Physical Science Data Infrastructure

Statement of Need

V1.0 - 15th February 2021

1. Ambition and Vision

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 subsequently becomes 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 in and across the physical sciences but are currently hampered by technical and social problems related to data discovery, access, integration, processing, curation and publication. Physical science is crying out for a socio-technical data infrastructure that connects existing experimental and computational facilities.

As research strives towards the fabrication of new materials and devices that benefit society, the need to address multi-physics problems requires the integration of data from many different techniques and a wide range of length and time scales, from atomic resolutions through to complete devices. While some advances can be made through relatively small-scale target-driven initiatives, accelerating the rate of discovery requires exploitation of accumulated data, for example through machine learning, and simulations to inform and focus the experimental elements of research. Physical science must transition to become a digitally driven and fully digitally enabled discipline.

This need is being addressed in parts, for example, the *Materials 4.0* Big Idea highlights the need for a "data centric framework for materials" and cyber-physical systems for materials discovery and manufacture. In the US, the multi-agency *Materials Genome Initiative* (MGI) aims to discover, manufacture and deploy advanced materials at twice the pace and a fraction of the previous cost. In the 5 years to 2016, the DoE, DoD, NSF, NIST and NASA invested over $500 million addressing themes of 'integrate experiment, computation & theory' and 'make digital data accessible'. This established infrastructures that have led to new materials and devices. A 2018 analysis calculated that a technology infrastructure to support the MGI could deliver $123-270 billion annually.

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 achieving global competitiveness in the 21st century. An example is the Catalysis Hub, a challenge-led centre bridging academic institutions, central facilities and high-performance computing, the first phase of which was hugely successful in publications and specialist training. However, it was recognised that its 2nd phase requires the support of a data infrastructure to capitalise on its results and it has invested in support for tracking its publications, and in virtual research environments to better support data management and analysis. Collaborations such as this would benefit from additional capability to share and integrate data pipelines in a systematic and sustainable manner.

Progress in physical science research not only requires leading-edge experimental research infrastructures, such as appropriately equipped laboratories, instruments and facilities, but also computational resources, such as data handling and analysis tools, modelling software and computing resources close to the data origin. Data from experimental infrastructures that support common enabling techniques such as diffraction, microscopy and spectroscopy, are often manually linked with complementary computational simulations. Support for workflows with effective data handling, that routinely link data from different sources with analytic resources, would reduce this time-consuming manual effort. A tailorable integrated environment, aggregating and automating the processing of data, and combining experimental and simulation data with effective machine learning methods, would drive integrated, multi-scale physical science to new levels.

The Physical Science Data Infrastructure (PSDI) would address these challenges. It would:

- Support multiscale modelling and multimodal research
- Leverage simulation data to drive experimental science and vice versa
- Surface data from many sources
- Provide reference-quality data
- Standardise, normalise and aggregate data and metadata
- Enable data to be exploited by AI methods
- Support workflows that automate data processing
- Provide a common platform to run models and codes from different sources
- Seamlessly access performance compute for scaling up
- Enable software curation and publication
- Be a place for curation of legacy beyond individual projects

In addition to the *ascending route of data-intensive research* that proceeds from microscale through mesoscale and constructive elements to devices on macroscale, it is notable that practical exploitation of materials and engineering solutions is also data-intensive. Ubiquitous sensors and Internet of Things, coupled with massive edge and back-end computing, have enabled *digital twins* for industrial equipment, smart cities, and civil infrastructures which are also rich sources of data. Feeding such real-world data back to physical science can be called the *descending route of data-intensive research*. Whilst initially focussing on the ascending route, PSDI will be designed for a future match with the descending route where new grand challenges and opportunities in data-intensive physical science are likely to emerge, making a virtuous circle of research and innovation complete.

In short, the PSDI will significantly improve UK physical science researchers' ability to share and combine data, analysis tools and computation models. It will complement existing institutional and cross-institutional facilities, offering a common data and computation space with multi-directional connections between data, tools, models and computation.

2. Strategic Importance and Context

Today, each physical science research infrastructure, from laboratory to large facility, has essentially its own data infrastructure. In contrast, in other domains data-centric infrastructures for collecting and reusing data act as community hubs and drivers of new methods and discoveries. The NERC data centres have driven their sub-disciplines for decades and are now aggregating into one service to further leverage their data assets. The UK-driven European Bioinformatics Institute (EBI) aggregates, coordinates and structures heterogeneous data, forming a foundation of the new field of data-driven biology. The current pandemic accentuated a need for further integration of data resources and EBI's COVID-19 Data Platform rapidly became a hub for the global response. Conversely, today's diversity of data infrastructures enables each platform to be tailored to the specific needs of its field and so it is critical that a shared resource does not become a master of none. The need is thus for an additional infrastructure layer to share existing resources whilst ensuring that each can remain dedicated to its specific application.

There is a need to preserve and exploit outputs from past research while keeping pace with the increasing rate of data generation, the latter posing the greatest challenge and potential for innovation. New chemicals, materials and devices are key to a sustainable future, both environmentally and financially. The UK needs to invent its way out of seemingly conflicting targets of maintaining economic growth whilst making unprecedented strides towards an imminent net zero carbon output. The proposed infrastructure will enable the wider community to do more with existing resources and build an ecosystem for discovery by augmenting them with a new integrating layer.

The PSDI will address a key element of the UK Research & Development Roadmap. It will *"develop our digital research infrastructure capability … by building an internationally leading national digital research infrastructure"* and will include a *"UK platform for experimental data analysis" as part of "a diverse network of infrastructure: internationally competitive, high-quality and accessible facilities, resources, data and services."*

Our vision also aligns closely with other Roadmap goals, particularly to *invest in new and existing infrastructure more flexibly and efficiently;* to *make the most of our capabilities to ensure they truly become national assets;* to *ensure these assets are on a sustainable financial footing;* and to provide *a step change in the capability that research and innovation infrastructure enables to promote economic growth and discovery across sectors and disciplines.* PSDI will also provide key infrastructure for UK's Exascale programme, ExCALIBUR, in its strategy to increase science productivity, bring about transformational change in capability and increase the skills base of interdisciplinary research software engineers.

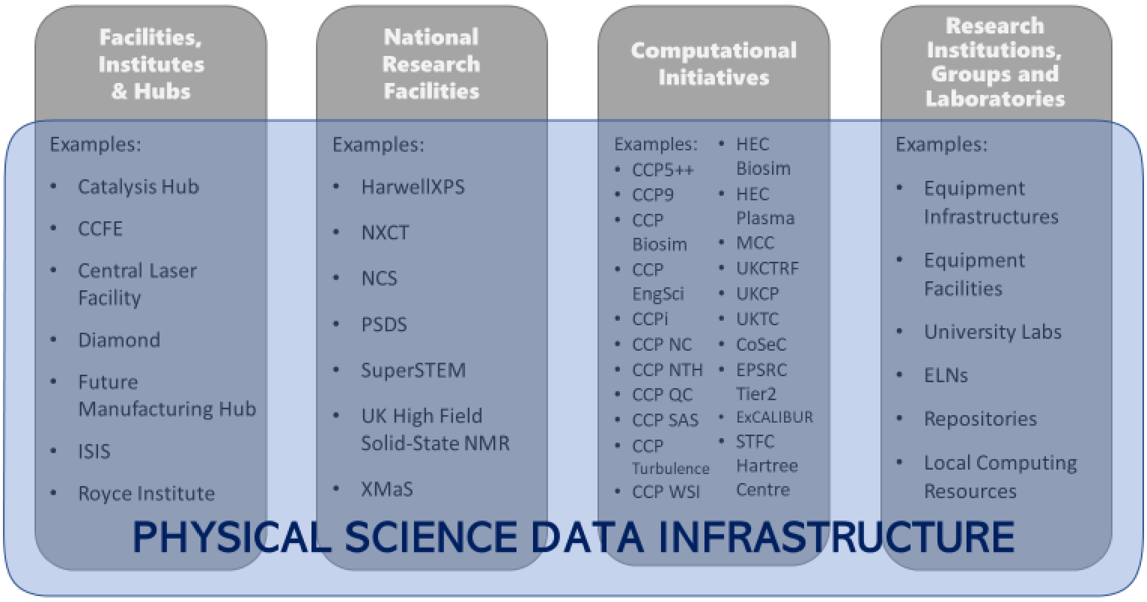
Initiatives outside the UK provide useful baselines for comparison, cooperation and lessons learned. PSDI will connect with the US Materials Genome Initiative and the NIMS initiatives in Japan particularly around mutual data sharing for wider benefit and common data standards and policies for automated exchange and reuse. European data infrastructures, such as E-CAM, MaX and NOMAD are also related; NOMAD is data-focussed, while E-CAM and MaX offer data capabilities as add-ons to computational science software and HPC. The PSDI, alternatively, offers a common data space across the pillars described in Sections 3 and 5, rather than highly specialised resources tightly coupled to particular fields or technologies. The German National Research Data Infrastructure (NFDI) and the Australian Research Data Commons (ARDC) are national multi-disciplinary programmes to harmonise management of research data that bring together multiple stakeholders to establish common data handling procedures, cross-disciplinary metadata standards and interoperable data services to increase reusability of data within and across subject boundaries. The European Open Science Cloud (EOSC) aims to provide a web of FAIR data which `federates the gems' of European research infrastructures. The EOSC Strategic Research and Innovation Agenda (SRIA) identifies fourteen areas of activity which will be pursued to achieve its vision. These provide a template of potential actions from which PSDI can select. The PSDI will differ from NFDI, ARDC and EOSC as its initial focus will be on linking and sharing across a specific set of existing physical science resources, thus enabling more focussed solutions to be developed.

The PSDI will link the data handling capability of UK world leading infrastructures with each other and internationally. It will provide a route to align physical science research outputs with UK Industrial Strategy programmes (ISCF, SPF, SIP) and other industrially focussed initiatives eg. Innovate UK-funded Catapult and Accelerator Networks and STFC's Hartree Centre and campus initiatives. It will cross-fertilise between existing infrastructure resources and so enable researchers to merge results from different areas of science to achieve breakthroughs in knowledge and understanding. It will facilitate the application of that knowledge to maximise the economic, environmental and societal impact of research, promoting innovation across the economy. New chemicals, materials and devices are key to an environmentally and financially sustainable future. By augmenting existing resources with a new integrating layer, the proposed infrastructure will become an ecosystem for such discoveries.

3. Users and Community Engagement

An integrated, national data infrastructure would support all sizes of research team from the single doctoral student-supervisor, through small and medium sized teams, to large multidisciplinary communities addressing grand challenges. The current focus is mainly on EPS researcher needs; future expansion, particularly with the providers of national compute platforms and data initiatives, and into other research domains, would lower the barriers to interdisciplinary research so that, in the longer term, the PSDI will serve users beyond the physical sciences.

Our consultations (see section 5) confirmed a widespread consensus in the community that investment in research data infrastructure is lagging behind investment in data sources and an urgent need for integration of data and computational infrastructures. In preparing this SoN therefore, identifying a community need for PSDI was straightforward; rather, the challenge is how to articulate the ubiquitous need in a way that will enable support to grow effectively and efficiently through targeted investments. We therefore propose an



incremental approach, that initially brings together a critical kernel of data sources in the following four pillars.

Pillar 1. Facilities, Institutes and Hubs – significant centralised national facilities and activities that serve a large number of 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 – the community of institutions with strong research profiles and a degree of data management and computing capabilities.

Examples of community demand are:

Pillar 1

Diamond, ISIS and the Central Laser Facility together have a user base of ~10,000. Facilities users often begin a study informed by reference quality data in a university lab or computer system, expand it with experimental techniques at the Facilities and/or complement by performing complex simulations using High Performance Computing. Each part generates data needing to be managed and integrated and often extracting knowledge from experimental data requires computational modelling. Finally, data must be curated and available to the research community for re-use and re-discovery. PSDI will build on the Facilities-specific data infrastructure being developed by them and the Ada Lovelace Centre to integrate with other experimental and computing initiatives.

Digital twinning is key to building more complex devices. Aggregation of quality data sets is necessary to build an actionable model of a fusion power plant proposed by CCFE STEP which maintains integrity of the plasma and associated systems, vital for successful commercialisation. Twins are key in the Brunel Challenge to deliver transformational changes to UK industry, requiring marriage of diverse data sets from different disciplines, integrating modelling data with streaming data from sensors and using AI to modify vessel behaviour.

The Materials Innovation Factory (MIF), an £82M investment collocating robotic equipment, 200 academics and 100 industrialists, produces significant data. Focussing on computational and experimental/robotic automated materials discovery, MIF exemplifies a need for data infrastructure in a facility spanning multiple academic and industrial users. The ERC ADAM project with the MIF generates very large amounts of computational and experimental data, often on the same materials / molecular systems and at different sites. Centralised storage and data analysis would benefit ADAM and the community by facilitating the opening of data sets for additional analysis.

Pillar 2

There is strong demand for making data discoverable. Highly valuable reference-quality data e.g. Cambridge Structural Database, Inorganic Crystal Structure Database, exist but are not ubiquitous and the majority of data is undiscoverable and unpublished. Despite the high cost of acquiring data, it is often discarded after fulfilling its primary use e.g. quality control data immediately discarded, academic research data lost at project end. Imaging communities particularly highlight these issues e.g. NXCT & SuperSTEM produce huge quantities of data typically used to derive only small amounts of information for a specific purpose and the remaining information stays hidden. Royce is developing digital fingerprints to turn GB of data into simple representative principal components to make features recognisable, cross-analysable and enable ML methods. If PSDI provided resource and know-how, as a first step the NRFs listed (~2000 research groups supported) would make available all experimental data in publications and PhD theses as a precursor to engaging wider communities.

Pillar 3

A large part of the UK computational researchers in the EPS domain are organised in recognised communities e.g. CCPs and HECs, which together count well over 3,500 members. The data needs of these communities range from data management and curation to integration with national compute and experimental Facilities (examples in Sec 5). Several of these communities have started to build specific data infrastructures e.g. CCP-NC, MCC, and collaborate with international efforts eg. OPTIMADE, NOMAD. PSDI would also link to wider HPC including petascale facilities such as DiRAC to maximise data sharing and exploitation.

Pillar 4

The Royal Society of Chemistry's Digital Futures Report illustrates the importance of infrastructure *"in harnessing data and digital technologies for discovery and innovation in* *the physical and life sciences"*. This provides additional insight from researchers which mirrors the requirements gathered from the community in support of PSDI.

4. Description of the Infrastructure and what it will enable

The PSDI will be a coordinating data infrastructure. It will provide a means to connect existing mature infrastructures as well as supporting key laboratories and facilities that have little data infrastructure at present. PSDI will not only be a hardware platform, but a distributed range of tools, services and environments with user support, that enhances the functionality of the existing components. The essential components of PSDI include a *service environment; repositories & catalogues; data curation capability; Open Science capability; Machine Learning environments; compute & storage resources – all facilitated by a community support programme*. Section 6 provides more detail. This environment will deliver:

- **Access to reference quality data** through commercial and open sources with data curation services; the current PSDS provides a foundation.

- **Data, software and model sharing** including experimental, computer simulation and long tail data; centrally catalogued so it is findable and able to move data to computational resources or compute to data as required.

- **Combining data from different sources** across research data lifecycle including data collection, data reduction, data analysis & data sharing, giving provenance.

- **Close-to-data computation and containerisation of data and software**. This allows orchestration of data and computation components and their assembly in loosely coupled workflows for specific use cases, while maintaining code independence.

- **Artificial Intelligence in research and innovation.** Machine Learning has huge potential in physical science but is critically dependent on access to learning data. We will work with AI centres, such as Turing, to ensure PSDI fully supports this approach.

- **Open Science.** Common data capabilities enable Open and FAIR Science. PSDI will make data much more findable, accessible, interoperable and reusable, facilitated via policy, standards, licensing, environment and guidelines.

The choice of pillars has a wide reach, not only to research but also for innovation with industry. These additional features will provide immediate benefits to those currently using one or more of the four pillars. Further, the championing of FAIR data standards will mean wider data accessibility across all sectors.

Historical investment in data support has been localised and piecemeal. This has not, and will not, achieve the full potential of our data. Evidence, such as the US approach to materials discovery, shows that inclusive integration and coordination through infrastructure is a way to make enormous gains. Experience in other areas of UKRI's portfolio, such as DAFNI for Infrastructure Systems research and in environmental and biological science, illustrate the benefits of the approach.

If the PSDI is not built, continued investment will not only be financially suboptimal but, more importantly, will miss opportunities that arise from "collection". Existing pockets of excellence will operate with low levels of interoperability, but the collective leverage and capability of assets will not be achieved. The lack of a multi-scale common data infrastructure would limit the potential for automated, or machine-assisted, data exchange, analysis and reuse. Interaction between physical equipment and software, or between experimental and simulated studies, would remain essentially manual, thus delivering more slowly. Other nations, including the US and EU countries that are making considerable investments in this area will direct global research practice and cause the loss of UK leadership.

An incremental and flexible approach is necessary to build such a large and diverse ecosystem. PSDI will not replace the current excellent infrastructures that need to continue to function and develop whilst they are weaved into a greater whole. However, it is imperative to begin with a sufficient critical mass to demonstrate the benefits and assess its impact.

This approach has driven the choice of the four pillars. They have common requirements, where PSDI can provide considerable leverage, whilst each has its own characteristics that can drive the overall vision. Once PSDI is established and the benefits quantified, further expansion to include other facilities can be considered.

A core consortium with participation from subsets of each of the four pillars will be sufficient for the first stage. This core will operate with the clear purpose of delivering a platform which will demonstrate benefits from its current configuration, whilst also being ready for further expansion.

The PSDI will be governed by a management structure with three bodies ensuring strong engagement and effective delivery. These bodies are:

Steering Panel including representatives from each participating facility and the funders. The steering panel is responsible for setting priorities and approving the workplan.

Management Board including representatives from the four pillars and the delivery team. The Management Board is responsible for ensuring the effective delivery of the workplan.

Community Forum including user representative of each facility and other experts from across the physical science research community. The forum will have ongoing engagement with user meetings, feedback channels, etc. It is advisory to the Steering Panel and Management Board.

Major Risks	**Mitigations**

Data security	Application of data security and access policies, practices and technology
Lack of adoption	Adopt current standards and practices. Community involvement in prioritisation and design
Degradation of existing functionality	Awareness that one size does not fit all. Continuation of dedicated funding for participant facilities
Reluctance to support data sharing	Promote the cultural shift to FAIR data sharing, while respecting IP rights.

5. Authors and Community Engagement

Scientific Computing Department, STFC

- Barbara Montanari, Head Computational Science and Engineering Division, Director of Computational Science Centre for Research Communities, CoSeC
- Juan Bicarregui, Head Data Division, Leader in Open Science policy, EOSC & RDA
- Brian Matthews, Leader Data Science and Technology and Co-I &Technical Lead PSDS
- Vasily Bunakov, Computational Scientist

University of Southampton

- Simon Coles, Prof. of Structural Chemistry, NCS Director, PSDS Director, GO-FAIR community lead, IUPAC FAIR Data Task Force, IUCr CommDat
- Jeremy Frey, Prof. of Physical Chemistry, PI AI3SD Network+
- Nicola Knight, PSDS Training and Engagement

The following submitted contributions and support.

Pillar 1

Name	Affiliation	Support for Need
Richard Catlow	UK Catalysis Hub	Catalysis Hub strongly supports this initiative. Data infrastructure developments are of growing importance and priority in the field
Shaun de Witt	CCFE	Aggregation of quality datasets is a key step to production of digital twins for complex experimental devices

Luke Benson-Marshall	Royce Institute U. Sheffield	There is a need to increase use of emerging technologies in research
Iain Todd	U. Sheffield Director EPSRC Future Manufacturing Hub Director Royce Translation Centre	This SoN has my support
Robert McGreevy	Director ISIS Facility	I strongly support this SoN
John Collier	Director Central Laser Facility	I am happy to support this

Pillar 2

Tom Hase	U. Warwick Co-director XMaS (ESRF)	Our NRF supports the PSDI
Steven Brown	U. Warwick PI UK High-Field Solid-State NMR NRF Warwick Analytical Science Centre & AS CDT	We would benefit from set up of repository of experimental NMR data associated with publications, starting with NRF data
Quentin Ramasse	U. Leeds SuperSTEM	Enabling local centres to contribute data into PSDI repositories and sharing code/software
Philip Davies	Cardiff U. HarwellXPS Director	I support this SoN
Philip Withers	U. Manchester Chief Scientist Royce Institute PI NXCT PI CCPi	PSDI would aid making data discoverable and allow more value extraction
Mark Wilkinson	U. Leicester DiRAC HPC Facility	Sharing data across communities, to facilitate comparison of simulations with observations and experiments is essential to maximise the scientific returns on investment in the DiRAC HPC resource

Pillar 3

Mark Savill	Cranfield U. CCP Steering Panel CCPEngSci UKTC UKAAC UKCOMES	The proposed PSDI would be of timely value to EPS, especially as exascale computing enables growth of multi-scale and multi-disciplinary simulation and design. There is a need to catch up with collaborating communities and unite data activities (big data, ML and AI) more broadly and uniformly across UKRI. Benefits will quickly emerge in data, modelling, and discovery through knowledge exchange, and multi-fidelity simulation campaigns that are thus enabled routinely
Neil Chue Hong	U. Edinburgh SSI ExCALIBUR Steering Committee	PSDI can aid collaboration by publishing curated data, models and code. Ability to exploit exascale HPC and AI/ML platforms by improved data access and portability. Enable open science through guidance and policies to aid reuse and exploitation of data, models and software
Arash Mostofi Scott Woodley	MMM Hub	Materials and Molecular Modelling crosses disciplines in physical sciences and beyond. Helping collaborations across the MMM community, experimental and industrial partners is key
Stewart Clark	CCP9	Electronic structure methods and their application to materials science requires data management, benchmarking, sharing of data, methods and codes, and useful data search across databases
Shuisheng He	CCP NTH	The nuclear community are interested in data management, sustainability, reproducibility and sharing and will benefit from storing and sharing large benchmarking datasets from HPC and experiments
Deborah Greaves	U. Plymouth CCP-WSI	Computational fluid dynamics and structural mechanics communities need tools to tackle complex simulations for wave structure interaction in coastal and offshore engineering
Paola Carbone	U. Manchester CCP5++	Network members need access to large, curated databases of experimental or computed data
Paul Hodgkinson	CCP-NC	CCP-NC has a repository of calculated NMR data which could directly benefit from PSDI

Andrew Morris		
Viv Kendon	CCP-QC	PSDI could streamline data access by quantum computers (eg NQCC) allowing more value extraction from datasets
Sarah Harris	CCPBioSim	Our community produces large simulation datasets but increasingly uses large experimental datasets as calculation inputs. Linking datasets and integrating experimental data would massively aid understanding of biological processes and biomaterials
Syma Khalid	HECBioSim	
Robin Pinning	Hartree Centre	PSDI will maximise exchange of ideas between EPSRC researchers and industry by increasing re-use, visibility and integration of research data in industry and vice-versa, supporting ISCF, SPF, SIP programmes and UK's Industrial Strategy
Scott Woodley	MCC	The materials discovery theme makes large-scale use of HPC generating data from many simulation techniques. Central data storage facilitates wider use of high-quality computational data
Nilanjan Chakraborty	UKCTRF	Increasingly complex simulations require an infrastructure for their expanding size
Matt Probert	UKCP	One UKCP theme is materials prediction generating large amounts of data, with members continually developing new ML approaches requiring large datasets for training. PSDI will be of great benefit, joining and sharing data at both experimental and HPC centres
Sylvain Laizet	UKTC CCP Turbulence	There is a need for a more practical solution for large scale data sharing

Pillar 4

Graeme M Day	U. Southampton ERC Synergy Grant ADAM Materials Lead MCC	Centralised storage and data analysis would benefit the community, facilitating open-sourcing of datasets for additional analysis by others
Richard Kidd	Royal Society of Chemistry	Increase data sharing in the chemical sciences, by links to publications and data resources such as ChemSpider and linkage

		through Open Science & Open Data activities

6. Indicative Resources

The PSDI augments rather than replaces, existing resources. It provides common platform and governance for the data infrastructure, establishes centres with enhanced resources for data curation and computation, and supports research across the community. A key aspect of the approach is community participation in the core activities and building consensus for a federated infrastructure. The PSDI has 3 programmes, with an indicative 5-year budget of £70M as follows.

Community Research: ~£5M/year

The community research programme would fund the integration and exploitation of the PSDI within physical science research projects, including:

- **Research Software Engineering:** supporting RSEs to build specialist data infrastructure and data analytics within research projects.
- **Community environments:** development of workspaces to access data and resources via the PSDI for research domains.
- **Pilot programme:** use cases to exploit the PSDI with physical science research projects.
- **Skills and Training:** developing data skills, including in CDTs.
- **Community network:** liaising with the community in the UK and internationally.

PSDI Centres: ~£6M/year

Centres provide support for data curation and computation within domain focussed areas of Physical Science. These would fund access to computational and data storage resources directly via hardware provision or via research or commercial cloud and provide dedicated expertise on data curation and computation within that domain. This would give a distributed repository of open-source data.

Computational resources would support the execution of codes from the community, via a containerized environment**.** Provision would include data exchange components, and support AI/ML. The ML would access high-data-throughput and GPC systems in cases when the ML needs are not covered by standard environments.

Core Programme: ~£3M/year

The Core programme would support common infrastructure development and deployment. It would include a **PSDI Coordination Hub**, which would work with the governance boards to identify common standards, policies and tools across the PSDI Centres and Community programmes. It would support software development and integration activities and provide common services across all the Centres, using existing best practise where possible (eg. UKAF, IRIS, DAFNI). The Core would support:

- **Platform software**: common tools for management of users, computation and data infrastructure across a distributed environment.
- **Repositories and catalogue tools**: repositories for datasets, software and containerised executables, including catalogues for data and software and other assets and cross-references to publications.
- **Data curation standards, tools, processes and expertise:** data ingest, quality, and integrity with processes to preserve data for the long-term.
- **Open Science standards, tools, processes and expertise:** data integration, exchange, searching, metadata & ontology standards, and licence management.