1st Conference on Research Data Infrastructure

Connecting RDM

https://doi.org/10.52825/CoRDI.v1i.339

Published: 07 Sept. 2023

Connecting Infrastructures: The Physical Sciences Data Infrastructure (PSDI) in the UK

Juan Bicarregui[1][https://orcid.org/0000-0001-5250-7653], Simon J Coles[2][https://orcid.org/0000-0001-8414-9272], Brian Matthews[1][https://orcid.org/0000-0002-3342-3160], Jeremy G Frey[2][https://orcid.org/0000-0003-0842-4302], Barbara Montanari[1][https://orcid.org/0000-0001-8654-9181], Vasily Bunakov[1][https://orcid.org/0000-0003-3467-5690], and Nicola J Knight[3][https://orcid.org/0000-0001-8286-3835]

[1] Scientific Computing Department, Science and Technologies Facilities Council, UK

[2] School of Chemistry, University of Southampton, UK

Abstract. In this presentation we discuss the activities undertaken in the UK through the Physical Sciences Data Infrastructure (PSDI) initiative, part of the wider Digital Research Infrastructure (DRI) Programme. We will present the aims of the PSDI initiative, our initial scoping work and trials, and suggest how this project can and should interact with other related initiatives on a national and global scale.

Physical Scientists are crying out for a socio-technical data infrastructure that connects existing experimental and computational facilities. We believe that a cross-discipline and cross-technique digital infrastructure that builds on and bridges across existing initiatives, while these continue to serve their particular fields, is crucial to, and the best way to achieve global collaborations in the 21st century.

Keywords: Infrastructure, Physical Science

1. Introduction

In this presentation we discuss the activities undertaken in the UK through the Physical Sciences Data Infrastructure (PSDI) initiative, part of the wider Digital Research Infrastructure (DRI) Programme. We will present the aims of the PSDI initiative, our initial scoping work and trials, and suggest how this project can and should interact with other related initiatives on a national and global scale.

The data needs of research are growing at previously unimaginable rates and the need for collaboration around data has never been clearer. Data is not simply an output of research but is itself a driver of further discovery. Experiments, observations, computations and simulations all generate data. From simple manual annotations to complex simulations and terabytes of measurements from bespoke equipment, data flows are the very fabric of research but are currently hampered by technical and social problems related to data discovery, access, integration, processing, curation and publication.

Physical Scientists are crying out for a socio-technical data infrastructure that connects existing experimental and computational facilities. We believe that a cross-discipline and cross-technique digital infrastructure that builds on and bridges across existing initiatives, while these continue to serve their particular fields, is crucial to, and the best way to achieve global collaborations in the 21st century.

The aim of PSDI is to enable researchers in the physical sciences to handle data more easily by connecting the different data infrastructures they use. PSDI will connect and enhance existing infrastructure in Physical Sciences.

Through PSDI researchers will be able to:

- Find and Access to reference quality data from commercial and open sources
- Combine data from different sources
- Share data, software and models including experimental and simulation data
- Use AI to explore data
- Learn how to make the results of their research open and FAIR

2. Statement of Need

The PSDI project was initially discussed as part of the EPSRC Large Infrastructure Investments Statement of Need (SoN) call, which was submitted in early 2021[1]. During this SoN exercise a project team from STFC and the University of Southampton developed the outline plan for the PSDI. This included commentary on the ambition of the project and the strategic importance of investment in infrastructure for the physical sciences. This SoN exercise was well supported across the physical sciences community. Contributions and backing from a wide range of projects and initiatives demonstrated a community need and support for such an initiative.

The SoN confirmed a widespread consensus in the community that investment in research data infrastructure is lagging behind investment in data sources and identified an urgent need for integration of data and computational infrastructures. It identified four major ‘pillars’ of user communities in the UK that would benefit from the proposed PSDI. (Figure 1)

- Pillar 1. Facilities, Institutes and Hubs – significant centralised national facilities and activities that serve many researchers based on a common need.
- Pillar 2. National Research Facilities – medium-scale centralised facilities operating at a world leading level to perform research that cannot be addressed in a standard laboratory.
- Pillar 3. Computational Initiatives – uniting performing simulations with the communities and tools required to do so.
- Pillar 4. Research Institutions, research groups and laboratories.

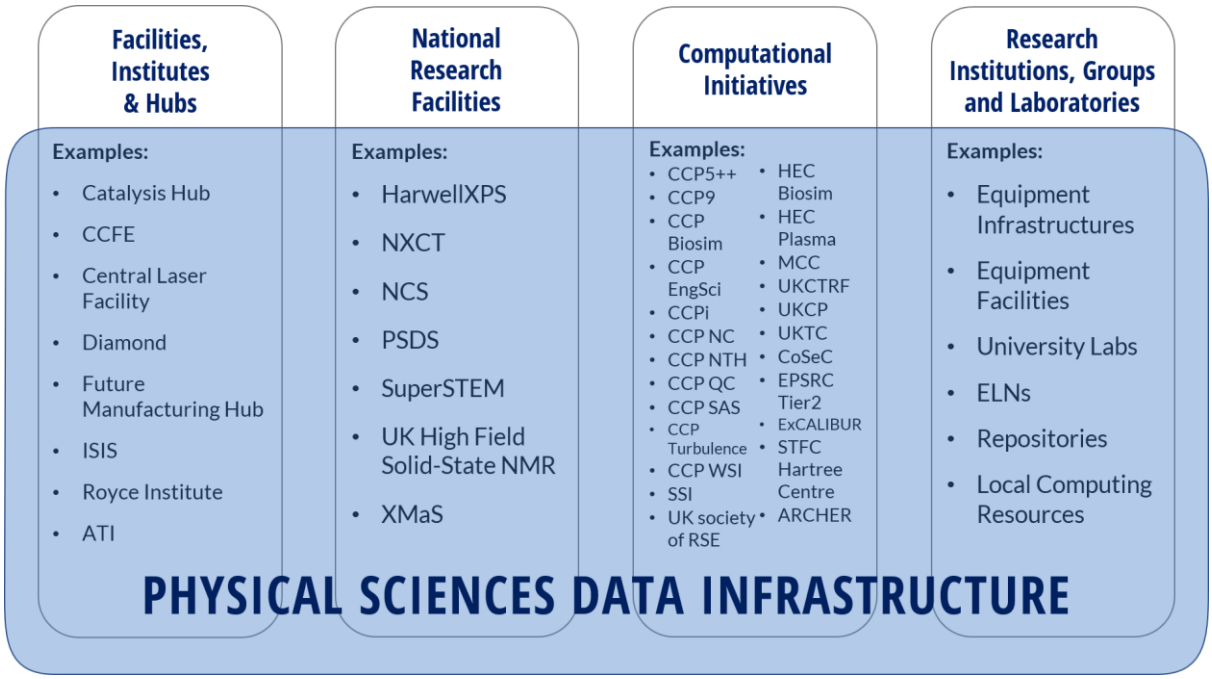

Figure 1 4 Pillars of UK User Communities producing and using data in UK Physical Science

3. PSDI Pilot

Following on from the SoN application, the PSDI team were requested by EPSRC to complete a proposal for a short phase pilot project, funded through the UKRI DRI programme. The PSDI is a large undertaking, involving a wide range of stakeholders within our proposed pillars and the wider community including data managers, data providers, system architects and many more roles that underpin our national data landscape. This pilot ran for 5 months in late 2021 – early 2022 and expanded on the ambitions of the project from the SoN. In this pilot PSDI undertook a wide range of community consultation on the scope and requirements of the PSDI, and planning for the future of PSDI. (Figure 2)

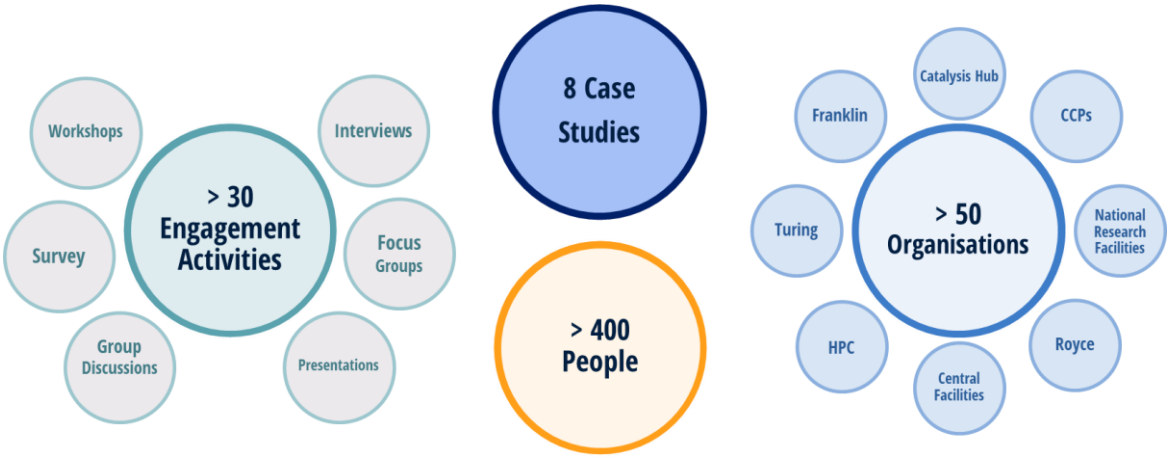

Figure 2 A summary of the PSDI Pilot phase activities

As a result of the PSDI pilot a series of reports [2,3] and recommendations [4] were published. There were 13 recommendations across 4 areas, summarised below:

- Connecting existing infrastructure: connecting existing research data services, support beyond the lifespan of individual projects, co-operation and co-creation between all stakeholder organisations
- Best Use of Data: developing a toolkit for publishing, access to provenanced data, tools for reproduceable data processing, support for transforming data to knowledge
- Best Use of People: co-ordination for community activities and input, community training and support, professionalisation for data roles, governance structure for PSDI
- Best Use of Technology: services to connect existing provision (data and services), adopt existing technologies

4. Current Phase

In 2023 PSDI has undertaken further scoping and prototype development work towards the realisation of the PSDI vision. As part of this phase 5 pathfinder activities are underway in the areas of: Catalysis, Process recording, Data collections, Bio-molecular simulations and Data to Knowledge.

As PSDI develops beyond into the future, interaction with other data initiatives across UKRI, and the wider international sphere, will continue to be a central focus. The components that drive science forward, such as data, standards, and research developments are not limited to a single research community or country but require connection and integration on a global scale.

The life sciences have had considerable financial input to create global infrastructure to support data sharing. Examples include the Protein Data Bank (PDB), Gene sequences, etc. The ability to access the protein structures was a key aspect in the success of DeepMind's development of AlphaFold2 and the need to share sequence information was key in the WHO's work in tracking the COVID-19 global pandemic. Particle Physics has developed a global computational infrastructure to support the specific types of data produced by organisations such as CERN. However, the wider, longer tail, smaller research group based Physical Sciences are ca, 20 years behind these efforts and PSDI together with other international initiatives seeks to make a start on bringing our community the benefits.

Data availability statement

The content of this presentation is not derived directly from specific datasets. However, all reports from the pilot phase are available through the PSDI zenodo community https://zenodo.org/communities/psdi

Underlying and related material

N/A

Author contributions

All authors were involved in conceptualization, funding acquisition and project administration. NJK prepared the original draft, and all authors contributed to the review & editing of the submission.

Competing interests

The authors declare that they have no competing interests.

Funding

PSDI is funded by EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1

Acknowledgement

We acknowledge those members of the community who contributed to our Statement of Need exercise, who carried out work in the PSDI pilot (https://www.psdi.ac.uk/the-pilot/team/), or current phase of our work (https://www.psdi.ac.uk/people/).

References

1. "Physical Science Data Infrastructure Statement of Need." PSDI. https://www.psdi.ac.uk/wp-content/uploads/2022/01/PSDI_SoN-V1.0_OneDocument_LargeInfrastructureInvestments.pdf (accessed: 21/04/2023)
2. N J Knight, J Bicarregui, B Montanari, S J Coles, J G Frey, B Matthews, & V Bunakov. (2023). Physical Sciences Data Infrastructure Phase 1 Pilot Report. Zenodo. https://doi.org/10.5281/zenodo.7684860
3. "Physical Sciences Data Infrastructure Community." Zenodo. https://zenodo.org/communities/psdi (accessed: 21/04/2023)
4. "Pilot Recommendations." PSDI. https://www.psdi.ac.uk/the-pilot/recommendations (accessed 21/04/2023)