Data Sharing Workshops

20 May and 6 September 2024

Workshops Report

20 December 2024

Dr Joanne M Leach
joanne.leach@ucl.ac.uk

Silent streams of code,
Whispers of knowledge unfold,
Truth in bytes bestowed.

Haiku generated by Copilot on
17 December 2024 using the prompt:
“Can you create a haiku on data?”

Table of Contents

1. List of Acronyms	**3**
2. Background and Context	**4**
3. Methodology of the Workshops	**5**
4. Outcomes	**7**
4.1 Workshop 1: The challenges of and opportunities for data sharing between industry and academia	7
4.1.1 Current Practices	*7*
4.1.2 Benefits of Data Sharing	*8*
4.1.3 Barriers to Data Sharing	*8*
4.1.4 Specific examples mentioned from the workshop	*9*
4.1.5 Recommendations	*10*
4.2 Workshop 2: The challenges of and opportunities for data sharing in urban observatory settings	12
4.2.1 Current Practices	*12*
4.2.2 Benefits of Data Sharing	*12*
4.2.3 Barriers to Data Sharing	*13*
4.2.4 Specific examples mentioned from the workshop	*14*
4.2.5 Recommendations	*14*
5. Appendices	**16**
5.1 Appendix 1: Session brief for workshop 1 – the challenges of and opportunities for data sharing between industry and academia	16
5.2 Appendix 2: Session brief for workshop 2 – the challenges of and opportunities for data sharing in urban observatory and living lab settings	18
5.3 Appendix 3: List of Attendees	20

1. List of Acronyms

ADR UK	Administrative Data Research UK
AI	Artificial Intelligence
CReDo	Climate Resilience Demonstrator
DAFNI	Data and Analytics Facility for National Infrastructure
DINI	Data Infrastructure for National Infrastructure
DISD	Digital Information and Smart Data
DSIT	Department of Science, Innovation and Technology
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
GHG	Greenhouse Gas
HMG	His Majesty's Government
ICT	Information and Communication Technology
IES4	Information Exchange Standard Number 4
NERC	Natural Environment Research Council
NIC	National Infrastructure Commission
NISTA	National Infrastructure and Service Transformation Authority
NUAR	National Underground Asset Register
ONS	Office for National Statistics
UKCRIC	UK Collaboratorium for Research on Infrastructure and Cities
UKEOF	UK Environmental Observation Framework
UKRI	UK Research and Innovation

2. Background and Context

In March 2023 the Department of Science, Innovation and Technology (DSIT) launched the National Science and Technology Framework which set out the government's approach to making the UK a "science and technology superpower" by 2030[1]. The approach included a two-year pilot of a 'national research cloud'. Delivered in partnership with UK Research and Innovation (UKRI), the pilot tests different ways of pooling information, encouraging collaboration and facilitating solving data-driven research challenges.

Four pilot projects have been funded through the Framework and UKRI. Their objective is to understand the need for a national research cloud through a series of interventions designed to remove data sharing barriers. The Data and Analytics Facility for National Infrastructure (DAFNI) was one of the four recipients and used the funding to create the Data Infrastructure for National Infrastructure (DINI) project. DINI is designed to explore the challenges and opportunities in data sharing within the domain of national infrastructure systems research.

As part of DINI, the UK Collaboratorium for Research on Infrastructure and Cities (UKCRIC) was commissioned to run two workshops. UKCRIC is a multidisciplinary network of UK universities connecting research with policy and practice in infrastructure and urban systems. It works with stakeholders to better understand and address complex infrastructure challenges. UKCRIC's activities are underpinned by its four scientific missions[2]:

1. Infrastructure and urban systems for one planet living
2. Transformational infrastructure and urban systems for a changing world
3. Infrastructure and urban systems as drivers of equity, inclusion and social justice
4. Innovative ownership, governance and business models for infrastructure and urban systems

The two workshops addressed, respectively (1) the challenges of and opportunities for data sharing between industry and academia, and (2) the challenges of and opportunities for data sharing in urban observatory settings. This report summarises the outcomes of these workshops.

Both the industry representatives who participated in the first workshop and the academics who participated in the second workshop face pressures that influence their ability to share data, but they also recognise the opportunities afforded by sharing data and see the potential in taking a 'data-sharing first' approach to their work. They see value in establishing a national data cloud and believe that the challenges to doing so are surmountable and now is the right time to tackle them.

[1] https://www.gov.uk/government/publications/uk-science-and-technology-framework/the-uk-science-and-technology-framework
[2] https://www.ukcric.com/about-ukcric/scientific-missions/

3. Methodology of the Workshops

The first workshop was held on the 20th of May 2024. The delegates were members of UKCRIC's Stakeholder Advisory group, comprising professional practitioners in the infrastructure and urban systems sectors. It was chaired by Mark Enzer, Chief Technical Officer, Mott MacDonald.

The second workshop was held on the 6th of September 2024. The delegates were or had been involved in some way with urban or infrastructure observatories or living labs, comprising academics specialising in data, urban systems and infrastructure. It was chaired by Joanne Leach, Executive Manager, UKCRIC.

Participation in both workshops was by invitation only. The workshops were delivered online to maximise attendance given the geographical spread of the participants.

Participants were provided with information prior to the workshops (see Appendices 1 & 2). This included the context within which the workshops were being held, the purpose of the workshops, and potential discussion points and questions to be asked during the workshop.

The following questions were used to guide discussion. Each workshop lasted for approximately one hour.

1. This discussion is predicated on the need for a 'national research cloud' that supports research data sharing. Do you recognise this need?
2. In your experience, what motivates data sharing?
3. What are the benefits and barriers you have experienced when sharing, exchanging and reusing data?
4. Can you recommend any data sharing best practices? Examples might be for data policies; data sharing agreements; and data annotation, terminology and ontologies.
5. What services do you think are needed to support data sharing? Examples might be for cataloguing, providing access and making data available for interoperation and reuse.

Between the commissioning of the workshops and the first workshop Onward, a new-to-the-scene, non-profit think tank, recommended the following to Government to support the UK's AI sector:[3]

> *"The Government should establish a British Library for Data – a centralised, secure platform to collate high-quality data for scientists and start-ups.*
>
> *The library should work with public services to make their data AI-ready and bring Government-held datasets together. It should include language and multimodal data with robust privacy-preserving mechanisms.*
>
> *The library should be open to contributions from archives, universities, and private companies. Starting with NHS data, the library would create a potent resource for AI advancement, particularly for new powerful, tailored foundation models."*

Between the first and second workshops the UK Labour Party published its 2024 election manifesto, which included a pledge to "create a National Data Library to bring together existing research programmes and help deliver data-driven public services, whilst maintaining strong safeguards and ensuring all of the public benefit"[4].

[3] https://sciencesuperpower.substack.com/p/lets-get-real-about-britains-ai-status
[4] https://labour.org.uk/change/kickstart-economic-growth/

An early vehicle for the Data Library is the Digital Information and Smart Data (DISD) Bill. This Bill is broadly designed to improve data sharing. It does this within particular spheres, one of which is science, enabling scientists to make better use of personal data. Another is the subsurface, providing location information for underground assets such as water pipes and electric cables to those who need it.

Notwithstanding these developments and the change in language from a ‘national research cloud’ (the terminology that will be used throughout this report) to a ‘national data library’, both workshops focussed upon the same central question: How can the UK transform its data into research assets that can be used to benefit society? Specifically, the workshops sought to identify the benefits and enablers of and the barriers to data sharing, exchange and reuse.

Initial findings were presented at a DAFNI webinar on the 30th of October 2024. DAFNI will use this report to inform a final project report on the DINI project, which will be submitted to DSIT.

4. Outcomes

4.1 Workshop 1: The challenges of and opportunities for data sharing between industry and academia

4.1.1 Current Practices

Increasing digitalisation has led to an increased need to critique the nature of data. Data is not a clean thing. It comprises, amongst other things: quality, context and accuracy. It cannot be assumed that data measures what it was designed to measure. Data may or may not exist within a context that makes sense of the data. Data on its own is meaningless unless there is a way of contextualising it and benchmarking it. If sharing relationships are to be productive, those involved must continually ask: what is data, what is it going to be used for, and how is it going to be contextualised.

The nature of data is different from the nature of data sharing. Data sharing incorporates metadata and comprises, amongst other things: the purpose of the data, trusting the data, trusting the source of the data, and provenance[5] of the data.

Current best practice focuses upon enabling FAIR (findable, accessible, interoperable, reusable) data practices.

There is a seemingly ever-present discussion about the intricacies and difficulties of data sharing:

- Who is liable if poor-quality data results in harm?
- If data is shared once, is the sharer committed to sharing future versions of the data?
- Can shared data be un-shared?
- What legal agreements are needed?
- Who bears the burden of transforming data for use? The user? The supplier? The data cloud?
- How is access to shared data controlled?

Proportionality and purpose are established cornerstones of data gathering. The proliferation of digital data and big data are testing these and there is increased recognition that the question being asked of data is not always known at the time the data are collected. Without a purpose though, the sharing of data is unlikely to succeed.

The importance of data descriptions and purpose is increasing, driven by data proliferation and big data. Traditionally, data would not be collected without a good data description and knowing the data structure. However, data ontology is constructed by a user coming from a singular viewpoint and is reliant upon subsequent users understanding that viewpoint. For Artificial Intelligence (AI) and machine learning, having an unstructured data pool can be more useful than a structured data pool, especially in terms of looking for the patterns that are outwith any structure that might be imposed. It is also true that the structure of data is not always knowable beforehand and that AI can expose data structures over time.

Current centralised data sharing means the data creator loses sight and control of their data. This is viewed variously as a concern and an opportunity. The compliance responsibility of the data sharer isn't always clear. For example, is a data sharer responsible if data are used inappropriately or in ways that are illegal (e.g., not GDPR compliant)? What are data

[5] See for data provenance standards https://dataandtrustalliance.org/work/data-provenance-standards

sharers' responsibilities to keep shared data up to date? More recently though, traditional perspectives of data sharing as a linear process are being challenged by new perspectives that frame data sharing as the creation of new data incarnations[6].

4.1.2 Benefits of Data Sharing

A distinction is to be made between pre-commercial and post-commercial data sharing. Pre-commercial data sharing addresses issues faced by an entire group or sector, the solving of which benefits the whole group or sector. Post-commercial data sharing has the potential to deliver market advantage to a specific organisation. Each has a different benefit profile, but the distinction between the two is not always clear or considered in data sharing paradigms.

Data sharing from commercial and government entities to researchers is often pre-commercial (for example, leading to better engineering standards, decreasing risk and, by extension, decreasing overengineering). This appears to be the primary focus of the proposed national data cloud. This approach has the potential to increase the impact of data and the nation's data analysis capacity and capability as well as to enable data to be used in new and innovative ways and, through these, contribute to better government policy.

The benefits of the reverse of this, data sharing from researchers to commercial and government entities, has had less consideration, seems to occur less often, and seems to align with post-commercial data sharing. Underpinning assumptions of data sharing in this direction include that the commercial and government sectors do not have the capacity to consider research data, and that researchers are not set-up to deliver commercially.

Post-commercial data sharing between organisations within the same or synergistic sectors has the potential to support better decision-making, and whole-system and system-of-system approaches to problems and opportunities. However, commercial sensitivities are a paramount concern and the mechanisms for addressing them can be arduous and lengthy.

4.1.3 Barriers to Data Sharing

The willingness of practitioners and those in industry to share their data is currently low. Conversely, their willingness to use shared data is high.

Who owns data is not always clear. For example, who owns the data produced by arms-length bodies to the UK Government? Is it the body itself or the Government? The truth may vary depending upon the body. Additionally, different parties may take different views regardless of the truth and those perceptions will determine how the data are managed.

The potential for commercialising what comes out of data sharing can lead to less sharing, although this is not the case if it is considered beforehand and formal or contractual agreements are put in place.

Data confidentiality (such as commercially sensitive data and data that has the potential to compromise national or organisational security) can also be addressed through formal and contractual agreements, but the policing of these can be difficult.

Data sharing places a burden on those sharing the data that includes time and financial costs. Simply handing over data is not sufficient for data sharing. The format, quality,

[6] Kitchin R, Davret J, Kayanan C, Mutter S (2024) Data mobilities: Rethinking the movement and circulation of digital data. Data Stories, Maynooth University Social Sciences Institute. https://mural.maynoothuniversity.ie/id/eprint/18716/1/DS%20WP3%20Data%20mobilities.pdf

context, provenance, limitations and appropriateness of the data must be put into forms that can be received and understood by the user, and the user must take the time to understand them. Questions from both sides may need to be answered and these can arise immediately as well as over time. The financial costs associated with sharing data and using shared data include data usage fees, data storage and analysis equipment, and trained/training personnel to analyse the data. These costs increase when multiple versions of data are created (for example, sharable versions that are aggregated or redacted to protect sensitive information). For data sharing to work, organisations must be able to justify the expense and create business cases that demonstrate value for money.

Data sharing also places a burden on the planet. Digital data, digitally sharing data, cloud computing, AI, machine learning, digital and cloud backups, and so on all have a carbon cost. It is estimated that the current greenhouse gas (GHG) footprint of the energy demands of the information and communication technology (ICT) sector[7] is between 1.5% and 4% of global GHG emissions[8] (roughly equivalent to that of the aviation sector). The implication is that collecting, sharing and analysing data must have a purpose and a benefit that outweighs the cost not just in time and money, but also the cost to the planet.

Finally, there already exist organisations whose business model is based around analysing data and that could consider a national data cloud unwelcome competition.

4.1.4 Specific examples mentioned from the workshop

Existing data stores and libraries seem to share a community of interest or practice, or an area of research focus.

- The Biobank has an umbrella purpose of public health which covers a broad range of more specific uses. It gates access to approved researchers, quality checks and anonymises all data stored, stores the data centrally, and it provides a platform to analyse the data.
- The UK Environmental Observation Framework (UKEOF) is a coordinating body across the public sector specifically for the environmental observation community. It focuses upon coordinating observational evidence, building a network, providing a neutral discussion space, and providing a central source of advice and information.
- Administrative Data Research UK (ADR UK) focuses upon public-sector data. It does up-front work on data governance, cleaning, and linkage, so that de-identified, research-ready, curated datasets can be maintained over time.
- Information Exchange Standard number 4 (IES4) governs and facilitates data sharing between a federated set of knowledge stores within His Majesty's Government (HMG). It explicitly recognises that the stores may use different terminologies, formats and schemas, that these can be important to the data owners, and that changing them to facilitate data sharing could degrade the usefulness of the data to the data owners. "Being able to exchange and share information effectively and efficiently is imperative and needs to be achieved without the need for collaborating organisations to: (1) develop

[7] For more information about the ICT sector see https://www.sciencedirect.com/topics/social-sciences/communication-technology-sector

[8] Bieser JCT, Hintemann R, Hilty LM, Beucker S (2023) A review of assessments of the greenhouse gas footprint and abatement potential of information and communication technology. Environmental Impact Assessment Review, Volume 99, https://doi.org/10.1016/j.eiar.2022.107033

numerous and bespoke bilateral interchange mechanisms; and (2) make costly and highly disruptive changes to their individual knowledge stores."[9]

- Data.gov.uk publishes data from central government, local authorities and public bodies.
- The European Union operates a data portal to access data published by its member states.
- The Office for National Statistics (ONS) Integrated Data Service provides gated access to data and assures the quality and provenance of the data. It creates data products (linking national data assets), and carries out its own data analysis to support decision making.
- The ONS has an Accredited Researcher Scheme that enables researchers to access anonymised and unpublished data.
- ArcGIS provides a software tool for geographic data analysis and curates compatible datasets for use with the software.
- The Connected Place Catapult has developed a toolkit for local authorities sharing non-personal data.
- The National Underground Asset Register (NUAR) enables underground utility owners and operators to share asset location data for the purpose of increasing efficiencies and derisking utility operation, maintenance, and planning.
- The UK Data Service is a repository for economic, population, and social research data.

4.1.5 Recommendations

Four 'parts of the solution' were identified, with all four needing to be incorporated into the proposed data cloud in order for it to be successful. They are:

1. Governance
2. Process (top down and bottom up)
3. Semantics (e.g., ontologies)
4. Software / technical

Sitting alongside these are seven 'watchwords' that must be woven through the design and operation of the data cloud if it is to be useful and sustainable into the future. They are:

1. Resilience
2. Scalability
3. Security
4. Provenance
5. Purpose
6. Transparency
7. Trust

[9] https://github.com/dstl/IES4/blob/master/introduction.md

Seven components of functionality of the data cloud were identified:

1. Custodianship. Being not simply a keeper and controller of data, but a caretaker of it.
2. Signposting and curating data, including protecting against poor-quality data, and removing lower-quality and out-of-date data as better, more recent data become available.
3. Enabling services to support easier ways for users to: find and access data, establish benchmarks, and conduct insights and analytics.
4. Supporting the sharing of data, including brokering data sharing agreements and supporting dialogue between data suppliers and data users to address whether the data is being used in a way that is valid and robust.
5. Horizon scanning for future data and data needs and making missing data explicit.
6. Conducting insights and analytics on the data, including trend analysis.
7. Advancing the practice of data sharing, including developing best practices and standards for sharing data, establishing benchmarks, and shaping policy, regulatory and other drivers.

There is substantial work still to be done to establish confidence in data. This is more than data ontology (a description of data and its structure) and more than metadata[10] (which is often bespoke to specific user communities). The emerging field of 'computational epistemology' has a role to play. This term has been coined to describe the data needed about data that provides confidence in the data. It includes when, where, and how data are collected, who asserts the data to be true, whether there are real-life examples of the data, and so on.

Understanding the data (and data about the data) to be brought into a national data cloud is the first step in developing the data cloud. However, it cannot be assumed that improved access to data will lead to 'the right answer'. It could simply increase and amplify spurious outcomes.

The data cloud could take the form of a centralised data depository, a federation of knowledge stores (favoured by the workshop participants), or a catalogue of available data and pointers to their locations. Whichever is the case investment in hardware, software and people will be needed.

The proposed national data cloud, and data sharing infrastructure more generally, must be recognised as a class of infrastructure that falls under the remit of the National Infrastructure Commission (NIC) (soon to be the National Infrastructure and Service Transformation Authority (NISTA)). Data sharing infrastructure cannot be owned by one group. It must be a shared resource.

[10] See for metadata standards https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/what-are-metadata-standards

4.2 Workshop 2: The challenges of and opportunities for data sharing in urban observatory and living lab settings

Some benefits, barriers and recommendations identified in Workshop 2 were the same as those identified in Workshop 1. Where this is the case, these have not been repeated.

4.2.1 Current Practices

For the purpose of this workshop, urban observatories were defined using the model devised by UKCRIC. UKCRIC created six urban observatories in the UK in Birmingham, Bristol, Cranfield, Newcastle[11], Sheffield, and Manchester. For each, research teams placed static and mobile sensors throughout the city to gather data that spoke to the sustainability and resilience of the urban environment. Sensors gathered continuous and near-continuous data on such things as air quality, people and vehicle movements, water quality, and the weather. These data were made publicly available via a website and users could see, analyse and download the data.

A lot has changed since UKCRIC set up its urban observatories. Cities are now used to having sensors on and in the fabric of the built environment. Local Authorities are placing sensors themselves (with air quality being an obvious example, driven by reporting and regulation). These sensor networks are more expansive than those the research teams implemented and they are maintained by the Local Authority, providing consistent data over time. There has also been a change in the willingness of Local Authorities to allow researchers to access data from their sensors and this is allowing for new and novel collaborative research.

Urban observatories contain a second type of infrastructure sensors, although the data from these may or may not be made publicly available. These data are from sensors that monitor specific pieces of infrastructure. For example, sensors in bridges to monitor the condition of the bridge. The wireless sensor network deployed to monitor the Clifton Suspension bridge is a good example of such an 'infrastructure observatory'[12].

4.2.2 Benefits of Data Sharing

There are benefits to sharing data and specifically to sharing continuous and near-continuous data. These include:

- Sharing data makes others aware of the data you have and this can prevent duplication of effort and lead to new collaborations.
- Coming together in a geographic space such as an urban observatory can lead to new partnerships.
- Bringing data from multiple sources together into a single computational structure enables multidisciplinary research.
- The perceived risks of sharing data often don't materialise in practice (when not dealing with personal data).
- Shared data can be used in test beds. For example, in determining how the data are useful for AI, and in identifying the advantages of edge compute[13].

[11] Newcastle's urban observatory is the most advanced. See https://newcastle.urbanobservatory.ac.uk
[12] See https://research-information.bris.ac.uk/ws/portalfiles/portal/140702178/Full_text_PDF_final_published_version_.pdf
[13] Edge compute is reducing latency between a data source and the storage/computation of the data through physical proximity

4.2.3 Barriers to Data Sharing

There are two headline challenges associated with sharing urban and infrastructure observatory data that can create barriers to data sharing: (1) providing consistent data over time, and (2) sharing continuous and near-continuous data.

1. The challenges of providing consistent data over time:

- There is a high turnover of sensor equipment. Sensors have short lifespans and require regular replacement.
- Sensors and the equipment and software that support them require periodic maintenance, including hardware and software updates. These can cause changes to the data that can be compounded by sensor manufacturers not always publishing how their sensors work.
- It can be difficult to get permission to place sensors into the public environment. Local Authorities, for example, may not have procedures or regulations that support sensor installation or the budget to support sensor maintenance. Even with willing partners, establishing the needed contractual agreements can be a lengthy process.

2. The challenges of sharing continuous and near-continuous data:

- One of the key characteristics of data in an urban observatory setting is the amount of data produced. Data is continuous in some cases – for example, video – and near to continuous in other cases – taking readings every x number of seconds or minutes. There are multiple financial, equipment, computer and time costs in capturing, storing and analysing very large amounts of data. The benefit of having the data must outweigh these costs.
- Data collected by urban observatories, by definition, is not collected to answer specific research questions. Whereas in Workshop 1 the participants were divided as to whether a research purpose should be an uncompromised prerequisite to data collection, this was not the case for the participants in Workshop 2 although they acknowledged the tension. In the second workshop all the participants supported collecting data irrespective of there being clear research questions, but caveated this with the need to understand the data's limitations. As one participant put it, "any data is better than no data at all".

Infrastructure research has two characteristics that can be barriers to data sharing.

1. Infrastructure research is largely based upon data from commercial organisations (as opposed to prior research).
2. The outcomes from infrastructure research are shared with commercial organisations and government (as opposed to solely other researchers).

Other barriers to data sharing:

- Lack of interoperability.
- Different interpretations of, or lack of, standards, especially in authentication and authorisation.
- The value of sharing data with academia is not fully understood, quantified or communicated.

4.2.4 Specific examples mentioned from the workshop

The Natural Environment Research Council (NERC) Environmental Data Service comprises five data centres[14]:

- British Oceanographic Data Centre (marine)
- Centre for Environmental Data Analysis (atmospheric, earth observation, and solar and space physics)
- Environmental Information Data Centre (terrestrial and freshwater)
- National Geoscience Data Centre (geoscience)
- UK Polar Data Centre (polar and cryosphere)

The Climate Resilience Demonstrator (CReDo) project brings together data from competing stakeholders onto the DAFNI platform. A core team can see and use the data. The stakeholders cannot see each other's data. This allows the stakeholders to collaborate without compromising their competitive confidentiality.

The NERC Digital Solutions Programme is developing a digital hub and associated toolkits including real-time data architecture, for which there is a testbed. It works with continuous and near-continuous data under a policy of storing such data for 30 days to enable prediction and analysis. After 30 days the data becomes a static data set.

The Data Sharing Playbook captures data sharing experiences and recommendations from healthcare professionals for the purpose of informing and improving the sharing of health data.

4.2.5 Recommendations

The following two watchwords must be woven through the design and operation of the data cloud if it is to be usable and successful:

- Simplicity
- Accessibility

Seven components of functionality of the data cloud were identified:

1. Federation. A federated data cloud keeps data within the control of those who use, work with and understand the data.
2. Supporting publicly available data and standardising data-sharing protocols. The data cloud's default position should be for shared data to be made public. From this starting point questions can be asked about why certain data can't be made public, what would need to be put in place to make them public, and if they can't be made public, what is the next least-restrictive sharing model that applies.
3. Enabling services to support easier ways for users to find, access and understand data. The data collected by urban observatories has little or no metadata attached to it and the burden is firmly on the user to ensure they understand the data. This has implications for the easy discoverability and use of data stored within the proposed national data cloud.
4. Data curation. Skilled people are essential to the successful curation of continuous and near-continuous data (and this will require a viable business model).

[14] A full list of UKRI facilities and resources is available at https://www.ukri.org/councils/nerc/facilities-and-resources/find-a-nerc-facility-or-resource/

5. Edge systems / edge computing. Reliable and sustainable monitoring systems run by experts are required at the edge in order to assure the quality of the processes that generate and extract data before they are sent to the data cloud.
6. Supporting data standardisation and interoperability. Urban observatories are not just about capturing continuous and near-continuous data, they also carry out data analysis and modelling. An important feature of the national data cloud is for it to enable modelling across datasets and scales (e.g., local, regional and national scales). This requires data are standardised and interoperable. The required groundwork to make data interoperable, available and knowable (e.g., via metadata) is fundamental to the data cloud's long-term sustainability. Without it the data within the data cloud will not be sharable.
7. Project funding. The data cloud should have its own budget to fund projects that support filling data gaps, updating and improving existing data sets, and advancing the science of data sharing.

A national policy framework for infrastructure reporting would drive the standardisation of data collection and formats.

The quantity of data and the costs of creating a national data cloud that includes continuous and near-continuous data should not be underestimated.

5. Appendices

5.1 Appendix 1: Session brief for workshop 1 – the challenges of and opportunities for data sharing between industry and academia

Session Brief: Data Infrastructure for National Infrastructure

The Data Infrastructure for National Infrastructure (DINI) project is exploring the challenges and opportunities in data sharing within the domain of national infrastructure systems research.

Background

This is one of a number of activities on data sharing commissioned by DAFNI (the Data and Analytics Facility for National Infrastructure) on behalf of DSIT (Department for Science, Innovation and Technology).

DAFNI represents an £8 million investment from UKCRIC to provide world leading infrastructure systems research capabilities and enhance the quality of outputs. The DAFNI platform supports better sharing and use of data, exploitation of simulation and optimization techniques, and engagement with stakeholders through visualisation.

In March 2023 DSIT launched the National Science and Technology Framework which sets out the government's approach to making the UK a "science and technology superpower" by 2030, and announced a 2-year 'national research cloud' pilot in partnership with UKRI that will allow UK researchers to test different ways of pooling information more effectively and collaborate to solve data-driven research challenges. In March 2024, the National Science and Technology framework funded four pilot projects with UKRI for a national research cloud. The overall objective is to understand the need for a national research cloud through a series of interventions designed to remove data sharing barriers (to which this session will speak). Future work will identify potential models and options for national scale initiatives, support the building of an investment case, and set out the role of Government.

Purpose of the session

An exploration of the challenges and opportunities of data sharing between industry and academia.

Situation analysis: identifying the benefits of and barriers to data sharing, exchange and reuse between industry and academia – paying particular attention to water, energy and transport.

Data access: recommending best practices to enable the FAIR (findability, accessibility, interoperability, reusability) publication of and access to infrastructure systems data across the industry-academia interface – including data policies, data sharing agreements, data annotation, terminology, ontologies and the use of digital object identifiers.

Enabling services: assessing the usefulness of and identifying the key characteristics of the services needed to support the sharing of data between industry and academia – including cataloguing, providing access and making data available for interoperation and reuse.

Outputs from the session

We will disseminate the findings in a webinar and a report to DAFNI. DAFNI will use these to inform a final project report for DSIT.

Format of the session

- Summary of the purpose of the DINI project (DAFNI) and the session (Brian Matthews)
- Facilitated discussion (Mark Enzer)
- Next steps

Discussion points

Discussion points may include, but are not limited to: data capture, curation and sharing (including of third-party data); data storage, processing, and analysis (including large quantities of data); and the challenges of using data on sharing and computational platforms (such as DAFNI).

Questions to guide the discussion

1. This discussion is predicated on the need for a 'national research cloud' that supports research data sharing. Do you recognise this need?
    a. National research cloud (UK) - giving researchers greater access to data from a range of sources (through the Office for National Statistics Integrated Data Service).
2. In your experience, what motivates data sharing between industry and academia?
3. What benefits and barriers have you experienced when sharing, exchanging and reusing data with academics?
4. Can you recommend any data sharing best practices? Examples might be for data policies; data sharing agreements; and data annotation, terminology and ontologies.
5. What services do you think are needed to support the sharing of data between industry and academia? Examples might be for cataloguing, providing access and making data available for interoperation and reuse.

5.2 Appendix 2: Session brief for workshop 2 – the challenges of and opportunities for data sharing in urban observatory and living lab settings

Session Brief: Data Infrastructure for National Infrastructure

The Data Infrastructure for National Infrastructure (DINI) project is exploring the challenges and opportunities in data sharing within the domain of national infrastructure systems research.

Background

This is one of a number of activities on data sharing commissioned by DAFNI (the Data and Analytics Facility for National Infrastructure) on behalf of DSIT (Department for Science, Innovation and Technology).

DAFNI represents an £8 million investment from UKCRIC to provide world leading infrastructure systems research capabilities and enhance the quality of outputs. The DAFNI platform supports better sharing and use of data, exploitation of simulation and optimization techniques, and engagement with stakeholders through visualisation.

In March 2023 DSIT launched the National Science and Technology Framework which sets out the government's approach to making the UK a "science and technology superpower" by 2030, and announced a 2-year 'national research cloud' pilot in partnership with UKRI that will allow UK researchers to test different ways of pooling information more effectively and collaborate to solve data-driven research challenges. In March 2024, the National Science and Technology framework funded four pilot projects with UKRI for a national research cloud. The overall objective is to understand the need for a national research cloud through a series of interventions designed to remove data sharing barriers (to which this session will speak). Future work will identify potential models and options for national scale initiatives, support the building of an investment case, and set out the role of Government.

Purpose of the session

An exploration of the challenges and opportunities for data sharing in urban observatory and living lab settings.

Situation analysis: identifying the benefits of, the enablers of, and the barriers to data sharing, exchange and reuse – paying particular attention to water, energy and transport.

Data access: recommending best practices to enable the FAIR (findability, accessibility, interoperability, reusability) publication of and access to infrastructure systems data – including data policies, data sharing agreements, data annotation, terminology, ontologies and the use of digital object identifiers.

Enabling services: assessing the usefulness of and identifying the key characteristics of the services needed to support the sharing of data and the use of shared data in urban observatory and urban living lab setting – including cataloguing, providing access and making data available for interoperation and reuse.

Outputs from the session

We will disseminate the findings in a webinar and a report to DAFNI. DAFNI will use these to inform a final project report for DSIT.

Format of the session

- Summary of the purpose of the DINI project and the session (Brian Matthews)
- Facilitated discussion (Joanne Leach)
- Next steps

Discussion points

Discussion points may include, but are not limited to: data capture, curation and sharing (including of third-party data); data storage, processing, and analysis (including large quantities of data); and the challenges of using data on sharing and computational platforms (such as DAFNI).

Homework prior to the session

Please bring your answers to the following four questions to the session:

1. What is your 'top benefit' to sharing data?
2. What is your 'top barrier' to sharing data?
3. What is your 'top benefit' to using shared data?
4. What is your 'top barrier' to using shared data?

Questions to guide the discussion

1. This discussion is predicated on the need for a 'national research cloud' that supports research data sharing. Do you recognise this need?
    a. National research cloud (UK) - giving researchers greater access to data from a range of sources (through the Office for National Statistics Integrated Data Service).
2. In your experience, what motivates data sharing?
3. Are there any benefits and barriers that are unique to urban observatories and urban living labs?
4. Can you recommend any data sharing best practices? Examples might be for data policies; data sharing agreements; and data annotation, terminology and ontologies.
5. What services do you think are needed to support data sharing? Examples might be for cataloguing, providing access and making data available for interoperation and reuse.

5.3 Appendix 3: List of Attendees

The following attendees have agreed to the inclusion of their names and affiliations in this report.

Haris Alexakis	Aston University
Lucy Bastin	Aston University
Sergio Cavalaro	Loughborough University
Katie Cartmell	Science and Technology Facilities Council (STFC)
Lee Chapman	University of Birmingham
Brian Collins	University College London
Jim De Waele	Keller Group plc
Tom Dolan	University College London
Mark Enzer	Mott MacDonald
Ann Holden	Cranfield University
Kat Ibbotson	WSP
Phil James	Newcastle University
Jens Jensen	Science and Technology Facilities Council (STFC)
Simon Jude	Cranfield University
Ben Kidd	Arup
Joanne Leach	University College London & University of Birmingham
Brian Matthews	Science and Technology Facilities Council (STFC)
Chrissy Mitchell	Environment Agency
Andy Moores	Construction Industry Research and Information Association (CIRIA)
William Powrie	University of Southampton
Phil Proctor	National Highways
David Richards	University of Southampton
Christopher Rogers	University of Birmingham
Bridget Rosewell	Atom Bank and M6 Toll
Anil Sawhney	Royal Institution of Chartered Surveyors
Sarah Sharples	University of Nottingham
Theo Tryfonas	University of Bristol
George Tuckwell	RSK Group
Liz Varga	University College London
Steven Yeomans	Manufacturing Technology Centre (MTC)