National Infrastructure Data

Literature Review

Prepared for
Data Analytics Facility for National Infrastructure (DAFNI)
Data Infrastructure for National Infrastructure (DINI)

Author: Ceri Stanaway
Contributors: Emily Judson, Charlotte Horler, Caroline Fraser

Document: IB1-DAFNI-DINI-2024-LITREVIEW
Version: v2024-10-14

Contents

Introduction **2**
Project scope and methodology **3**
Report structure 5
Section 1: Benefits of sharing infrastructure data **6**
Socio-economic benefits 7
Environmental benefits 8
Benefits to industry 9
Direct benefits to those sharing the data 9
Section 2: Good practice examples of data sharing **10**
Cross-sector 12
Energy 14
Water 16
Transport 18
Section 3: Barriers to data being shared for research **21**
Legal and privacy barriers 23
Security barriers 25
Commercial barriers 26
Cultural barriers 27
Technical barriers 30
Section 4: Recommendations for better data sharing **34**
Conclusion **35**
Appendix 1: Cross-sector initiatives to support data sharing **38**
Appendix 2: Detailed recommendations **42**
Appendix 3: Barriers to researchers sharing the data outputs of their research **48**
References **51**

Introduction

Across the globe and across all industries, vast amounts of data are generated on a daily basis. It's estimated that for every person on earth, 1.7 MB of data is created every second[1]. Data, used effectively, has the potential for huge socio-economic and environmental benefits. But to maximise its potential, it has to get into the right hands, and in the right way.

Sharing data for research is an important part of the equation. Researchers that are external to an organisation in which data originated can offer fresh perspectives to maximise data value, enable new forms of analysis through dataset linking, and offer important insights into the quality of data. Data held and published by public bodies is valuable, but data held by private companies is equally crucial, potentially offering 'powerful insights into the behaviors of individuals, communities, organizations, systems, and the physical environment—as well as into the interactions among these levels' (The Royal Society; National Academy of Sciences, 2024, p6). However, accessing such data can be challenging, with researchers – and other users – facing issues around legalities, privacy, commercial resistance and cultural reticences, as well as technical challenges.

While this report is intended to focus on sharing data – and specifically infrastructure data – for research, it is worth noting that many published examples in grey literature (of both good practice and barriers) are not specifically about sharing infrastructure data for research. There is literature concerning sharing in infrastructure generally, and literature about data sharing for research in other sectors, but very little about sharing infrastructure data for research. Additionally, a significant number of papers have a focus on open data rather than shared data, further limiting the scope of the existing evidence base.

Where there are insights into private organisations sharing data with researchers, whether related to infrastructure or otherwise, their approach to doing so can vary significantly. Some companies have formal initiatives to engage researchers, including active outreach or research challenges. Some offer internships, where data doesn't leave the company. Others share data with researchers on a more ad-hoc basis (Harris & Sharma, 2017). Examples of private organisations proactively sharing data with research networks, rather than individual researchers, are sparse.

However, even where examples provided are not specifically about researcher access to infrastructure data, parallels can be drawn and lessons learned.

[1] https://digitaldecarb.org/the-figures/

Two key examples illustrate just why improving access to data is so important. Transport for London (TfL) stands as a notable example of data sharing success, and is highlighted as such in several sources. Its open publication of large amounts of non-personal data – such as timetables and service status – has led to the development of new products and services that have benefited commuters in London and the wider London economy. As Vernon Everitt, Managing Director at TfL, puts it in a report about the value of TfL's open data: 'Our guiding principle ever since [we started] has been to make non-personal data openly available unless there is a commercial, technical or legal reason why we should not do so' (Deloitte, 2017b, p2).

Conversely, examples of data sharing failures – such as Network Rail accidentally hitting a telecoms cable during infrastructure works – illustrate the risks of insufficient data exchange between sectors (Black and Veatch, 2019).

It is clear from this review that striking a balance between the commercial and societal value of data, customer privacy, and academic interests poses a challenge for companies (Harris & Sharma, 2017). Meanwhile, there can be quite different perceptions of how industry and academia use data, when in practice there isn't always a clear distinction between the two, especially when academics partner with commercial organisations (The Royal Society; National Academy of Sciences, 2024).

This report explores the documented benefits and good practice around sharing infrastructure data (particularly in relation to energy, water and transport), and highlights the numerous, interacting barriers that have been recognised in existing grey literature, as well as the recommendations that have been made to address these barriers.

Project scope and methodology

Icebreaker One has been commissioned to undertake a landscaping exercise to consider the current state of the art in data sharing in energy, water and transport infrastructure, particularly in regard to the support of research, with a focus on gathering evidence from the private and government sectors.

This literature review report on published opportunities for and challenges in infrastructure data sharing forms the first part of the overall landscaping analysis, covering:

- Identification and review of relevant pre-existing studies of data availability, sharing and integration in the domain of water, energy and transport

- Investigation on accessing data from private sector

Alongside findings from surveys, workshops and interviews, these literature review findings will feed into Icebreaker One's ultimate conclusions from the analysis to support recommendations for best practices in data publication and sharing in the sector.

This literature review was conducted during the period spanning August to September 2024, and included literature published between 2017 and 2024 (inclusive). The review focused primarily on available 'grey literature'[2], largely produced by industry, policy, and third-sector bodies. The review focused on publications made available to the general public, without paywalls or other access barriers which restricted the thoroughness of search and review processes within the available time and budget. Future studies may benefit from reviewing additional access-controlled sources where possible, for example including the outputs of any industry fora which are not openly published, or longer pieces of investigative journalism which are only available through subscription.

The review was informed by two methods of desk research and search:

1. **Online searches** performed in September 2024 using Google. Search terms included:
    a. [all of these words] data sharing good practice transport/energy/water [region] UK
    b. [all of these words] data sharing transport/energy/water "data for research" [region] UK
    c. sharing data "with academics"
    d. sharing data "with researchers" AND transport/energy/water
2. **Literature recommendations** made primarily by project partners, including DAFNI, supplemented by suggestions from a limited range of non-academic domain experts known to Icebreaker One.

Literature was reviewed by the researcher, aiming to draw out evidence, analysis and good practice examples of: 1) benefits of sharing infrastructure data, and 2) barriers to sharing infrastructure data. This analysis focused where possible on instances of data sharing with publicly funded researcher audiences (academic and publicly-funded adjacent research). However, significant literature gaps around this topic were identified during the research process. Accordingly, where appropriate, the following analysis makes some inferences about data sharing practices which are assessed in the

[2] Information produced and published outside of standard academic, peer-reviewed publication platforms such as academic journals.

literature in terms of wider audiences (e.g. including data sharing with commercial or government bodies).

Additionally, while this review focused on the target sectors of energy, water and transport, it was necessary to incorporate some literature assessing data sharing across the wider economy. This is because data sharing in infrastructure sectors is broadly influenced by cross-economy factors such as government policy and strategy (e.g. Industrial Strategy, Data Strategy), law and regulation (e.g. the UK Data Protection Act, 2018), and concerns regarding security. Furthermore, it is valuable to consider literature from cross-economy spaces when exploring some of the less formal or codified socio-technical factors that can have influence on data sharing practice, for example the impact of well-known data sharing precedents (e.g. Open Banking).

Report structure

This report is structured into four main sections. The first section outlines the benefits of sharing infrastructure data for public research purposes. Benefits are discussed thematically in terms of: 1) socio-economic benefits, 2) environmental benefits, 3) benefits to industry, and 4) benefits directly to data sharing entities. The second section follows on from this, providing a selection of good practice examples for data sharing in cross-sector and sector-specific spaces.

The third section examines barriers to infrastructure data sharing for public research purposes, particularly highlighting the linkage and complexity between socio-economic and technical challenges. Barriers are discussed thematically concerning: 1) legal and privacy, 2) security barriers, 3) commercial barriers, 4) cultural barriers, and 5) technical barriers. The fourth section presents a short discussion of recommendations for improving data sharing practices which are presented within the current literature base. This section particularly highlights repeated themes and areas of contention that are presently known. The report concludes with a set of summary remarks and implications for future research.

Section 1: Benefits of sharing infrastructure data

In 2017, a Deloitte report estimated that promoting greater data sharing in infrastructure could generate up to £15 billion in annual benefits (equivalent to nearly £20 billion in today's money, after accounting for inflation) (Deloitte, 2017a). The benefits highlighted included improved efficiencies, such as demand-side response solutions for energy supply management, increased competition and innovation (with Transport for London (TFL) open data cited as an example), and enhanced network

planning and resilience. In the water sector, for instance, smart meter data could help identify leaks. The report noted that there was even more significant potential from cross-sector data sharing; for example, data sharing between the energy sector and transport companies, which are high users of energy, could help energy companies to better plan for spikes in energy demand (Deloitte, 2017a).

The Deloitte report addressed data sharing in general, rather than data sharing for research purposes. However, a blog post on acquiring corporate data for academic research, from US-based firm Dewey, notes that through better access to large datasets – 'researchers can identify patterns, correlations, and causal relationships that may not be evident in smaller samples they've collected themselves. This can help to validate existing theories or generate new hypotheses that can be tested in future research' (Dewey, 2023).

Deloitte's report on data sharing in infrastructure is now seven years old, but the general benefits it highlighted have been echoed in more recent publications, such as the potential of data to help develop solutions to critical issues (Dewey, 2023), with time series data noted as being of particular value to enable innovation (Evans & Johnston, 2023).

Meanwhile a report summarising the findings from a recent US-UK Scientific Forum on Researcher Access to Data notes that privately held data, that is not necessarily generated for research, can have enormous potential: 'These data can produce powerful insights into the behaviors of individuals, communities, organizations, systems, or the physical environment, as well as the interactions among these levels' (The Royal Society; National Academy of Sciences, 2024, p35). Such data can include, for example, digital footprint data – data generated with digital systems, devices, and sensors. An example was given of how a combination of commercial satellite and social media data has delivered early warning detectors for conflict zones.

The same report recognises that in order to be of maximum benefit, data must be easily usable and interoperable so that multiple datasets can be combined. 'Tools such as federated systems, artificial intelligence–driven analytic methods, and secure infrastructure and methods can derive valuable findings from large, complex, and heterogeneous datasets. Standards, shared specifications, and trusted brokers can enable and enhance data sharing. Suites of tools and methods can mix protections according to different use cases' (The Royal Society; National Academy of Sciences, 2024, p6).

The benefits of data sharing identified in this analysis can broadly be divided into four categories: socio-economic benefits; environmental benefits; benefits to industry as a whole; and benefits to the specific organisations sharing their data.

Socio-economic benefits

Government and policy sources reviewed were most likely to highlight socio-economic benefits of data sharing that included, but reached beyond, direct benefits to data users. The National Data Strategy states that: 'By embracing data and the benefits its use brings, the UK now faces tangible opportunities to improve our society and grow our economy' (UK Government, 2020, Section 1). This quote summarises the overarching perception, echoed more implicitly across the wider grey literature, of the substantial benefits that data sharing can bring to the UK economy and society.

By leveraging data and data-driven technologies, such as AI, the strategy anticipates that the UK can boost productivity, improve job quality, transform public services, enhance health outcomes, and support crime reduction and decarbonisation efforts. Improved public sector access to data enables better decision-making at scale, guiding housing developments and enabling a reduction in infrastructure disruptions. Data can also contribute to creating a more inclusive society (UK Government, 2020). Meanwhile a recent report by the Office for Statistics Regulation notes that sharing and linking data drives socio-economic development by fostering policy development and innovation (Office for Statistics Regulation, 2024).

In the energy sector, data sharing has been identified to support key government objectives such as energy security and equality, promoting innovation, and ensuring system flexibility and resilience (Arup, Catapult Energy Systems, University of Bath, 2023). While environmental aspects are addressed in more depth in the section below, both data sharing and wider energy system digitalisation have also been identified as potentially holding a role in supporting a socio-economically 'just' or 'fair' transition to Net Zero (Judson, 2023). For example, the Smart Data Communications Company (DCC) launched their 'Data for Good' work program in 2021, which continues to support research into how smart metering system data could be used to address socio-economic concerns such as fuel poverty, in the context of the transition towards a decentralised, flexible, net zero energy system.

A further key benefit noted of cross-sector data sharing in several reports is its necessity for the development of large-scale digital twins which rely on multi-directional information exchange, often in close to real time. Such tools are essential for a sustainable future by, for example, improving decision-making relating to critical

infrastructure and enhancing productivity (Centre for Digital Built Britain (CDBB), 2020; Hetherington & West, 2020).

Environmental benefits

In 2017, the Conservative Government launched an industrial strategy which positioned data sharing, and the associated developed of wider digital products, services and markets, as essential to underpinning an economy-wide net zero transition that achieved co-benefits for economic development, international competitiveness, and regional equality (UK Government, 2017, withdrawn 2023). The strategy was subsequently replaced by the national 'Plan for Growth', which continued to advocate for the importance of data sharing and digitalisation in creating economic prosperity through the Net Zero transition (UK Government, 2021). While the new Labour Government continues to build their digital and environmental policies, their manifesto commitments concerning climate, water and energy governance suggest that continuity can reasonably be expected with regards to ongoing prioritisation of data and digital policy development (Labour, 2024).

Beyond the policy space, creation of large-scale digital twins highlighted in the previous section were also recognised for their potential to reduce environmental impact, for example by predicting (and helping develop solutions for) climate risks (Centre for Digital Built Britain (CDBB), 2020; Hetherington & West, 2020).

In relation to sharing infrastructure data in specific sectors, in the energy sector environmental benefits highlighted included the UK government - and regulated entities - being better able to meet strategic and legal objectives around net zero (Arup, Catapult Energy Systems, University of Bath, 2023). Ongoing initiatives in the water sector, such as Stream and H2Open, similarly identify how data sharing (initially focusing on open data publication) can contribute to the achievement of wider environmental goals such as water quality improvements (Ofwat, 2023, Stream, no date). In the transport sector, TfL's publication of open data contributes to reduced emissions and improved air quality through customer-facing products that encourage walking or cycling over the use of private or public vehicles (Deloitte, 2017b).

Cross-sectorally, it has been recognised that increasing access to engineering data can reduce impacts on the environment by reducing waste in construction or engineering processes, or through better management of energy and water resources (Dodds et al., 2019). Separately, R&D project Perseus is trialling how granular resource consumption data from SMEs (e.g. electricity, gas, water) can be shared with carbon accounting providers in order to improve access to green finance products which require the

provision of greenhouse gas emissions data (Icebreaker One, 2024). While currently patchy, it is expected that grey literature concerning the environmental benefits of cross-sector data sharing will increase as further initiatives are trialled.

Benefits to industry

In the energy sector, literature indicates that data sharing can lead to lower system costs through efficiencies (Arup, Catapult Energy Systems, University of Bath, 2023). This holds potential to both benefit industry players and end consumers. In the transport sector, open data in London has been found to foster a virtuous circle that benefits both users and providers of transport networks (Deloitte, 2017b). For example, by giving external users access to their data, Transport for London has gained access to insights that stimulate new ways of thinking and can support operational improvements.

More generally across the engineering sector, sharing data has been recognised as having the potential to offer significant benefits to businesses, employees, and clients. These include enhanced safety through monitoring and improving working conditions, increased productivity in construction and engineering by fostering better collaboration, and driving innovation by involving more communities in service development. Additionally, data sharing can spur economic growth by enabling new business models and reducing costs across sectors like engineering and construction (Dodds et al., 2019).

Direct benefits to those sharing the data

Research carried out in the US shows that reasons for companies to share data with academics include gaining insights to support their corporate missions, demonstrating the value of and increasing trust in their data, and (to a lesser extent) supporting their philanthropic missions by sharing data for the public good (Harris & Sharma, 2017). This range of motivations suggests that for some organisations, particularly those already at a high level of data maturity, the socio-economic or environmental benefits may be enough of an incentive for them to share their data. Some others, particularly in the private sector, may be more driven by direct benefits to them. This finding is significant for any organisations carrying out engagement or communications work concerning data sharing, as it implies that there may be value in segmenting motivations for engagement.

While benefits to data providers are often framed primarily in terms of financial gain (e.g. paid for data services as a revenue stream), the range of benefits far exceeds this and can also apply to free or open datasets. The benefits that TfL has reaped from making many of its datasets openly available may offer some motivation. As Deloitte's

report into the value of TfL's open data and digital partnerships highlights: 'Through partnerships with major data and software organisations, TfL receives back significant data on areas it does not itself collect data' (Deloitte, 2017b). A similar view is reflected in a 2021 Ofwat report on open data in the water industry. This notes that companies generating data may not always find the best uses for it, highlighting the value to data owners of broader data sharing (Ofwat, 2021).

The benefits to private data providers of sharing data specifically with researchers were reflected in several reports. Firstly, collaboration between industry and researchers offers significant business value, including improving data quality and providing actionable business insights – to support R&D and deliver products and services, for example (Harris & Sharma, 2017; Office for Statistics Regulation, 2024). Secondly, sharing data with researchers can also have reputational benefits for private companies, increasing transparency and potentially improving their brand PR image (Harris & Sharma, 2017; The Royal Society; National Academy of Sciences, 2024).

Finally, a knock-on benefit of fostering mutually beneficial relationships between researchers and industry through data sharing is the creation of a pipeline of highly valued skills that enables companies to recruit talent (Deloitte, 2017; Office for Statistics Regulation, 2024).

Section 2: Good practice examples of data sharing

Many of the benefits of improving access to data sharing in infrastructure highlighted in published literature are largely theoretical, as robustly measured and documented examples of the benefits of data sharing are relatively limited. Additionally, this review identified a gap in the literature concerning publicly available measurement and reporting of the benefits of sharing data that has restrictions ('Shared Data'[3]); many of the published good practice case studies are based on Open Data alone and/or are fairly small scale in nature.

Some examples of data sharing driving measurable benefits can be drawn from non-infrastructure sectors, in particular finance and health.

[3] Shared Data is shared with a preemptive licence between multiple parties, subject to certain restrictions.

Open Banking is often highlighted as a successful example of data-sharing initiatives, offering valuable lessons for other sectors. Initially driven by regulatory pressure[4] on incumbent banks, it has fostered increased competition and spurred innovative services (Dodds et al., 2019). The UK government also recognizes Open Banking as a model of good practice in data sharing, particularly of consent-managed personal data, demonstrating how regulation can play a crucial role in facilitating such advancements (UK Government, 2020).

Meanwhile health data played a crucial role in helping the UK understand and respond to the COVID-19 pandemic, exemplified by the OpenSAFELY platform. OpenSAFELY is a set of software tools designed to run analyses on electronic health records while maintaining data security. Researchers write code using synthetic data and submit it to be executed on real patient data in a secure environment. The outputs are reviewed by the OpenSAFELY team to ensure patient confidentiality. This system addresses privacy concerns, facilitates code sharing, and allows transparency in the reuse of previous analyses (The Royal Society; National Academy of Sciences, 2024).

It is also worth noting, however, that a cautionary tale can be drawn from the health sector. The General Practice Data for Planning and Research (GPDPR) programme[5] was designed to enable flow of pseudonymised data from GP surgeries into a central NHS database, to aid more rapid and effective health research and planning. But concerns by leading medical bodies about a lack of transparency and public information, combined with public concerns about transparency and reidentification resulting in mass opt-outs, led to the programme being halted (Yalamanchili, 2024). This emphasises the importance that any data sharing that involves personal data is supported by appropriate measures to safeguard privacy and trust (as will be discussed further in the section on barriers to data sharing).

The sub-sections below explore specific published examples of good practice in infrastructure data sharing, cross-sectorally and specifically for each of the target sectors. The review includes both initiatives to facilitate and support data sharing, and real-world examples of data sharing in practice.

[4] https://www.fca.org.uk/publications/policy-statements/ps17-19-implementation-revised-payment-services-directive

[5] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research/about-the-gpdpr-programme

Cross-sector

Initiatives to facilitate and support data sharing

In September 2020, the UK Government published the National Data Strategy, which identifies five 'concrete and significant opportunities' for data to positively transform the UK. The third opportunity is increasing the speed, efficiency and scope of scientific research (UK Government 2020). In itself, this strategy – alongside sector-specific strategies highlighted later in this report – is an example of good practice, as it recognises the need for a focus on better data sharing in the UK.

Strategies have also been developed that focus specifically on researcher access to data, such as the Smart Data Research UK Strategy – a UKRI initiative to develop research partnerships that allow researchers to access and use smart data (data generated through engagement with digital systems, devices and sensors) (Smart Data Research UK, UKRI, 2024). Such data might include, for example, navigation systems, public transport data, and connected vehicle data. Part of the strategy that is of high relevance for this analysis is around federated data – building partnerships with data owners that allow sharing with multiple accredited researchers.

Alongside such strategies, the literature reviewed identifies a number of cross-sector platforms, programmes and other initiatives that have a role to play in facilitating and supporting better data sharing in the UK, including the UK Government's Building Information Modelling (BIM) programme, the National Digital Twin programme, data.gov.uk, and the Integrated Data Service (IDS). Summaries of these and other relevant initiatives identified in the literature have been provided in Appendix 1.

Effective data sharing in practice: cross-sector case studies

UK examples

- **National Underground Assets Register (NUAR).** This register, which covers electricity and phone cables, gas and water pipes across England and Wales, was recognised in several reports as being a high-value infrastructure. Findings from pilots indicate significant economic and social value would be unlocked by creating a national register and ensuring field operatives have accessible data to carry out their work safely and effectively. At present available as a Minimum Viable Product (MVP) to planners and excavators, the NUAR is scheduled to be fully operational by the end of 2025, and work is currently underway to explore the benefits and risks of widening access. According to Ofwat, once complete the NUAR is expected to deliver £350m annual benefits through avoiding accidental asset strikes (Sources: UK Government, 2020; Ofwat, 2021).

- **Central hub for roadworks data.** Elgin aggregates roadworks data from various infrastructure bodies onto a centralised, publicly accessible portal [now the Causeway One.network platform[6], which can be shared via an API]. This portal collates data shared by Local Highways Authorities, metropolitan transport authorities, Network Rail, and utility companies in the gas, water, electricity, and telecom sectors. The centralised data hub can improve coordination and transparency across these sectors, and help to manage roadworks more efficiently (Source: Deloitte, 2017a).

International examples

- **A central repository for asset data in Belgium.** This example reinforces the importance of initiatives such as the NUAR. After a 2004 explosion caused by construction workers piercing a gas main in Belgium, the Flanders region implemented regulations requiring all 300 utilities companies to provide digital representations of their infrastructure using a common data model. This information is shared via a single repository to enhance the safety of construction work. By modernising regulations to adapt to new technology, Flanders has achieved a reduction in accidents, as well as significant cost and time efficiencies (Source: Dodds et al., 2019).

- **Emergency planning in Puerto Rico**. In the aftermath of Hurricane Maria in 2017, which caused widespread devastation and a water crisis in Puerto Rico, researchers developed new data infrastructure to aid scientific investigation and data-driven decision-making for disaster recovery. They are now utilising the HydroShare data platform to share data, models, and code, ensuring these resources are easily findable, accessible, reusable, and interoperable. This initiative supports more efficient responses and recovery efforts in the wake of natural disasters (Source: Dodds et al., 2019).

- **Supporting the use of building information models in Singapore.** A joint collaboration between Ordnance Survey (OS) and the National University of Singapore, aimed at making BIM model data more accessible to stakeholders such as city planners or regulatory bodies, resulted in the development of the *IFC2CityGML transformation engine*, a software tool that can automate the transfer of detailed building model information for various geospatial use cases. This project not only streamlined the use of building information models (BIM) for mobility, energy, and urban planning but also fostered greater engagement and collaboration between the BIM and geographic information systems (GIS) communities in Singapore (Source: UK Government, 2020).

[6] https://uk.one.network/

Energy

Initiatives to facilitate and support data sharing

Of the three infrastructure sectors that Icebreaker One has been asked to focus on in its landscaping analysis, the energy sector appears to be implicitly recognised as leading the way compared with water and transport. This is likely in large part down to, as noted in Ofwat's reports on open data in the water industry, the introduction of regulatory requirements for network operators to publish Open Data to support decarbonisation and encourage behavioural changes in the industry. Ofgem's requirements are supported by its Data Best Practice Guidance, introduced in 2021, which aims to enhance data interoperability across the sector to support decentralisation and decarbonisation and meet the UK's 2050 Net Zero target. This guidance includes principles ensuring reliable, secure, and exploitable data, with regular updates to ensure it delivers the desired outcomes (Ofgem, 2023). In parallel, Ofgem also introduced new regulatory requirements for energy networks to regularly publish a Digitalisation Strategy and Action Plan, enabling more transparent recording of industry strategy, and associated progress against both data and wider digitalisation targets (Ofgem, 2021, updated 2023).

Meanwhile the introduction by government and Ofgem of The Energy Data Task Force, which is dedicated to reducing costs and facilitating innovation by improving data transparency and availability, was recognised by several sources as helping drive better access to energy sector data alongside its associated project on Modernising Energy Data Access (MEDA) (Black & Veatch, 2019; UK Government, 2020). One positive outcome of the MEDA project highlighted in a number of reports was the Icebreaker One's Open Energy Data Portal, which makes it easy to search, access, and securely share energy data (Ofwat, 2021; PricewaterhouseCoopers, 2023; Evans & Johnston, 2023). The Open Energy initiative continues to develop data sharing governance via the Energy Sector Trust Framework (Icebreaker One, 2024).

Other beneficial developments in data infrastructure include the National Grid Electricity Distribution's Connected Data Portal (CDP) and the government's Automatic Asset Registration (currently in development) (PricewaterhouseCoopers, 2023; Evans & Johnston, 2023). Such initiatives can reduce data-sharing costs through consistent approaches to data licensing (reducing legal overheads), interoperability of data (reducing the need to transform data to meet requirements) and the people capacity to manually cater for every actionable request for data. Efforts to link CDP's data into the Open Energy solution have improved the visibility of the data available to data users (Evans & Johnston, 2023).

The value of data to support energy planning by local authorities was noted in summary reports of workshops commissioned by the UK Energy Research Centre (UKERC). The Energy Systems Catapult has also developed a Local Area Energy Planning (LAEP) roadmap, to provide modelling that supports the needs of local authorities across numerous localities, in recognition of the fact that: 'Whilst local data is critical to robust modelling, there is a need to triage data; scenario preparation is resource intensive and data priorities should focus on areas of highest impact' (Britton et al., 2023, p10).

Meanwhile a 2024 UKERC report on local energy planning to accelerate net zero emphasised the need for accurate, standardised data for effective energy planning, with accessible tools to support local authorities in their energy modelling efforts. In this regard, it was highlighted that there are notable differences in approaches to planning across the UK's devolved administrations. For instance, the Welsh Government funds Local Area Energy Plans (LAEPs) for all local authorities, while it remains optional in England. Meanwhile in Scotland, there is now a statutory obligation to publish Local Heat and Energy Efficiency Strategies (LHEES) by the end of 2023, supported by standardised data (Britton & Webb, 2024).

Effective data sharing in practice: energy case studies

- **Energy modelling in Greater Manchester.** The Greater Manchester Combined Authority has five year Environment Plans and carbon budgets feeding into its strategy for meeting its 2038 net zero targets, and plays an enabling and empowering role for local authorities. Since 2016, its experience in energy modelling and Local Area Energy Planning (LAEP) has led to the Greater Manchester (GM) Local Energy Market trial and the creation of LAEPs for each local authority in the region in 2022. The first phase of the project showed that results were often similar across different areas, suggesting the potential for adopting 'cheaper, faster modelling techniques with the main model used for validation and calibration'. (Source: Britton et al., 2023, p8-9)

- **Forecasting future demand response.** A Doctoral Researcher used historical time series data from the NGED's Connected Data Portal, particularly that related to low voltage, to forecast future demand response and extrapolate future trends. (Source: Evans & Johnston, 2023)

- **Short term predictions for solar energy.** Smart home technology firm Passiv is responsible for managing the performance of a large number of solar panels in the UK, and shares solar panel data with a team at Sheffield University. The University uses the data to forecast solar electricity generation in half-hourly intervals from sites across the portfolio. At the time of the report, the data was also shared with Open Climate Fix, a non-profit which was developing

machine-learning algorithms to make more accurate predictions for solar electricity generation, with a view to helping tackle challenges around renewable energy supply and demand. (Source: Ofwat, 2021)

Water

Initiatives to facilitate and support data sharing

In 2021, Ofwat published H2Open – Open data in the water industry: a case for change. Its report highlighted its expectations for water companies to make better use of data, including by embracing open data. At the time, it stopped short of placing specific regulatory requirements on water companies in this regard.

However, despite the introduction in 2021 of Stream, a collaborative industry-wide data initiative led by Northumbrian Water to explore how the benefits of open (and, more recently, shared) water data can be unlocked, in a 2023 review of progress Ofwat recognised that things had not developed as far or as fast as it had hoped. It therefore decided to 'start work to develop new powers that would enable us to take action against water companies that are not making sufficient progress in [making data open]. We will commence work on a licence condition that would require water companies to deliver the outcomes and expectations we are seeking on behalf of customers and the environment' (Ofwat, 2023).

Effective data sharing in practice: water case studies

- **Creation of an open data marketplace.** Wessex Water has developed a data marketplace, a central hub that brings together and provides open access to different datasets including APR and EDM data. It has used the marketplace as a basis for working with the wider market on specific datasets and challenges/hackathons to develop greater learning opportunities (Source: PricewaterhouseCoopers, 2023).

- **Advanced emergency weather warnings.** The British Red Cross combines open data published by the Environment Agency (EA) on flooding with data from the Met Office to provide advanced emergency weather warnings on its app. The EA's flood data is also used by the Pang Valley Flood Forum to display live dashboards for local communities (Source Ofwat, 2021).

- **Use of leakage data for predictive analytics.** Alongside datasets on drinking water quality and customer meters, Yorkshire Water opened a leakage through Data Mill North, a collaboration between Leeds City Council and other stakeholders. Ideas generated using leakage data included new methods of leak detection (Ofwat, 2023).

- **Detecting pipework defects.** Dŵr Cymru used CCTV labelled images captured from sewer surveys to train an AI model to detect pipe defects. The automated solution identified in this project would improve efficiencies by freeing up employees to work on more important tasks. Since the project, Dŵr Cymru has been working with the Water Research Centre (WRC) on a project to establish a sector wide repository of sewer CCTV labelled images that water companies can use to train their own AI models (PricewaterhouseCoopers, 2023).

- **Managing the risk of lead service pipe contamination.** Hafren Dyfrdwy have worked with Sheffield University to use property data, housing records and other geographical information to predict the likelihood of properties being at risk of contamination (PricewaterhouseCoopers, 2023).

Transport

Initiatives to facilitate and support data sharing

Following on from the earlier development of data strategies for energy and water, in 2023 the Department for Transport (DfT) published its own Transport Data Strategy: Innovation through data. This covers the areas of sharing, discoverability and access (including the development of a transport data catalogue); data standards and quality; skills, culture and leadership; user needs and communication; and governance, protection and ethics. Almost all of the examples of good practice around facilitating data sharing in transport have been drawn from this strategy (Department for Transport, 2023).

Key services delivered by the DfT include Street Manager[7], which provides data on planned and live street works to support better planning and management, and Bus Open Data[8], improving customer access to information on bus trips (PricewaterhouseCoopers, 2023). The National Public Transport Access Nodes (NaPTAN) service[9], which is crucial for supporting accessibility needs at public transport access

[7] https://www.gov.uk/guidance/plan-and-manage-roadworks
[8] https://www.bus-data.dft.gov.uk/
[9] https://www.data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan

points, is currently being redeveloped and modernised (PricewaterhouseCoopers, 2023). And the Rail Data Marketplace[10] enables third parties such as app developers, tech firms and academia to use rail data to drive innovation and enhance services and user experiences. This marketplace, along with the overarching Rail Technical Strategy[11] that highlighted the transport sector's ambitions to improve the sharing of data across the railway industry, was also recognised as an example of good practice in an assessment of open data in the water industry (PricewaterhouseCoopers, 2023).

Other initiatives include the Office for Zero Emission Vehicles (OZEV)'s work with industry to make public chargepoint data openly available, which will help drivers plan journeys, give Distribution Network Operators the information they need to plan interventions, and enable operators to optimise electric vehicle charging networks. The Transport Infrastructure Efficiency Strategy (TIES) team has developed living labs[12] to use data for more efficient infrastructure project management. And following on from user research carried out as part of its strategy work, the DfT plans to create Local Authority Transport Data Guidance to improve data use across the sector. It's anticipated that this guidance will help local authorities use and share their transport data for social good, for example by enabling people to better plan their journeys through improved roadwork data.

Finally, from the Transport Data Strategy, while the DfT promotes open data by default, it recognises that not all data can be made open; The Centre for Connected and Autonomous Vehicles (CCAV) and Future Transport Zones (FTZs), as well as Street Manager, are exploring secure data sharing solutions (Department for Transport, 2023).

Effective data sharing in practice: transport case studies

- **Transport for London's (TfL) open transport data.** TfL has been publishing large amounts of open data about its transport network for more than a decade, and is cited in several reports as an example of good practice in data sharing. As of 2017, the opening up of its datasets was estimated to have contributed up to £130m per year to the London economy in savings for commuters, generated several hundred jobs, and delivered millions of pounds a year to companies using the data. While TfL's open data is available to all, rather than specifically to researchers, a Deloitte report outlining its value notes that 'Businesses…as well as a large number of academics and

[10] https://raildata.org.uk/
[11] https://railtechnicalstrategy.co.uk/
[12] Living labs enable collaboration between a range of stakeholders at local, national or international level to address various challenges, with a focus on co-creation and live testing. Source: https://www.surrey.ac.uk/sites/default/files/2023-01/what-is-a-living-lab.pdf

professional developers partner with TfL and use this data to create new commercial and non-commercial customer-facing products and services.' (Sources: Dodds et al., 2019; Deloitte, 2017b; UK Government, 2020)

- **Increasing safety at sea.** Two examples from the maritime and offshore sector highlight the role that data can play in enhancing safety. The EfficienSea2 project facilitates automated reporting on safety issues, weather conditions and smarter navigation through improved data infrastructure, providing better connectivity for ships. Meanwhile the HiLo (High Impact, Low Frequency) project facilitates the sharing and combining of datasets by multiple organisations, thus enabling the development of better predictive models for asset failures. According to the source, the project has already substantially reduced risks related to lifeboat accidents (Source: Dodds et al., 2019).

- **Supporting sustainable transport policy design.** One notable example of good practice in data linkage is the collaboration led by the University of Leeds, in partnership with the Department for Transport (DfT), DVLA, DVSA, ONS, RAC Foundation, and the University of Bristol. This project connects administrative vehicle data, linking vehicle attributes and MOT data from the DVLA with vehicle standards from the DVSA. The resulting dataset covers all light vehicles in Great Britain and includes information on vehicle type, mileage, emissions, and location at the level of Lower Layer Super Output Areas (LSOAs). This dataset plays a crucial role in supporting sustainable transport policy design at both local and national levels (Source: Yalamanchili, 2024).

- **Sharing car accident data to improve safety.** In 2019, Volvo announced that it would begin sharing insights from 40 years of car accident data with other car manufacturers to enhance vehicle safety. This data-sharing initiative allows manufacturers to better understand and address biases and gaps in existing data, including those related to the predominance of testing based on "average" men, a longstanding issue in vehicle safety design. By making this data accessible, Volvo aims to help the industry produce safer vehicles that are suitable for a wider range of individuals[13] (Source: Dodds et al., 2019).

- **Local Authority parking and traffic data sharing.** As part of its Local Transport Plan, the Department for Transport (DfT) funded eleven local authorities to open their parking and traffic data, which were identified as highly valuable during the Discovery phase. With a relatively small investment,

[13] While not strictly an infrastructure-related example, it is a good example of a car manufacturer sharing data that could be relevant for infrastructure. For example, if a car's smart functionality recorded data about the specific time and location of accidents that could indicate a high-risk area of road. However, such data would likely be of more value if all car manufacturers made this type of data available, rather than only Volvo.

these projects are delivering a range of benefits (Source: Department for Transport, 2023).

- **Heat mapping for improved critical asset access.** A small-scale 2018 academic project, supported by a mini-project award from the Centre for Digital Built Britain, demonstrated how heat mapping using large data sources such as Google Maps could inform transportation-related asset management decisions. The project, focused on solving commuting issues in Cambridge and Hertfordshire, focused on understanding access to critical assets such as research centres and hospitals. It confirmed that large-scale data repositories like Google Maps could support asset management and criticality analysis. For example, the findings indicated that adding a bus route in a specific area had the potential to improve journey times to a critical asset. The project underscored the importance of good quality data for effective heat mapping (Source: Salvador Palau et al., 2018).

- **Easing traffic congestion in China**. Finally, an international example illustrates the benefits of combining private and public datasets. In China, analysis of data from a leading ride-hailing app in combination with data from smart traffic signals allowed authorities to optimise traffic light changes in real time, easing congestion in a high-traffic area of China (Source: Deloitte, 2017a).

Section 3: Barriers to data being shared for research

As outlined earlier in this report, the theoretical benefits of sharing data are widely recognised and there are a number of notable examples of good practice that demonstrate data having been shared with tangible benefits and/or of initiatives supporting better access to data. However, literature analysed for this review reveals an overarching finding that the UK has not yet reached the point where data sharing (at all, let alone specifically for research purposes) is a default activity for most organisations.

The nature of the barriers identified through this analysis have changed little from the earliest reports reviewed (2017) to the most recent (2024), indicating that there is still much work to be done to address these challenges. Future research could potentially benefit from further exploration of the range of sociotechnical barriers to achieving change, enabling stronger recommendations to be made in this regard.

Many of the barriers highlighted in the sections below are generalisable to data sharing for any audience type since, as noted earlier in this report, there is limited literature

available which specifically addresses barriers to data sharing for research purposes. However, it is likely that researchers face similar challenges to other audiences in many respects. Please note that this section focuses on the barriers to data sharing between data owners and researchers. Some literature also covered barriers to researchers subsequently sharing the outputs from their data analyses; these have been captured in Appendix 3.

Points with particular relevance to sharing infrastructure data, or to sharing data for research purposes, have been highlighted. For example, accessing data that was not designed with research in mind can be costly, controversial and unreliable (The Royal Society; National Academy of Sciences, 2024). Much of the data that researchers need, such as information about social behaviour and economic activity, is held by the private sector – which might face commercial or other barriers to sharing it, as outlined below (Harris & Sharma, 2017; Office for Statistics Regulation, 2024). Even when data is initially shared by private firms, there can be a fear by researchers that data owners may revoke access to essential datasets (The Royal Society; National Academy of Sciences, 2024).

While some barriers apply most strongly to data held by private firms, public bodies are not immune from criticism in the literature reviewed. In its follow-up report on data sharing and linkage for the public good, the Office for Statistics Regulation noted that: 'Despite welcome pockets of innovation, there continues to be a failure to deliver on data sharing and linkage across government, alongside many persisting barriers to progress. Linking datasets for research, statistics and evaluation – both across government and among external researchers – is not yet the norm in the UK statistical system' (Office for Statistics Regulation, 2024, p3). Meanwhile a 2017 report commissioned by the National Infrastructure Commission flagged the difficulty of accessing data – such as transport-related data – from local government (Deloitte, 2017a).

It is also worth noting that barriers to accessing a single dataset can be exacerbated when a researcher needs to link datasets from multiple sources together. For example, in a project designed to investigate the value of using data from multiple sources to improve road safety, the researcher found that the complexity of combining four different datasets from different sources proved challenging, including time lags, incomplete data/no common identifier, ethico-legal issues, data silos, and administrative processes (Yalamanchili, 2024).

Meanwhile, one report, assessing the extent of open data in the water industry, noted that there can be a lack of feedback loops between data providers and data users. This can mean that data providers may not prioritise access to the right datasets, or may not

do so in a way that makes it easy for the data to be used (PricewaterhouseCoopers, 2023).

The barriers to data sharing can broadly be broken down into five categories: legal and privacy, security, commercial, and technical. However, there can be strong overlap between categories, with barriers in one category often causing further barriers. For example, ingrained cultural resistance to data sharing, combined with outdated legal practices, can make organisations reluctant to make the commercial decision to invest in technology to support data sharing. In combination, the range of barriers identified can contribute to 'business as usual' no-change attitudes.

Legal and privacy barriers

Privacy and other legal concerns are often cited as reasons given for organisations' reluctance to share data. Researchers wishing to access data must often navigate the ownership and control both of data-collecting organisations and of the individuals or communities involved (The Royal Society; National Academy of Sciences, 2024).

Key repeated legal challenges include GDPR compliance, determining when consumer consent is required, maintaining consumer trust, and protecting intellectual property (IP) rights. Solutions such as data aggregation and anonymisation are often suggested to address these issues (Deloitte, 2017a). However, confusion surrounding data rights and licensing, which involve overlapping regulations related to IP, data protection, and freedom of information (FOI) or Environmental Information Requests (EIR), can complicate the process (Dodds et al., 2019).

Legal barriers to data sharing

Privacy and data protection concerns are not the only legal barriers raised in relation to data sharing. Companies handling data from corporate or government clients face limitations due to data provenance, customer expectations, IP risks, and regulatory constraints (Harris & Sharma, 2017). According to the Department for Transport's data strategy, legal and contractual uncertainties often leave organisations unsure about what data can be shared and at what level. The Transport Data Strategy notes that this barrier can be exacerbated by outdated contracts, often established before data was considered a valuable asset (Department for Transport, 2023).

The time it takes to negotiate the legal maze can also be prohibitive. A US-based report exploring practices for sharing corporate data with researchers found that negotiating data-sharing agreements can take months, and each researcher often has to rebuild trust with every agreement (Harris & Sharma, 2017). And a 2024 report by the Office for

Statistics Regulation highlighted lengthy legal reviews of complex, non-standard data-sharing agreements, which can delay or deter researchers from accessing government data (Office for Statistics Regulation, 2024).

Meanwhile, even if legal approval for data sharing is eventually granted, restrictions on how researchers can use or further share that data are common (The Royal Society; National Academy of Sciences, 2024). Concerns about IP implications (as well as privacy) often lead companies to limit data use (including sharing with others) or impose restrictions on how data can be combined with other datasets (Dewey, 2023; Harris & Sharma, 2017; The Royal Society; National Academy of Sciences, 2024).

Privacy barriers to data sharing

Concerns about breaching privacy legislation was a strong theme across the majority of reports that covered barriers to data sharing, relating to infrastructure or otherwise. There can be a focus by organisations on protecting privacy over enabling research utility (Harris & Sharma, 2017). Indeed, in an (admittedly small-scale) survey of 11 commercial organisations across five European countries, GDPR was identified as the biggest organisational barrier to the creation and use of data-driven energy services (Psara et al., 2022). In general, concerns over the misuse of personal and confidential data, including re-identification risks, are prevalent (Black and Veatch, 2019; Harris & Sharma, 2017).

Resistance to data sharing can often arise from an deep-rooted conservative approach by organisations, particularly when there is uncertainty about whether consent covers the intended use of the data (The Royal Society; National Academy of Sciences, 2024). Many organisations struggle with the legal boundaries concerning the use of personal data, and may lack the expertise or knowledge to anonymize data or make informed data-sharing decisions (Dodds et al., 2019). Although not extensively explored, it can be anticipated that these concerns disproportionately affect smaller or more cost-constrained research initiatives, potentially acting as an ‘invisible’ barrier to data use within subsections of the research community. For example, this may disproportionately affect the use of personal or consented data by Doctoral or Early Career Researchers to whom legal resources may be less readily available.

To an extent, concerns about breaching privacy legislation are easy to understand, especially given a perception that individuals are generally reluctant to share their data unless it is clearly anonymized or they are remunerated (The Royal Society; National Academy of Sciences, 2024). The same report noted that the lack of an empirical test for GDPR privacy standards further complicates the sharing of medical data, leading organisations to err on the side of caution (The Royal Society; National Academy of

Sciences, 2024). While this example is drawn from the health sector, the same issues are likely to apply to much personal data, particularly if it is sensitive in nature either in isolation or via data combination (e.g. inferences drawn from combining energy data with health data).

A specific example from the health sector sheds further light on company conservatism. Privacy concerns halted the General Practice Data for Planning and Research (GPDPR) program in 2021 due to fears of re-identification and lack of transparency (Yalamanchili, 2024). The COVID-19 pandemic has exacerbated public fears, particularly around foreign firms accessing sensitive data (Yalamanchili, 2024). The public may feel more confident when allowing their data to be shared with accredited academics in the UK for research that is clearly for the public good, provided they have assurances about the security of their data.

Meanwhile a 2019 report about the use of energy smart meter data for research and innovation highlights the challenges of accessing personal data from the researcher perspective. In this report, it was noted that researchers wanting to access smart meter data must get consent from every participant in a study and, the highly secure systems needed to access smart meter data via the DCC Gateway were regarded as 'expensive to develop and implement' (Webborn et al., 2019).

It is widely recognised that protecting sensitive data and safeguarding public trust is essential for sharing personal data with researchers (Office for Statistics Regulation, 2024). However these barriers must be addressed for the full benefits of such data to be recognised. While this theme poses several knotty problems, this review highlights the necessity of addressing these meaningfully in order to enable data sharing to achieve the widest possible range of benefits.

Security barriers

Security concerns can be divided into two broad themes.

Firstly, concerns over the security of systems which collect, store, and transfer data are becoming more prevalent as more devices become interconnected, increasing the potential for breaches. Data sharing can increase security risks, and centralised data hubs and platforms can provide access to large amounts of sensitive data if compromised. The healthcare sector has been cited as an example of where data sharing is commonplace, but there have been a number of security breaches in the UK and overseas. Concerns may be particularly heightened when sharing dynamic data (such as sensor data) which may be harder to make secure (Deloitte, 2017a). Legacy

systems in sectors such as energy and water are especially vulnerable to cyber-attacks, with hacks becoming more frequent as smart technologies and IoT evolve, exemplified by cyberattacks in Ukraine causing power outages (Deloitte, 2017a). According to a Guardian article, a similar attack on Britain's energy system in 2020 was warded off thanks to 'robust security measures'[14].
Meanwhile, risks to national security are a particular concern in relation to infrastructure data. This can make data owners cautious and limit the availability of many big datasets (The Royal Society; National Academy of Sciences, 2024; Lieberman, 2019). These risks, such as exposing the locations of combustible energy sources or critical water supplies, discourage organisations from sharing data (Dodds et al., 2019).

However, it is also acknowledged that excessive concern about data security can result in potentially bigger issues. In disaster planning, for example, the lack of standardised and integrated data across strategic and street-level utility infrastructures leaves jurisdictions reactive and poorly prepared for disaster events (Lieberman, 2019). As critical national infrastructures come under further anticipated stress from climate change, there may be requirements for some change in risk management and prioritisation to balance security requirements with evolving resilience challenges.

Commercial barriers

High costs, lack of incentives, and other commercial sensitivities can make organisations, particularly in the private sector, reluctant to make their data available. In some cases, if companies are willing to share data in principle, they may only be willing to do so if they can feel confident that their initiative will be reciprocated by others. Or, even if a minority of commercial organisations within a sector are willing to act as trailblazers, the data they provide may be of limited value without the same data being made available industry-wide (Ofwat, 2021). Commercial barriers can be broadly categorised into financial barriers, and a reluctance to share data that is regarded as commercially sensitive.

Several reports cited a lack of commercial incentive for organisations to share data. Particularly when organisations do not perceive direct benefits to themselves from the data sharing, they may struggle to make a compelling business case to unlock the funding and resources to do so (Department for Transport, 2023; UK Government, 2020).

14

https://www.theguardian.com/business/2020/may/14/lights-stay-on-despite-cyber-attack-on-uks-electricity-system

The lack of commercial incentive can lead to cost being a crucial barrier on both sides of the data provider/data user relationship. The high up-front costs of making data ready to share (including data cleaning, processing, storage and hosting), as well as building organisational data skills, can feel prohibitive for data providers. Cost and resourcing issues can be exacerbated by the absence of clear processes for researchers to access and organisations to provide data. If the same data is requested multiple times by different researchers, the lack of a streamlined way to process such requests can result in wasted time and, potentially, duplication of effort (Smart Data Research UK, UKRI, 2024). This could result in some important requests being refused on the grounds of lack of capacity. In the water industry, Ofwat cited costs as a particular barrier for smaller companies, though this is likely to apply equally across all sectors.

Cost reservations may, however, be linked to a poor understanding of the benefits of data sharing (see cultural barriers), as Ofwat's report also noted that initial costs should be outweighed by operational efficiencies gained (Deloitte, 2017a; Ofwat, 2021). Meanwhile, from a data user perspective, the cost of accessing private data can be prohibitive for academics, as price points for data are usually aimed at other businesses (Dewey, 2023).

Organisations often hesitate to share data due to concerns about commercial sensitivity. Many datasets that could benefit the scientific community are generated by private-sector companies that prefer to keep this information confidential for commercial reasons (The Royal Society; National Academy of Sciences, 2024). For example, certain datasets may be critical for business models and funding decisions, creating fear about the potential negative impacts of data sharing (Dodds et al., 2019). Additionally, there are apprehensions about the possible reduction in the value of intellectual property if data is shared (Harris & Sharma, 2017). While not addressed explicitly in the current literature, this may be a particular barrier to commercial data sharing in the context of research projects for which funding or other conditions stipulate requirements to openly publish research results in a manner which presents some risk of reverse engineering and/or inferences being made about commercially sensitive data inputs. Given that it is important not to discourage open publishing where possible, consideration should be given up-front on ways to mitigate these concerns.

Meanwhile researchers in one US study noted that proposed research often has to align with interests of corporate researchers, leading to a risk that companies rather than academics set the research agenda and data is not used to its full potential (Harris & Sharma, 2017). This holds potential to create conflict with the principles of academic independence and ethical integrity which UK researchers must abide by.

Cultural barriers

Organisational culture plays a significant role in data sharing practices. In some cases, a vicious cycle can emerge where reluctance to share data leads to underinvestment in making data shareable, resulting in concerns about data quality, which further reinforces the reluctance to share (Deloitte, 2017a).

Cultural barriers can include a lack of leadership, low risk appetite, poor data literacy, limited awareness of the benefits of data sharing, and challenges related to staff retention and skills (Office for Statistics Regulation, 2024).

In some cases, even identifying the right person to ask for data can be a challenge for researchers. A blog post published in 2023 by Dewey, a US firm that provides a gateway for academics to access corporate data, noted that companies' marketing is often aimed at their target audience, which is usually not academics. This can result in academics having to go through business procurement processes that might not be fit for purpose (Dewey, 2023).

The cultural barriers identified through this literature review can be broken down into several themes.

Low understanding of the benefits of data sharing

The private sector may fail to recognize the potential of data for broader social and scientific advancement (Harris & Sharma, 2017). More generally, uncertainty regarding the value of data sharing can hinder organisations' willingness to invest resources into collaborative efforts to develop shared data assets for the long term; where data sharing does take place, it might be limited to contained projects that reap short-term benefits (Dodds et al., 2019; Psara et al., 2022). The private sector also frequently fails to recognize the potential of data for broader social and scientific advancement (Harris & Sharma, 2017).

Risk aversion

In some cases, rather than simply not understanding the benefits, organisational aversion to data sharing is more overt. Risk aversion might be as a result of issuing poor-quality data, or sharing it at all. Some reports highlight a general culture of risk aversion in the UK (Gov.uk, 2020), especially in comparison with the US, Australia, New Zealand and much of Europe, particularly when data pertains to safety (Dodds et al., 2019).

Concerns about the potential for reputational damage may be a root cause of risk aversion. These can stem from the risk that issuing poor-quality or old data, or inconsistencies in its application, could have an adverse impact on the data provider's reputation (Harris & Sharma, 2017; Dodds et al., 2019; Black and Veatch, 2019). In one report reviewed, this was noted as being a deterrent for local authorities, who were loath to expose themselves to scrutiny (Deloitte, 2017a).

Organisations may also fear the potential misuse of data that they share, and may either fail to supply or, in some cases, revoke data access as a result (Dodds et al., 2019). This fear may be heightened if the consequences of misuse are severe (in terms of reputational damage, risks to national security, or other incidents).

Reputational risks were not always highlighted prominently as a barrier across all literature. This may reflect the fact that reputational concerns are malleable depending on the external environment (such as media coverage or the electoral cycle), and are therefore less predictable. For example, in Ofwat's 2021 H2Open report, reputational risks were only acknowledged as a minor concern (Ofwat, 2021). However, this predated the media criticism that Thames Water experienced when it was the first water company to make its Event Duration Monitoring (EDM) data fully available via an API.

Lack of governance and strategic leadership

A lack of strategic leadership and focus on data governance in organisations poses a significant barrier to data sharing in general (Psara et al., 2022). In the transport sector, it has been indicated that insufficient leadership results in a failure to prioritise and value data sharing, as well as a lack of clear purpose (Department for Transport, 2023). Separately, while governance of Open Data in the energy sector has seen significant advances in recent years, barriers still exist with regards to tackling the governance of Shared Data, particularly where this may be more resource intensive or sensitive (Judson, 2023).

In the water sector, strategic oversight and having a roadmap for delivery have been identified as areas of weakness; as of 2023, only seven water companies assessed had a general data strategy (PricewaterhouseCoopers, 2023). And in its 2023 review of industry progress on making water data open, Ofwat noted that there has been an 'overemphasis on the enabling development of data platforms and an underemphasis on strategic leadership, planning and skills development' (Ofwat, 2023).

Meanwhile an independent review of the UK Statistics Authority found significant systemic and cultural barriers to data sharing across government, exacerbated by

insufficient prioritisation of data literacy among senior leaders (Office for Statistics Regulation, 2024).

Skill gaps

A shortage of data skills and capabilities was noted across all three infrastructure sectors that are the focus of this analysis. While the energy sector is typically deemed to be further ahead in its data sharing maturity than the other sectors, as recently as 2023 a need was identified to address the shortage of necessary skills by engaging and supporting the academic community and other development programs (Arup, Catapult Energy Systems, University of Bath, 2023).

Similarly, the transport sector faces a lack of data literacy, with both data providers and users often struggling to understand how to share data and extract its value. This skills gap was underscored in research informing the Transport Data Strategy (Department for Transport, 2023). Many water companies also lag behind in developing data capabilities, with those that are less advanced on this front typically lagging behind in their work on open data (Ofwat, 2021).

Siloed thinking

A typical outcome of the cultural barriers already identified so far can be data being held in silos, which can in turn present significant challenges to effective data sharing (The Royal Society; National Academy of Sciences, 2024). Additionally, data monopolies[15] pose a problem by controlling markets and creating barriers to entry and innovation; addressing this requires regulatory action to promote safe and secure data sharing and to support the development of necessary infrastructure (Arup, Catapult Energy Systems, University of Bath, 2023). Even where effective data sharing does take place within sectors, data sharing between sectors – which is essential to maximise the value of data – may be less advanced due to perceived risks, high costs, and closed organisational cultures (Black and Veatch, 2019).

Technical barriers

Some data sharing challenges arise because data were not originally collected for research purposes, and little consideration being given up-front to how and why data could be used. This can result in issues around data quality that can necessitate

[15] Identified to comprise a combination of: 1) the data 'layers' of physical assets monopolies such as networks, 2) the data 'layer' of incumbent companies such as suppliers or generators, 3) secondary data lakes combining - and possibly transforming - multiple primary data sources, 4) digital intellectual property and data generated in relation to this, such as data outputs of algorithms, and 5) data elements of 'back office' (Glachant and Rosetto, 2018) processes and digital infrastructure (Judson, 2023, p.205-210).

extensive “wrangling and engineering”. On the research end, data users may struggle with the software and engineering required to use the data, particularly where the scale of data may far exceed that of other research projects (The Royal Society; National Academy of Sciences, 2024).

Meanwhile the technical infrastructure required is not always adequate, if it exists at all. Data infrastructure must meet stringent security requirements and be able to handle diverse, complex data types, including large volumes at speed. While the capability to support researchers with this data exists in the UK, resources need better connectivity and additional capacity to fill existing gaps (Office for Statistics Regulation, 2024).

Technical barriers typically fall into at least one of the following categories.

Data discoverability and accessibility

If a data user is unable to find out what data exists, they will struggle to access and use it, especially when pertinent data may be held by multiple organisations and across multiple sectors. Issues pertaining to the discoverability of data have been identified as a significant barrier (UK Government, 2020). Decarbonising the transport sector, for example, requires data from a range of sources to be combined (including electric vehicle chargepoint, emissions, traffic and active travel data). Even honing down on one type of data can be challenging; Traffic Regulation Orders data, for example, is held by over 150 LAs in different formats and not all data is stored digitally (Department for Transport, 2023). The dispersed nature of data, which often requires sourcing from a large number of different data providers, can also make practical access to data difficult and time-consuming.

Compounding the challenge of finding the data is poor quality metadata. As a comment in one report on researcher access to data put it, ‘the metadata meant to provide context often are not enough to make effective use of the data’ (The Royal Society; National Academy of Sciences, 2024). Portals that bring together multiple datasets in one place are generally welcomed, but even when they exist they need to be user-friendly to be of value (a 2019 report on infrastructure sharing by the UK Regulators Network noted that Data.gov.uk, for example, was not deemed to be terribly user friendly – at least at the time).

A key point raised in a US report in relation to corporate data being shared with researchers is that, even where data is shared with researchers, it may only be made available to a small group of ‘trusted’ researchers: ‘It is not that [corporate] data is never available to some researchers, but it is unlikely to be made widely available to all researchers’ (Harris & Sharma, 2017, p4). This point was echoed in the Smart Data Research UK Strategy, which noted that ‘Data access agreements are often exclusive to

projects or institutes' (Smart Data Research UK, UKRI, 2024, p15) There can be a reluctance by companies to share data with anyone other than carefully vetted and trusted individual researchers; they may be unwilling to surrender control of their data by sharing it with academic research networks, and are likely to require substantial reassurance about the integrity of potential users.

Data quality and granularity

Variability in the quality of data across sectors poses challenges. This was noted in reports about both the water sector, in which many water companies have only recently digitised their reporting processes (Ofwat, 2021), and the transport sector, where concerns about data quality render organisations hesitant to share imperfect data (Department for Transport, 2023). Such concerns can be exacerbated if companies fear that being honest about data quality may risk them facing penalties (Black and Veatch, 2019).

Granularity of data can also be an issue. In the energy sector, energy consumption data has historically been made available at too high a level and with little contextual information, due to record anonymisation. Smart meter data offers a solution, but its use is currently limited by privacy concerns which must be handled appropriately to ensure consumer trust. And, as noted in relation to data not being collected for research purposes, when the processes behind data collection are unclear, so too is the quality of the data itself (The Royal Society; National Academy of Sciences, 2024).

Meanwhile if the data supplied to users, including researchers, represents only a snapshot in time, this may limit its use; for example if an organisation provides only current data, without equivalent historic data, this may restrict the ability to back-test forecasting models (Evans & Johnston, 2023).

Data formatting

Inconsistent data formatting poses significant challenges across sectors, including the public sector (UK Government, 2020). Some data is stored in formats that are difficult to use or, as relevant for this analysis, not commonly adopted in academia (Dewey, 2023). And a lack of machine-readable open data persists, with many interactive maps not offering downloadable options (PricewaterhouseCoopers, 2023).

As an example, although data exchange arrangements between local authorities and DNOs have improved, formats and access remain inconsistent across the country (Britton & Webb, 2024).

Interoperability

Organisations encounter difficulties maintaining data in interoperable formats (UK Government, 2020), which can often be due to a lack of common frameworks and standards (Dodds et al., 2019; Ofwat, 2021). Researchers often face challenges of Integrating data across misaligned spatiotemporal resolutions, fusing qualitative and quantitative structures, and assimilating real-time data (The Royal Society; National Academy of Sciences, 2024).

Interoperability is a widespread issue, with calls for establishing standards and mechanisms, as well as policy or regulatory changes to address the problem (Arup, Catapult Energy Systems, University of Bath, 2023). There is also a need for flexible and scalable digital infrastructure, supported by sector-wide governance frameworks and open-source tools to assist smaller players (Arup, Catapult Energy Systems, University of Bath, 2023, Icebreaker One 2021, Judson, 2023).

Interoperability challenges can spill over from data into models. A survey of industry data models and reference data libraries carried out to support the development of the National Digital Twin Information Management Framework identified around 30 models, ontologies and reference libraries of potential relevance (Leal et al., 2020). Some models are concerned with things, some with processes. some with activities. This illustrates the complexity of a landscape that creates interoperability difficulties that are not easily solvable and, potentially, a tendency to try and solve by creating yet more models – which only makes the situation more complex. Meanwhile, not all models are open source, which can be challenging for an academic audience (or cash-constrained academic partners, such as local authorities or the third sector) that doesn't necessarily have the money to buy them (or, necessarily, the time to work out if and how they can be integrated with other models without the provision of additional information).

Even an analysis purely of three prominent built environment standards – CityGML, IFC and LandInfra – found issues of interoperability born out of the tools being originally developed for different purposes; in this case as a result of geospatial and built environment domains having traditionally been seen as distinct (Gilbert et al., 2020). As infrastructure digitalisation and data flows continue to blur the borders between traditionally distinct sectors, reasonable assumptions can be made that, without further attention, such interoperability challenges are likely to grow.

The value of ontologies also came up in reference to dataset interoperability. In the context of a study about the importance of ontologies to ensure interoperability in infrastructure, it was noted that: 'Data sets are regularly implemented without ontological consideration. Without explicit top level ontological commitment it is

difficult to: automate reasoning; develop inference (through logic); know the precision of data; differentiate between continuants and occurrents; be certain of data provenance (and connections to the semantic world); and, in general, achieve interoperability' (Varga et al., 2022, p9).

Legacy systems

The technical barriers highlighted above are often compounded by the legacy systems that remain in place in many organisations. For example, in energy, rail and construction sectors it has been found that substantial data still exists in analogue form or in outdated IT systems. This can make the data difficult to access or extract (Department for Transport, 2023). Closed ICT systems (systems comprising hardware, software, data and the people that use them) were also referenced as a barrier in a report on barriers to data-driven services in energy across five European countries (Psara et al., 2022). Meanwhile data collected under proprietary formats, or processed with proprietary algorithms, rather than using open-source alternatives can have limited availability and may have limited value without access to those formats or algorithms (The Royal Society; National Academy of Sciences, 2024).

Section 4: Recommendations for better data sharing

This section summarises the key points drawn from the literature reviewed, noting areas of thematic commonality. A more detailed list of recommendations for better data sharing is included in Appendix 2. A recurring theme emerging from much published literature is that data sharing – in particular by the private sector – will not happen naturally in the current environment. Making it happen tends to require a firm steer from government and regulators. At the time of Ofwat's original H2Open report in 2021, for example, it noted that 'no [water] company has opened large numbers of data sets beyond what is required by Government or a regulator for regulatory purposes' (Ofwat, 2021, p11).

However, it is also acknowledged that 'mandates to treat data in particular ways are of little practical value unless accompanied by the resources to carry out those mandates, including assuring compliance. Sharing data requires infrastructure, frameworks, standards, trained people, methods of curation, and metadata generation' with a need for 'long-term funding of data generation, analysis, and stewardship' (The Royal Society; National Academy of Sciences, 2024, p7).

On a technical level, the need for more standardisation, of both data and processes – within and across sectors – is recognised as essential for the interoperability that is so essential if different datasets are to be effectively combined for maximum value, for example for the creation of a National Digital Twin. Tools to enable the discovery of data are noted as being of high value (Black and Veatch, 2019; The Royal Society; National Academy of Sciences, 2024); it is difficult for data users, including researchers, to make best use of data if they are unaware of what data exists. To make progress in this field, the literature also implies that improving open publication of metadata is a key tool in supporting discoverability.

Privacy concerns were widely identified across the literature reviewed as a significant barrier to data sharing, indicating that measures to address these concerns are essential. This will involve developing technical tools to support this, such as the privacy-enhancing technologies (PETs) referenced in the Office for Statistics Regulation's 2024 report on data sharing and linkage for the public good (Office for Statistics Regulation, 2024), the provision of clear guidance on the legalities of sharing data for research, and ensuring that members of the public are engaged in decisions on the use of their data.

Commercial barriers to sharing data for research must be addressed with corporate data owners, who may need the organisational benefits of making their data available to be spelled out.

Finally, and crucially, there needs to be a cultural shift across governments, regulators, industry bodies and the private sector from an inherently risk-averse approach to data-sharing, to an open culture that sees data sharing as 'business as usual'. Given that the Office for Statistics Regulation has highlighted that there are still 'significant systemic, often cultural, barriers limiting data sharing across government' (Office for Statistics Regulation, 2024, p6), with data literacy not being prioritised, there is a clear need for government bodies to lead by example.

Conclusion

This report has reviewed and discussed grey literature published between 2017 and 2024, concerning the landscape for infrastructure data sharing for publicly funded research.

Findings have been presented in four main sections: benefits of sharing infrastructure data, good practice in sharing infrastructure data, barriers to sharing infrastructure

data, and a discussion of the recommendations present in the existing literature concerning how data sharing might be improved in future.

The benefits of data sharing have been further broken down into:

- Socio-economic benefits, including an increase in productivity, greater innovation, reductions in infrastructure disruptions, and security and flexibility in the energy system.
- Environmental benefits, including reduced emissions, better management of energy and water resources, and enabling the UK to meet its net zero objectives.
- Benefits to industry, including greater efficiencies and the identification of new business models.
- Direct benefits to those sharing the data, including reputational benefits from being seen to be transparent, and receiving back valuable insights about their data that can both improve data quality and facilitate greater use of data by the originating organisation.

Good practice examples of initiatives to facilitate data sharing have been identified, including national and sector-specific data strategies, data-sharing platforms such as the Rail Data Marketplace, and programmes such as the National Digital Twin programme. Examples of effective data sharing in practice have also been drawn out, such as the National Underground Assets Register, which aims to help avoid accidents such as that highlighted in the introduction of when National Rail accidentally hit a telecoms cable, and a number of small scale research projects that illustrate the potential for data if used at scale.

Despite strong acknowledgement of the benefits of data sharing, literature analysed for this review reveals an overarching finding that the UK has not yet reached the point where data sharing (at all, let alone specifically for research purposes) is a default activity for most organisations – across both public and private sectors.

The barriers to data sharing can broadly be broken down into five categories: legal and privacy, security, commercial, cultural, and technical. However, there can be strong overlap between categories, with barriers in one category often exacerbating (or even causing) further barriers. In combination, the range of barriers identified can contribute to ‘business as usual’ no-change attitudes. It’s worth noting, too, that barriers to accessing a single dataset can quickly multiply in number and complexity when a researcher needs to link datasets from several sources together.

Finally, this report rounds up the key recommendations made in the literature reviewed to address the barriers identified, including a firm steer from government and

regulators, backed with the tools, guidance and other resources to fulfil legal and regulatory requirements for data sharing. This must happen alongside a cultural shift to a position where all organisations see data sharing as something to be encouraged, rather than avoided. As one stakeholder summarising key messages from the 2024 US-UK Scientific Forum on Researcher Access to Data put it: 'In all disciplines and at all organizational levels, data should not be seen as something that can be owned but as something that can be released and reused (The Royal Society; National Academy of Sciences, 2024, p54, Richard Sever).

For this to happen, the tone of conversations may need to change, as noted by the Office for Statistics Regulation in its follow up report on data sharing and linkage for the public good: 'The conversation around data sharing and linkage continues to focus on the risks – from the individual privacy risks to the reputational risks to data owners and government departments...It remains important to acknowledge the potential risks associated with sharing data and linked datasets... However, we would also like to see an emphasis on potential benefits for the public good (Office for Statistics Regulation, 2024, p7).

This literature review is only the first part of a suite of research planned to analyse the current landscape of sharing infrastructure data for research. As noted in the methodology, there are gaps in literature specifically covering sharing infrastructure data for research purposes; planned surveys, workshops and interviews are likely to be valuable to focus down on the specific benefits of and barriers to infrastructure data being shared with researchers.

It is also worth reiterating that the nature of the barriers identified through this analysis have changed little from the earliest reports reviewed (2017) to the most recent (2024), indicating that there is still much work to be done to address these challenges. Future research could potentially benefit from further exploration of the range of sociotechnical barriers to achieving change, enabling stronger recommendations to be made in this regard.

Appendices

Appendix 1: Cross-sector initiatives to support data sharing

This appendix summarises key initiatives (referenced in Section 2) identified in the literature review that facilitate and support better data sharing across multiple infrastructure sectors in the UK.

UK Government programmes

- **The National Digital Twin programme:** this aims to develop an ecosystem of connected digital twins which can enable system optimisation and planning across sectors and organisations (Centre for Digital Built Britain, 2020). For example, the National Digital Twin could model the effect of staggered working hours on transport and energy networks at a national and local level. Plans for the National Digital Twin are based on the Gemini Principles[16] of purpose (e.g. it must be used for public good and create value); trust (elements of trustworthiness include security, openness, and quality) and function (e.g. clear ownership, governance and regulation, plus the ability to evolve). Work has also begun on the development of an Information Management Framework for the National Digital Twin, which will 'seek to establish a common language by which digital twins can communicate securely and effectively (UK Government, 2020), with recommendations that this should comprise (Hetherington & West, 2020):
    - A foundation data model: a consistent, clear understanding of what constitutes the world of digital twins
    - A reference data library: the particular set of classes and properties used to describe digital twins
    - Integration architecture: protocols for the managed sharing of data (including discovery, authorisation, transformation and validation).

  A key recommendation made is the need to draw on existing, openly-available tools rather than creating new, bespoke tools.
- **The Building Information Modelling (BIM) programme:** This is an integral part of the joint industry-government industrial strategy for construction. It means, for example, that all government funded projects (i.e. HS2 and Crossrail) must

[16] https://digitaltwinhub.co.uk/download/the-gemini-principles/

follow mandatory data specifications and are constructed in a similar way (Black and Veatch, 2019)

Platforms

- **Data.gov.uk:** Creation of the data.gov.uk site, containing open data published by central government, local authorities and public bodies (Black and Veatch, 2019)
- **Integrated Data Service[17]:** A government multi-cloud platform designed to provide government analysts, statisticians, data scientists, researchers, academics, educators, and scientists with straightforward access to data that can be used for the public good (The Royal Society; National Academy of Sciences, 2024)

Other initiatives

- **The Digital Framework Task Group (DFTG):** The DFTG reports to the Centre for Digital Built Britain. The group brings together government, industry, academia and regulators with the goal of improving the quality and openness of infrastructure data. Publications include Gemini Principles[18] and the Roadmap for delivering an Information Management Framework. (Black and Veatch, 2019)
- **Ontologies**: One report reviewed explores the value of ontologies in relation to infrastructure and cities, recognising several existing ontologies that aim to ensure interoperability (Varga et al., 2022). Examples cited include SEMANCO, which aims to make urban planning and management more energy efficient, and WatERP, which aims to coordinate the management of supply and demand in order to reduce water usage and associated energy consumption. The report notes that 'Ontologies have the potential for system clarity, exposing biases, overcoming narrow perspectives, rewarding pluralism, and enabling stakeholder engagement' (Varga et al., 2022, p8). However, the sheer number of ontologies – which may not be interoperable – can be problematic.
- **Local Resilience Forums (LRFs):** These are multi-agency partnerships formed of representatives from local public services, supported by the Highways Agency and public utilities. To prevent and reduce any impact of local incidents and catastrophic emergencies, the members of the partnerships must coordinate their efforts, sharing relevant information and data as and when required. (Black and Veatch, 2019)

[17] https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/researchexcellenceandpartnerships/integrateddataservice

[18] https://www.cdbb.cam.ac.uk/DFTG/GeminiPrinciples

- **Five safes framework**: This framework, established by the Office for National Statistics (ONS), is designed to safeguard the public interest when sharing ONS data, including with researchers. The Five Safes are: safe people, safe data, safe outputs, safe settings, and safe projects (The Royal Society; National Academy of Sciences, 2024). It is worth noting, however, that a Office for Statistics Regulation (OSR) report highlights the need for the five safes to be reviewed and potentially updated to ensure they are not unintentionally limiting the beneficial use of data. One potential change suggested is to 'broaden use cases for data by considering a change in focus from "safe projects" to "safe programmes"'(Office for Statistics Regulation, 2024, p6).
- **Smart Energy Research Laboratory (SERL):** This was a five year project funded by the Engineering and Physical Sciences Research Council to 'deliver a high quality multi-disciplinary research programme, facilitated by a smart energy research portal' (Webborn et al., 2019). SERL recruited 13,300 households to share energy smart meter data, accompanied by a survey about the occupants and type of dwelling. SERL's goals were to carry out research to support government policy, kick start new products, provide solutions to the 'energy trilemma (security, affordability, environmental sustainability), and promote better research. While SERL appears to still be operational, it also appears to be closed to new households joining the initiative, which may limit its long-term value.
- **Energy Demand Observatory and Laboratory:** The Energy Demand Observatory and Laboratory (EDOL) is a 5-year EPSRC funded research programme (Jan 2023 – Jan 2028). It collects and shares energy use data for research and policy, aiming to accelerate the zero carbon transition (*Energy Demand Observatory and Laboratory*, n.d.).
- **Office for Statistics Regulation recommendations for data sharing and linkage for the public good.** Since 2023, OSR has been making recommendations and engaging with key stakeholders to try and maintain momentum on its recommendations (Office for Statistics Regulation, 2024). Positive impacts include 'influencing the strategic approach taken by the Department for Science, Innovation and Technology (DSIT) to reviewing cross-government data sharing policy; developments in the Data Marketplace led by the Central Digital and Data Office (CDDO); the implementation of Wave 2 of the Public Engagement in Data Research Initiative (PEDRI); and technical innovation by the ONS Data Science Campus in developing new privacy-enhancing technologies (PETs)' (Office for Statistics Regulation, 2024, p5).
- **Better public engagement:** Public engagement activity and support resources have been developed, led by National Institutes and programmes such as - cites programmes such as Administrative Data Research UK (ADR UK), Data and

Analytics Research Environments UK (DARE UK) and Health Data Research UK (HDR UK). (Office for Statistics Regulation, 2024).

- **The SIM4NEXUS research project.** This project in the South West of England involved multi-stakeholder collaboration to improve understanding of the inter-relationships and interdependencies of water, energy and land management. (Black and Veatch, 2019).

Appendix 2: Detailed recommendations

This appendix expands on the recommendations summarised in Section 4. Please note that recommendations are recounted descriptively from the available literature only and do not constitute recommendations produced by the researcher's own analysis.

General

- **Better support and guidance from government and regulators:**
    - More support and guidance from government and regulators on how to share data would be beneficial, with consistent rules across regulators. Because data sharing is generally not mandated, the cost of doing so is typically not factored into price control processes. (Black and Veatch, 2019)
    - Tackle the lack of common data sharing practices by establishing best practices, encouraging collaboration and partnerships, and creating regulatory frameworks to determine minimum requirements for sharing data, security and privacy (Arup, Catapult Energy Systems, University of Bath, 2023).
    - Without government oversight to ensure coordination, collaboration and careful resource allocation, there is the risk of duplication of activities across large-scale programmes such as the National Digital Twin, Virtual Energy System, and Open Energy (Arup, Catapult Energy Systems, University of Bath, 2023).
- **Better identification of researcher needs.** A 'clearinghouse' identifying the types of data desired by academics would be of value (Harris & Sharma, 2017).
- **Trust Frameworks.** At a 2024 scientific forum on researcher access to data, Icebreaker One's Gavin Starks spoke about the value of Trust Frameworks[19] for data-sharing which can address many data issues. According to Starks, a Trust Framework is 'an effective way to implement and automate the adoption of rules for data providers, aggregators, and users' (The Royal Society; National Academy of Sciences, 2024, p38). It is:
    - Cohesive—It has common applications.
    - Interoperable—It uses common processes and systems.
    - Legal—It provides framework data rights, liability, and redress.
    - Controlled—It is assured and secure and manages consent.
    - Universal—It is open to the whole market.

[19] https://ib1.org/definitions/trust-framework/

To address legal/privacy barriers

- **Clarity on relevant legislation**. In its follow up report on data sharing and linkage for the public good, the Office for Statistics Regulation recommends that a single organisation in each nation should produce guidance to help researchers understand the legislation that is relevant to data sharing and linkage. (Office for Statistics Regulation, 2024)
- **Public engagement**. Several sources highlight the importance of meaningful public engagement and co-creation when considering projects that involve sharing personal data (Office for Statistics Regulation, 2024; The Royal Society; National Academy of Sciences, 2024; Psara et al., 2022). The public needs not only to be told that their data is being used safely and responsibly, but involved in discussions about how to make sure this is the case. This can not only increase public confidence in how their data will be used, but also help them understand how its use could improve their lives (The Royal Society; National Academy of Sciences, 2024). It is also recognised that public interest can be a legitimate reason for releasing data, and that there is general public support for the use of their data in research and statistics as long as they are fully informed and the data is properly safeguarded (Office for Statistics Regulation, 2024). The Office for Statistics Regulation recognises the Public Engagement in Data Research Initiative (PEDRI)[20] as key for supporting public engagement work.
- **Internal data-sharing policies.** In its review into data sharing under the Digital Economy Act 2017, the Information Commissioner's Office recognises that good internal data sharing policies are as important as external data sharing agreements when dealing with personal data (Information Commissioner's Office, 2023).
- **Strong de-identification resources and controls.** If personal data is to be shared with researchers, it should be de-identified so that the researchers do not have access to information that could directly identify individuals. Safeguards such as Privacy Enhancing Technologies may be required (Harris & Sharma, 2017; Office for Statistics Regulation, 2024).

To address security barriers

- **Effective collaboration.** According to Deloitte's report on data sharing in infrastructure, a 'Collaborative and coordinated approach across organisations and sectors' is key to minimising security risks (Deloitte, 2017a, p26).
- **Trusted data infrastructure.** This is necessary to ensure that private data holders feel confident that commercially sensitive or personally identifiable

[20] https://www.pedri.org.uk/

information will not be made public (The Royal Society; National Academy of Sciences, 2024).

To address commercial barriers

- **Incentives or mandates for data sharing by private companies**. Two suggestions put forward in a 2024 scientific forum on researcher access to data were that either the public sector could pay the private sector for research access to data, or that it could be made a mandatory requirement that such data was made available to statistical agencies for public use or research data products (The Royal Society; National Academy of Sciences, 2024).
- **Principles that govern stewardship.** In recognition of a concern that private data owners could withdraw access to their data at any point, it was observed in the 2024 scientific forum on researcher access to data that 'A better option than designating a single group as a gatekeeper would be principles that govern stewardship. In that way, organizations specifically interested in enabling research would be responsible for data stewardship while providing access to anyone with a good research question' (The Royal Society; National Academy of Sciences, 2024, p23).

To address cultural barriers

- **The development of capability and skills.** This is regarded as essential to create a strong data culture within organisations (Ofwat, 2021; The Royal Society; National Academy of Sciences, 2024; Office for Statistics Regulation, 2024; Smart Data Research UK, UKRI, 2024). One source posits that improving data culture internally increases the likelihood of organisations being willing to share data with academics (Harris & Sharma, 2017).
- **Data sharing targets.** For example, could consistent data quality measures become part of company Key Performance Indicators (Black and Veatch, 2019).
- **Emphasise the reputational benefits of data sharing.** Given that poor organisational understanding of the benefits of data sharing is common, reinforcing the positive PR that can be gained by organisations sharing their data with researchers for the public good may help address some reservations (Harris & Sharma, 2017).
- **Help organisations understand the value of even lower-quality data.** Organisations may be reluctant to share data if they are concerned about its quality, but it has been noted that 'Quality and value should not be confused. Data can be of high quality but little value, whereas low-quality data can be valuable—for example, if the data are the starting point for an important line of investigation' (The Royal Society; National Academy of Sciences, 2024, p19).

To address technical barriers

- **Data discoverability resources**
    - The value of better tools for finding data were recognised in a number of sources, with one article suggesting a 'Google Scholar' for data (The Royal Society; National Academy of Sciences, 2024), while another recommended a 'central data portal that holds the "what, where and who" for all assets in the infrastructure industry' (Black and Veatch, 2019, p2).
    - A report by the OSR indicated a need for greater transparency by government organisations about how data they hold can be accessed, noting that 'The CDDO [Central Digital and Data Office] is developing a data marketplace that should drastically improve the discoverability of data held across government, but we would encourage the CDDO to make this resource publicly available to support external researchers' (Office for Statistics Regulation, 2024, p42).
    - Metadata and documentation for data held within organisations are seen as key to improving the efficiency of data sharing and linking across government, which will then support the delivery of public services and policy decisions (Office for Statistics Regulation, 2024). While this recommendation relates specifically to data sharing within government, the principle is equally applicable to wider data sharing.
- **Standardisation and consistency**
    - Data does not need to be held centrally – but it needs to be created and structured consistently and in a way that makes it for multiple users to query simultaneously (Hetherington & West, 2020).
    - There is a need for a common data standard, definitions and shared framework, ideally developed by industry with support from regulators (Black and Veatch, 2019). A 2024 report on local energy planning noted the need for a consistent framework, including the standardisation of data protocols and access (Britton & Webb, 2024).
    - There must also be a consistent approach to standardisation across industries (Office for Statistics Regulation, 2024); it was recognised in the recently published Digital Spine feasibility study that a data sharing infrastructure created only for energy would risk creating further siloes (Arup, Catapult Energy Systems, University of Bath, 2023). In relation to modelling, one report referred to this approach as 'systems thinking'; combining datasets requires interconnectedness (The Royal Society; National Academy of Sciences, 2024).
- **Trusted intermediaries**

  The report on the 2024 scientific forum on researcher access to data, in

particular, highlighted the value of intermediaries to facilitate access to data by researchers.

  - Intermediaries, whether automated or not, could link data to resources in ways that help downstream users make good use of those data (The Royal Society; National Academy of Sciences, 2024, p18).
  - Data sharing can be facilitated by trusted brokers, who set standards and shared specifications (The Royal Society; National Academy of Sciences, 2024).
  - Data institutions, including archives, statistical agencies, data repositories, federated data systems, and data commons, play critical roles in supporting scientific research and enabling access to data…. How data are stewarded by institutions, disciplines, groups, and individuals ultimately affects what types of products, services, and insights data can be used to create; what decisions data can inform; and which activities data can support' (The Royal Society; National Academy of Sciences, 2024, p6).

- **Avoid reinventing the wheel**

  In some reports, it was recognised that, just as data can exist in siloes, there can be a tendency to try and develop solutions in siloes. This can result in a multitude of tools, terminologies, ontologies and standards that may not lend themselves to interoperability or easy access across, or even within, sectors.

  - 'The benefits of using common data ontologies and terminologies could be better communicated to communities that use data rather than creating new ontologies and terminologies.' (The Royal Society; National Academy of Sciences, 2024, p19)
  - In a report on the development of an Information Management Framework, it's noted that 'We will emphasise the use of openly available tools and standards… By proposing and standardising open protocols we will minimise barriers to participation…arising from vendor lock-in. We must be able to freely distribute new content…, and new actors will be able to contribute data sets and digital twins to the NDT with zero cost of entry' (Hetherington & West, 2020, p38).

- **Support for less stakeholders that are less mature in their data use**
  - Local authorities could benefit from Government and industry support via challenge funds that local authorities and data users or re-users can bid for, to support the opening up of specific datasets for wider use' (Deloitte, 2017a, p10).

- **New technologies**
  - Many organisations still rely on static PDFs or Excel files for data sharing, which can result in outdated data or multiple versions of the same

dataset being in circulation. APIs are regarded as a more useful data sharing method (Britton & Webb, 2024).

- ‘Machine learning and AI may increasingly be valuable to detect tampered or invalid data, which can give additional protection to systems from attacks based on data manipulation. For example, machine learning algorithms can automatically learn relationships between various parameters – such as the pH and chlorine levels of water – and warn if parameters deviate from expected values’ (Deloitte, 2017a, p25).
- Privacy-Enhancing Technologies (PETs) are tools for enabling safe sharing of people’s data, and can in theory provide a way for organisations to more confidently provide access to such data. However, many PETs are still developmental and costly, and cannot guarantee there will be no data leakage. This can make it hard to justify using them (Office for Statistics Regulation, 2024).

Appendix 3: Barriers to researchers sharing the data outputs of their research

Although this literature review has focused on the challenges of data being shared with researchers, several reports analysed highlighted the challenges faced by researchers around sharing the new data generated by their research. A key motivation for data owners to share data with researchers is knowing that the data will be used to benefit society as a whole; in order for this to happen, it is vital that valuable research outputs are made as widely available as possible. This is therefore an important issue to address alongside data being shared with researchers in the first place, in order for research findings to have the maximum benefit.

Lack of time and resources

- 'Providing metadata and good quality data indicators takes time. Managing data across multiple institutes, ethics teams and collaboration agreements can be complex. The different disciplinary domains common to energy consortia may have different standards that need to be met. All of these require expertise, attention and resourcing' (Colechin & Quigley, 2023, p2).
- There is a need for specialists in data management and sharing to support researchers carrying out data projects, but this type of role can be hard to recruit and retain (CREDS and UK Energy Research Centre (UKERC), 2023).

Low incentivisation for researchers

- Sharing energy research data is often seen as a burden, and can be perceived as 'time-consuming, difficult and frankly not interesting' (Colechin & Quigley, 2023, p10).
- The way that research projects are set up may not recognise or reward researchers for the effort they put into creating high-quality datasets or for making them available for sharing (Colechin & Quigley, 2023; The Royal Society; National Academy of Sciences, 2024). There is a view that 'Institutions need to take the value of data more seriously, funding activities effectively, rewarding individuals for taking an active role, and recognising the importance of workload management' (Colechin & Quigley, 2023, p15).

Varying data abilities

- Data management skills and knowledge in the area of data management can vary widely; researchers need effective training on the value of data sharing and how to manage data. While this point was recognised in relation to the energy community, it is likely to apply across all infrastructure sectors (Colechin & Quigley, 2023).

Challenges of working with (sometimes multiple) partners

Issues around sharing research data can be exacerbated by:

- Working with commercial partners, which might have Non Disclosure Agreements in place that restrict data-sharing (Colechin & Quigley, 2023; CREDS and UK Energy Research Centre (UKERC), 2023).
- Funder requirements, for example on how data is collected, stored and shared (Colechin & Quigley, 2023; CREDS and UK Energy Research Centre (UKERC), 2023).
- Working in large consortia, which can involve multiple projects across multiple domains. This can make it difficult to set standards and expectations that work for all parties and in all circumstances (CREDS and UK Energy Research Centre (UKERC), 2023).

Inadequate planning

- Data management is often not seen as a priority when starting a project, and can be left till the end, at which point it may be too late. Data management plans and guidance on data management and sharing should be implemented at the start of projects, not least to make sure that anyone joining a project (or consortium) later is fully aware of requirements (Colechin & Quigley, 2023).

Lack of common approaches

- 'The energy community is a large producer and user of models in a wide variety of areas and common standards for what to archive to enable FAIR data and reproducibility have not yet been agreed. Such protocols would be helpful to discuss. The energy research specific issues for sharing the outputs of energy models should continue to be highlighted' (Colechin & Quigley, 2023).
- In relation to work with confidential data, 'reproducible research practices (e.g. sharing workflows, models and algorithms) have not been broadly established (Smart Data Research UK, UKRI, 2024. p15).

Difficulties in sharing certain outputs

- More work is needed on how to effectively share and use models, including in combination with other forms of data. Models are a common output from energy research (CREDS and UK Energy Research Centre (UKERC), 2023).
- 'Energy researchers are adopting new forms of research outputs including Jupyter notebooks which are more of a challenge to deposit and curate; demonstrating that data sharing and management is a constantly evolving field and needs to be supported by expert data stewards' (CREDS and UK Energy Research Centre (UKERC), 2023, p21).

- Lots of researchers still save data on hard drives (rather than in the cloud) due to the expense of transitioning to the cloud, with little indication that this is going to change in the near future. (The Royal Society; National Academy of Sciences, 2024)

Environmental considerations

- 'The environmental consequences of gathering, analyzing, storing and sharing data' are identified as a key reason not to carry out these activities for the sake of it; it's important to be clear about reasons and value of doing so. 'The global greenhouse gas emissions of data centers are an estimated 100 megatons of carbon dioxide equivalent per year, which is about the same as the emissions of U.S. commercial aviation' (The Royal Society; National Academy of Sciences, 2024, p28).

References

Arup, Catapult Energy Systems, University of Bath. (2023, September). *Digital Spine Feasibility study: Developing an energy system data sharing infrastructure* [Commissioned by Department for Energy Security and Net Zero].

Black and Veatch. (2019, September). *UK Regulators Network: Infrastructure Data Sharing*.

Britton, J., & Webb, J. (2024). *Planning Works: Local Energy Planning to Accelerate Net Zero*.

Britton, J., Webb, J., Hawker, G., Broad, O., & Chaudry, M. (2023, November). *UKERC - Energy Modelling Across Scales Stakeholder Workshop Summary* [Commissioned by UK Energy Research Centre (UKERC)].

Centre for Digital Built Britain (CDBB). (2020, May 29). *The approach to delivering a national digital twin for the united kingdom - Summary report*.

Colechin, M., & Quigley, C. (2023, November). *CREDS and UKERC - Improving Data Sharing in Energy Consortia: Summary of Workshop Outputs* [Commissioned by: CREDS and UK Energy Research Centre (UKERC)].

CREDS and UK Energy Research Centre (UKERC). (2023). *CREDS (Centre for Research into Energy Demand Solutions) Data sharing workshop briefing note*.

Deloitte. (2017b, July). *Assessing the value of TfL's open data and digital partnerships* [Commissioned by Transport for London].

Deloitte. (2017a, November). *New Technologies Case Study: Data sharing in infrastructure* [Commissioned by National Infrastructure Commission].

Department for Transport. (2023). *Transport Data Strategy: Innovation through data*.

Dewey. (2023, April 2025). *Acquiring Corporate Data for Academic Research*.

Dodds, L., Maddison, J., Yates, D., & L'Henaff, P. (2019, October). *Insight report on sharing engineering data - Using data for the public good* [Commissioned by Lloyd's Register Foundation and The Open Data Institute].

EDOL. *Energy Demand Observatory and Laboratory*. (n.d.). Retrieved September 30, 2024, from https://edol.uk/Research/Observatory/index.html

Evans, E., & Johnston, G. (2023, January). *NGED Value of Data* [Commissioned by: National Grid].

Gilbert, T., Rönsdorf, C., Plume, J., Simmons, S., Nisbet, N., Gruler, H.-C., Kolbe, T. H., van Berlo, L., & Mercer, A. (2020, March 2). *Built environment data standards and their integration: an analysis of IFC, CityGML and LandInfra*.

Harris, L., & Sharma, C. (2017, November). *Understanding corporate data sharing decisions: practices, challenges, and opportunities for sharing corporate data with researchers* [Commissioned by Future of Privacy Forum and FPF Education and Innovation Foundation].

Hetherington, J., & West, M. (2020). *The pathway towards an Information Management Framework - A 'Commons' or Digital Built Britain* [Commissioned by National Digital Twin Programme].

Icebreaker One. (2021, July). *Open Energy Modernising Energy Data Access Phase 3 Report*.

Icebreaker One. (2024). *Perseus: automating emissions reporting to unlock finance for SMEs Icebreaker One*. Icebreaker One. Retrieved September 29, 2024, from https://ib1.org/perseus/

Information Commissioner's Office. (2023, March). *The ICO's review into data sharing under the Digital Economy Act 2017*.

Judson, E. (2023, October). *Governing the GB digital energy revolution to support net zero by 2050*.

Labour Party. (2024). *Labour Party Manifesto 2024*.

Leal, D., Cook, A., Partridge, C., Sullivan, J., & West, M. (2020). *A Survey of Industry Data Models and Reference Data Libraries - To identify requirements for, and provide input to, a Foundation Data Model* [Commissioned by Construction Innovation Hub].

Lieberman, J. (Ed.). (2019, November 25). *Model for Underground Data Definition and Integration (MUDDI) Engineering Report* [Commissioned by Open Geospatial Consortium].

Office for Statistics Regulation. (2024, July). *Data Sharing and Linkage for the Public Good: Follow-Up Report*.

Ofgem. (2021, updated 2023). *Digitalisation Strategy and Action Plan Guidance*.

Ofgem. (2023, August 07). *Decision on Data Best Practice guidance and Digitalisation Strategy and Action Plan guidance*.

Ofwat. (2021, October). *H2Open – Open data in the water industry: a case for change*.

Ofwat. (2021, October 13). *H2Open – Open data in the water industry: a case for change, case study appendix*.

Ofwat. (2023, June). *Open data in the water industry: making the change*.

Open Energy in the UK Icebreaker One. (2024). Icebreaker One. Retrieved September 29, 2024, from https://ib1.org/tf/estf/

PricewaterhouseCoopers (PwC). (2023, May). *Open data assessment report*.

Psara, K., Papadimitriou, C., Efstratiadi, M., Tsakanikas, S., Papadopoulos, P., & Tobin, P. (2022, March 17). *European Energy Regulatory, Socioeconomic, and Organizational Aspects: An Analysis of Barriers Related to Data-Driven Services across Electricity Sectors*.

The Royal Society; National Academy of Sciences. (2024). *Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data (2024)*.

Salvador Palau, A., Roozenbeek, J., & Kumar Parlikad, P. A. (2018). *Exploiting traffic data to improve asset management and citizen quality of life*.

Smart Data Research UK, UKRI. (2024). *Smart Data Research UK Strategy*.

Stream. (No date). *Enabling innovation with water data*. Stream - Portal. Retrieved September 29, 2024, from https://www.streamwaterdata.co.uk/

UK Government. (2017 (withdrawn 2023)). *Industrial Strategy: Building a Britain fit for the future*.

UK Government. (2020, September). *National Data Strategy*.

UK Government. (2021, March 3). *Build Back Better: our plan for growth*.

Varga, L., McMillan, L., Hallett, S., Russell, T., Smith, L., Truckell, I., Postnikov, A., Rodger, S., Vizcaino, N., Perkins, B., Matthews, B., & Lomax, N. (2022, July 26). *Infrastructure and Cities Ontologies*.

Webborn, E., Elam, S., McKenna, E., & Oreszczyn, T. (2019). Utilising smart meter data for research and innovation in the UK.

Yalamanchili, S. (2024, February). *Data Linkage in Road Safety - Bridging the divide to support better health outcomes*.