EUROPEAN
PLASMA RESEARCH
ACCELERATOR WITH
EXCELLENCE IN
APPLICATIONS

EuPRAXIA
Preparatory Phase

WP7 activities and Open-Science challenges for EuPRAXIA

Ricardo Fonseca (IST/Portugal) for WP7

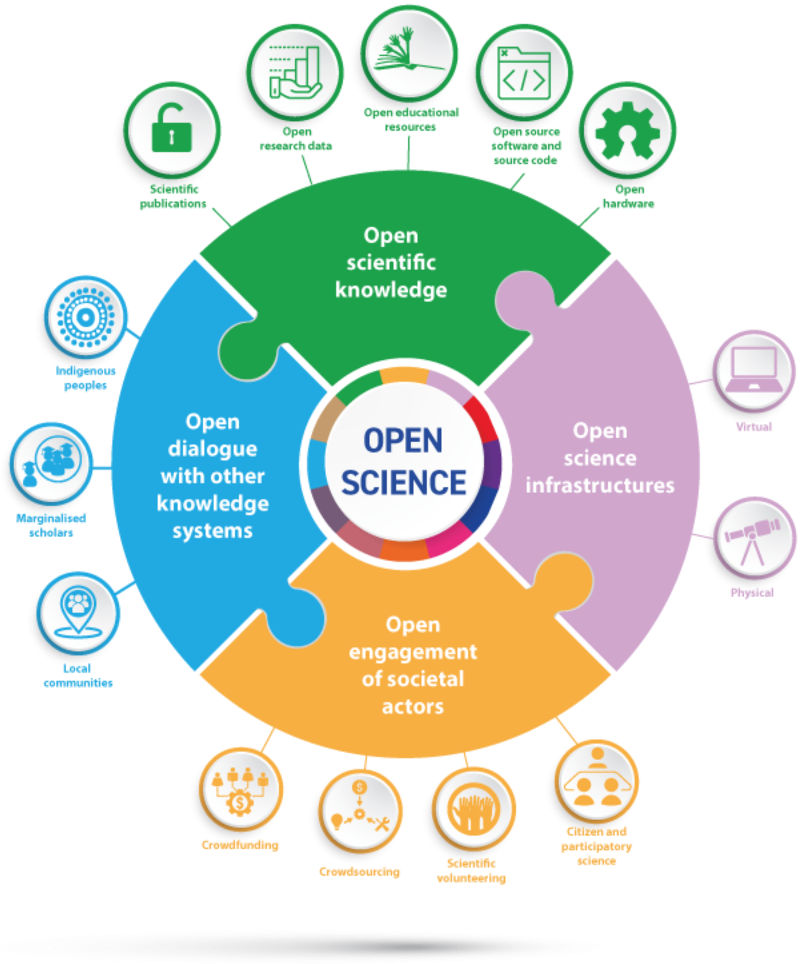
This project has received funding from the European Union´s Horizon Europe research and innovation programme under grant agreement No. 101079773

EuPRAXIA and Open-Science

Funded by the European Union

- **Open-Science**
  - Aims to make **scientific research** (including publications, data, physical samples, and software) and its dissemination **accessible to all** levels of society, amateur or professional

- **From the EuPRAXIA Grant Agreement:**
  - “EuPRAXIA implements **open science** and open innovation practices, including **research data management**”


Scientific publications
Open research data
Open educational resources
Open source software and source code
Open hardware
Open scientific knowledge
Indigenous peoples
Marginalised scholars
Local communities
Open dialogue with other knowledge systems
OPEN SCIENCE
Open science infrastructures
Virtual
Physical
Open engagement of societal actors
Crowdfunding
Crowdsourcing
Scientific volunteering
Citizen and participatory science

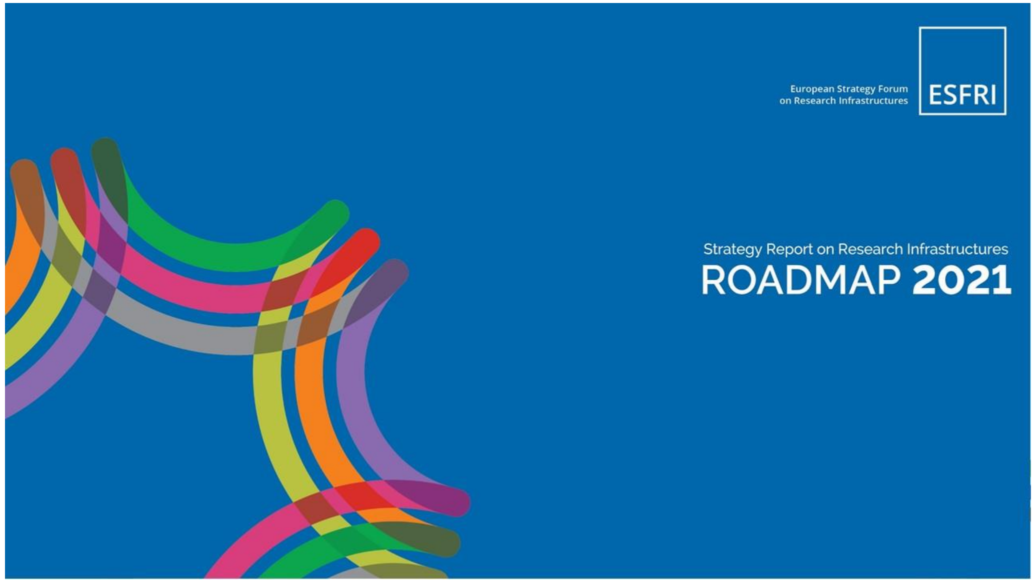
Pillars of the Open Science according to UNESCO's 2021 Open Science recommendation

E-needs (from ESFRI roadmap)

Funded by the European Union

- **Design level**
  - Vision on e-infrastructure requirements, including access policy and security measures ready
  - Interfacing with communication networks or distributed calculation or HPC/HTC

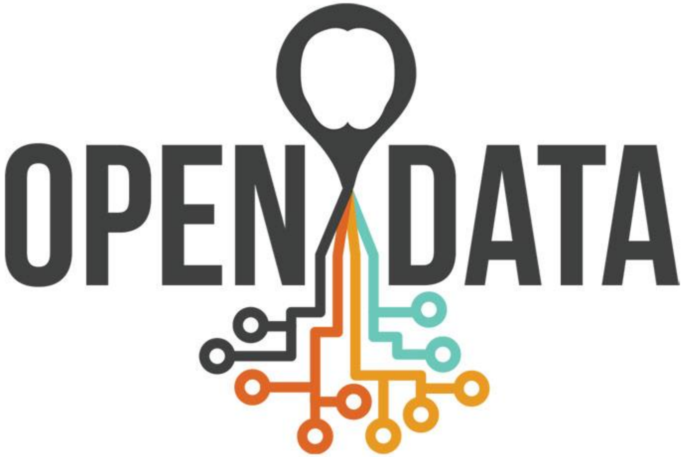
- **Preparation level**
  - Conceptual design of e-infrastructure ready
  - Contributions of e-infrastructure resources at all levels (institutional, regional, national, international) described
  - Access policy and Data Management Plan (DMP) outlined
  - Compliance with FAIR principles

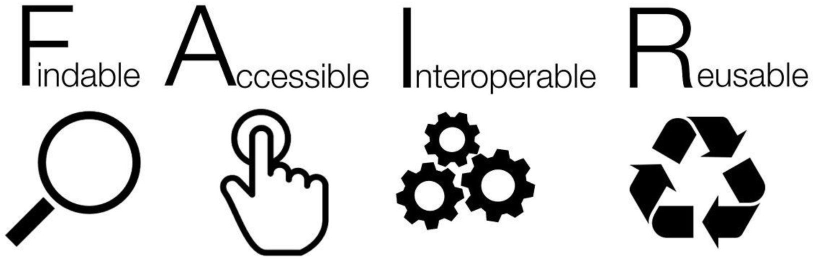

European Strategy Forum on Research Infrastructures
ESFRI
Strategy Report on Research Infrastructures
ROADMAP 2021

Work-package focus

Funded by the European Union

- **Implementing Open-Science for EuPRAXIA research data**
  - Establish access to the external data networks, archives, HPC and HTC resources needed for the full exploitation of the research data produced.
  - Develop a data management plan ensuring EuPRAXIA research data is findable, accessible, interoperable and reusable (FAIR), as well as ensuring data preservation and compliance with data protection legislation





Outline

- Workpackage Activities
- Operational Phase Data-management plan
- Benchmark with existing facilities
  - Data Policy
  - E-needs
- On-going and Future work
- Overview
- Discussion

Deliverable Report

EuPRAXIAPP
Grant Number 101079773
DELIVERABLE: WP7-DVL7.1

EuPRAXIA
Preparatory Phase

Work package Activities

WP7 – e-needs and data policy

Work package members

- **Staff effort**
  - IST / Portugal lead (30 PM)
  - INFN / Italy co-lead (6 PM)
  - EMPA / Switzerland (24 PM)

- **Connection to other WP**
  - WP5 – User strategy and services
  - WP8 – Theory & Simulation

TÉCNICO LISBOA

INFN
Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali di Frascati

Empa
Materials Science and Technology

EuPRAXIA WP Objectives

- **Objectives**
  - Define E-Needs and Data Policy
  - Connect to ongoing integrating projects in EU
  - Connect to funding sources
  - (If possible) involve European Open Science Cloud (EOSC)

EUROPEAN OPEN
SCIENCE CLOUD

panosc
photon and neutron
open science cloud

Work package Tasks

- **Project Tasks**
  - **T7.1** - Benchmarking of E-Needs of similar facility and infrastructure in Europe.
  - **T7.2** - Integrate the EuPRAXIA E-needs and requirements for data policy into the European Facility landscape. In particular this task will be connected with user and access strategy work packages.
  - **T7.3** - Finalize the E-needs and data policy for the EuPRAXIA distributed facility and excellence centres in the EOSC framework.

Consortium Agreement

HORIZON EUROPE GRANT AGREEMENT N. 101079773

Deliverables / Milestones

Funded by the European Union

- **Deliverables**
  - **D7.1** - Data management plan (M12) [completed]
  - **D7.2** - Report integration into European facility landscape and standards (M30) [in progress]
  - ***D7.3** - Final report E-Needs and Data Policy (M42)*

- **Milestones**
  - **MS18** - Benchmark within existing facilities on e-needs and data policy (M16) [completed]

Deliverables Report

EuPRAXIA PP
Grant Number 101079773
DELIVERABLE: WP7-DVL7.1

Activities / Progress

- **Coordination by IST / Portugal**
  - Joined the project after proposal approval
- **Most work coordinated by email / online meetings**
  - Team is very small
- **Work on tasks T7.2 and T7.3 require interfacing with other WP**
  - EuPRAXIA workshop(s)
  - Specialized meetings later in the project
- **No workshops or additional activities planned**

TÉCNICO LISBOA

EuPRAXIA
Preparatory Phase

Operational Phase Data-management plan

WP7 – e-needs and data policy

Operational Data Management Plan

- **Due date**
  - Delivered M12
- **Description**
  - A report will be produced together with WP1 that describes the plan and the tools for managing the data that is generated in the EuPRAXIA-PP project. The EuPRAXIA data management shall allow for efficient collaboration and documentation among the institutes and organisations involved, at the technical, financial and planning levels.

Operational Data Management Plan

Funded by the European Union

- **Operation phase**
  - Initial version of the DMP for the EuPRAXIA operation phase
  - Based on the **Horizon Europe DMP template**
  - Document will be reviewed regularly as other parts of EuPRAXIA project come to fruition
  - Reinforces EuPRAXIA commitment to **Open Science** and publishing research data according to **FAIR principles**

Horizon Europe

Data Management Plan Template

Version 1.0
05 May 2021

Operational Data Management Plan

Funded by the European Union

- **Data**
  - **Data is to be published in a trusted repository**; we are exploring options within the project framework.
  - We will strive to have published **data available for 10 years** after the end of the experimental campaign.
  - The tentative choice for the main format for data is the **NeXus/HDF5** data format.
  - We will **publish only raw data** and related software/workflows.

Operational Data Management Plan

- **Metadata, quality and management**
    - Metadata and additional information will be included with the published data to **improve data findability and useability**.
    - EuPRAXIA will implement **data quality assurance** procedures focusing on both **fitness-for-use** and **fitness-for-purpose**.
    - Data management during the operation phase will be the responsibility of a team hired for this purpose.
    - A data management committee to be nominated will handle all policy decisions during the operational phase.

Operational Data Management Plan

- **Open questions**
  - Many choices depend on outcomes from other WP, in particular the choice for second site
  - The data volumes / rates presented are very crude estimates, as they depend on many design decisions not yet made
- **The main open question pertains the use of a (external) trusted repository for long term storage**
  - Initial conversations with the Bologna INFN - EOSC node and the foreseen EuPRAXIA Hungarian National Node for this purpose
  - Additional collaborations (e.g. PaNOSC) will be pursued

EuPRAXIA Preparatory Phase

Benchmark within existing facilities on e-needs and data policy

WP7 – e-needs and data policy

Milestone 7.1

Benchmark within existing facilities on e-needs and data policy

- **Description**
  - Report benchmarking within existing facilities on E- needs and data policy. Summarizes the work done by the WP surveying the landscape of research facilities that operate comparable sources and/or perform similar activities to the future EuPRAXIA sites.

LEAPS

LENS

Chosen Facilities

Funded by the European Union

ESRF
The European Synchrotron

ess
EUROPEAN SPALLATION SOURCE

European XFEL

CERIC
Central European Research Infrastructure Consortium

ill
NEUTRONS FOR SOCIETY

PAUL SCHERRER INSTITUT
PSI

ISIS Neutron and Muon Source

SOLEIL
SYNCHROTRON

eli

diamond

Elettra Sincrotrone Trieste

SLAC

Argonne
NATIONAL LABORATORY

Data Policy

Funded by the European Union

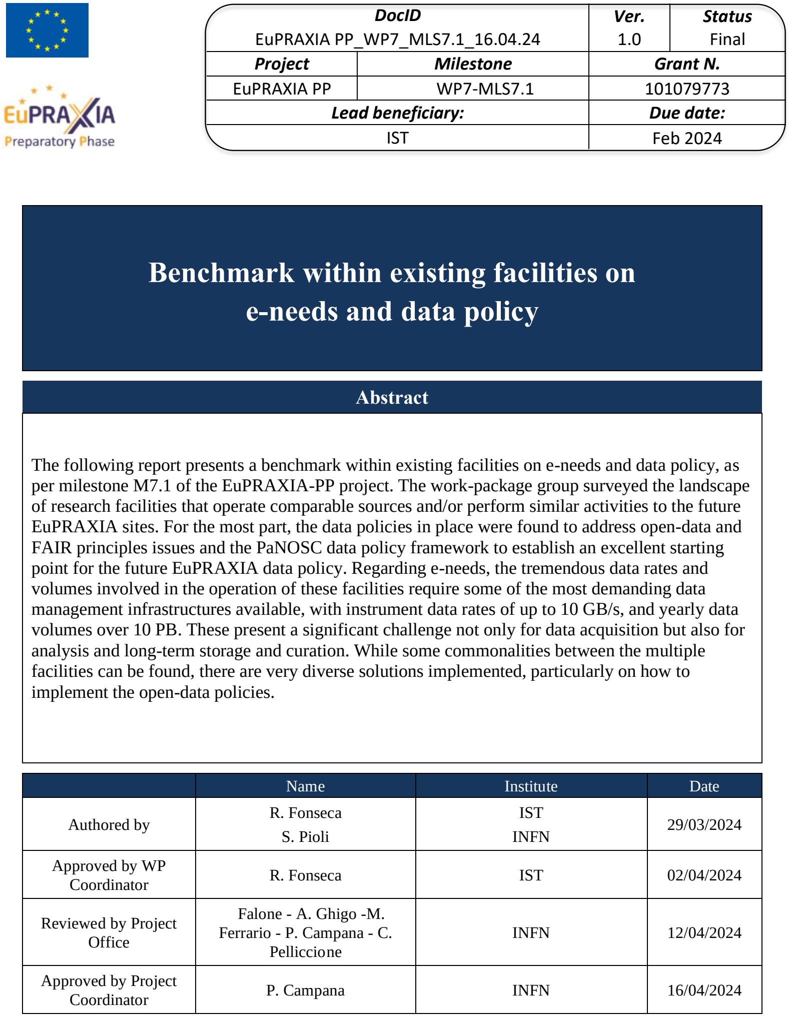
- **Key points**
  - The PaNOSC data policy framework covers in detail the data policy requirements for EuPRAXIA and similar facilities, in particular concerning **open data** and **FAIR principles**:
  - It makes **data openly available after an embargo period**, using an "appropriate license";
  - It establishes the **facility as the curator** of the data;

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

Deliverable: D2.1 - PaNOSC data policy framework

Data Policy

- **Data publication**
  - It proposes that the facility **publishes the metadata in an online catalog** 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**.



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.

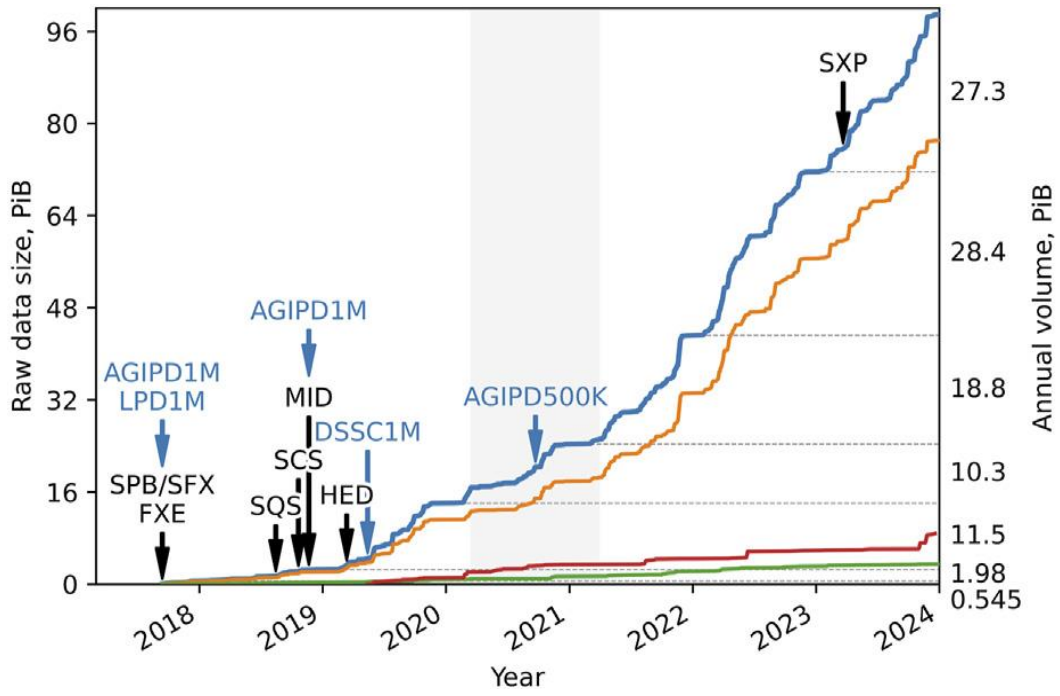
	Name	Institute	Date
Authored by	R. Fonseca S. Pioli	IST INFN	29/03/2024
Approved by WP Coordinator	R. Fonseca	IST	02/04/2024
Reviewed by Project Office	Falone - A. Ghigo -M. Ferrario - P. Campana - C. Pelliccione	INFN	12/04/2024
Approved by Project Coordinator	P. Campana	INFN	16/04/2024

E-needs

Funded by the European Union

- **Data storage**
  - Data rate 1 - 10 GB/s (per instrument), ~ 10 - 20 GB/s (complete experimental setup), typically for 30 minutes acquisition;
  - 5 – 10 PB total data volume per year (typical operation values). Storing for 10 years requires ~ 50 - 100 PB storage space;
  - Data reduction techniques may reduce these numbers by ~ 10x;
  - Storage system 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;

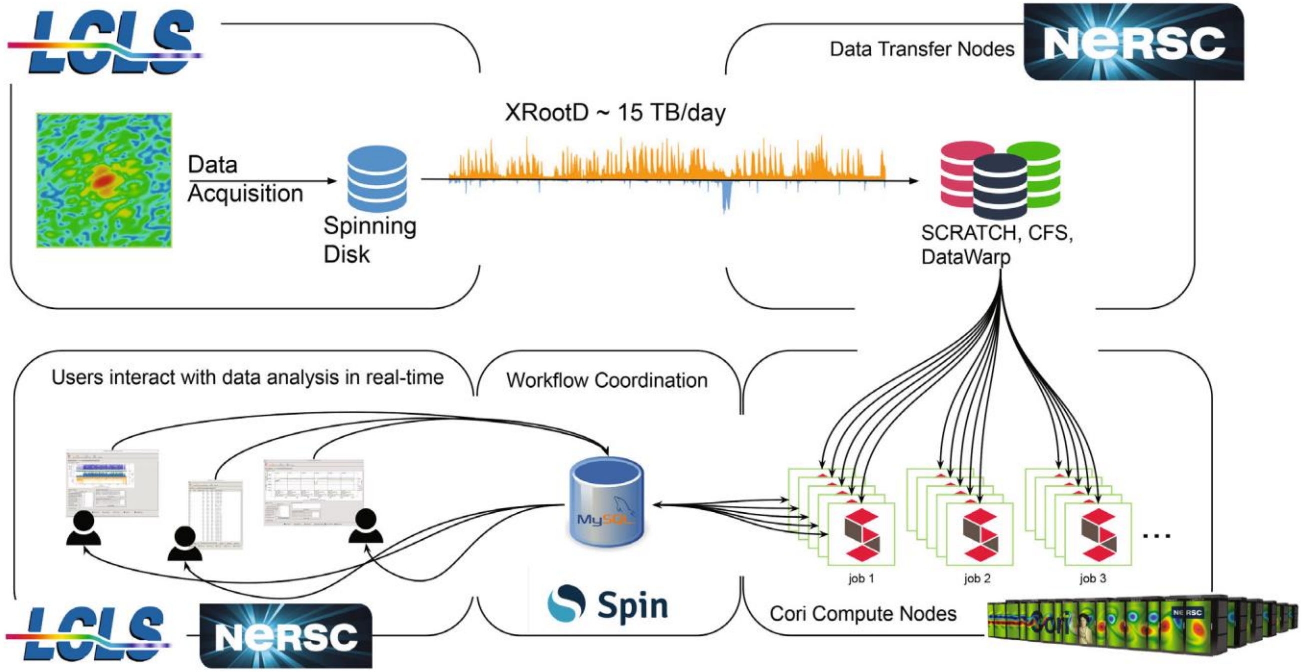
Collected data @ EuXFEL



Sobolev et al., Frontiers in Physics, 2024

E-needs

- **Data processing**
  - Two categories of data processing: Real-time data processing (~ seconds), and offline data processing (saved data);
  - Electronic logbooks are implemented in several facilities, in connection with the data policies;
  - Most facilities have an on-site computing cluster for offline data processing. Estimated requirements are on the order of 1 - 10 PFLOPS.
  - To lower this requirement EuPRAXIA sites may collaborate with external supercomputing facilities.
  - Most systems use some form of integrated systems control and data acquisition. EPICS and Tango controls are the two most popular choices.



Blaschke et al., Concurrency Computation Practice and Experience, 2024

EuPRAXIA Preparatory Phase

On-going and Future work

WP7 – e-needs and data policy

Report integration into European facility landscape and standards

- **Due date**
  - Due M30 – In progress

- **Description**
  - The existing European e-data structures and tools are analysed with respect to the needs of the EuPRAXIA-RI, when it enters full user operation. The basic EuPRAXIA e-data strategy and its alignment with the existing landscape is proposed.

Funded by the European Union

Deliverable 7.2

- **Data related standards**
  - Data interoperability standards
  - Data policy standards
  - User DMP standards

- **Connection to European facility landscape**
  - PaNOSC
  - LEAPS
  - External HPC centres?

Final report e-needs and Data Policy

- **Due date**
  - M42 – not started yet

- **Description**
  - The e-needs and plans for the acquisition, handling, processing and storage of EuPRAXIA data are described. The distributed nature of the facility, its various types of users and data exchange needs are taken into account, as well as long term storage and the FAIR policy. Structures are defined at institutional, regional, national and European levels.

Final report e-needs and Data Policy

Funded by the European Union

- **Plans for deliverable**
  - Prepare a Data Policy proposal based on the PaNOSC data policy Framework
    - To be coordinated with the access policy for EuPRAXIA sites
  - Prepare a conceptual design of e-infrastructre
    - Connection with trusted repositories required
    - Will depend on the instruments / beamlines
  - Review the data management plan with input from all relevant WP
    - Re-inforce compliance with FAIR principles

Overview

WP7 – e-needs and data policy

Workpackage Activities Overview

- **Progress**
  - WP completed 1 deliverable (operational DMP) and 1 milestone (benchmark with similar infrastructures)
  - Next deliverable is due M30
  - No foreseeable issues, should be delivered in time

- **Concerns**
  - WP work sometimes “out-of-phase” with other WP; most questions are answered as “too soon to tell”
  - The decision on the second site will greatly influence many options
  - Some WP7 decisions have too much overall impact to be handled at WP level

Open-science challenges

- **Data rates and volumes present an outstanding challenge**
  - Operation of similar facilities require some of the most most demanding data management infrastructures available
  - Instrument data rates of up to 10 GB/s producing over 10 PB/y
- **Not just a data acquisition problem**
  - Data analysis requires HPC level resources
  - Long-term storage and curation has the largest demands in terms of hardware, policy and management
- **As whole, the community is embracing open-data and FAIR practices**
  - The PANOSC data policy recommendations are gradually being adopted
  - Very diverse solutions are implemented, and not all facilities have fully embraced open-science

EuPRAXIA
Preparatory Phase

Discussion

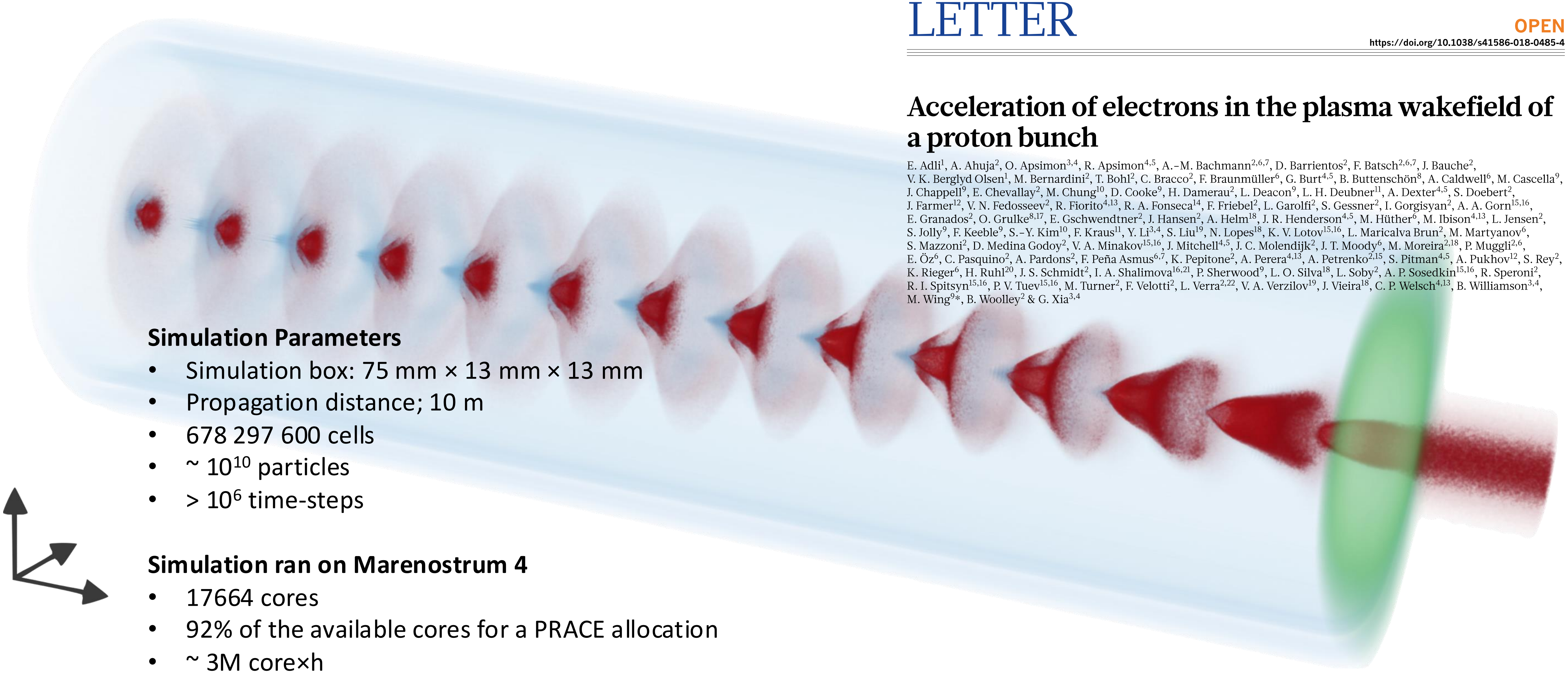
WP7 – e-needs and data policy

Discussion

- **Open-science challenges for EuPRAXIA**
  - Are EuPRAXIA users ready to make their data open and who will pay for this?
  - Where would we host the data / metadata?
  - Should our strategy change to being “towards open-science”? (i.e. how should we schedule the implementation of full open-science practices)

- **Integration with European landscape**
  - (When) Should EuPRAXIA try to join PANOSC?
  - Should we use an external data repository, and if so which one?

Full-scale modelling of the AWAKE* experiment



LETTER

OPEN

https://doi.org/10.1038/s41586-018-0485-4

Acceleration of electrons in the plasma wakefield of a proton bunch

E. Adli[1], A. Ahuja[2], O. Apsimon[3,4], R. Apsimon[4,5], A.-M. Bachmann[2,6,7], D. Barrientos[2], F. Batsch[2,6,7], J. Bauche[2], V. K. Berglyd Olsen[1], M. Bernardini[2], T. Bohl[2], C. Bracco[2], F. Braunmüller[6], G. Burt[4,5], B. Buttenschön[8], A. Caldwell[6], M. Cascella[9], J. Chappell[9], E. Chevallay[2], M. Chung[10], D. Cooke[9], H. Damerau[2], L. Deacon[9], L. H. Deubner[11], A. Dexter[4,5], S. Doebert[2], J. Farmer[12], V. N. Fedosseev[2], R. Fiorito[4,13], R. A. Fonseca[14], F. Friebel[2], L. Garolfi[2], S. Gessner[2], I. Gorgisyan[2], A. A. Gorn[15,16], E. Granados[2], O. Grulke[8,17], E. Gschwendtner[2], J. Hansen[2], A. Helm[18], J. R. Henderson[4,5], M. Hüther[6], M. Ibison[4,13], L. Jensen[2], S. Jolly[9], F. Keeble[9], S.-Y. Kim[10], F. Kraus[11], Y. Li[3,4], S. Liu[19], N. Lopes[18], K. V. Lotov[15,16], L. Maricalva Brun[2], M. Martyanov[6], S. Mazzoni[2], D. Medina Godoy[2], V. A. Minakov[15,16], J. Mitchell[4,5], J. C. Molendijk[2], J. T. Moody[6], M. Moreira[2,18], P. Muggli[2,6], E. Öz[6], C. Pasquino[2], A. Pardons[2], F. Peña Asmus[6,7], K. Pepitone[2], A. Perera[4,13], A. Petrenko[2,15], S. Pitman[4,5], A. Pukhov[12], S. Rey[2], K. Rieger[6], H. Ruhl[20], J. S. Schmidt[2], I. A. Shalimova[16,21], P. Sherwood[9], L. O. Silva[18], L. Soby[2], A. P. Sosedkin[15,16], R. Speroni[2], R. I. Spitsyn[15,16], P. V. Tuev[15,16], M. Turner[2], F. Velotti[2], L. Verra[2,22], V. A. Verzilov[19], J. Vieira[18], C. P. Welsch[4,13], B. Williamson[3,4], M. Wing[9]*, B. Woolley[2] & G. Xia[3,4]

Simulation Parameters

- Simulation box: 75 mm × 13 mm × 13 mm
- Propagation distance; 10 m
- 678 297 600 cells
- ~ 10^{10} particles
- > 10^{6} time-steps

Simulation ran on Marenostrum 4

- 17664 cores
- 92% of the available cores for a PRACE allocation
- ~ 3M core×h
- includes everything: diagnostics, smoothing, …

* E. Adli et al, **Nature** 561, 363-368 (2018)