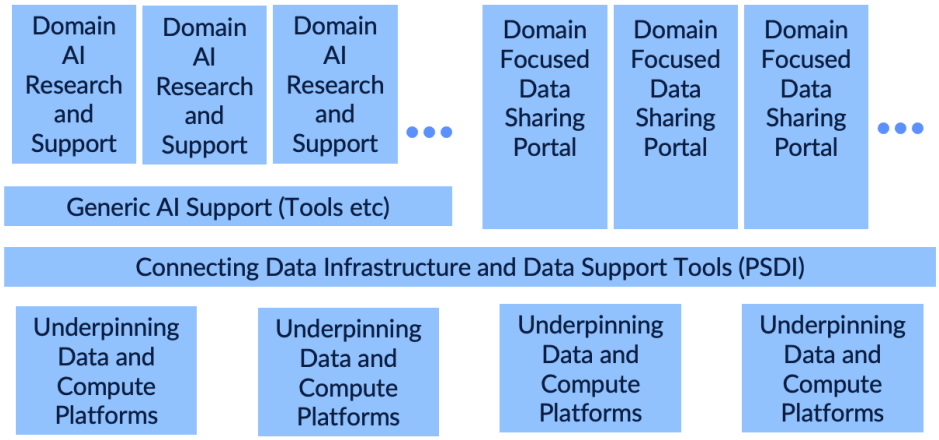
# PSDI and AI Hubs Discussion

## Summary of Meeting held 2 May 2023

## Background

PSDI aims to provide a connecting layer that enables sharing of data across different data infrastructures. As such it could also provide a platform that supports data management for the AI Hubs. A high-level view of how this might work is show in the following diagram. The purpose of the meeting was to explore whether this is a workable approach and how this idea could be made more specific.




*An abstract view of how PSDI could support the AI Hubs*

Following the announcement of the proposals that were successful at the first round of EPSRC's Artificial Intelligence Hubs call[1], PSDI invited representatives of all proposals to a meeting to discuss how PSDI can support the Hubs, and how the AI Hubs can work collaboratively.

In response to this invitation PSDI received replies from 19 of the 31 proposals, including all 12 of the science and engineering proposals, 4 of the foundations proposals, and 3 of the real data proposals. This demonstrated a high level of interest and engagement from the AI Hubs, particularly within PSDI's focus area of the Physical Sciences. A meeting was held on 2nd May with 31 people in attendance, representing 17 of the Hub proposals.

[1] Artificial intelligence hubs for real data and for scientific and engineering research - https://www.ukri.org/opportunity/artificial-intelligence-hubs-for-real-data-and-for-scientific-and-engineering-research/
Mathematical and computational foundations of artificial intelligence - https://www.ukri.org/opportunity/mathematical-and-computational-foundations-of-artificial-intelligence/

Given the short timescale between the announcement of the successful first round applicants and the submission deadline for the full proposals, discussion was focused on seeking and agreeing high-level of areas where PSDI could work with the AI-Hubs, rather than developing specific plans for individual collaborative activities.

A number of areas emerged where it was agreed that it would make sense for the AI-Hubs and PSDI to work together are described below. Guidance will be sought from the funders as to how this collaboration could be supported.

# Topics discussed in the meeting

To seed the discussion, a list of 14 topic areas were presented and an opportunity given for attendees to vote for the topics they thought would be useful to work on cooperatively. These areas are listed below, ordered by the number of votes received (highest first).

1. Data sources – training data
2. Data preparation, data cleaning, data transformation
3. Data standards
4. AI model development & adoption
5. Model evaluation & Benchmarking
6. Data storage
7. AI model management (e.g. versioning, metadata, workflows)
8. Application of AI to data management (e.g. cleaning, curation, querying)
9. Compute resources
10. Standards for models
11. Output data (where applicable)
12. Training
13. Building the skills pipeline
14. Publication of models

The following additional topics were also suggested by the attendees:

- Ethical standards and approval processes for data collection and curation
- Legal issues and liability

# Points raised in the discussion

These topics were then used to structure a discussion about possible collaboration, beginning with the highest voted topics, and then more generally on any of the other topics. This discussion is summarised below.

## Data sources – training data

- Much of the discussion on data sources centred around sharing experimental data so that it can be used for AI without the need to repeat costly experiments.
    - Given that AI often requires large datasets, experimental data from some of the larger facilities may be a source of data, with any necessary additional curation undertaken by the facilities.
- How to bring the legacy data that was not created with AI in mind into a state where it is AI ready. What additional data curation activities would be needed?

    - AI applications will want to utilise data from different projects so it would be helpful to facilitate linking data from different sources (e.g., Digital footprints and other ESRC projects.)
- There is a need to incorporate labels on data from multiple different perspectives.

## Data preparation, data cleaning, data transformation

- Increasing transparency requires accurate documentation of the full pipeline of data transformation.
    - Want to be able to build different aggregations of data to be used for models.
    - What techniques can be applied to data more entirely across the domain.
- AI can be used to clean data prior to analysis, not just for data analysis
    - However, there are concerns around the veracity of this
    - There is the worry that outlying data might be lost, that could have been important.

## Data standards

- Adoption of agreed standards is required to enable cross-application.
    - Development of standards requires significant community engagement collaboration.
    - Ontologies are and important aspect, but not the only area.
- Standards are also required for defining how the training data was collected.

## Non-open data

- The sharing of data that cannot be made open was also discussed.
    - Examples of non-open data include sensitive personal data and data that cannot for shared for security or commercial reasons.
    - Tools and techniques that are being developed for Trusted Research Environments may also be useful for other non-openable data.
    - There is a potential confusion of terminology here as the term “trusted” can also be used to refer to data with a high level of validation and curation, which could be open.

## Model evaluation & benchmarking

- Scenarios and contexts for benchmarking data collections were discussed. A challenge here is to understand what information about a trial needs to be captured in order for benchmarking to give meaningful results.
- Metrics for benchmarking are often domain-specific and vary across application domains.
    - To what extent can application specific benchmarks be made more applicable across domains?
    - How do we know whether a benchmark is appropriate for an application?
    - How would these metrics be adequately recorded?

## Skills and Training

- Many (maybe all) Hubs are likely to require a platform for training. There are some core skills that will overlap between different Hubs. It may well be efficient and effective to adopt a single training platform across the Hubs so that common training materials could be more easily shared and co-developed where appropriate.
- **Building the skills pipeline**: there is also a significant need for a much longer-term view to be taken as to how we can created substantial skilled workforce for AI.

## Other Topics raised during the meeting

- The individual Hubs should be the place where model development and training should take place. However, there are very many topics that should be worked on cooperatively between the different AI hubs as this could enable faster innovation and less duplication. Some of these are relevant to PSDI also. Examples where co-development may be appropriate are typically the components around the model, (e.g. inputs, model sharing, standards)
- Another common area relates to issues around the ethical standards for data collection and curation, and the legal issues surrounding data and AI. These are very important issues, but these were not discussed further in the scope of this meeting.
- There are a lot of workflows and log details that are produced in scientific experiments, but these are hidden behind the experiments. There is an opportunity to mine these for more information. Thinking about the full signature of an experiment, with much more richness. How can this detailed experimental data be adequately described and explored?

# Current PSDI Activities relevant to the points discussed

Several current PSDI activities align with topics discussed in the meeting. Areas for further investigation include:

- **Experimental Data**: collaboration with the STFC facilities and Ada Lovelace Centre around how access to data from the facilities, that is currently performed with human data consumption in view, could be extended to also cater for AI agents.
- **Data annotation** is a central topic in the Process Recording Pathfinder (PF2) and is also relevant to some of the other domain pathfinders. Developing and promoting standards and best practices for data annotation/labelling for AI model training and benchmarking would be beneficial.
- **Workflow recordin**g is a key topic explored in several current PSDI Pathfinders. Supporting workflows for training AI models may, however, require different approaches supported by different technologies than are is currently being considered by PSDI.
- **Skills** training is being explored under Workpackage 2 of PSDI, and will be a key component of PSDI going forward. PSDI could additionally provide training events and workshops specifically tailored to suit the needs of the AI Hubs.
- **Federated data sharing** (as opposed to centralised data sharing) is the approach favoured by PSDI. This may also be the most appropriate data sharing model for the AI Hubs so that individual Hubs maintain their ownership of data and take

responsibility for its curation, whilst other Hubs can get access via tools and standards provided by PSDI.

## Conclusions

The meeting enabled a constructive conversation to begin about how data and technology can be shared across the AI hubs, and with other initiatives, and several areas were identified where collaboration could be beneficial as described above. These areas included not only data Infrastructure in the sense of hardware and software to support the management of data, but also included many other aspects of support for the AI research lifecycle, including standards, processes, training, and the promotion of best practices, etc.

It is clear that PSDI can help with many of these areas, such as the creation and use of appropriate standards and processes to ensure that experimental and computational data are presented in a way as to ensure they will be available in machine processible form, along with high quality metadata making it suitable for inter/multi-disciplinary research within and beyond the remit of the physical sciences.

The meeting provided an opportunity for an initial conversation that will help set the agenda for future meetings. Many more detailed conversations will be required to identify and prioritise specific the areas of work to be undertaken by PSDI in conjunction with the funded hubs. The meeting also identified some areas where the AI hubs could work together, that potentially lie outside the scope of PSDI, as in the diagram presented above.

Given that the AI Hubs will be established in early in 2024 and run for 5 years, and that the next phase of PSDI is currently being planned, there is now an opportunity to build the work required to support collaborative activities described above into the overall PSDI programme. Guidance is being sought from EPSRC as to the most appropriate way to enable this.

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

# Attendance at the meeting

The meeting was attended by the following people

| Name | Institution | AI Hub |
|---|---|---|
| **Adham Hashibon** | University College London | Novel AI for Accelerated Materials Design & Discovery (AI4MD) |
| **Ana Basiri** | University of Glasgow | AI Hub for Real Data: Intelligence from New Forms Of data (INFO Hub) |
| **Barbara Shollock** | King's College London | Artificial Intelligence for a Changing Economy - enabling efficient, effective and sustainable products (AICE) |
| **Ciara Pike-Burke** | Imperial College London | AI Hub in Foundational Reinforcement Learning (RLHub) |
| **David Barber** | University College London | AI Hub in Generative Models |
| **Duc Pham** | University of Birmingham | The National Hub for EMBodiED AI in Healthcare (EMBED-AI) |
| **Feng Li** | Imperial College London | AI for Chemistry: AIchemy |
| **Henry Reeve** | University of Bristol | MSAI: Mathematical and Statistical underpinnings of AI |
| **Ian Fairlamb** | University of York | AI Hub for Digital Chemistry - Trustworthy Discovery & Insight |
| **Ian Nabney** | University of Bristol | AI Hub for Digital Chemistry - Trustworthy Discovery & Insight |
| **Jean-Baptiste Cazier** | University of Birmingham | Building AI Communities For Real WorlD Knowledge ImplEmentation In Health & Social Care |
| **Jonathan Hirst** | University of Nottingham | AI for Molecule Making |
| **Jose Miguel Hernandez Lobato** | University of Cambridge | AI for Molecule Making, AI Hub in Generative Models |
| **Konstantinos Tsavdaridis** | City, University of London | AI for Megacity Building Lifecycle Management Hub |
| **Matthew Harrison** | The University of Manchester | ML4Science: The AI Hub on Machine Learning for Science |
| **Miguel Bravo Haro** | City, University of London | AI for Megacity Building Lifecycle Management Hub |
| **Matthew Gaunt** | University of Cambridge | AI for Molecule Making |
| **Paul Fearnhead** | Lancaster University | Probabilistic AI Hub |
| **Phil Kitson** | University of Glasgow | Open Network for Artificial Chemical Intelligence & Discovery Hub (ON-ACID) |
| **Praminda Caleb-Solly** | University of Nottingham | The National Hub for EMBodiED AI in Healthcare (EMBED-AI) |
| **Safa A. Sway** | Imperial College London | AI Hub in Foundational Reinforcement Learning (RLHub) |
| **Tanya Roumelioti** | University College London | New AI-guided Earth Simulations (NewAGE): leading-edge Uncertainty Quantification of complex computer models |
| **Themis Prodromakis** | University of Edinburgh | AI for Productive Research & Innovation in eLectronics (APRIL) |

| | | |
|---|---|---|
| **Serge Guillas** | University College London | New AI-guided Earth Simulations (NewAGE): leading-edge Uncertainty Quantification of complex computer models |
| **Barbara Montanari** | STFC | PSDI |
| **Vasily Bunakov** | STFC | PSDI |
| **Juan Bicarregui** | STFC | PSDI |
| **Nicola Knight** | University of Southampton | PSDI |
| **Simon Coles** | University of Southampton | PSDI |
| **Jeyan Thiyagalingam** | STFC | STFC SciML |
| **Susmita Basak** | STFC | STFC SciML |