: https://confluence.slac.stanford.edu/display/PCDS/Data+Retention+Policy

[33] H. Goerzig *et al.*, 'Active DMPs for Photon and Neutron RIs', Dec. 2022, doi: 10.5281/ZENODO.7223438.

[34] M. Bodin *et al.*, 'Data Management Plans for the Photon and Neutron Communities', *Data Science Journal*, vol. 22, p. 30, Aug. 2023, doi: 10.5334/dsj-2023-030.

[35] R. Pergl, R. Hooft, M. Suchánek, V. Knaisl, and J. Slifka, '"Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning', *Data Science Journal*, vol. 18, p. 59, Dec. 2019, doi: 10.5334/dsj-2019-059.

[36] European Commission. Directorate General for Research and Innovation., *Supporting the transformative impact of research infrastructures on European research: report of the High Level Expert Group to assess the progress of ESFRI and other world class research infrastructures towards implementation and long term sustainability.* LU: Publications Office, 2020. Accessed: Mar. 28, 2024. [Online]. Available: https://data.europa.eu/doi/10.2777/3423

[37] ESFRI - European Strategy Forum on Research Inftrastructures, ‘Strategy Report on Research Infrastructures, Roadmap 2021, Public Guide’, Sep. 2019. [Online]. Available: https://www.esfri.eu/sites/default/files/ESFRI_Roadmap2021_Public_Guide_Public.pdf

[38] C. Wang, U. Steiner, and A. Sepe, ‘Synchrotron Big Data Science’, *Small*, vol. 14, no. 46, p. 1802291, Nov. 2018, doi: 10.1002/smll.201802291.

[39] R. Garoby *et al.*, ‘The European Spallation Source Design’, *Phys. Scr.*, vol. 93, no. 1, p. 014001, Dec. 2018, doi: 10.1088/1402-4896/aa9bff.

[40] W. Decking *et al.*, ‘A MHz-repetition-rate hard X-ray free-electron laser driven by a superconducting linear accelerator’, *Nat. Photonics*, vol. 14, no. 6, pp. 391–397, Jun. 2020, doi: 10.1038/s41566-020-0607-z.

[41] J. Galayda, ‘The LCLS-II: A High Power Upgrade to the LCLS’, *Proceedings of the 9th Int. Particle Accelerator Conf.*, vol. IPAC2018, p. 6 pages, 3.180 MB, 2018, doi: 10.18429/JACOW-IPAC2018-MOYGB2.

[42] B. Rus *et al.*, ‘ELI-Beamlines laser systems: status and design options’, presented at the SPIE Optics + Optoelectronics, J. Hein, G. Korn, and L. O. Silva, Eds., Prague, Czech Republic, May 2013, p. 87801T. doi: 10.1117/12.2021264.

[43] S. Kühn *et al.*, ‘The ELI-ALPS facility: the next generation of attosecond sources’, *J. Phys. B: At. Mol. Opt. Phys.*, vol. 50, no. 13, p. 132002, Jul. 2017, doi: 10.1088/1361-6455/aa6ee8.

[44] G. Blaj *et al.*, ‘X-ray detectors at the Linac Coherent Light Source’, *J Synchrotron Rad*, vol. 22, no. 3, pp. 577–583, May 2015, doi: 10.1107/S1600577515005317.

[45] A. Mozzanica *et al.*, ‘The JUNGFRAU Detector for Applications at Synchrotron Light Sources and XFELs’, *Synchrotron Radiation News*, vol. 31, no. 6, pp. 16–20, Nov. 2018, doi: 10.1080/08940886.2018.1528429.

[46] D. Khakhulin *et al.*, ‘Ultrafast X-ray Photochemistry at European XFEL: Capabilities of the Femtosecond X-ray Experiments (FXE) Instrument’, *Applied Sciences*, vol. 10, no. 3, p. 995, Feb. 2020, doi: 10.3390/app10030995.

[47] M. Sikorski *et al.*, ‘First operation of the JUNGFRAU detector in 16-memory cell mode at European XFEL’, *Front. Phys.*, vol. 11, p. 1303247, Nov. 2023, doi: 10.3389/fphy.2023.1303247.

[48] R. Mokso *et al.*, ‘GigaFRoST: the gigabit fast readout system for tomography’, *J Synchrotron Rad*, vol. 24, no. 6, pp. 1250–1259, Nov. 2017, doi: 10.1107/S1600577517013522.

[49] S. Feister *et al.*, ‘Control systems and data management for high-power laser facilities’, *High Pow Laser Sci Eng*, vol. 11, p. e56, 2023, doi: 10.1017/hpl.2023.49.

[50] J. Thayer *et al.*, ‘Data systems for the Linac coherent light source’, *Adv Struct Chem Imag*, vol. 3, no. 1, p. 3, Dec. 2017, doi: 10.1186/s40679-016-0037-7.

[51] J. B. Thayer *et al.*, ‘Building a Data System for LCLS-II’, in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Atlanta, GA: IEEE, Oct. 2017, pp. 1–4. doi: 10.1109/NSSMIC.2017.8533033.

[52] J. Thayer *et al.*, ‘Data Processing at the Linac Coherent Light Source’, in *2019 IEEE/ACM 1st Annual Workshop on Large-scale Experiment-in-the-Loop Computing (XLOOP)*, Denver, CO, USA: IEEE, Nov. 2019, pp. 32–37. doi: 10.1109/XLOOP49562.2019.00011.

[53] S. Veseli, N. Schwarz, and C. Schmitz, ‘APS *Data Management System*’, *J Synchrotron Rad*, vol. 25, no. 5, pp. 1574–1580, Sep. 2018, doi: 10.1107/S1600577518010056.

[54] N. Schwarz, S. Veseli, and D. Jarosz, ‘Data Management at the Advanced Photon Source’, *Synchrotron Radiation News*, vol. 32, no. 3, pp. 13–18, May 2019, doi: 10.1080/08940886.2019.1608120.

[55] R. Dimper, A. Götz, A. De Maria, V. A. Solé, M. Chaillet, and B. Lebayle, ‘ESRF Data Policy, Storage, and Services’, *Synchrotron Radiation News*, vol. 32, no. 3, pp. 7–12, May 2019, doi: 10.1080/08940886.2019.1608119.

[56] R. Dimper *et al.*, ‘ESRF Information Technology Data Strategy 2022-2026’, European Synchrotron Radiation Facility, Oct. 2020. [Online]. Available: https://www.esrf.fr/files/live/sites/www/files/about/information-material/ESRF_IT_Strategy.pdf

[57] J. Malka *et al.*, ‘Data Management Infrastructure for European XFEL’, p. 6 pages, 1.023 MB, Dec. 2023, doi: 10.18429/JACOW-ICALEPCS2023-WE1BCO02.

[58] E. Sobolev *et al.*, ‘Data reduction activities at European XFEL: early results’, *Front. Phys.*, vol. 12, p. 1331329, Feb. 2024, doi: 10.3389/fphy.2024.1331329.

[59] J. P. Blaschke, F. Wittwer, B. Enders, and D. Bard, ‘How a Lightsource Uses a Supercomputer for Live Interactive Analysis of Large Data Sets: Perspectives on the NERSC-LCLS Superfacility’, *Synchrotron Radiation News*, vol. 36, no. 4, pp. 10–16, Jul. 2023, doi: 10.1080/08940886.2023.2245700.

[60] J. P. Blaschke *et al.*, ‘Real-time XFEL data analysis at SLAC and NERSC: A trial run of nascent exascale experimental data analysis’, *Concurrency and Computation*, p. e8019, Feb. 2024, doi: 10.1002/cpe.8019.

[61] V. Nikitin, P. Shevchenko, A. Deriy, A. Kastengren, and F. De Carlo, ‘Streaming Collection and Real-Time Analysis of Tomographic Data at the APS’, *Synchrotron Radiation News*, vol. 36, no. 4, pp. 3–9, Jul. 2023, doi: 10.1080/08940886.2023.2245693.

[62] P. Schmidt *et al.*, 'Turning European XFEL raw data into user data', *Front. Phys.*, vol. 11, p. 1321524, Jan. 2024, doi: 10.3389/fphy.2023.1321524.

[63] D. Damiani *et al.*, 'Linac Coherent Light Source data analysis using *psana*', *J Appl Crystallogr*, vol. 49, no. 2, pp. 672–679, Apr. 2016, doi: 10.1107/S1600576716004349.

[64] wwPDB consortium *et al.*, 'Protein Data Bank: the single global archive for 3D macromolecular structure data', *Nucleic Acids Research*, vol. 47, no. D1, pp. D520–D528, Jan. 2019, doi: 10.1093/nar/gky949.

[65] T. Friedrich, 'Overview of the Integrated Control System', European Spallation Source, ESS-0297798, Jun. 2018. [Online]. Available: https://europeanspallationsource.se/sites/default/files/files/document/2018-09/IntegratedControlSystemOverview.pdf

[66] 'Elog'. PSI - Paul Sherrer Institute. [Online]. Available: https://elog.psi.ch/elog/

[67] 'ESRF e-logbook'. European Synchrotron Radiation Facility. [Online]. Available: https://gitlab.esrf.fr/icat/E-DataPortal/

[68] B. E. Granger and F. Perez, 'Jupyter: Thinking and Storytelling With Code and Data', *Comput. Sci. Eng.*, vol. 23, no. 2, pp. 7–14, Mar. 2021, doi: 10.1109/MCSE.2021.3059263.

[69] 'HZDR e-logbook'. HZDR – Helmholtz-Zentrum Dresden-Rossendorf. [Online]. Available: https://www.hzdr.de/db/Cms?pOid=67705&pNid=0&pLang=en

[70] D. Bard *et al.*, 'LBNL Superfacility Project Report', None, 1875256, ark:/13030/qt6tm8m61q, Jun. 2022. doi: 10.2172/1875256.

[71] L. R. Dalesio *et al.*, 'The experimental physics and industrial control system architecture: past, present, and future', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 352, no. 1–2, pp. 179–184, Dec. 1994, doi: 10.1016/0168-9002(94)91493-1.

[72] A. Götz *et al.*, 'TANGO a CORBA Based Control System'.

[73] S. Hauf *et al.*, 'The Karabo distributed control system', *J Synchrotron Rad*, vol. 26, no. 5, pp. 1448–1461, Sep. 2019, doi: 10.1107/S1600577519006696.

[74] 'ICAT Project'. [Online]. Available: https://icatproject.org

Annex 1

List of Acronyms

Acronym	Definition
ALCF	Argonne Leadership Computing Facility
AMI	Analysis Monitoring Interface
APS	Advanced Photon Source
CC	Creative Commons
CERIC	Central European Research Infrastructure Consortium
DESY	Deutsches Elektronen-Synchrotron DESY
DLS	Diamond Light Source
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
ELI	Extreme Light Infrastructure
EOSC	European Open Science Cloud
EOSCA	European Open Science Cloud Association
ESRF	European Synchrotron Radiation Facility
ESS	European Spallation Source
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
HEC	High-end computing center

Acronym	Definition
HPC	High-Performance Computing
HTC	High-Throughput Computing
HZDR	Helmholtz-Zentrum Dresden-Rossendorf
ILL	Institut Max Von Laue – Paul Langevin
KM	Knowledge Model
LBNL	Lawrence Berkeley National Laboratory
LCLS	Linac Coherent Light Source
LEAPS	League of European Accelerator-based Photon Sources
LENS	League of Advanced European Neutron Sources
NERSC	National Energy Research Scientific Computing Center
OECD	Organisation for Economic Co-operation and Development
PaNOSC	Photon and Neutron Open Science Cloud
PI	Principal Investigator
PSI	Paul Scherrer Institute
RDM	Research Data Management
UV	Ultra-Violet
XFEL	X-ray free electron laser

Annex 2

Creative Commons Licenses

In the framework of this benchmark, several types of Creative Commons Licenses were referred to for publishing scientific data. In particular, the PaNOSC data policy framework references the CC-BY, CC-BY-NC, and CC0 licenses, and some facilities (e.g. ELETTRA, PSI) also reference the CC-BY-SA license.

In this context, the suffixes to the Creative Commons (CC) licenses have the following meaning:

- **BY** - credit must be given to the creator;
- **NC** - Only non-commercial uses of the work are permitted;
- **SA** - Adaptations must be shared under the same terms.

The main differences between the licenses are as follows:

- **CC0** - Enables “reusers” to distribute, remix, adapt, and build upon the material in any medium or format, with no conditions (public domain).
- **CC-BY** - Enables “reusers” to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.
- **CC-BY-NC** - Enables “reusers” to distribute, remix, adapt, and build upon the material in any medium or format for non-commercial purposes only, and only so long as attribution is given to the creator.
- **CC-BY-SA** - Enables “reusers” to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

Please refer to the Creative Commons website for the complete text of the licenses and detailed legal information.

References

- Creative Commons, https://creativecommons.org
- License CC0, https://creativecommons.org/publicdomain/zero/1.0/
- License CC-BY, https://creativecommons.org/licenses/by/4.0/deed.en
- License CC-BY-NC, https://creativecommons.org/licenses/by-nc/4.0/deed.en
- License CC-BY-SA, https://creativecommons.org/licenses/by-sa/4.0/deed.en