

# DARE UK

## Paving the way for a coordinated national infrastructure for sensitive data research

A summary of findings to date from Phase 1 of the UK Research and Innovation DARE UK programme

August 2022



# DARE UK

## DARE UK Phase 1 Delivery Team

**Hans-Erik G. Aronson** Director

**Gerry Reilly** Design Authority

**Susheel Varma** Technical Lead

**Fergus McDonald** Senior Programme Manager

**Elizabeth Waind** Senior Communications  
and Engagement Manager

**Michelle Amugi** Programme Manager

## Acknowledgements

DARE UK is funded by UK Research and Innovation (UKRI) as part of its [Digital Research Infrastructure](#) portfolio of investments. Phase 1 of the programme – Design and Dialogue – is led by Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK) with a total investment from UKRI of just over £3 million from May 2021 to August 2022, just over £2 million of which was allocated directly to a programme of nine exploratory Sprint Exemplar Projects.

A wealth of input has contributed towards the co-design of the recommendations set out on this report. As such, we would like to express our gratitude to all those who have supported their development, including the many researchers, technologists, members of the public and others who have engaged with us throughout DARE UK Phase 1.

The DARE UK Phase 1 Delivery Team would like to thank all those who participated in the initial DARE UK landscape review, and Carnall Farrar for their support delivering the review. We would also like to thank the 44 members of the public who gave their valuable input to the public dialogue, as well as members of the Public Dialogue Oversight Group and Kohlrabi Consulting for supporting its design and delivery. Further, our thanks extend to all those who attended our six thematic workshops held in March, which sought initial views on our emerging recommendations; and the 24 individuals who attended the persona and user journeys workshop held at the Hartree Centre, as well as the Science and Technology Facilities Council (STFC) for hosting that workshop. We also thank all those who shared their input in response to the open call for views on a draft version of this report.

Our thanks and appreciation also extend to the DARE UK Programme Board, Scientific and Technical Advisory Group, Oversight Group and Public Contributors for their invaluable guidance and input throughout Phase 1.



# Foreword

**Today, in the era of the fourth industrial revolution, technological advancement and increased interconnectivity are leading to rapid change across the globe<sup>1</sup>. The power of data has never been greater and stands to transform every aspect of our lives.**

Prudent data analysis holds immense promise to benefit society, accelerating research and informing and enhancing public policy. Often, the data in question is of a sensitive nature, dictating strict data handling procedures and significant restrictions on access, analysis, linkage and use – measures that are essential to ensure the privacy of individuals and prevent data misuse. The DARE UK programme was created to explore the potential for trusted research environments (TREs) to enable research and innovation involving all types of sensitive data, with the aim to create a community co-designed, next-generation national data research infrastructure.

DARE UK has a UK-wide remit and covers all UK Research and Innovation research council domains. As such, our goal is to smooth the waves for sensitive data linkage and analysis across a wide range of research needs. The vision is for a federated, interoperable infrastructure that facilitates cross-domain linkage and analysis of data – data about education, the environment, health, welfare and much, much more – at scale for public good. A measure of success is an infrastructure that accelerates discovery by lowering

barriers for researchers, whilst simultaneously maintaining appropriate safeguards and demonstrating trustworthiness to the public.

This first year of the DARE UK programme has been all about dialogue, driven by discussions with researchers, technologists, the public and others via workshops, interviews and more. The aim has been to learn from these groups what is needed to establish a more coordinated and trustworthy national data research infrastructure. But the conversation is far from over – it will continue throughout further phases of the programme to ensure the design and delivery of a novel and innovative infrastructure is geared towards meeting the needs of all those who rely upon it, including the wider public.

The power of data is increasing, arguably to an extent we did not quite foresee in the past. We have an opportunity to harness that power for good, and the findings and recommendations set out in this report represent an important step towards achieving this.



**Hans-Erik G. Aronson**  
Director of DARE UK Phase 1

<sup>1</sup> Schwab K. 2015. [The Fourth Industrial Revolution: What It Means and How to Respond](#). Foreign Affairs.

# Executive summary

**Data has the power to improve lives, and has been fundamental to the UK's response to the COVID-19 pandemic. It is crucial that the different components of the UK's data research infrastructure work in a coordinated, impactful and trustworthy way, to support research at scale for public benefit. They need to be able to support fast and efficient sharing, linkage and advanced analysis of sensitive data in an ethical and secure manner, whilst maintaining the confidence – and meeting the needs of – researchers, data guardians and the public.**

The UK Research and Innovation DARE UK (Data and Analytics Research Environments UK) programme has been established to design and deliver a coordinated and trustworthy national data research infrastructure to support research at scale for public good. DARE UK is a cross-domain programme – its scope covers all types of sensitive data, including data about education, health, the environment and much more.

This report sets out the emerging findings and recommendations so far from Phase 1 of the DARE UK programme – 'Design and Dialogue' – which began in July 2021 with the aim of establishing the key challenges across the data research landscape, and first steps on how to overcome them to better support data research at scale for the benefit of society. This was achieved via a programme of engagement with stakeholders from across the landscape including interviews, workshops and other discussions with researchers, technologists, the private sector, the public and more.



## Recommendations

Based upon the findings of engagement with stakeholders so far during Phase 1 of DARE UK, this report makes the following key recommendations, which are further detailed and evidenced in the main body of the report:

### Demonstrating trustworthiness

1. Consistently practice proactive transparency about what sensitive data is being used for research, how, why and by whom.
2. Conduct a UK-wide public information campaign to raise general awareness of how and why sensitive data is made accessible for research.
3. Publish and maintain standardised and accessible data use registers.
4. Drive a culture shift to recognise the crucial importance of public involvement and engagement and embed it throughout the sensitive data research lifecycle.
5. Investigate the requirements for establishing an independent coordinating function for public involvement and engagement with sensitive data research, either as a new entity or as an off-shoot of a relevant existing body.
6. Standardise, centralise and unify processes enabling access to sensitive data for research across the UK where appropriate and feasible.

### Researcher accreditation and access

1. Provide a unified user authentication capability to enable researchers to access services more easily across the entire sensitive data research ecosystem.
2. Provide a streamlined researcher accreditation framework to enable trustworthy researchers to access sensitive data for research in the public benefit in a timelier fashion.
3. Develop a standardised and streamlined – yet extensible – process for accredited researchers to request access to sensitive data from data guardians whilst maintaining appropriate levels of data privacy and security.

### Accreditation of research environments

1. Review and extend the existing standard, accreditation, and audit framework under the Digital Economy Act (DEA) to further establish it as the nationally recognised trusted research environment (TRE) standard, accreditation, and audit framework.

### Data and discovery

1. Enhance the data lifecycle to support effective cross-domain sensitive data research.
2. Explore the implications of new data types on approaches to making these data available for research.
3. Develop guidelines on privacy enhancing technologies (PETs) for use by TREs.
4. Establish a UKRI-wide metadata standard working group.
5. Leverage existing Digital Object Identifier (DOI) minting services to provide persistent identifiers for all UKRI discoverable assets at UKRI-wide and council levels.

### Core federation services

1. Develop reference architectures for TREs.
2. Assemble an API (application programming interface) library to support core federation services.
3. Run a competitive call for driver projects to utilise the new infrastructure services and validate that they are fit for purpose.
4. Establish an approach to business continuity and disaster recovery.

### Capability and capacity

1. Establish clear technical career pathways in data research infrastructure that can be adopted across the UKRI research domains.
2. Improve recruitment pathways for technical roles in data research infrastructure.
3. Improve the availability of resources and training for career development in data research infrastructure.
4. Use automation to ensure data research infrastructure services are reliably secure, auditable and reproducible.

### Funding and incentives

1. Develop a new type of grant tailored to addressing the costs for maintaining cross-domain, national sensitive data research infrastructure.
2. Determine the funding requirements to establish the first phase of a federated national infrastructure for sensitive data research, with a focus on enabling federation across existing infrastructure and complimenting existing investments.
3. Investigate, test and prototype the operational model(s) for a federated national infrastructure for sensitive data research. Critically, ensure federation lessons and insights from those outside of the sensitive data space are considered.

4. Investigate the cost implications for appropriate business continuity and disaster recovery requirements for a federated national infrastructure for sensitive data research.
5. Investigate the scope and funding requirements for the integration of large-scale compute availability in a federated national infrastructure for sensitive data research.
6. Building upon existing best practice, improve the availability of all data produced through publicly funded grants for reuse and investigate the funding requirements for provisioning such archival capability.
7. Support raising awareness amongst data collectors and data guardians regarding the legal framework around the use of data for research.
8. Dedicate greater resource to supporting data collectors and data guardians to routinely make their data accessible for research in the public benefit.



# Contents

<b>1 / Introduction</b>	<b>8</b>	<b>5 / Accreditation of research environments</b>	<b>37</b>	<b>8 / Capability and capacity</b>	<b>64</b>
Programme purpose, scope, and premises	8	Context	37	Context	64
Synergy with other initiatives	10	Existing challenges and opportunities	38	Existing challenges and opportunities	65
Glossary	11	Recommendation	40	Recommendations	70
List of acronyms	13				
<b>2 / Process and summary of input</b>	<b>14</b>	<b>6 / Data and discovery</b>	<b>41</b>	<b>9 / Funding and incentives</b>	<b>72</b>
Sprint Exemplar Projects	17	Context	41	Context	72
Structure of the report	19	Existing challenges and opportunities	42	Existing challenges and opportunities	72
		Recommendations	46	Recommendations	79
<b>3 / Demonstrating trustworthiness</b>	<b>20</b>	<b>7 / Core federation services</b>	<b>48</b>	<b>10 / Next steps</b>	<b>82</b>
Context	20	Context	48		
Existing challenges and opportunities	21	Existing challenges and opportunities	48	<b>11 / Appendices</b>	<b>83</b>
Recommendations	27	A federated network of cross-domain trusted research environments (TREs)	51		
<b>4 / Researcher accreditation and access</b>	<b>31</b>	Preparing for production deployment	60		
Context	31	Driver projects	61		
Existing challenges and opportunities	32	Partnerships and collaboration	61		
Recommendations	36	Recommendations	63		

# 1 / Introduction

## Programme purpose, scope and premises

**DARE UK (Data and Analytics Research Environments UK) is a programme funded by UK Research and Innovation (UKRI) to design and deliver a more coordinated national data research infrastructure for the UK.**

Data has the power to improve lives and has been fundamental to supporting an evidence-based response to the COVID-19 pandemic. It is crucial that the different components of the UK's data research infrastructure work in a coordinated, impactful and trustworthy way, to support research at scale for public benefit. Further, in line with the ['Levelling Up' agenda](#), driving the sensitive data research ecosystem (and the data ecosystem more broadly) towards greater cohesion and collaboration will provide more equitable secure access to sensitive data across the UK that in turn will encourage innovation and growth. The UK's data research infrastructure needs to be able to support fast and efficient sharing, linkage, and advanced analysis of sensitive data in an ethical and secure manner, whilst maintaining the confidence – and meeting the needs of – researchers, data guardians and the public.

DARE UK has been established to design and deliver – together with the different research communities – a novel and innovative data research infrastructure for the UK, with a specific focus on supporting cross-domain linkage and analysis of sensitive data. The programme is one of several initiatives funded by UKRI – the UK's largest public funder of research and innovation – under the [Digital Research Infrastructure portfolio](#), whose strategic vision is to deliver a coherent, state-of-the-art national infrastructure that will enable UK researchers and innovators to harness the full power of modern digital platforms, tools, techniques, and skills. A key theme within the Digital Research Infrastructure portfolio strategy is to provide secure and trustworthy data services for sensitive data and the appropriate tools that will enable researchers, innovators and decision-makers to derive benefit from this data.



In collaboration with the various communities of the UK research and innovation sector, DARE UK aims to:

- Design and deliver a novel and innovative UK-wide data research infrastructure that is coordinated, demonstrates trustworthiness and supports research at scale for public good.
- Establish the next generation of trusted research environments (TREs) across research domains that will enable fast, safe and efficient sharing, linkage and advanced analysis of data where it is legal and ethical to do so.
- Enable UK researchers and innovators to harness, securely and efficiently, the full power of linked datasets, modern digital platforms, tools, techniques and skills.
- Enable research and analysis on a broad range of potentially sensitive data from across the UK research and innovation spectrum.

The scope of DARE UK includes all research conducted by UKRI research councils that uses, or anticipates use of, sensitive data from across different domains. However, it is important to note that the programme scope does not include the use of data for algorithmic decision-making or predictive analytics – its focus is on the use of data for generating insights to inform policy and services. The DARE UK programme is being undertaken in an open and inclusive manner, with involvement from researchers, technologists, funders, the public and others embedded throughout.

### Key premises and assumptions

The DARE UK programme is premised on several assumptions:

- We are **not starting from scratch** – the research ecosystem already has established data infrastructures and there are opportunities for adoption of existing best practice across different data research domains.
- There is a **cross-domain need** – there is a need to enable high priority research that involves data from across different research disciplines, and this is expected to grow in the future. Further, there is a research community with the skills and know-how to do this cross-domain, data-enabled research.
- A federated infrastructure is **technically feasible** – we can technically make the data discoverable and understandable and align metadata and API (application programming interface) standards.
- It is **ethically and legally feasible** – there are existing governance best practices which can be adopted and scaled across research domains, and public involvement and engagement must be embedded throughout.
- It is feasible to do within the **funding and time envelope** set by UKRI – by leveraging existing established community expertise, technologies, architectures and standards, this could be deployed in a phased approach.

### DARE UK Phase 1

DARE UK is a multi-phase programme, with Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK) commissioned to oversee Phase 1: Design and Dialogue, which began in July 2021.

Phase 1 of the DARE UK programme is an extensive listening exercise. The goal has been to understand, through open dialogue with stakeholders – including researchers, technologists, funders, the public and others – what is needed to enable more efficient, coordinated, and trustworthy cross-domain research using sensitive data across the UK. By exploring stakeholder experiences and challenges of existing infrastructure, we aim to ensure that subsequent phases of DARE UK address the needs of the UK in making the best use of data at scale for public benefit. A short summary of the process and input received so far in DARE UK Phase 1 can be seen below in Chapter 2: Process and summary of input.

DARE UK Phase 1 has been supported by the valuable guidance and input of a dedicated Programme Board, Scientific and Technical Advisory Group and Oversight Group, including five Public Contributors. Membership of these groups can be seen in Appendix 1. You can find out more about Phase 1 governance [on the DARE UK website](#).

## Synergy with other initiatives

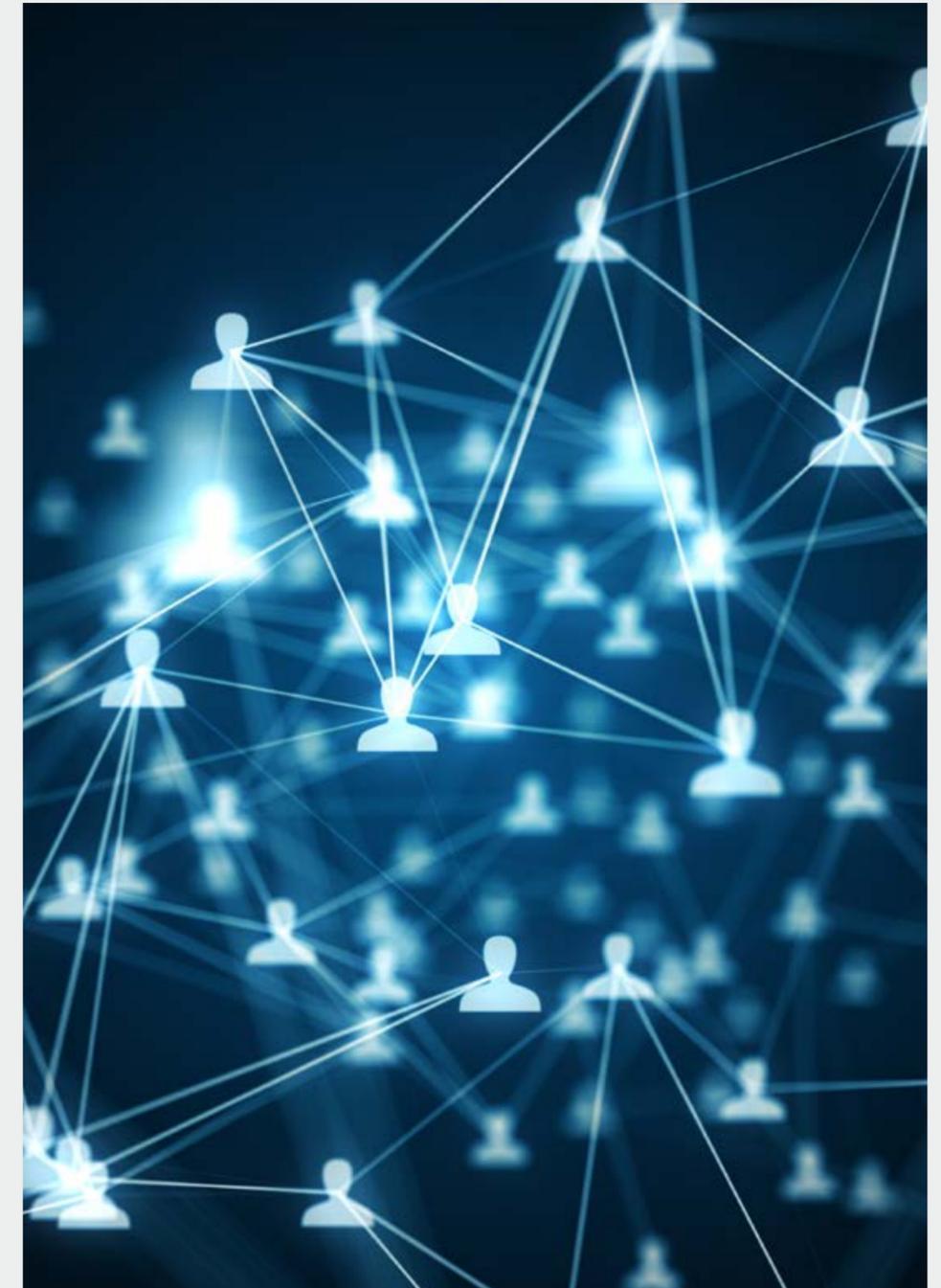
The DARE UK programme is operating in a complex and rapidly evolving data strategy landscape across disciplines and the four nations of the UK. These strategies share many common themes, and in general represent a trajectory towards the more efficient use of de-identified data for public benefit in a safe and secure way. It is critical that the delivery of the recommendations set out in this report is complimentary to the delivery of other activities, to avoid the duplication of effort and ensure any development in this area is aligned and does not lead to further fragmentation.

The development of the next phase of DARE UK will need to consider a broad range of initiatives, including but not limited to: the [UK National Data Strategy](#); the Scottish Government [Data Strategy for health and social care](#); the [Northern Ireland Department of Health Data Strategy](#); and the [Digital Strategy for Wales](#). It will also be important to align with initiatives towards the availability of open data to be used alongside sensitive data; an example of this is the already extensive portal of open datasets available from through the [OpenDataNI portal](#).

There have also been recent welcome developments in proposed approaches for the support of the ethical and secure use of NHS England data for research and development, with the recent publication of the NHS

England [Data Saves Lives strategy](#). The recent [Goldacre Review](#) has recommended an approach of using a small number of centrally delivered and managed national and sub-national trusted research environments (TREs – sometimes now referred to as secure data environments) to facilitate the use of NHS data for research, operational management and policy development.

It is important to note that the scope of sensitive data covered by DARE UK crosses all UKRI councils and all four nations of the UK, with cross-domain research being a key priority. Its scope therefore includes not only NHS health data, but also data about the environment, education, welfare and much more collected by a variety of government departments and public sector bodies, as well as health data beyond NHS England. Some of the recommendations outlined in this report – particularly around the development of TRE reference architectures and a federated network of TREs to support cross-domain research – are critical to this. These recommendations will also need to work to facilitate appropriate and secure interoperability with the infrastructure and services delivered through the [NHS England Data for Research and Development Programme](#).



## Glossary

The language in this area is evolving and would benefit from further discussion and agreement across the data research community to achieve alignment. Nevertheless, for the purpose of this report, the following definitions are used:

### Data collector

The data collector is the organisation responsible for the original collection of data, whether or not its collection was initially intended for research purposes. For example, schools collect data about pupil attendance, and hospitals collect data about their patients' health; this data is collected in the course of delivering the services these bodies provide and may later be made accessible for research.

### Data custodian

The data custodian is the organisation responsible for securely storing data and making it accessible to accredited researchers for analysis. In the context of this report, the data custodian is generally a trusted research environment (TRE – see definition below).

### Data fabric

Data fabric is an architecture that facilitates the end-to-end integration of various data pipelines and cloud environments through the use of intelligent and automated systems. This allows for more holistic, data-centric decision-making.<sup>2</sup>

### Data guardian

The data guardian is the body responsible for assessing and deciding whether data access should be granted, which is generally done on a case-by-case basis for research projects.

At the point of data collection, the data collector can be considered the data guardian. In some cases, the data collector may maintain the role of data guardian for the data they collect; in others, the data collector may delegate the role of data guardian to the data custodian or to an independent decision-making panel of experts and public representatives. In other scenarios, the role of data guardian may involve a process in which all or some of the data collector, data custodian and decision-making panel(s) play a part.

### Data research infrastructure

Data research infrastructure refers to the systems and processes in place to support research and analysis using sensitive data. It includes physical systems, such as the data centres where the data itself is held; computer software that researchers use to analyse data; governance processes, such as those guiding who can access what data and for what purposes; and the people who run the systems and do the research. It is everything that makes data research happen.

### Federation

A 'federated' network of trusted research environments (TREs) is one which would allow analysis of sensitive data to be conducted across different TREs. The TREs would follow mutually agreed and understood security and governance protocols, and the different systems used across the TREs would be able to work together coherently. This can occur in three different ways:

1. Analyse multiple datasets in the TREs in which they are and then bring together the results in one TRE for final review, sometimes called study level meta-analysis. This approach works well for analysis on comparable data, known as horizontally partitioned data, held in different TREs.
2. Temporarily combine the data, allowing approved researchers to access and analyse data within any TRE in the network, rather than only within the one where the data is held. This approach is appropriate for linking data held in different TREs together.
3. Analyse data held in different TREs as if the data had been combined and linked into a single environment. This approach can work in some cases, but in other situations can be technically very challenging.

<sup>2</sup> IBM. [What is data fabric?](#) Accessed 15.08.2022.

Approaches two and three address the more complex examples of analysing different linked datasets (also known as vertically partitioned data) rather than analysis on comparable datasets.

### The ‘Five Safes’ framework

Developed by the Office for National Statistics, the ‘Five Safes’ framework is designed to reduce the risks to individual privacy and the potential for data misuse when making sensitive data accessible for research via trusted research environments (TREs – see below). The Five Safes are:

1. **Safe data:** data is de-identified before being made accessible to researchers to protect the privacy of individuals
2. **Safe settings:** data is only made accessible via a secure, physical safe setting, or via a secure, approved connection to the safe setting. This way, no data ever leaves the TRE
3. **Safe people:** only approved researchers who have undergone training and assessment are granted access to sensitive data for research
4. **Safe projects:** sensitive data is only made accessible for research projects which have been assessed as being in the public benefit
5. **Safe outputs:** all research outputs are checked before they leave the TRE to make sure the identity of individuals is not disclosed

### Internet of things

The internet of things describes the network of physical objects that are embedded with sensors, software and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet. These devices range from ordinary household objects to sophisticated industrial tools.<sup>3</sup>

### Interoperability

Interoperability means different trusted research environments (TREs) – often managed by different organisations – can work together so that researchers can work across them. This is achieved by enabling the different systems or components of the TREs to successfully ‘communicate’ with one another, allowing them to connect and exchange information between one another and work together, with a common approach for identifying accredited researchers.

### Metadata

Metadata provides information about other data, including a description of the data. This includes information that provides context to the data – for example, how the data was collected, the coverage of the data, and licencing arrangements. Metadata can include information such as publication date, description and search keywords.<sup>4</sup>

Metadata can be held at a variety of levels, from administrative information about the dataset, through field level technical descriptions of the dataset, to overview statistics (for example, the number of participants included in the dataset).

### Reference architecture

A reference architecture is a document or set of documents that provides recommended structures and integrations of IT products and services to form a solution. The reference architecture embodies accepted industry best practices, typically suggesting the optimal delivery method for specific technologies. A reference architecture offers IT best practices in an easy-to-understand format that guides the implementation of complex technology solutions.<sup>5</sup>

<sup>3</sup> Oracle. [What is IoT?](#) Accessed 15.08.2022.

<sup>4</sup> Office for National Statistics. [Metadata policy.](#) Accessed 12.07.2022.

<sup>5</sup> Hewlett Packard Enterprise. [What is a Reference Architecture?](#) Accessed 15.08.2022.

## Sensitive Data

For the purpose of this report, the following simplified definition of sensitive data – which may expand and develop during future phases of the programme – is used:

Sensitive data includes data which contains personally identifiable information such as names, addresses and identifying numbers. This can still be sensitive once it has been de-identified (has had all personal identifiable information removed) if there is potential for re-identification, particularly when used with other data. Commercial data such as retail information, business details, IP (intellectual property) and Copyright information or confidential product details may also be considered sensitive data.

## Trusted research environment (TRE)

A trusted research environment (TRE) is a highly secure digital environment that provides access to sensitive data for analysis by approved researchers. A series of strict security measures protect the confidentiality of the data, significantly reducing the potential for data misuse or the possibility of re-identification of de-identified data.

## List of acronyms

AAAI – Association for the Advancement of Artificial Intelligence	HRA – Health Research Authority
ADR UK – Administrative Data Research UK	ICO – Information Commissioner’s Office
ARDC – Australian Research Data Commons	IP – intellectual property
AHRC – Arts and Humanities Research Council	IPU – Intelligence Processing Unit
AI – artificial intelligence	IRAS – Integrated Research Application System
API – application programming interface	MRC – Medical Research Council
BBSRC – Biotechnology and Biology Sciences Research Council	MVP – minimum viable product
BCS – British Computer Society	N8 CIR – N8 Centre of Excellence in Computationally Intensive Research
BHF – British Heart Foundation	NERC – Natural Environment Research Council
CDEI – Centre for Data Ethics and Innovation	NHS – National Health Service
CIS2 – Care Identity Service 2	NIH – National Institutes of Health
DARE UK – Data and Analytics Research Environments UK	OIDC – OpenID Connect Federation
DEA – Digital Economy Act	ONS – Office for National Statistics
DOI – Digital Object Identifier	PEDRI – Public Engagement with Data Research Initiative
eDRIS – electronic Data Research and Innovation Service	PETs – privacy enhancing technologies
ESRC – Economic and Social Research Council	RAM – random access memory
EPSRC – Engineering and Physical Sciences Research Council	RPO – recovery point objective
FAIR – Findable, Accessible, Interoperable and Reusable	RSE – Society of Research Software Engineering
GA4GH – Global Alliance for Genomics and Health	RTO – recovery time objective
GDPR – General Practice Data for Planning and Research	SAML – Security Assertion Markup Language
GPU – graphics processing unit	SLA – service level agreement
HA – high availability	STFC – Science and Technology Facilities Council
HNCDI – Hartree National Centre for Digital Innovation	TEHDAS – Towards European Health Data Space
HDR UK – Health Data Research UK	TRE – trusted research environment
HPC – high performance computing	UKRI – UK Research and Innovation
HTC – high throughput computing	UN – United Nations
	UPRN – Unique Property Reference Number
	WHO – World Health Organization

# 2 / Process and summary of input

Input from stakeholders across the sensitive data research community – as well as the public – that forms the basis of this report has been gathered through a mixed method of activities throughout DARE UK Phase 1 to date. These have been captured in the timeline below.

Additionally, the DARE UK Phase 1 Delivery Team – responsible for the day-to-day delivery of the programme objectives – have drawn on the collective knowledge and experience of our oversight partners in Health Data Research UK (HDR UK), ADR UK (Administrative Data Research UK) and the UKRI research councils in the development of this report.

## Initial landscape review, August-September 2021

The first engagement activity undertaken in DARE UK Phase 1 was an [initial review of the landscape](#) to establish a fundamental understanding of the context and overarching challenges within the sensitive data research ecosystem across the UK. During August and September 2021, the programme commissioned Carnall Farrar – a management consulting and data science company – to deliver this landscape review together with the Delivery Team.



There were two parts to this initial work, the first was a series of 60 interviews of approximately 1-1.5 hours each with stakeholders selected from across the spectrum of research disciplines. Broadly, these were intended as in-depth, 1:1 interviews, though in some cases it was appropriate to include multiple stakeholders in a single sitting. In total, 79 people including researchers, technologists, and funders were interviewed at this stage.

The second part to the review involved two open invite workshops aimed at researchers and technologists – with members of the public also invited to attend – of 1.5 hours each. These focused on reviewing the synthesis of the interviews to receive further input on the initial, broad framing of the key challenges and areas of unmet needs for the cross-domain sensitive data research landscape in the UK. Each of the workshops was attended by approximately 50 participants and provided valuable additional input, feedback, and constructive questions for further consideration. This provided an initial, broad framing of the key focus areas for the landscape that could be taken forward into further investigation, analysis, and engagement with stakeholders.



Over  
**60 hours**  
of interviews



Over  
**30 hours**  
of workshops



Over  
**500**  
participants

researchers, technologists, members of the public, third sector representatives and others

### Public dialogue, January-February 2022

In January and February 2022, 44 members of the public from England, Northern Ireland, Scotland and Wales were recruited to take part in a series of deliberative workshops. The dialogue aimed to deepen public conversation around data research practices on a national scale and capture tangible actions that could be taken forward by those holding and using sensitive data for research to address public views.

In January, two initial workshops of the same structure were held online over two full days, each with half of the total participants. In February, a single, half-day follow-up workshop was then held with a cross-section of 10 participants from the two initial workshops to verify that analysis of the initial workshops had accurately captured participants' views and expectations. The follow-up workshop also aimed to bring those expectations to life through discussion of tangible actions that could be taken forward by the sensitive data research community to address them. You can find out more about the methodology, findings and recommendations in the full [Public Dialogue report](#).



### Wider stakeholder workshops, March 2022

In the first quarter of 2022, initial drafting of early thinking around the potential directions of travel for future phases of the DARE UK programme was completed, structured around six broad thematic areas of focus as established in the initial landscape review: demonstrating trustworthiness; access and accreditation; data and discovery; core federation services; capability and capacity; and funding and incentives. This was then shared and discussed with stakeholders throughout the month of March 2022 via six open, virtual workshops of two hours each, with each workshop focused on one thematic area of focus.

The workshops were structured to provide participants the chance to understand the early thinking around recommendations, clarify their understanding, and provide specific feedback (for example, gaps, concerns, ratification and so on) through virtual breakout room discussions facilitated by the DARE UK Delivery Team. Overall, the six workshops were attended by 275 participants. The feedback received was compiled and summaries were [published on the DARE UK website](#) capturing the key messages from each workshop. This feedback has been incorporated into this set of recommendations.

### Initial creation of user personas, April 2022

Essential to the design and delivery of a coordinated national data research infrastructure is a clear understanding of the different stakeholder and user groups who have requirements or needs based on their use of or interaction with the infrastructure itself, or requirements based on the impact use of the infrastructure could have on their work or lives. To achieve this understanding, the DARE UK Phase 1 Delivery Team is working in collaboration with stakeholders to develop ‘user personas’ – fictional characters designed to broadly represent the interests and needs of different stakeholder and user groups. These personas are useful for visualising the needs of different groups when designing and testing solutions for a coordinated, efficient and trustworthy national data research infrastructure.

This work began in April 2022 with an in-person, half-day workshop hosted at the Science and Technology Facilities Council (STFC) Hartree Centre, facilitated and led by the STFC team. The DARE UK Delivery Team provided input around the scope of the workshops’ focus. The workshop was attended by 24 participants (plus six members of the DARE UK Delivery Team) from across different research domains, with approximately half of the participants coming from a variety of relevant private sector organisations. This workshop was only the start of what will need to be an iterative discussion with the wider research community to validate a set of representative user personas and

subsequently map out the user journeys these personas could take through the infrastructure that can then guide design choices and decisions. The user personas resulting from this initial workshop can be found in Appendix 2, though as stated this is only the first step and will need to be iterated and validated in future phases of the programme.

### Open call for views on draft recommendations, July 2022

In July 2022, after synthesising the input and feedback received so far during Phase 1 into a draft version of this report and recommendations, we sought feedback from the community – technologists, researchers, members of the public or anybody else interested in responding – via an open call for views. The call remained open for two weeks and stakeholders responded via set questions in a survey format. Specifically, the call asked respondents to comment on how accurately they felt the draft recommendations reflected current challenges across the existing data research infrastructure landscape, addressing any challenges they felt were missing from or not sufficiently addressed; whether they were aware of any initiatives not already mentioned in the report that are currently working on solving some of the issues covered; and whether they felt any of the recommendations set out should be prioritised. A total of 30 responses were received on behalf of organisations or individuals to help shape the final report.

## Sprint Exemplar Projects

As part of DARE UK Phase 1, UK Research and Innovation awarded just over £2 million to fund a programme of nine Sprint Exemplar Projects over eight months from January to the end of August 2022.

The work conducted by the project teams is exploratory, uncovering and testing early thinking in the development of a coordinated and trustworthy national data research infrastructure. Solutions explored in the projects may be taken forward for further exploration or wider adoption if they demonstrate secure, ethical, sustainable, trustworthy and useful working solutions to meet the needs of the wider sensitive data research community and the public. The funding was awarded following an open call for proposals which ran from September to November 2021. All applications received in response to the call were assessed by an independent panel of experts, with the nine successful projects scoring highest in terms of excellence, novelty, and diversity.

Applications were invited and selected in three broad areas:

- 1) **Driver use cases** based on real-example scientific problems, or the lessons learnt from programmes and projects previously funded.
- 2) **Technology demonstrators** that improve data discovery,

metadata management and API (application programming interface) development for: federated analytics; data visualisation; automation of trusted research environment (TRE) ‘Five Safe’ processes; and the development and use of privacy enhancing technologies.

- 3) **Establishing best practice** for information governance, ethics, standards, training and career development, and public involvement to enable the secure use of sensitive data in research.

### Funded projects

The nine Sprint Exemplar Projects include:

- TREEHOOSE: Investigating the use of trusted research environment enclaves for hosting open, original science exploration, led by researchers at the University of Dundee
- PRiAM: Exploring privacy risk assessment methodology, led by researchers at the University of Southampton
- STEADFAST: Education outcomes in young people with diabetes – innovative involvement and governance to support public trust, led by researchers at Cardiff University and Diabetes UK
- Creating a federated, cloud-based trusted research environment to facilitate collaborative research between existing institutions, led by researchers at the Francis Crick Institute
- Overcoming technical and governance barriers to support innovation and interdisciplinary research in trusted research

- environments, led by researchers at the University of Edinburgh
- FED-NET: Creating the blueprint for a federated network of next generation, cross-council trusted research environments, led by researchers at University Hospitals Birmingham
  - Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets, led by researchers at the University of Cambridge
  - FAIR TREATMENT: Federated analytics and artificial intelligence research across trusted research environments for child and adolescent mental health, led by researchers at the University of Cambridge
  - GRAIMatter: Guidelines and resources for artificial intelligence model access from trusted research environments, led by researchers at the University of Dundee

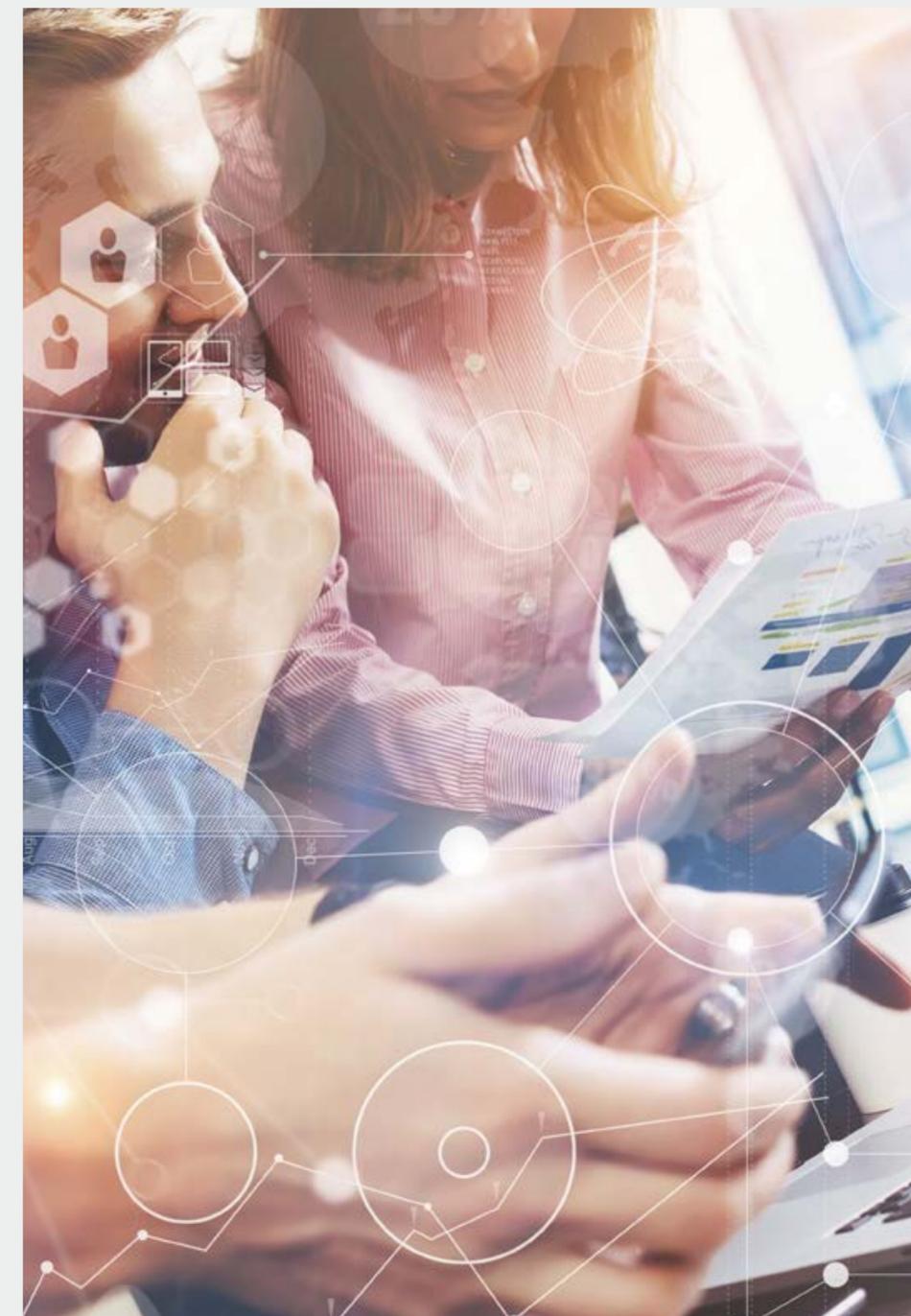
The end of the funding period for the nine Sprint Exemplar Projects coincides with the publication of this report – as such, the findings and recommendations to emerge from the projects will be published separately. Their learnings will then be considered in the next phase of DARE UK to

establish what could be explored further or taken forward more widely in the delivery of a coordinated national data research infrastructure.

Nevertheless, through regular connects with Sprint Exemplar Project teams throughout the funded period, interim insights from the projects have been derived by the DARE UK Phase 1 Delivery Team. Where applicable, these insights have informed the recommendations compiled in this report.

Each of the nine projects is included as a case study in this report, identifiable by the  symbol. You can also find out more about each of the projects [on the DARE UK website](#).

This report sets out a series of recommendations centred around seven broad thematic areas of challenges and opportunities within the sensitive data research landscape, identified through DARE UK Phase 1 to date.



## Structure of the report

Each of the following chapters covers a different theme:

### **Chapter 3: Demonstrating trustworthiness**

How those handling and using sensitive data for research should demonstrate trustworthiness to enable the confidence of the public.

### **Chapter 4: Researcher accreditation and access**

Streamlining access for researchers in the context of trusted research environments (TREs) to support improved governance frameworks and processes for enabling access to sensitive data for researchers.

### **Chapter 5: Accreditation of research environments**

Standards and frameworks for accrediting trusted research environments (TREs) which store, process, and manage sensitive data for analysis.

### **Chapter 6: Data and discovery**

Data and metadata standards, defining sensitive data, evaluating data privacy risks at increasing scale, and considerations for the lifecycle management of research data assets.

### **Chapter 7: Core federation services**

System requirements to begin enabling interoperability across a federated network of trusted research environments (TREs).

### **Chapter 8: Capability and capacity**

Staffing, training, and improved career structures to support an appropriately skilled workforce to underpin sensitive data research.

### **Chapter 9: Funding and incentives**

Long-term, sustainable funding and incentive considerations for a coordinated national data research infrastructure.

When considering the recommendations laid out in this report, it is important to note that timeframes and requirements for delivery will vary, and that they will be subject to prioritisation and resource constraints. While some of the recommendations set out could be actioned in the shorter-term, others will require longer time horizons for delivery or will require further scoping – with involvement from across the research community and the public – in future phases of the DARE UK programme or through other initiatives, as appropriate.

While many of the recommendations sit within the scope of the DARE UK programme and are recommended for adoption in future phases of the programme, others may be considered to fit more appropriately within the remit of other programmes, initiatives or organisations with existing subject-matter expertise. Where this is the case, this is either stated or will be established via further engagement and planning in future phases of DARE UK.

We believe that the recommendations set out in this report have relevance to stakeholders across the entire sensitive data research landscape of the UK, and hope that the synthesis of evidence provided is useful to a broad spectrum of individuals and organisations.

Ultimately, these recommendations are for UK Research and Innovation (UKRI) to consider and decide which elements to take forward based on relevant priorities. Further, it is critical to acknowledge and ensure that work already underway around some of the recommendations should be supported and complimented, to collaboratively and cohesively move the landscape forward.

# 3 / Demonstrating trustworthiness

## Context

DARE UK's focus is on sensitive data. Sensitive data is often data about people, or data which can affect people's lives, and public confidence in how it is handled and used is therefore crucial. We know from [previous public attitudes research](#) that when data is kept safe and secure and used only for purposes in the public benefit, the public are supportive of the use of their data in research. This is reflected in the findings of the [DARE UK Phase 1 public dialogue](#) carried out in early 2022, which explored views towards current sensitive data research practices and where improvements might be needed to enable public confidence. However, the public need to be able to trust that these conditions are met.

Rather than an attempt to 'build' or 'maintain' trust, however, recent discussions around trust in data research have emphasised the need to focus on demonstrating trustworthiness. As emphasised by philosopher Onora O'Neill in her [2013 TEDx talk](#), 'What we don't understand about trust' – when thinking about trust it is important to consider who is the 'giver' of trust and what must be done to receive it:

*“Trust, in the end, is distinctive because it's given by other people... You have to give them the basis for giving you their trust... we need to think much less about trust... much more about being trustworthy, and how you give people adequate, useful and simple evidence that you're trustworthy.”*

Various recent papers have expressed the importance of a focus on demonstrating trustworthiness in the context of research using sensitive data, while others have explored how trustworthiness can be demonstrated in the context of specific research topics<sup>6,7,8</sup>. A [2020 review](#) of research into public understanding and perceptions of, attitudes towards and feelings about data practices by the Living With Data programme found most existing research demonstrates “dissatisfaction with the current ways in which data is used and managed, and a desire for this to change”. Specifically, the authors found that the public want more honesty, transparency and genuine dialogue, as well as regulation, enforcing compliance, safeguards and accountability, and the right to redress. They also wanted more personal control.

These and many other public conversations over recent years have led to a host of recommendations for how the sensitive data research landscape can be more trustworthy. But more needs to be done to drive the implementation of these recommendations and make them a reality.

During our conversations with the public and others during DARE UK Phase 1, we have identified four key factors in demonstrating trustworthiness to the public when sensitive data is used in research:

- 1) **Proactive transparency** around all data research processes – including around what data is used, how it is stored and accessed, why and by whom.
- 2) **Meaningful and inclusive public involvement and engagement**, in which members of the public are involved in governance and decision-making processes.

<sup>6</sup> Aitken M. et al. 2016. [Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme](#). Science and Public Policy, 43:5.

<sup>7</sup> Sheehan M. et al. 2021. [Trust, trustworthiness and sharing patient data for research](#). Journal of Medical Ethics, 47:26.

<sup>8</sup> Milne R. et al. 2021. [Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries](#). Genome Medicine, 13:92.

- 3) **Strong and reliable data security systems and processes** across the entire sensitive data ecosystem, which remain fit for purpose.
- 4) **Public benefit established as the principal motivation** of all research using sensitive data, with involvement from the public in assessment processes.

This chapter addresses each of these factors and concludes with a set of recommendations regarding actions that should be taken by those handling and using sensitive data to better demonstrate trustworthiness.

## Existing challenges and opportunities

### Proactive transparency

Recent initiatives where data collectors and data custodians have aimed to increase the use of sensitive data for research – including the [Care.data](#) scheme in 2013 and the [General Practice Data for Planning and Research \(GPDPR\)](#) initiative in 2021, both of which aimed to enable greater use of general practice data for research – have demonstrated

the necessity of proactive transparency around initiatives involving sensitive data. Both programmes were paused following public outcry due to a lack of information around how the data would be handled and used, leading to privacy concerns.<sup>9, 10</sup>

During Phase 1 of DARE UK, our conversations with the public and other stakeholders identified proactive transparency as being the single most important factor in demonstrating trustworthiness in the use of sensitive data for research. ‘Proactive transparency’ involves making proactive efforts to reach out to the public with information about what is being done with their sensitive data, how and why. This means taking steps beyond putting information on websites for those who seek it out and using a variety of communications channels to actively go out to the public to raise awareness about data research initiatives.

A [2020 report](#) from the Centre for Data Ethics and Innovation (CDEI) found that: “A lot of personal data is shared across and outside the public sector. While this may be for beneficial purposes, public awareness of it is generally low. This gives rise to an environment of ‘tenuous trust’”.

A [public attitudes tracker](#) survey carried out by the CDEI in December 2021 found that existing uncertainty about current data practices, as well as perceived risks around data security, data control and accountability “*are barriers that must be overcome to build confidence in data use*”.

Initially, participants of the DARE UK Phase 1 public dialogue had low understanding of the ways in which sensitive data is used for research; throughout the dialogue, as their understanding grew, so did their sense of its importance for public good. These findings closely resonate with those of other previous studies<sup>11, 12, 13, 14</sup>. Participants felt their own experience during the workshops demonstrated how greater awareness can lead to greater trust and emphasised the need for those handling and using sensitive data for research to actively reach out to the public with information about how and why their data is being used.

A [2021 public dialogue](#) commissioned by the Geospatial Commission exploring the ethics of location data found that “*first, and most importantly, participants wanted transparency.*” The authors found that the public need to know what data is being collected and how it will be used to

<sup>9</sup> Trigg, N. 2013. [Care.data: How did it go so wrong?](#). BBC News.

<sup>10</sup> Which? 2021. [Around 20 million people unaware of plan to share GP medical records with NHS database, Which? finds](#). Which? Press Office.

<sup>11</sup> Aitken M. et al., 2016. [Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme](#). Science and Public Policy, 43:5.

<sup>12</sup> Aitken M. et al., 2016. [Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies](#). BMC Medical Ethics, 17:73.

<sup>13</sup> Cameron, D., et al. 2014. [Dialogue on Data: Exploring the public's views on using administrative data for research purposes](#). IPSOS Mori.

<sup>14</sup> Davies M. et al., 2018. [Public attitudes to data linkage](#). NatCen Social Research.

feel secure. A [public dialogue](#) commissioned by the National Data Guardian in the same year to explore public views towards how health and care data can be used to benefit people and society similarly found consensus on a desire for wide communication about data use for public benefit. Participants felt that, without transparent communications, wider society will think there is “*something to be hidden*”. A multitude of other studies have similarly highlighted the importance of transparency for demonstrating trustworthiness when using sensitive data for research<sup>15, 16, 17</sup>.

Public dialogues, however, typically involve the opportunity for participants to learn about the issues under discussion over several hours and ask questions directly to subject matter experts. Developing messaging which presents

sensitive data research in an understandable and easily digestible way via shorter interactions is therefore a challenge that needs to be overcome. Participants of the DARE UK dialogue and other stakeholders engaged with during Phase 1 also stressed the importance of education in schools about how data is used to generate insights, to give people the requisite knowledge to be able to understand and be part of discussions around the use of sensitive data from a young age.

### Public communications

Public communications are essential for achieving two key goals in demonstrating trustworthiness in the use of sensitive data in research:

- 1) **increased and ongoing proactive transparency** by those handling and using sensitive data for research, particularly data collectors, data custodians and data guardians; and
- 2) **increased general awareness of data research practices**, i.e. through a widescale, tailored public information campaign.

In addition to increased, ongoing proactive transparency, an ambitious public information campaign is essential to address a crucial gap in public awareness about sensitive data research, and the resulting ‘tenuous trust’ highlighted

by the CDEI. A campaign should focus on the use of all types of sensitive data for research – including data about education, welfare, health and care, the environment and more – and be tailored to reach different groups in society.

Participants of the DARE UK dialogue emphasised the need to reach out to different groups and communities via channels and messages that are accessible and pertinent to them. They particularly emphasised the need to reach people without access to the internet, those who don’t have much interaction with or trust of public services, and those who are geographically isolated.

A [2020 paper](#) by H. Kennedy et al. stresses the importance of social inequalities in informing perceptions of data practices. The authors found their research – which involved focus groups exploring views towards ‘datafication’ (the process by which subjects, objects, and practices are transformed into digital data<sup>18</sup>) – to challenge assumptions that understanding is the main pre-requisite to developing views about data practices, and suggest that feelings offer a way to engage people.

<sup>15</sup> Aitken M. et al. 2016. [Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme](#). Science and Public Policy, 43:5.

<sup>16</sup> Davies M. et al. 2018. [Public attitudes to data linkage](#). NatGen Social Research.

<sup>17</sup> Stockdale J. et al. 2019. [“Giving something back”: A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland](#). Wellcome Open Research, 3:6.

<sup>18</sup> Southerton C. 2020. [Datafication](#). Springer Link Encyclopedia of Big Data. Accessed 20.05.2022.

“*There’s a gap between the great work being done and what people actually know. People are badly informed.*”

DARE UK public dialogue participant

Based on their own experience, participants of the DARE UK dialogue suggested different groups could be reached via the following methods:

- through **practitioners** such as health service professionals and teachers, and through trusted community **leaders** such as faith organisations and local councils;
- via **social media and mainstream media** – print, television and radio;
- in **public areas** such as on community noticeboards; and
- at the **point of data collection** – i.e. when people connect with a public service.

The National Data Guardian dialogue found that participants felt communications should be widely distributed and displayed in on and offline spaces such as at GP surgeries, libraries, local authority websites and newsletters, and in community venues. There could also be learnings to take from public campaigns related to other issues, such as stopping smoking and reducing the spread of COVID-19.

Participants of the DARE UK dialogue felt that messaging should present complex issues in an honest, consistent and accessible format and not be overcomplicated, incorporating translations for those whose first language is not English. Although messaging should be tailored to specific groups, dialogue participants felt it should particularly focus on:

- **what sensitive data** is collected from the public, **how and where the data is stored and the security processes** in place to protect it – particularly de-identification and the existence of trusted research environments (TREs) – to reassure the public of how their privacy is protected;
- the **data access processes** for researchers wishing to use sensitive data; and
- the intended and actual **outcomes and impacts of data research projects** – ultimately, how the insights produced from data research could or have impacted people’s lives.

However, participants ultimately stressed the importance of involving members of the public from different communities in identifying channels and co-producing messaging that is accessible and relatable to them.

Other stakeholders engaged with during DARE UK Phase 1 also emphasised a need to raise public awareness of the important role of the private sector in supporting sensitive data research for public benefit and driving innovation.

### Data use registers

Participants of DARE UK’s public dialogue wanted accessible information about what sensitive data is collected from the public, how and where it is stored, the technologies involved in data privacy, how researchers access data, and

*“This is not so much a technical challenge, but a ‘hearts and minds’ issue about trust and trustworthiness.”*

*Researcher, university*

the intended findings and societal implications. Similarly, a recommendation of the [OneLondon citizens’ summit](#) – held in 2020 to explore Londoners’ views towards the use of their health and care data – was for the NHS to produce a publicly available annual report with details of who has accessed data for what purpose, the impact of the research and the distribution of any financial benefits to the NHS.

One solution for providing this information in an ongoing fashion is data use registers. A January 2022 [white paper](#) published by the UK Health Data Research Alliance recommends that all data custodians and controllers should publish and actively promote a data use register of approved research studies, projects and other data uses. The report

recommends that data use registers should be populated in near real time, have a consistent format based on the [‘Five Safes’ framework](#), provide links to research findings and other outputs and exist in both human-readable and machine-readable formats.

NHS Digital has already begun to use [‘Power BI’ data use registers](#), which include information about: the start and end date of the data sharing agreement; the organisation name(s) of the data custodian(s); the purpose for which the data was provided; information on the datasets approved; the legal basis under which data is released; and more. A Power BI report is *“a multi-perspective view into a dataset, with visuals that represent different findings and insights from that dataset”*<sup>19</sup>. The UK Statistics Authority (UKSA) also publishes and maintains a [data use register](#) for data accessed under the [Digital Economy Act, 2017 \(DEA\)](#) by DEA-accredited researchers. The register currently includes details of the data that has been accessed; the researcher(s) accessing the data; the environment in which the data has been accessed; and the date the project was accredited.

Stakeholders engaged with during DARE UK Phase 1 showed widespread support for the adoption of data use registers by the guardians of all types of sensitive data as a key aspect of maintaining transparency. There was a clear view that these registers should be standardised, accessible

in language, format and location and regularly updated and reviewed, with their existence proactively communicated to the public.

### Public involvement and engagement

Public involvement refers to activity which captures and addresses the views and concerns of the public. The primary goal of public involvement is for activities to be carried out ‘with’ or ‘by’ the public, rather than ‘to’, ‘about’ or ‘for’ them; it is to seek public input and make sure it is taken on board<sup>20</sup>. Public involvement can exist at different levels, from exploring views regarding a topic or issue, to involving the public in governance and decision-making processes.

Public engagement is the dissemination of information to the public in a forum in which questions can be asked and views expressed. The primary goal of public engagement is to offer a space for information sharing and dialogue.

Meaningfully involving and engaging the public in data research – particularly those whose lives may be most affected by it – is important for shaping research in a way that reflects and addresses the needs and concerns of society and ensures research outputs are as beneficial as possible. In recent years, public involvement and engagement in research has increasingly been acknowledged as crucial<sup>21</sup>. However, we have heard from those we engaged with during DARE UK

Phase 1 that public involvement and engagement often still appears to be a secondary concern in research using sensitive data. This is particularly the case for research concerning non-health data – such as administrative data relating to education, welfare, justice, social care and more – for which public involvement and engagement is currently far less routine.

A **culture shift** is needed in which all those handling and using sensitive data for research – data collectors and custodians, technologists, funders, researchers and others – fully acknowledge the necessity of public involvement and engagement. They must dedicate appropriate resources to embed it throughout the data research lifecycle and enable it to happen in a meaningful and inclusive way.

Participants of the DARE UK public dialogue felt it was important that members of the public sit on decision-making panels, whilst the National Data Guardian dialogue found public benefit to be undermined without authentic public engagement integrated into data assessment. Participants of the DARE UK dialogue particularly felt the public should be involved in decisions around what is in the ‘public benefit’, due to the subjectiveness of the phrase. They also expressed concern that the public’s views might not be used in a meaningful or

<sup>19</sup> Microsoft Build. [Reports in Power BI](#). Accessed 20.05.2022.

<sup>20</sup> National Institute for Health and Care Research 2021. [Briefing notes for researchers – public involvement in NHS, health and social care research](#). Accessed 10/08/2022.

<sup>21</sup> National Co-ordinating Centre for Public Engagement. [Why does public engagement matter?](#) Accessed 19.05.2022.

genuine way and that they might be involved in research purely as a ‘tick box’ exercise. They felt that, for it to be meaningful, participants in involvement and engagement activities need to be provided with sufficient knowledge and understanding to be able to give informed views and make decisions.

Participants of the OneLondon citizens’ summit similarly desired ongoing public engagement with the use of their health and care data, reflecting that, since their own views had changed during the course of the summit, *“any future public input [should] be equally well informed before it influenced decisions”*.

In addition, it is important that those designing and facilitating involvement and engagement activities have the skills needed to do so in a meaningful way. We heard from those we engaged with during DARE UK Phase 1 that greater support in developing these skills, such as via training and other resources, is crucial if researchers, technologists and others are to develop effective skills in this area. For example, the National Institute for Health and Care Research (NIHR) manages an [online portal](#) where training and resources for public involvement in research can be accessed and uploaded. The Imperial College London Patient Experience Research Centre also provides resources for professionals through its [Public Involvement Resource Hub](#). A similar portal for public involvement in data research relating to all types of data – not only health and care – could be set up. This would require

ongoing resource to ensure it is appropriately maintained. In addition, skills courses could be developed and run by involvement and engagement experts.

Participants of the DARE UK dialogue had a sense that there is not representation of all people living in the UK in engagement and involvement activities; for example, of all nationalities, ethnicities, ages, socio-economic positions and interests. They wanted to see diversity and inclusion in public involvement and engagement, and suggested a more inclusive public could be recruited via:

- **Offline communication channels**, with researchers coming directly into people’s communities – talking to people on the street, distributing fliers and using public noticeboards.
- **Social media**.
- Building relationships with **trusted community members** and visiting physical locations (for example, faith organisations).
- An **online database or portal** where people can sign up to receive newsletters and opportunities to get involved in data research.

Ultimately, it was felt a more proactive approach to recruitment for involvement activities was needed. Participants acknowledged that this would require time and effort, but felt it was crucial to trustworthiness. Ensuring there are mechanisms for public involvement and engagement professionals to share

### Sprint Exemplar Project



## STEADFAST: Education outcomes in young people with diabetes – innovative involvement and governance to support public trust

Diabetes is a common, long-term health condition affecting 40,000 children and young people in the UK. The four UK home nations have legal commitments to support young people with medical conditions in their education, but there are significant challenges in providing evidence to support interventions.

Public understanding and support are critical for the use of sensitive data in research. Researchers at Cardiff University, charity Diabetes UK and partners previously developed a data access framework and set up a Young People with Diabetes Panel to support research into education outcomes for young people with diabetes. The STEADFAST project built on this work by exploring the best ways to inform, engage and involve young people, their families and the wider public in issues around the use of their sensitive data for research. The findings have been developed into a toolkit for use across other health conditions and social impacts.

learning and resources – such as through an independent coordinating function (see Recommendation 5 of this chapter) – could also help to fill gaps in representation.

Each of the DARE UK Phase 1 Sprint Exemplar Projects – which are due to complete at the same time this report is published – have public involvement and engagement embedded throughout, to ensure their work meets public expectations and to test out novel approaches to involvement and engagement. The [STEADFAST project](#), for example, is exploring the best ways to inform, engage and involve young people with health conditions in the use of their sensitive data for research (see box on page 25 for further detail). The outputs of all nine Sprint Exemplar Projects will be useful for informing best practice approaches to involving the public in data research more broadly.

### Data security systems and processes

Participants of the DARE UK public dialogue were reassured by the security processes in place to protect their data and did not express a desire for these processes to be stronger. However, some expressed concern that the systems and processes in place across the UK to protect data varied and were not standardised or regulated; and many were surprised at how long it can take to access data, expressing concern that this may have an impact on the public benefits of research. To better demonstrate trustworthiness in sensitive data storage and access processes, the DARE UK public dialogue

recommends: ***‘Where feasible, processes enabling access to sensitive data for research should be standardised and centralised’***. This could include more standards setting and transparent auditing of TRE structures and processes – including, but not limited to, the setting of standards for what qualifies as a TRE, and adoption of unified user authentication for researchers (see Chapter 4: Researcher accreditation and access; and Chapter 6: Accreditation of research environments) – to ensure best practice is met across the entire ecosystem.

Participants of the dialogue wanted this standardisation to occur on a UK-wide level, with involvement from each of the four nations in agreeing these standards. They felt a more unified approach would be fairer and lead to greater benefits on the whole, whilst acknowledging that flexibility is needed to account for country-specific needs and differing legal frameworks. Centralisation (coordination and oversight by a single responsible body) of processes, such as TRE and researcher accreditation, could occur on either a devolved or UK-wide basis depending on the needs, existing best practice and legal bases of each nation.

More detail regarding DARE UK Phase 1 findings and recommendations for the standardisation and possible centralisation of processes can be found in Chapters 4, 5 and 7 of this report.

“  
*Why do the people who need the data for research have to go through all the different institutes to get the information they need? It seems to be a lot of red tape. I also think it’s a bit worrying that different institutions have different levels of security.*”

DARE UK public dialogue participant

## Research for public good

Public conversations over the last decade and more have [consistently found](#) that people want public good (or ‘public benefit’; ‘public interest’) to be the principal motivation of any research using sensitive data. Members of the public have widely been found to be against the use of sensitive data when it is motivated by financial gain over and above public good. Some participants of the DARE UK dialogue also expressed concern about government use of data to drive political agendas.

However, while participants of the DARE UK dialogue considered excessive financial profit from the use of sensitive data unacceptable, they also did not want financial profit to be a barrier to public benefit. They were comfortable with private sector access to data if the proposed research was independently assessed to ensure it is motivated by public benefit above all, with any other benefits existing in appropriate balance. Participants of [deliberative workshops](#) held with members of the Scottish public as part of [a DARE UK Sprint Exemplar Project](#) led by researchers at the University of Edinburgh (see box on page 34) had similar findings, with participants acknowledging the potential benefits of private sector researchers accessing health data. Participants of the Geospatial Commission dialogue recognised that *“for the benefits of location data to be realised, all kinds of organisations need to be involved, including those that may also have interests beyond only*

*benefiting the public”*. See also Chapter 4: Researcher accreditation and access.

Research accreditation under the DEA requires that public interest is the primary purpose of research projects, as set out in the [UKSA Research Code of Practice and Accreditation Criteria](#). The Code of Practice sets out the conditions required to demonstrate that public interest is the primary purpose of the work, which were developed with two rounds of public consultation.

Nevertheless, participants of the DARE UK dialogue and other recent public conversations have expressed a desire for further work to explore the concept of public good/public benefit in more depth, and what it might mean for different groups in society. Participants of the DARE UK dialogue expressed that the term ‘public benefit’ is subjective and wanted the public to be deciding what is in the public benefit, with public benefit assessments for data access requests being made on a case-by-case basis. Participants of the National Data Guardian dialogue wanted a clear, broad and flexible definition to be developed for the case-by-case assessment of public benefit relating to the use of health and social care data. At the time of writing, a public dialogue jointly led by ADR UK (Administrative Data Research UK) and the Office for Statistics Regulation (OSR) was underway to explore views towards the public good of data and statistics.

## Recommendations

In line with the above, to better demonstrate trustworthiness in the handling and use of sensitive data for research DARE UK Phase 1 makes the following recommendations:

### 1 Consistently practice proactive transparency about what sensitive data is being used for research, how, why and by whom.

Data collectors, data custodians and data guardians should consistently practice proactive transparency – in which **active efforts** are made to reach out to the public with information about what is being done with their sensitive data, how, why and by whom.

Sustain proactive transparency throughout the **entire data journey**, from data collection to research impacts.

Develop **honest, accessible and consistent** messaging and definitions to be adopted across the sector.

Tailor communications to different communities and groups to ensure a **diverse and inclusive** public is reached, with involvement from members of those communities and groups in developing messaging

that it accessible and relatable to them.

Dedicate **sufficient resources** to carry out proactive transparency in an ongoing fashion.

## 2 Conduct a UK-wide public information campaign to raise general awareness of how and why sensitive data is made accessible for research.

In line with the principles of proactive transparency as outlined under Recommendation 1:

- **Fund and resource** an ambitious campaign to raise general awareness across UK society.
- Bring data research into the **mainstream** via channels such as newspapers, television and social media.
- **Collaborate with relevant organisations and practitioners**, such as teachers and GPs, and **involve the public** in developing messaging and identifying communications channels.
- Focus messaging on **what data research** is and what the **public benefits** are, and the **security processes** in place to protect data.

The next phase of the DARE UK programme will involve further work to scope out how such a campaign should be conducted, and by whom – including what resources would be required – in

greater depth. This may build upon the work of the newly formed Public Engagement with Data Research Initiative (PEDRI), a working group including representatives from DARE UK, HDR UK, ADR UK, ONS, the CDEI, the Ada Lovelace Institute and others, which aims to take a more coordinated, cross-sector approach to public engagement with data research across the UK. At the time of writing, PEDRI was in the process of designing a pilot public information campaign to raise awareness of sensitive data research within specific communities.

## 3 Publish and maintain standardised and accessible data use registers.

Guardians of **all types of sensitive data** should publish and maintain data use registers.

Include information about **who** has accessed **what data, when** and for **what purpose**, and cross-reference any research outputs and impacts as they emerge. Regularly update registers as and when data access is granted.

Develop a **standard, clear and accessible format** for data use registers in collaboration with stakeholders from across the community.

Include requirements for data use registers as a **condition for TRE accreditation** (see Chapter 5: Accreditation of research environments), as well as mandating as a condition for UKRI funding.

**Centrally collate data use registers** – or direct people to them from a central place – alongside information describing them, so that people can find and access them easily.

**Widely promote** data use registers to raise awareness of where people can find out what data has been used and for what purposes. The maintenance of data use registers will require **dedicated resource** to ensure longevity.

## 4 Drive a culture shift to recognise the crucial importance of public involvement and engagement and embed it throughout the sensitive data research lifecycle.

Data custodians, data guardians and researchers should **embed public involvement and engagement** as a central component across the entire research lifecycle, and not as an afterthought.

**Involve members of the public in governance and decision-making processes** relating to the use of

sensitive data in research, particularly in relation to assessment of ‘public benefit’.

Involve people from across different groups, communities, backgrounds and identities and give them the time and resources needed to fully understand and respond to the issues being discussed to enable **meaningful and inclusive** public involvement and engagement.

Include **dedicated funding** for public involvement and engagement as part of research grant applications; hire **dedicated staff** within data research organisational structures; and provide **incentives** for researchers to conduct public involvement and engagement activities.

Embed **mandatory requirements** for public involvement and engagement within data access requests and research funding applications; include these activities in monitoring and reporting processes. Enable access to **training and resources** for researchers, technologists and others to support the delivery of public involvement and engagement activities.

Recommendation 5 sets out what is needed to drive forward and resource this culture shift across the sector.

5

### Investigate the requirements for establishing an independent coordinating function for public involvement and engagement with sensitive data research, either as a new entity or as an off-shoot of a relevant existing body.

This should involve engagement with relevant organisations from across the community and members of the public to establish what would be needed (in terms of both remit and resources), which organisations would be responsible for leading it, and how.

This function would need be appropriately and sustainably resourced and cover all types of sensitive data from across the different research domains.

The function could:

- Lead the consolidation and adoption a **best practice standards** public involvement and engagement with sensitive data research.
- Lead on **better understanding and documenting public attitudes** towards the use of sensitive data for research; and drive forward the **implementation of recommendations** to emerge from public conversations about sensitive data research.
- Provide a central point of **information sharing**

**and coordination** for public involvement and engagement professionals, to support collaboration towards shared goals and avoid duplication of effort.

- Develop and provide (or signpost to existing) **public involvement and engagement training and resources** for researchers, technologists and others working within the sensitive data research sector.

The purpose of this function should not be to replace the role of individual institutions or initiatives embedding public involvement and engagement into their own ways of working, but rather to drive a more coordinated approach across the sensitive data research ecosystem. Some of this work may build upon existing work, such as that of [Understanding Patient Data](#) (but covering all types of sensitive data, not only health) and the newly formed PEDRI group (see above).

6

### Standardise, centralise and unify processes enabling access to sensitive data for research across the UK where appropriate and feasible.

Consolidate and drive the adoption of **UK-wide best practice standards** for TRE structures and processes (for example, the setting of standards

for what qualifies as a TRE), with consideration of international best practice also (see also Chapter 4: Researcher accreditation and access; and Chapter 5: Accreditation of research environments for more detailed recommendations in this area).

Where appropriate and feasible, centralise (place coordination and oversight with a single responsible body) processes and systems either on a **devolved or UK-wide basis** depending on the individual needs, legal bases and existing best practice of each of the four nations.

**Streamline and unify** researcher accreditation and data access request processes to avoid delaying or discouraging important research in the public benefit.

Set the **same data access processes for all researchers** – including those from academia, government and the third and private sectors. Include rigorous assessment of public benefit and maintain transparency and stringent monitoring of researchers throughout the entire research lifecycle (see also Chapter 4: Researcher accreditation and access, Recommendation 2).

**Regularly review** processes for accessing sensitive data to ensure they remain fit for purpose as

technology advances; and regularly review researcher accreditation status to make sure researchers continue to meet standards over time.

Further detail regarding DARE UK Phase 1 recommendations in relation to accreditation and data access processes for researchers can be found in Chapter 4: Researcher accreditation and access.



# 4 / Researcher accreditation and access

## Context

Safe and timely access to sensitive data is crucial to enable research and innovation for public good at scale. However, one of the biggest challenges highlighted by stakeholders during DARE UK Phase 1 is the time it takes for researchers to move through accreditation and application processes for data access and have all approvals, checks and safeguards in place to do the analysis.

Ethical and secure access to sensitive data is often impeded by lengthy information governance processes that are labour intensive and often inconsistent – many data guardians have different processes for data access and collect information from researchers in an inconsistent way. Researchers are often left with doubts about what training to obtain and how to fill in applications, sometimes finding themselves having to provide the same information to multiple data guardians in a slightly different format.

This chapter focusses on three main aspects of streamlining access for researchers in the context of trusted research environments (TREs):

- The user accreditation process
- Unifying user authentication capabilities for accredited researchers to access TRE services
- Data access request standardisation

The processes, policies, and standards of data access panels and committees are outside the scope of this report, as they are often subjective on the basis of the combined requirements of the data guardian, the data custodian (for example, TRE provider) and the research project. This chapter therefore focuses on the underlying baseline standards and support required to enable complex governance procedures to take place in a more efficient way by streamlining how researchers are accredited within the network and the processes for subsequently accessing sensitive data for analysis.

However, it is clear from feedback received from the community throughout this first phase of the DARE UK programme that the challenges cannot be addressed through technical approaches alone – addressing the

cultural and behavioural factors that heavily influence data access decisions is crucial alongside any enabling technology approaches.

A consistent challenge that has been identified across all stakeholders during DARE UK Phase 1 has been a lack of or unclear and inconsistent application of researcher access and accreditation best practice standards that have evolved organically over time. Variations across the ecosystem can often lead to misinterpretations of policies, leading to undue delays in data access for research, or resources wasted upon ‘reinventing the wheel’ for each new research environment. A simple and clear example of this is user authentication – every TRE needs to provide some form of user identification and authentication for accredited researchers to access services within their environment. Despite the availability of many identity federations (for example, [UK Access Management Federation for Education and Research](#); [Geant](#)) and industry standards (for example, [OpenID Connect Federation](#); [OAuth 2.0](#)), each TRE implements user authentication in their own bespoke way, creating a

lock-in and friction in the system where a user needs to create new logins to access data from each TRE. Managing this in a piece-meal fashion further creates potential security risks as each implementation may vary.

This lack of standards is also seen in researcher accreditation (information governance and sensitive data handling training) requirements which place a large administrative burden on TREs to verify and validate each researcher separately for access to each TRE. User authentication and researcher accreditation standards must also have a global outlook and not focus only on a UK-wide approach. Science is global, so our approach must be the same.

## Existing challenges and opportunities

Alongside the input received as part of DARE UK Phase 1 as outlined in Chapter 2: Process and summary of input, this chapter summarises additional international input from the [Global Alliance for Genomics and Health \(GA4GH\)](#), the [Australian Research Data Commons \(ARDC\)](#), [Towards European Health Data Space \(TEHDAS\)](#), the [World Health Organization \(WHO\)](#) and the US [National Institutes of Health \(NIH\)](#). Though these are predominantly within the health sector, the maturity of sensitive data research in the health domain internationally provides valuable insights to learn from.

DARE UK's [recent public dialogue](#) also provided two key recommendations in this area, including: unifying processes and systems supporting data research across the four nations of the UK; and, where feasible, centralising and standardising processes enabling access to sensitive data for research as outlined in detail in Chapter 3: Demonstrating trustworthiness. Participants of the dialogue wanted the four nations of the UK to be unified in their approaches to the use of sensitive data in research, while acknowledging that there exist unique, country-specific needs and issues that require a level of flexibility. They also felt greater centralisation and standardisation would improve their confidence in the use of sensitive data as it would provide clearer oversight and would speed up research benefit for the public by streamlining processes. Centralisation of processes – in which coordination and oversight is placed with a single responsible body – could therefore occur on either a UK-wide or devolved level depending on the needs, existing best practice and legal bases of each nation. Involvement from members of the community from each of the four nations in setting best practice standards and agreeing nationwide processes where relevant is crucial.

### Federated identity and user authentication standards

There is a need to identify – in collaboration with stakeholders from across the landscape – and drive forward the adoption of a common user authentication protocol by

infrastructure providers. Conceivably, this would need to be coordinated and overseen by UKRI itself, as it has the appropriate remit to act as such an authority.

A transparent, cross-domain, national approach could remove the responsibility from individual groups and therefore improve consistency and increase efficiency across the sensitive data research ecosystem. Stakeholders engaged with during DARE UK Phase 1 have highlighted

*“When researchers realise the secure data requirement, they’re trying to avoid it. People just change the variables they request access to... There needs to be a level of flexibility... So many regulations and requirements make the use of data slow and difficult.”*

*Researcher, university*

that this is a prerequisite for all forms of federation to occur and will aid in the creation of a ‘research passport’ that is cross-linked to multiple regulatory bodies for verification and validation by data custodians.

Stakeholders also highlighted existing federations, for example the [UK Access Management Federation for Education and Research](#), which will need to either be expanded or linked to other federations being created, such as [NHS Care Identity Service 2 \(CIS2\)](#) or [GovRoam](#). The existence of modern industry and community standards of user authentication (for example, [SAML](#), [OIDC](#), [OAUTH2](#) and [Global Alliance for Genomics and Health \(GA4GH\) Passports](#)) were also highlighted. These existing standards should be leveraged as the basis for user authentication to allow for maximum interoperability at a national and international level. As user authentication is a crucial component of a national TRE standard, stakeholders also highlighted the need to support different forms of identity verification and have logging and auditing embedded across the system.

### Researcher accreditation

A key requirement highlighted by stakeholders has been the need for a streamlined approach to researcher accreditation. While there are a number of existing training modules for sensitive data handling (for example, [those provided by ONS](#)), many of these trainings are duplicative without allowing for equivalence or mutual recognition between modules.

Those engaged with during Phase 1 highlighted the need to develop a shared standard with service users and providers towards a federated approach to training content. Modularisation was also highlighted as important to allow for flexibility to cater for specific data modalities or sensitivities, for example through ‘core’ modules as a standard foundation for all accreditation courses with the possibility of ‘extended’ modules in specific cases or contexts as needed.

Stakeholders affirmed that work to standardise and streamline the researcher accreditation process was sorely needed, along with reciprocal or mutual recognition of accreditation by different TRE providers. Providers should aim to offer a consistent researcher experience across data access points, and ideally make the process feel as though the researcher were accessing data on their own machine when this is not the case. Training could be made portable across TREs through standard accreditation for researchers acting as a TRE ‘passport’. The [Digital Economy Act, 2017 \(DEA\)](#) already works as a passport in some respects, with shared accreditation existing across certain TREs.

Stakeholders considered it best practice for TREs to maintain teams of individuals to support researchers, data collectors and data guardians, including the ability to maintain transparency around what sensitive data is being used for and by whom. A surge in people using TREs would need to be prepared for and staffed, as discussed in

Chapter 8: Capability and capacity. A key recommendation is therefore to provide online training modules that can be delivered at a national and international scale with on-site drop-ins to scale delivery, and regular maintenance of researcher accreditation.

### Private sector and international researcher accreditation

Currently, private sector researchers can apply to become accredited researchers under the DEA, and therefore apply for access to data held within DEA-accredited research environments once accredited, via the same process as academic researchers. In the context of UKRI-funded research, private sector researchers can also participate in sensitive data research in the public good as part of consortia led by a UKRI-approved research organisation.

However, there was widespread feedback from stakeholders engaged with during DARE UK Phase 1 that improving the ability for private sector researchers to collaborate on sensitive data research is important. Participants of the DARE UK public dialogue wanted sensitive data to be made securely accessible to private sector organisations and did not see a need for data access requirements to differ for these organisations, as long as the research is motivated by public benefit over financial profit and there is transparency throughout the research lifecycle (see Chapter 3: Demonstrating trustworthiness).

“ Profit should not be a barrier to research which is valuable to the public. ”

DARE UK public dialogue participant

A DARE UK Phase 1 Sprint Exemplar Project – ‘Overcoming technical and governance barriers to support innovation and interdisciplinary research in trusted research environments (TREs)’ (see box to the right) – has also [explored public perspectives](#) on private sector access to health data in Scotland. They found that participants acknowledged the potential benefits of ‘non-traditional researchers’ (including private sector organisations) accessing health data and were comfortable with access so long as certain conditions are met – most of which align with the conditions currently required of all types of researchers accessing sensitive data. Some stakeholders engaged with during Phase 1, however, have suggested the possibility of a specific accreditation framework for private sector researchers and organisations (including, for example, guidance around fair corporate use, fair compensation, and prohibition of outright monetisation).

Our recommendation at this time is that private sector organisations should continue to be subject to the same (stringent) data access and accreditation processes as other types of researchers, with a strong focus on maintaining public benefit as the primary research purpose. This is in line with public expectations as outlined above, and a general consensus amongst stakeholders engaged with during DARE UK Phase 1 that the ability for private sector researchers to collaborate on sensitive data research in an ethical and secure way needs improving rather than curtailing.

The accreditation of international researchers is also an important element identified by stakeholders. Currently, accredited researchers require a link to a UK institution, primarily to ensure that from a legal jurisdiction perspective there is the possibility of legal recourse should they breach their obligations under the accreditation framework. This is an important element in providing confidence to the public and data guardians that there are appropriate consequences for any misuse of their sensitive data. Given that TREs operate in a global context, connectedness with global partners is essential, including those in low-resource settings. The accreditation of international researchers is therefore a topic for further investigation.

### Sprint Exemplar Project



## Overcoming technical and governance barriers to support innovation and interdisciplinary research in trusted research environments (TREs)

Led by the DataLoch team at the University of Edinburgh in collaboration with Public Health Scotland (PHS), this project has explored the barriers to the use of TREs by researchers from different disciplines – for example, those from the third and private sectors. The project team have delivered a prototype solution and user training module for DataLoch operating alongside the National Safe Haven in Scotland.

The project has investigated public perspectives around different types of researchers accessing health and social care data, and is producing a lessons learned report with recommendations for TREs across the UK. This has included exploration of what is expected from different types of researchers to be considered trustworthy and credible; what additional technical security is required from a TRE to support research and innovation projects from different disciplines or organisations; and what different information governance is required to support TRE access from different types of researchers.

### Data access request standardisation

Stakeholders engaged with during DARE UK Phase 1 highlighted standards in information security, platform specifications and service descriptions, as well as a centralised approach to data access requests and licensing. Working with data custodians and data guardians to agree standards with governance teams to navigate the interpretations of legal positions would be an important step forward in normalising data access requests and licensing. Developing and setting standards alongside work with higher-level government bodies to set policy would

address the desire from stakeholders, particularly those within research councils, to ensure platforms are accessible to researchers across disciplines and that they work for everyone, not just select groups of researchers.

Stakeholders were keen to convene a ‘research data alliance’ to help consolidate the data access request standard and align with international efforts to minimise duplication. They also highlighted the need to learn and leverage existing work, such as the [HDR UK Gateway’s Five Safes form](#) or [HRA’s Integrated Research Application System \(IRAS\)](#), as well as from consortiums such as the [BHF Data Science Centre](#) and the [SAIL Databank](#).

Leveraging and harmonising existing data access request processes into a single baseline procedure around the Five Safes that can be instituted and maintained by a centralised service would be a substantial step forward. Furthermore, to improve public transparency and system-wide intelligence, the use of cross-links of a data access request with other entities such as people, project, grant and datasets that can then be used to publish data use registers should be supported and mandated (see also Chapter 3: Demonstrating trustworthiness).



“Custodians have different requirements for what they consider is needed to be accredited. This can be clunky in terms of validation times and differing training requirements.”

Workshop participant

## Recommendations

DARE UK Phase 1 makes the following recommendations for researcher access and accreditation, with delivery in collaboration with the wider community and existing initiatives in both the UK and internationally:

- 1 Provide a unified user authentication capability to enable researchers to access services more easily across the entire sensitive data research ecosystem (see also Chapter 7: Core federation services).**

Leverage existing identity federations to develop a **framework for identity brokerage services** to allow them to be cross-linked, drawing on existing industry and community standards as the basis to allow for maximum interoperability nationally and internationally.

Pilot a **test case of identity federation and authentication** nationally and internationally.

- 2 Provide a streamlined researcher accreditation framework to enable trustworthy researchers to access sensitive data for research in the public benefit in a timelier fashion.**

Leverage existing work from regulatory authorities and TREs to institute a **federated approach to accreditation**.

Develop **consistent guidance** for stakeholders to undertake accreditation, including private sector researchers.

Develop **accreditation online training modules** that can be delivered at a UK-wide scale with on-site drop-ins to scale the delivery and maintenance of accreditation.

- 3 Develop a standardised and streamlined – yet extensible – process for accredited researchers to request access to sensitive data from data guardians whilst maintaining appropriate levels of data privacy and security.**

Leverage and harmonise existing data access request procedures and processes into a **single baseline procedure** that can be instituted by providing **centralised support** for research institutions in adopting and implementing a common protocol – with common tooling to manage it.

**Align data access request forms** using the Five Safes framework.

**Develop consistent resource descriptors** – for datasets, tools, funders/sponsors, people, and project/grant identifier – that can be queried and linked across TREs to ensure data access request procedures can leverage system-wide intelligence.

Publish **data use registers** transparently for all approved data access requests flowing through the network (see also Chapter 3: Demonstrating trustworthiness, Recommendation 3).

# 5 / Accreditation of research environments

## Context

The focus of this chapter is around the accreditation of trusted research environments (TREs). Standard and transparent accreditation of TREs is crucial if both data guardians and the public are to feel confident that sensitive data is securely held and appropriately managed for research in the context of a federated network of TREs (see Chapter 7: Core federation services). Critically, the recommendations outlined in this chapter consider the need for a coordinated infrastructure that supports sensitive data linkage and analysis across research domains, especially when data components of the linkage may fall under different legal frameworks.

As a starting point for the context around this chapter, it is important to state that reference to research environments is specifically related to the technical infrastructure (both hardware and software) that stores, processes, and manages sensitive data for research – in this case TREs. Further, it should be made clear at the outset that, in principle, there should not be more than a single standard for TREs across the UK. The accreditation of processors of sensitive data (which encompasses TREs) that falls

under the remit of the [Digital Economy Act, 2017 \(DEA\)](#) is well established, with the authority for that accreditation assigned to the [UK Statistics Authority \(UKSA\)](#). Where new or additional accreditation criteria are deemed necessary for the processing of data which is outside of the scope of the DEA (for example, certain health data), these should draw upon the existing DEA accreditation framework. This would reduce the duplication of effort and make sure there is alignment across the ecosystem, which would aid the creation of a federated network of TREs.

It is also crucial to acknowledge that accreditation and audit is a significant commitment of time and staffing for TRE operators and there is therefore a need to ensure processes are not duplicative and that mutual recognition exists between accreditation frameworks (if truly more than one is necessary). TREs are fundamentally composed of people, processes, policies, and technologies which together enable efficient and safe access to sensitive data for research. Heterogeneous and often incompatible TREs are being created almost like a cottage industry in response to the need to manage secure access to sensitive data

for research. Beyond hindering the need for clarity and confidence for data guardians and the public, there are interoperability challenges created when the management processes of different TREs are not aligned. Streamlined and harmonised management of data access is a foundational requirement for any TRE alongside interoperable data and interoperable systems (which are addressed in Chapters 6 and 7 respectively). Furthermore, the governance frameworks for managing data access and enabling researchers to conduct analysis must also be aligned to achieve a federated network of TREs.

Feedback throughout this first phase of the programme – including from the public (see Chapter 3: Demonstrating trustworthiness) – has been consistent in the need for a more standardised approach towards what defines a TRE – in other words, an accepted TRE standard – alongside an accreditation framework which is aligned to that standard and includes independent audit, serving as an accepted authority and providing vital clarity and confidence to data guardians and the public.

## Existing challenges and opportunities

### TRE standards

TREs, while a relatively new terminology, have existed to varying degrees for some time and certainly the standalone characteristics of a TRE are not conceptually new – there are long-standing examples such as [UK Biobank](#), the [SAIL Databank](#), the [Scottish National Data Safe Havens](#) (facilitated through the [electronic Data Research and Innovation Service \(eDRIS\)](#)) and [Genomics England](#), to name just a few. It is the emergence of greater demand (and visibility) for access to sensitive data for research, the increasing sensitivity around the risks of large-scale data analysis, and the increasing scale of the data itself that have jointly driven the landscape towards the TRE as a solution.

Most – if not all – existing TREs within the UK operate in line with the ‘Five Safes’ framework developed by the Office for National Statistics: safe people, safe projects, safe data, safe settings and safe outputs<sup>22</sup>. Although a simplified definition of a TRE has been provided for the purpose of this report, a key challenge that has not been adequately addressed is the establishment of an *agreed* definition of a TRE across the data research community.

This should not be as an abstraction of the Five Safes as this is widely understood and agreed, but rather a definition at a level of detail that can be effectively structured into a standard against which accreditation can be executed and linked to. It should also feature an appropriate, independent audit process to ensure compliance, covering details such as administrative setup, access management processes, security and privacy management processes, federation outlook, technical capability and maturity. Fundamentally, while there is clear consensus across the community that all TREs should adhere to the Five Safes framework, there need to be proportionate approaches to applying the Five Safes based on the sensitivity of the data and the related risk (and impact) of disclosure.

Data sensitivity is a spectrum; accordingly, how that sensitivity is managed should vary. However, there should be a minimum or baseline threshold of defined characteristics and requirements that define a TRE. Varying interpretations of what a TRE is add additional complexity to the information and data governance decisions that data guardians need to make, which in turn slows down research and the public benefits of that research, and hinders clarity for researchers and the public on how these decisions are made consistently. Ultimately, a TRE standard with appropriate flexibility to cater for the varying

tiers of data sensitivity and resultant tiers of environments and capabilities is necessary to provide clarity across the ecosystem of what constitutes a TRE.

Some participants of the [DARE UK Phase 1 public dialogue](#) expressed the view that, as long as the public are aware data is kept safe and secure, different types of data should not be subject to differing access requirements for researchers. However, the issue was not explored in significant depth with participants, with the implications of a tiered versus non-tiered approach to sensitive data not having been discussed, and deeper public conversation regarding the issue – as well as public involvement in the development of any tiered approach to data sensitivity – is therefore recommended.

Some important considerations from stakeholders during DARE UK Phase 1 have been around:

- A TRE standard and accreditation framework should look to draw on the range of certifications that already exist (for example, ISO27001/27002/27701/27017/27018<sup>23</sup> and the [NHS DSP Toolkit](#), among others), as a basis for enhancing the existing DEA standard as a baseline that could be extended with plugins or extensions to cater for data not within the scope of the DEA – for example, specific use cases or role-based access models.

<sup>22</sup> Desai, T. et al. [Five Safes: Designing Data](#). University of the West of England.

<sup>23</sup> International Organization for Standardization. [ISO - Standards](#). Accessed: 09.08.2022

- A TRE standard and accreditation framework should not only consider how a TRE operates in isolation, but critically how TREs interoperate, further providing transparency through a central register of accredited TREs.
- A key focus should be **UK-wide** recognition of a TRE standard and accreditation framework, acknowledging and leveraging expertise across the four nations of the UK, as desired by participants of the DARE UK Phase 1 public dialogue (see Chapter 3: Demonstrating trustworthiness) and others.
- Compliance to a TRE standard and accreditation framework should be incentivised through funding opportunities, with careful consideration of how this contributes to the fiscal sustainability of TREs and an interoperable network of TREs. However, funding should be linked with minimum levels of service in terms of staffing, compute and speed of disclosure control.

As mentioned above, the standalone characteristics of a TRE are not new and there is significant expertise within the UK research community that must be leveraged effectively in defining what a flexible TRE standard should be.

### TRE accreditation

It is important to acknowledge that while there should not be a proliferation of different standards for what is defined as a TRE, not all TREs will be the same – nor should they be – so

long as they adhere to the Five Safes framework. There are myriad factors influencing how a TRE could be established in accordance with the Five Safes framework, such as the sensitivity, type, volume and velocity of data, the purpose of the research and the tools (software or hardware) needed to carry out the research.

As such, there are two very important principles that need to be considered in the accreditation of TREs. The first has been covered in the need for a **flexible** standard that can provide an agreed baseline threshold of defined characteristics and requirements that define a TRE across the spectrum of data research domains – including how TREs interoperate. The second is that of mutual recognition; that is, the mutual recognition across the landscape that, while there may be supplementary or additional standard and accreditation requirements in certain areas (for example, in specific data domains), there are core components that are largely equivalent across all TREs.

Acknowledging and accepting that there are components of a flexible TRE standard and accreditation framework that are broadly equivalent regardless of the data research domain, the focus should be on expanding and extending the existing standard and accreditation framework that exists under the DEA administered by UKSA as the responsible independent authority. This is critical to ensuring that TRE

operators can manage the significant time and staffing costs associated with maintaining their accreditation status. The principle of mutual recognition is important in the DARE UK context of cross-domain sensitive data research and especially important in reducing the burden to enabling interoperability between a federated network of TREs across the academic, public, third and private sectors for research in the public good. There should not be more than a single standard and accreditation process for TREs across the UK research ecosystem; working to further develop the established standard under the DEA together as a research community so that it is adequately flexible to meet the broad range of research requirements is the sensible approach.

There are likely to be many emerging research use cases that will dictate the need for dedicated or specialised TREs with design principles that differ from existing TRE capabilities. These use cases will become clearer over time and in future phases of the DARE UK programme. Furthermore, there is a need for a TRE landscape based on open standards where federation is incentivised and TRE operators can compete on service delivery to drive innovation. We therefore do not make recommendations for the number of TREs that could or should exist in a cross-domain, federated infrastructure for sensitive data research. We will maintain an ongoing dialogue with the community on this topic throughout further phases of the programme.

## TRE audit

Alongside any accreditation process should be an authority and audit process to ensure the standards that are being accredited against are adequately adhered to. In the context of TREs, an independent authority and process should be established to effectively accredit and audit TREs against the relevant standard. This is of course important to ensure compliance with the relevant standard but critically also to provide researchers, data collectors and data guardians with clarity about how each TRE is set up and the opportunity to browse and compare a central register of TREs and their capabilities. This would enable them to make an informed decision on the most appropriate TRE for their purpose or inform information governance decisions related to making sensitive data available for research. This will also allow TREs to demonstrate trustworthiness through a consistent TRE accreditation process that can be independently verified, providing a strong foundation for the confidence of the public, data collectors and data guardians in TREs handling sensitive data.

It is important to acknowledge that the existing UKSA DEA accreditation framework and process is operating today, and as such should be considered as a strong foundation on which to build, with the opportunity to revise the existing framework and processes of the DEA audit framework. It should be re-iterated that some of the foundational components of what constitutes a TRE across different

research domains will be the same, and accordingly the accreditation (and related audit) framework should reflect this. Equally, it must be acknowledged that there will be many differences, and appropriate, independent audit – including work on extending the existing framework to cater for these differences – is critically important.

## Recommendation

The following recommendation is made for investment as part of UKRI's broader remit to support cross-domain research on sensitive data across the four nations of the UK, with delivery in collaboration with the wider community and existing initiatives in both the UK and internationally:

- 1 **Review and extend the existing standard, accreditation, and audit framework under the Digital Economy Act (DEA) to further establish it as the nationally recognised trusted research environment (TRE) standard, accreditation, and audit framework.**

In collaboration with the UK Statistics Authority (UKSA) – as the independent authority – and with involvement from stakeholders across the UK research community (including the private sector) and the public, **develop a working definition of a TRE.**

Iterate to achieve a consistent, standard definition of a TRE with appropriate flexibility, ensuring this is harmonised with existing standards and pulling in rather than reinventing what already exists (for example, ISO27001/27002). This standard definition should cover the broad range of research domains as well as the minimum standard for interoperability between TREs.

Review and, where necessary, extend the existing **approved processor accreditation and audit framework** under the DEA to ensure it covers the broad range of sensitive data research domains and TREs specifically.

Investigate a **tiered approach to data sensitivity** and the implications for a TRE standard, accreditation, and audit framework, in consultation with the public. Develop a **searchable central registry of accredited TREs** with transparent summaries of capabilities. **Implement and test** the accreditation process with a set of TREs from separate domains to refine and consolidate the process.

Establish a **consistent cadence for review** of the standard, accreditation, and audit framework to ensure it remains fit for purpose as the research, data, and technology landscape evolves.

# 6 / Data and discovery

## Context

This chapter summarises the input from a wide range of stakeholders, as well as previous work from research data lifecycle management efforts from funders, universities and research organisations.

Data and metadata lifecycle management underpin every trusted research environment (TRE) and every project within a TRE. Lifecycle management for data and metadata allows TRE operators to ensure the right data is shared with the right people, for the right purposes, with appropriate permissions and governance applied in the right settings. The [Findable, Accessible, Interoperable and Reusable \(FAIR\) principles](#) is the accepted solution to efficient data sharing practices.

Metadata provides information about other data, including a description of the data. This includes information that provides context to the data – for example, how they were collected, the coverage of the data, and licencing arrangements. Metadata can include such information as publication date, description and search keywords. Metadata can be held at a variety of levels from

administrative information about the dataset, to field level technical descriptions of the datasets to overview statistics of the datasets (for example, the number of participants included in the datasets).

According to the FAIR principles, research data should be:

- **Findable** – data should include metadata and a persistent identifier (a link that continues to provide access to the dataset into the indefinite future<sup>24</sup>) to make it discoverable.
- **Accessible** – metadata should be freely accessible in open formats and standards, with documented routes to request access to sensitive data.
- **Interoperable** – metadata should use controlled vocabularies, be machine-readable and include references to other metadata.
- **Re-usable** – metadata and data should conform to existing standards for greatest reusability.

Making data FAIR is critically important and benefits all participants in the research ecosystem. However, it requires consistent effort and should continue to be supported both

within and across research domains. Increased visibility and documented routes for the availability of sensitive data available within TREs would allow more research to be conducted and, importantly, would improve the efficiency with which this increasing scale of research takes place (see Chapter 4: Researcher accreditation and access).

Interoperable metadata facilitates innovative research through novel linkage of datasets, not only within research domains themselves but increasingly between different research domains. This should be extended to also cover open datasets and commercial datasets that are available for research. Reusable data and metadata ensure the outputs of innovative research can be reused and built upon by others, though this is a real challenge when considered from a cross-domain perspective.

Most research funders across the research and charitable sector require the development of research data management plans in support of research projects and other investment. This has improved the findability and

<sup>24</sup> Kunze J. 2003. [Towards Electronic Persistence Using ARK Identifiers](#). California Digital Library.

accessibility of data outputs but has not addressed other challenges in interoperability and reusability.

It must be acknowledged, as we heard from stakeholders throughout DARE UK Phase 1, that individual research domains hold tacit knowledge (on multiple levels of granularity) that by nature cannot be easily interpreted outside of the domain itself without specialist support. Naturally, there are many technical standards for data that have emerged as a result. However, this proliferation of standards leads to a challenge in enabling the interoperability of data even within research domains, let alone across them. Data from different sources is recorded in variable ways, using a variety of data standards and common data models, and is also described in different ways using a variety of metadata standards. Even if the same data standard has been used, other features of data can differ, leading to a reduction in interoperability and reusability. As with the advance in research, data standards frequently become outdated and need to be maintained, creating an additional administrative burden on the data collection and maintenance phases of the data lifecycle.

Efficient recording of metadata is often left as an after-thought, leading to the decreased utility of the data. Simple metadata attributes such as missingness and spatial and temporal coverage can lead to significant improvements in the utility of the data. A lack of consistent metadata for



datasets can lead to data being misunderstood and under-utilised for research projects; or worse, data collection being performed again, leading to wasted resources.

Metadata catalogues are available in some domains – ONS recently launched a [new metadata catalogue](#) for data held in its Secure Research Service and HDR UK maintains the [Innovation Gateway](#) for discovery of datasets related to health, and these two catalogues interoperate with one another. However, metadata catalogues do not exist in all domains and do not always interoperate with each other, both within and across research domains. Lack of such

catalogues also prevents datasets from being assigned persistent identifiers that would improve the clarity of ownership and responsibility for maintaining the metadata. Lack of visibility also prevents the ability to understand who is using what data for what purpose, reducing collaborations and transparency (see Chapter 3: Demonstrating trustworthiness).

## Existing challenges and opportunities

Engagement during DARE UK Phase 1 has helped identify several challenges related to data and discovery that need to be addressed as the infrastructure and ecosystem evolve to meet the needs of cross-domain research on sensitive data. Further, the DARE UK Phase 1 [public dialogue](#) highlighted the need for a standardised approach for access to sensitive data, which would help ensure adherence to data security and ethics best practices and improve efficiency of data access so that research is not unduly delayed by differing data management processes (see Chapter 3: Demonstrating trustworthiness).

### Data management lifecycle

Stakeholders engaged with during DARE UK Phase 1 expressed overwhelming feedback for more streamlined data management lifecycle approaches across data custodians and TREs, with the need for more data

stewardship capacity to help manage data collection, curation, harmonisation and linkage being a recurring theme (see Chapter 8: Capability and capacity). With the volume of data being generated, a key concern for stakeholders has been deciding what data to keep and what not to – there is a need for a more consistent approach to data archiving within and across research council domains, as currently required for ESRC funding with final artefacts being archived with the [UK Data Service](#) (see Chapter 9: Funding and incentives, Recommendation 6).

Many stakeholders also highlighted the lack of cross-TRE approaches across all stages of the data management pipeline, from access requests through multiple data guardians, coordinating data preparations and linkage, to data provisioning (securely bringing data into TREs to be made accessible for research), especially when linkage is required. This causes the administrative burden of coordinating the provisioning of data across TREs to be significantly greater than provisioning individual datasets. Similarly, stakeholders also highlighted significant challenges in obtaining approvals from multiple data guardians for access. Trusted third parties to support data provisioning and linkage have been proposed in the past, though concerns have been noted on the fairness, security and costs associated.

### Data standards

The challenges of dealing with the volume and variety of sensitive data with uses in research was highlighted in our engagement during Phase 1, with a particular emphasis on streaming (continuously produced), wearable and near-real-time data. Beyond these specific types of data, stakeholders also highlighted the need to consider the storage of new data types, such as imaging data. Imaging data will consume more storage than is practical, with some data being processed only at the point of production and the raw data being discarded – for example, the compression of sequencing data before storage or the pre-processing of internet of things data (such as data from wearables) at the edge. These emerging data requirements are already impacting the volume, velocity and variety of data within the sensitive data research ecosystem.

Data standards, metadata standards and interoperability standards were of importance to stakeholders. However, there was clear feedback from some that introducing additional standards would not serve to alleviate all of the challenges, since some of the key challenges are around governance and the capacity to curate and maintain data and metadata. Stakeholders also highlighted the need to adopt standard approaches for describing data utility and quality. Tools will need to be developed that can be used to

automate the collection of information related to data utility and quality, to be shared through metadata. Phase 2 of the programme could include an open call to evaluate the wide range of open and commercial tools available.

### Data governance and privacy

Following the initial development of a data lifecycle management approach, stakeholders engaged with during DARE UK Phase 1 highlighted the need for a streamlined approach to data governance and improvements in privacy risk management. In the absence of an appropriate risk assessment tool with endorsement from the research community and the public, data guardians have traditionally tended to take a cautious approach to granting access to sensitive data for research. While this may have been well-suited to the management of limited amounts of data in the past, the advent of new techniques to assessing privacy risk – and, importantly, proportionate privacy risk – can support improved management of data governance processes at scale. The DARE UK Phase 1 [Privacy Risk Assessment Methodology \(PRiAM\) Sprint Exemplar Project](#) has explored solutions to a privacy risk assessment framework (see box on page 44). The project outputs will provide useful insight for the development of a framework for wider adoption across the sensitive data research landscape.

**Sprint Exemplar Project****PRiAM: Privacy Risk Assessment Methodology**

Organisations responsible for data protection must demonstrate that sharing data for research does not put individuals' privacy at risk. Although best practice privacy management principles such as the 'Five Safes' framework are used, there is no standard privacy risk assessment approach. This leaves organisations to make their own choices about risk management.

Personal data may be held by many organisations. Often, research requires combinations of data – for example, studying patients' journey from hospital to recovery may involve combining medical data with data from social care and digital health applications. With no standard risk assessment approach, it's hard for multiple organisations to assess and manage risk consistently.

Led by researchers at the University of Southampton, PRiAM has produced a way of assessing privacy risks for data managed by multiple organisations. Engaging experts and members of the public in research use cases, a privacy risk assessment framework has been developed.

Alongside this, new techniques to minimise privacy risk – such as privacy enhancing technologies (PETs), where there is already work underway by the Information Commissioner's Office (ICO), the United Nations (UN), the Royal Society, the Alan Turing Institute and the Centre for Data Ethics and Innovation (CDEI) – provide new opportunities to develop approaches that not only minimise privacy risk but also support data custodians and data guardians in their decision-making around data governance. This enables better efficiency and consistency of decisions relating to data access, and also allows risk management to be more proportionate. However, considerable work remains to translate PETs from research into production use, including identifying those technologies that together would be most effective in supporting the privacy of sensitive data in a federated network of TREs (see Chapter 7: Core federation services).

**Sensitive data taxonomy**

Throughout discussions with stakeholders and members of the DARE UK Programme Board and Scientific and Technical Advisory Group during Phase 1, it has been clear that there is no existing simple definition of sensitive data, nor a 'taxonomy' – or classification – that could be used to describe such data. The development of a sensitive data taxonomy is important to support work on privacy, linkage and the approaches that might be applied to different

data for federation. For example, for data that is typically horizontally partitioned (logically equivalent data held in different locations which can be analysed locally and then aggregated for meta-analysis) or vertically partitioned (data in different locations that need to be logically linked to be analysed as if a single dataset)<sup>25</sup>.

A taxonomy for sensitive data will need to encompass not only existing structured and unstructured data, but also emerging types of data. For example, data from wearables (for example, heart monitors and smart watches) and other emerging technologies, for which delivery may not be through conventional datasets but potentially through approaches such as publish/subscribe distribution (an interaction pattern that characterises the exchange of messages between publishing and subscribing clients)<sup>26</sup>. This work would also assist the development of cross-domain synthetic data to support a variety of needs, such as early development of models and training of analysts to work on cross-domain sensitive data research.

It is therefore proposed that the next phase of the DARE UK programme should include work with the sensitive data research community – including members of the public – to develop a cross-domain taxonomy for sensitive data.

<sup>25</sup> Towards Data Science. [Database Terminologies: Partitioning](#). Accessed 14.07.2022.

<sup>26</sup> Jacobsen H. 2009. [Publish/Subscribe](#). Encyclopedia of Database Systems. Accessed 14.08.2022.

## Metadata and discoverability

Making data discoverable, for example through the publication of metadata, was highlighted by multiple stakeholders as a first step in the direction of federation. Stakeholders engaged with during DARE UK Phase 1 highlighted the availability of many existing standards and warned not to invent a new standard. They also highlighted the challenge of creating terminologies and controlled vocabularies across domains.

There is a clear opportunity for UKRI to assist in making recommendations for use of certain metadata standards or convening groups to collaborate on developing, enhancing and/or adopting data standards. Consortia of bodies (such as universities) could be best placed in the implementation of standards, particularly those with similar interests, to share their learnings. One of our recommendations is to survey the UKRI-council landscape on metadata usage for different modalities and current approaches. Further to this, there is a need to define a minimally acceptable metadata standard across all UKRI councils with opportunities for individual councils to extend the minimum standard for capturing metadata specific to their domain. Stakeholders expressed that the focus should be on ensuring the approach to standardising metadata is straightforward. One approach would be to concentrate the effort on agreeing cross-discipline, high level descriptive metadata and then agree common document formats for

data modalities that would facilitate sharing across TREs. It will also be important to look at approaches that have worked across the different research domains and in the private sector.

Some stakeholders also suggested that metadata should include indications of the level of curation of the underlying dataset. This would be an indication of the ‘data curation debt’ that is associated with a dataset – in other words, the work that might be needed to make the dataset ‘research-ready’. For consistency, a standard, cross-disciplinary framework to describe data utility should be adopted, building on existing work<sup>27</sup>.

The DARE UK programme should also seek opportunities to collaborate with those already active in the area, such as the [FAIRsharing](#) organisation and the Alan Turing Institute-led [AI for Multiple Long-Term Conditions Research Support Facility](#).

Researchers need user-friendly metadata to describe datasets (or objects). UKRI could support this with more tools or capabilities for data-holders to register their datasets and their terminologies. If increased transparency enabled a member of the public to see where their data is being used, this could also help to demonstrate trustworthiness in the use of sensitive data for research. In addition, to facilitate cross-council discovery and reuse, we recommend the creation of a federated registry to hold

a list of available catalogues, standards, vocabularies and terminologies as held and used by different TREs across the ecosystem.

UKRI could also support the creation of infrastructure that allows for the sharing of metadata, browsing services for different types of data and pointing towards places or groups that could provide feedback on the data.

Understanding quality, missingness and how a dataset was generated requires collaboration. We recommend the development or selection of a reference implementation of a metadata catalogue that can support one or more minimally



<sup>27</sup> Gordon B. et al. 2021. [Development of a data utility framework to support effective health data curation](#). *BMJ Health & Care Informatics*. 28:e100303.

acceptable metadata standard and the use of Digital Object Identifiers (DOIs) – unique character strings which can be assigned to datasets to identify their location<sup>28</sup>.

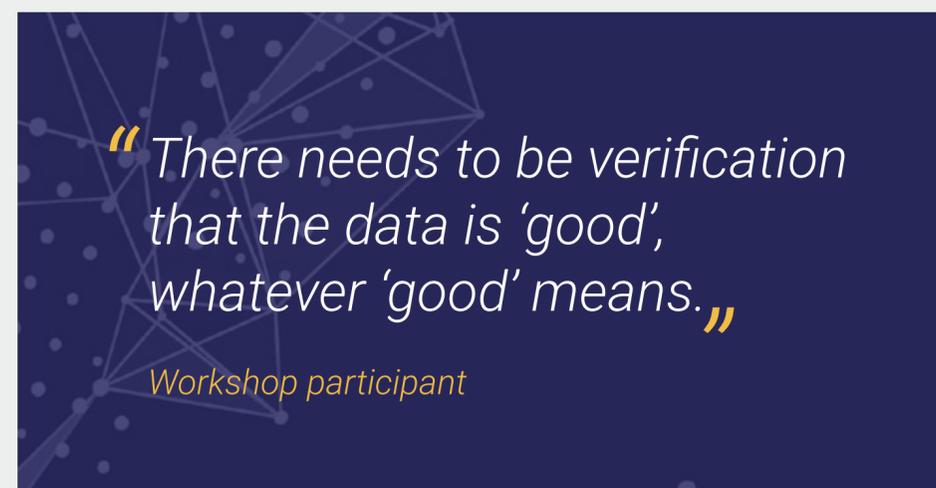
Stakeholders engaged with during DARE UK Phase 1 highlighted the challenge of identifiers of data being made available across the existing fragmented landscape. Each release of a dataset, and revisions of those datasets, needs a unique DOI that can be cross-linked with other resources. It was expressed that a DOI service needs to be rich enough to provide all the technical details necessary for discovery, cross-linkage and distribution. It will also need continuous investment for maintenance.

A number of data custodians also highlighted the need to invest in the management and maintenance of metadata, noting that this is an ongoing requirement and can be a significant cost for them. This has not been highlighted as a specific recommendation of DARE UK Phase 1 but does need attention from funders.

### Advanced discoverability

Observations were made by several stakeholders that it should be possible to extend the concept of cohort discovery from the health domain, to having cross-domain utility. This would allow distributed discovery to be run securely against the raw data. The security of such an approach can be achieved by ensuring that the sensitive

data remain in the host TRE and the queries are shipped to be run locally, with just statistical information returned to the research for meta-analysis. This is functionality allows the research to understand whether the demographics within a dataset (or datasets) match the requirements of their research study before committing to a potentially long data access request process. This could be explored further in Phase 2 of the DARE UK programming build from experiences such as the [COVID CO-CONNECT programme](#).



## Recommendations

The following key recommendations are made for investment in future phases of DARE UK to support data and discovery, with delivery in collaboration with the wider community and existing initiatives both within the UK and internationally:

### 1 Enhance the data lifecycle to support effective cross-domain sensitive data research.

Develop a common, cross-domain **taxonomy (classification) for sensitive data**.

Pilot cross-council, **automated provisioning pipelines** for sensitive and open data, with analysis being conducted in TREs.

Develop an approach for **data provisioning between federated TREs** to support use cases where federated analytics is not technically feasible.

Evaluate existing **data utility frameworks** and identify appropriate options that can be extended to cover cross-domain and multimodal data.

<sup>28</sup> University of Strathclyde. [Digital object identifiers](#). Accessed 04.08.2022.

## 2 Explore the implications of new data types on approaches to making these data available for research.

Review emerging **data requirements for new data types** (for example, data collected by wearables such as smart watches), delivery models beyond datasets, such as streaming data, and requirements for near-real-time access to data.

Develop a **data management lifecycle model** to address the requirements of new, emerging data modalities.

To facilitate the development and support of streaming data modalities, develop a minimum viable product (MVP) service to support **near-real-time flow** of internet of things data, such as data from wearables, using streaming technology.

## 3 Develop guidelines on privacy enhancing technologies (PETs) for use by TREs.

In collaboration with existing initiatives (for example, the ICO, UN, Royal Society and Turing), develop guidelines on the **deployment of PETs** alongside TREs.

Develop a **risk model for the linkage of cross-council data**, and provision of linked data, to support guidelines on the usage of PETs, building on some of the work of the DARE UK Phase 1 PRiAM Sprint Exemplar Project.

Run a **focused research call** to demonstrate effective use of PETs for federated analysis on sensitive data. Develop **training** for research and technical teams on the effective use and deployment of PETs.

## 4 Establish a UKRI-wide metadata standard working group.

**Survey each UKRI council landscape** on metadata usage for different data modalities and current approaches.

Define a **minimally accepted metadata standard** across all UKRI councils, extending existing standards to define UKRI-council minimal metadata standards and pilot the metadata standard.

Explore options for **search and browse** over a network of federate metadata catalogue and extending this capability to **securely analyse** the statistical characteristics of the associated datasets.

Develop or select a reference implementation of a metadata catalogue that can support the metadata specifications and use of DOIs.

Define a **federated registry** to hold a list of available catalogues, standards, and vocabularies/terminologies.

## 5 Leverage existing Digital Object Identifier (DOI) minting services to provide persistent identifiers for all UKRI discoverable assets at UKRI-wide and council levels.

Provide a central **UKRI-council level service and guidance** for UKRI data custodians.

Review recommending that **all UKRI investments have metadata registered** for all appropriate outputs, to support findability and reuse.

# 7 / Core federation services

## Context

The requirements discussed in this chapter are central to delivering the core infrastructure elements in support of DARE UK's aim to design and deliver a national data research infrastructure that is coordinated, demonstrates trustworthiness and supports research at scale for public good.

To enable efficient and trustworthy research using sensitive data, researchers require access to data within a secure context that can support a wide range of analytical and computational capabilities. Trusted research environments (TREs) have emerged as the preferred solution for providing highly secure digital environments that enable remote access to information and analytical tools. Whilst there is no single definition of a TRE, much less an established approach for accreditation (see Chapter 5: Accreditation of research environments), the consensus is that the Office for National Statistics (ONS) 'Five Safes' model – safe people, safe projects, safe settings, safe data and safe outputs – represents an appropriate framework for minimising the risk of data misuse or the disclosure of personal information.

The UK's TRE landscape is developing rapidly with the addition of new capability and the deployment of new environments. There has also been development of successful new models of TREs – for example, to provide a separation between the researcher and the raw data, with analysis developed on synthetic or sample data and then deployed on the full data without the researcher having any direct data access. The [OpenSafely](#) project showed the potential for this approach with several impactful COVID-19 studies. This approach has been shown to extend to other data modalities – for example, through the work by the [London Medical Imaging & AI Centre for Value Based Healthcare](#).

It should be anticipated that the approaches adopted by TREs will continue to develop across the domains and the evolving requirements of different use cases. However, there is a real risk that, in the absence of carefully guided, strategic investment, the current evolution will result in an even more fragmented landscape without a trustworthy model for accreditation or effective interoperability. DARE

UK needs to address this by providing leadership to ensure that a federated network of TREs that builds on existing investments can be established, ensuring security and privacy are appropriate and proportionate for a range of data and levels of sensitivity. This will be the focus of this chapter.

It is also important to consider the technical recommendation in this chapter in the wider context of culture changes, skills development and governance, as many of the existing challenges across the sensitive data research landscape will not be addressed purely through technology solutions. An example of this is the need to align with the ICO's [Regulatory Sandbox programme](#).

## Existing challenges and opportunities

Engagement with stakeholders during DARE UK Phase 1 has identified several consistent challenges with evolving the current infrastructure to meet the future needs of cross-domain research on sensitive data.

First, there are many physical and software infrastructures existing across the UK research landscape, but with **very limited integration**, which has resulted in siloed (isolated) working. This is particularly the case when crossing research organisations and disciplines, limiting the scope of research and the questions that can be asked and answered despite an abundance of data. There is an increasing need for cross-disciplinary research to answer questions of importance; for example, the impacts of climate change on infectious diseases. Researchers wanting to access data from multiple environments face hurdles in terms of multiple data access request applications and often this results in longer delays (see Chapter 4: Researcher accreditation and access). [ADR UK \(Administrative Data Research UK\)](#) is doing important ongoing work to enable the creation of legacy data linkages for the wider research community; for example, cross-domain linkages relating to health and education, and education and childhood vulnerability. However, in most cases, researchers wishing to carry out cross-disciplinary research must work out from scratch how to access the data for each project. Federation of TREs is critical to linking different sources of data and facilitating deeper cross-disciplinary research.

There is also a high degree of scepticism in the research community, especially in the health data domain, about the **efficiency and effectiveness of research within TREs**.

It will therefore be important that investment is focused to provide enablement services for federated analytics that have been co-designed with the research community and can then be effectively deployed across a network of TREs. In addition, the research community identified the need to be able to access large-scale compute on-demand – several research communities expressed a requirement for intermittent access to high performance computing (HPC) and high throughput computing (HTC) capability. For this to be provisioned within the UK in a cost-effective way, it also needs to integrate with a federated infrastructure that encourages sharing. This applies equally where provision is on-premise or through public cloud. Additionally, as the UK moves to a high level of dependency on provisioned environments for research, the availability characteristics will become more critical. The prolonged loss of capability from a single key national infrastructure could impact research across a wider range of projects.

Some input was received from stakeholders around supporting ‘**safe return**’, i.e. the ability to provide a service that would securely reidentify records so that identifiable information can be returned to a data collector in exceptional circumstances – for example, if a researcher were to identify someone as being at risk of an adverse health outcome. It is important to note that this reidentification would not be undertaken by the researcher

but through a service, possibly a third-party linkage service that would allow the data collector to reidentify the record from the de-identified record flagged by the researcher.

An alternative approach to addressing ‘safe return’ would be to develop best practices and approaches for researchers to share analysis code with the data collectors to allow the identification from within the original datasets. This would avoid the legal, technical and governance challenges of securely implementing safe return while ensuring compliance with, for example, the DEA. It is recommended that this

“  
Researchers often access data and need compute resource in an episodic way. They might suddenly need compute and storage infrastructure to process images or run models.  
”

*Technologist, research council*

approach is explored before a more complex technical and legal approach is considered – piloting a proof-of-concept implementation is therefore not recommended at this time.

There was strong support among stakeholders for the development of an open but formally governed, community-led project to build a **TRE reference architecture** (or blueprint/template) that could be deployed using cloud native technologies<sup>29</sup>, perhaps using an existing framework such as from the [Apache Software Foundation](#). This would then support integration with core federation services without the need to install additional services. A number of existing projects were identified that could act as a starting point for this, including work by the Alan Turing Institute<sup>30</sup> and Microsoft<sup>31</sup>, as well as outputs from the DARE UK Phase 1 Sprint Exemplar Projects – for example, the [TREEHOOSE project](#) (see box to the right).

It was clear that this isn't a 'silver bullet' that solves all problems, but would need to be extensible, built to allow evolution as new technologies become available and integrated with an open portfolio of governance and operational processes. This would also need to align with work on accreditation. Some concern was raised that organisations would still need to be fully aware of

the staffing and skills needed to run a secure production environment; use of a TRE reference architecture would not eliminate this requirement. It is therefore important that any work to provide a TRE reference architecture is supported with guidance on appropriate cybersecurity and operational processes and procedures.

There was also interest in the **provision of a 'sandpit' environment** where researchers could explore potential cross-domain use cases using synthetic and open data. This would enable early testing of the viability of a project prior to potentially lengthy and costly applications for access to the datasets themselves.

Stakeholder input on **business continuity and disaster recovery** showed a wide range of differing opinions. Some considered this to be a key issue and expressed that current approaches are often no more sophisticated than ensuring there are offsite backups. The move to federation, with more significant use of public cloud, was seen as an opportunity to partially mitigate some of the risks of site failure or malicious attack. Others took the view that infrastructures could be recovered with a move to a more software-defined infrastructure model. In addition, some felt that data custodian holds responsibility for re-provisioning primary data and researchers for re-provisioning their research artefacts, and any investments would therefore likely be disproportionate to risk.

<sup>29</sup> InfoWorld 2021. [What is cloud native? The modern way to develop software](#). Accessed 04.08.2022.

<sup>30</sup> The Alan Turing Institute. Data safe havens in the cloud. Accessed 04.08.2022.

<sup>31</sup> GitHub. Microsoft/AzureTRE. Accessed 04.08.2022.

### Sprint Exemplar Project



## TREEHOOSE: Trusted Research Environment and Enclave for Hosting Open Original Science Exploration

There is currently little standardisation of trusted research environment (TRE) infrastructure or deployment between operators, leading to duplication of effort and hindering service improvement. Led by researchers at the University of Dundee, TREEHOOSE has built on the research team's experience of migrating a TRE hosting anonymised patient data over to 'public cloud' – meaning the data is accessible over a secure internet connection rather than in a local data centre – for the benefit of other operators.

The project has explored a new capability of 'enclave computing' to TREs, which add a layer of software encryption to protect intellectual property and code in addition to the data itself. Secure enclaves go beyond the traditional TRE infrastructure by adding additional barriers to prevent software algorithms from leaking data.

The researchers will release open-source software to streamline building and operating TREs on public cloud infrastructure whilst maintaining security and trust.

Two other observations to note were that the costs of replication for some data, such as imaging or geonomics, would be prohibitive; and that replication would also need to have appropriate governance to ensure data protection requirements were covered across mirrored repositories. There was more general support for providing greater resilience for the ‘crown jewels’ of data and research – though it is not clear how these would be identified – and suggestion that this could be through a centralised service. There was also more consensus that further study was needed and that there should be a strategy to cover business continuity and disaster recovery requirements for a federated network of TREs. It will also be important that any approach is aligned with the requirements of the data custodians, as these may impose restrictions on the back-up or replication of data.

The clear view from most stakeholders was that there needs to be a **common, open library of APIs (application programming interfaces) and services** and that this needs to be funded and community led. It must be open source to avoid proprietary lock in, and there should be both a definition of the services and APIs and reference implementations with sample usage. There was a view also expressed by several stakeholders that the emphasis should be on supporting federated and machine learning APIs. Strong emphasis was placed on the need to assemble rather than reinvent, building from existing projects and focusing

on the needs of identified use cases, and not on building interesting APIs just for technical curiosity.

There was strong support for validating all development with **driver projects**; it was felt that a few outstanding projects during DARE UK Phase 2 would demonstrate the need, and potentially would be far more impactful than any number of plans and documents. However, concern was raised that short projects might not be viable, especially given lead times on getting access to sensitive data. It was felt that there would need to be a range of driver projects to cover the breadth of data and the need to validate both essential and edge requirements. A key concern raised by a few people was the need for a funded service and support infrastructure for these services, including a help desk and consultancy on how to make effective use of the infrastructure and services.

The public input – including from the recent DARE UK Phase 1 [public dialogue](#) (see Chapter 3: Demonstrating trustworthiness) – has been overwhelmingly in favour of research on sensitive data, provided that it is undertaken securely, transparently and with clear intended public benefit. As the scale of research in TREs increases, it will be important to look to how **trustworthiness** can continue to be demonstrated. This could involve greater automation of key processes to support the Five Safes model, including partial automation of federated statistical disclosure control;

data pipeline management; and policy-driven approaches to accreditation and data access request management.

A number of these challenges and opportunities are being explored by the DARE UK Phase 1 Sprint Exemplar Projects, the interim outputs of which have informed these recommendations.

## A federated network of cross-domain trusted research environments (TREs)

The following section will review the key technical requirements that need to be met to address the challenges and opportunities identified during DARE UK Phase 1. These include the recommended investment decisions to be made for Phase 2 of the DARE UK programme to prepare the foundations for Phase 3 – deployment of a world-class, federated infrastructure to support cross-domain research on sensitive data across the four nations of the UK.

There is a clear opportunity to create a distributed and federated infrastructure that will be more effective in supporting research using linked data from different disciplines. The federated approach has huge potential benefits across UKRI-funded research. Federation of data and analytics could solve a number of unmet needs – particularly related to health and administrative data – by

providing capability to securely link, for example, health, crime, housing, education, environmental, consumer and retail data. Federation could also fulfil specific research needs across the UK nations, including cross-disciplinary research into the environment, human movement and economic opportunity.

However, there are several key principles that should be followed in the design and delivery of a federated infrastructure. First, any development should build from and integrate with existing capability, involve co-design across the community, be based on well-governed and open-source practices and avoid re-invention or duplication. Co-design should involve engagement with both the public and private sectors, with public involvement and engagement embedded throughout (see Chapter 3: Demonstrating trustworthiness), and open-source projects should be shared through a managed and sustainable framework (for example, [Apache Software Foundation](#)). All requirements and designs should be tested with use cases and driver projects.

In addition, this should be a peer-network of TREs with no central coordination, and all services should be deployable in any TRE that is implemented on an appropriate technology stack. This is not about building a centralised infrastructure, but supporting a collaborative approach

that democratises access to data and infrastructure and addresses the needs of areas of historic underinvestment. Ultimately, delivery must support a more flexible and efficient model of research that aligns with [UKRI's net zero objectives](#).

Delivering these capabilities through an open-source approach should provide capability that can be consumed effectively by both public-funded and private infrastructures – such as the [AIMES Trustworthy Research Environment](#) and the [Canon Safe Haven Artificial Intelligence Platform](#). The governance and licencing approaches for the open-source delivery will need to allow for the adoption by both public and privately funded infrastructure without jeopardising the intellectual property of private sector providers.

Consensus amongst stakeholders engaged with so far during DARE UK Phase 1 was that a federated infrastructure should not be delivered through a new infrastructure – except where there are technology gaps that cannot be addressed with existing assets or complimentary investments – but through the gradual provisioning of new API-enabled services that integrate with existing and novel infrastructures. Importantly, this is not about the deployment of new, national level TREs nor of significant additional deployments of storage or compute. It is also important that this programme of work aligns with other aspects of the [UKRI Digital Research Infrastructure programme](#), especially

“  
Do not aim to not re-invent the wheel. There are already good solutions in place. Be a place for consensus of best practice.”

*Workshop participant*

around capacity building and a carbon neutral future for research and research infrastructure. This work should also align closely with other major investments – such as the [NHS Federated Data Platform](#) – to ensure interoperability.

Phase 2 of the DARE UK programme should design and deliver proof-of-concept deployment of a scalable set of API-enabled core federation services to integrate existing and future TREs with appropriate information governance processes to provide security and demonstrate a robust approach to trustworthiness. It is proposed these services be delivered using cloud-native technologies and approaches, such as containerisation, to provide flexibility and support reuse. The focus for DARE UK should be on the integration of services to provide a consistent common framework for federated analysis across research domains.

“  
 I think there's a need for environments where there can be linkage of... data, and permitting more free analysis within those environments. It's difficult to share by virtue of its bulk and also the sensitivity.  
 ”

Researcher, university

### Identity federation

There is a need to provide authenticated, authorised and auditable access to federated resources using standardised, single sign-on and identity federation. This should integrate with the research accreditation capabilities discussed in Chapter 4: Researcher accreditation and access.

There are a wide range of existing initiatives in this space, such as: the UKRI-funded [JISC National Association for the](#)

[Advancement of Artificial Intelligence \(AAAI\) Framework for Researchers](#); open-source projects including the widely adopted [Keycloak platform](#); and private sector offerings such as [Mvine](#), which has been successfully deployed at scale in the telecoms industry to manage access to mobile applications. It should not therefore be necessary to implement new core services, but rather drive consensus on an approach and then provide enablement and access to a managed deployment of the service. Stakeholders also viewed that a broker capability would be essential to federated identity services to integrate with existing services such as [Eduroam](#) and the NHS network.

### Enablement to support federated analytics

To facilitate a federated approach to analytics will require the ability to deploy a wider range of tools with standardised workspaces using cloud-native open technologies (for example, [Docker](#) and [Kubernetes](#)). The use of a container-based approach will allow for the deployment of capability across TREs, reuse of best practice and tools, and support reproducibility and improved high availability characteristics. The focus should be on cross-community, cross-vendor and cross-tool capabilities such that an independent technical framework can drive open innovation. The infrastructure adopted should support deployment to existing on-premise infrastructures, as well as future hybrid and public cloud TREs.

There should be consideration of developing an open repository (based perhaps on [Docker Hub](#)) to encourage the reuse of workflows and best practices, and to enhance trustworthiness by the open sharing of these workflows. This would also enhance reproducibility and reduce the level of duplicative rework across projects. In DARE UK Phase 2, it is recommended that a pilot be run on creating a portable container-based workspace, with more extensive development in Phase 3.

The approach should align with a containerised model for the deployment of core federations services including tools, workflows and integrated analytics environments, building upon work from other initiatives such as the [GA4GH Cloud workstream](#), which shares the same ambition of ‘bringing the algorithms to the data’ by creating standards for defining, sharing and executing portable workflows.

The DARE UK Phase 1 Sprint Exemplar Projects – such as that demonstrating [federation of genomic data](#) held across TREs (see box on page 54), and the [FAIR TREATMENT](#) project (see box on page 55) – will provide valuable insights in this area. The DARE UK programme should not, however, look to deliver the AI research or algorithms themselves, but focus on enabling capabilities.

## Metadata federation

There is a need to provide services to support the federation of metadata – including enabling data custodians to control the publication of metadata; and consumers to discover, analyse and visualise metadata as appropriate to their requirements (see also Chapter 6: Data and discovery). This should not be dependent on a centrally coordinated approach to metadata management; however, it will need to support integration with existing catalogue services such as the [HDR UK Innovation Gateway](#). The programme should also look at emerging open-source metadata distribution projects such as the Linux Foundation [Egeria Project](#), and novel discovery and visualisation, an example of which is the [Linked Data Explorer](#) from Agrimetrics.

This must have the capabilities to support both managed and open data sources. Metadata federation is fundamental to providing a comprehensive, cross-discipline data discovery service.

## TRE reference architecture

There are a number of established TRE environments that have operated securely and effectively for many years, and there is also a regular cadence of new environments being commissioned and deployed. This potentially risks fragmentation and could hinder efforts to share best

### Sprint Exemplar Project



## Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets

Many organisations have their own trusted research environments (TREs), but they cannot currently 'talk' to each other. The ability for TREs to talk is known as federation. Even where researchers are allowed to use data held in separate TREs, analysing them together would still require their combination within a single TRE, which is challenging, costly and can delay new discoveries. Led by researchers at the University of Cambridge, this project has created a UK first demonstration of federation of genomic data by bridging the TREs of the NIHR Cambridge Biomedical Research Centre and Genomics England. After querying the genomic data within the two separate TREs, a joint analysis was run within both environments and the results combined in a separate secure cloud environment – no original data moved, only results.

Learnings from the project will unlock unprecedented possibilities for collaborations with clinical-genomic data across the UK Research and Innovation research councils.

practice on implementation, operation, and integration. It is proposed that DARE UK Phase 2 should develop a standard reference architecture (or blueprint/template) – a 'TRE in a box' – that can be used for the development of these new environments. It should be possible to use this for standalone TREs, but it should already be fully integrated with the core federation services discussed in this chapter to lower the barrier of entry to integrations with other infrastructures.

The reference architecture should follow the [Infrastructure as Code model](#) to ensure reproducible environments that can be easily extended, customised and shared. It should be supplemented with open-source, cloud-native architectures building on learnings from the DARE UK Phase 1 Sprint Exemplar Projects such as [FED-NET](#) (see box on page 56), and work by partners and the private sector. Any reference architecture should be built to provide an abstraction layer above the services of specific cloud providers to ensure that it can be effective at targeting particular providers, but still provide portability across environments.

The reference architecture should include both technical capability and an integrated governance framework. This should build on existing work underway by the Alan Turing Institute – the '[Data safe havens in the cloud](#)' project; work by Microsoft to provide an [open-source TRE on Azure](#); and other

**Sprint Exemplar Project****FAIR TREATMENT: Federated Analytics and Artificial Intelligence Research across Trusted Research Environments for Child and Adolescent Mental Health**

Negative aspects of a young person's life can lead to poor mental health. However, services are stretched so often intervene late. It is possible to spot patterns showing where professional help is needed early, but this is difficult as the information needed is secured in different places – for example, across health, education and social care records.

Predictive models aren't accurate enough: there are difficulties linking different types of data together, potentially resulting in many important risk or resilience factors being missed. Furthermore, models built in one place may not be effective in others.

Led by researchers at the University of Cambridge, FAIR TREATMENT has combined new technologies to demonstrate linking cross-domain data and analysis across different trusted research environments (TREs) while preserving individual privacy. The team have consulted with patients, the public, organisations contributing data and legal/ethics experts to agree the best way to oversee data use.

initiatives such as the DARE UK Phase 1 TREEHOUSE Sprint Exemplar Project (see box on page 50). The architecture should fully support the Five Safes model with appropriate controls, and should be containerised to allow for deployment within existing infrastructures as well as to be deployed onto public cloud. It should also include fully technical and process documentation to allow for consistent deployment and alignment with the requirements of accreditation.

This TRE reference architecture should then be used as the basis for developing a shared TRE capability ('pop-up TREs'). This will cover cases where multiple TREs need to temporarily aggregate their data into a single environment for linkage to enable access to specialist analytics or computational capabilities. The pop-up TREs – which would be built within an existing TRE but with data from multiple different TREs – would require new approaches to allow multisite governance so that data custodians could continue to exercise their responsibilities over the aggregated data and enable a shared approach to statistical disclosure control. Findings from a DARE UK Phase 1 Sprint Exemplar Project led by researchers at the Francis Crick Institute which is exploring cloud-based TRE federation (see box on page 57) will provide useful insights in this area.

These reference architectures will provide assets to support the scale-out of federation and ensure the barrier to participation is low for both new and existing

infrastructures. This work should be undertaken in collaboration with equivalent activities underway across the community to avoid the duplication of effort.

**Data fabric management and linkage service**

A core capability that will need to be defined during DARE UK Phase 2 will be a federated data fabric management service (see glossary) that covers the whole of the data pipeline, from provisioning from data collectors and data guardians through to deployment within the TRE network and onward where appropriate to archival. These services will need to encompass a wide range of data – quantitative and qualitative – from across the different research domains, as well as many different approaches to data governance from those typical within health, administrative and open data. Increasingly, this will also need to cope with internet of things and near real-time data (such as from wearables like smart watches), which may be different in both structure and rate of change compared with other existing datasets.

The data pipeline should leverage existing approaches to secure data management, transfer and sharing and integrate into more recent technologies like event streaming. This will require new methods to integrate traditional approaches with those that follow a publish/subscribe pattern (an interaction pattern that characterises the exchange of messages between publishing and subscribing clients)<sup>32</sup>, for example over [Kafka](#) or [MQTT](#) infrastructure.

<sup>32</sup> Jacobsen H. 2009. [Publish/Subscribe](#). Encyclopedia of Database Systems. Accessed 14.08.2022.

**Sprint Exemplar Project****FED-NET: Creating the blueprint for a federated network of next generation, cross-council trusted research environments**

Solving society's complex challenges requires experts working together, studying data collected for different purposes and from different sources. However, combining data is challenging: data governance is critical and there are technical challenges in combining data with different 'data languages'.

Led by researchers at University Hospitals Birmingham, working in collaboration with teams from the University of Birmingham, the University of Nottingham and Nottingham University Hospitals, FED-NET builds on the research team's success in setting up and running PIONEER, the HDR UK data hub for

acute care. Working with patients, the public, analysts and clinicians, the team have co-designed a secure way to combine sensitive health data with other data, working across five NHS hospitals.

FED-NET has scaled existing trusted research environments (TREs) using 'federated analytics', where the data stays put and the analysis moves. It has tested how different data languages can be translated into a common standard using a study of asthma, and has tested governance solutions through workshops with members of the public and experts.

Linkage capability will be a major aspect of the data fabric. This will need to support all data modalities and different approaches to linkage, from manual curation and automated linkage on 'well-known' common data elements, such as NHS number or UPRN (Unique Property Reference Number), through to probabilistic linkage (a method which makes explicit use of probabilities for deciding when a given pair of records is a match or not<sup>33</sup>). Linking de-identified datasets from different domains was identified as a challenge by some of the Phase 1 Sprint Exemplar Projects. This is not easily solved, but could be addressed through a third-party

linkage service coupled with policy-based access control services, such as from [Privitar](#). This would allow different research domains to use common, tokenised identification to allow for linkage whilst ensuring that the privacy of both the initial and linked data is preserved. The successful development of these capabilities will be central to enabling efficient cross-domain research. This service could also be the basis for later support for safe return (see above) where this is legally and ethically appropriate.

<sup>33</sup> Eurostat 2019. [Probabilistic record linkage](#). CROS. Accessed 15.08.2022.



The final area that will be important will be the integration into the fabric of privacy enhancing technologies to augment the security provided in TREs. There are learnings from some of the DARE UK Phase 1 Sprint Exemplar Projects in this area, though probably not sufficient to establish a programme of work for Phase 2. A further study will therefore be required across the UKRI research community to understand this area, and in particular the linkage requirements for cross-domain research.

**Threat modelling**

In establishing a federated network of TREs utilising open API libraries, and the support for services such as 'pop-

*“ We have recently had dilemmas about sharing data with other TREs and struggled to find a standardised system of ensuring their TRE at least met the standards of our TRE. ”*

*Workshop participant*

**Sprint Exemplar Project****Creating a federated, cloud-based trusted research environment to facilitate collaborative research between existing institutions**

A significant practical barrier to research collaboration is the fact that the needs of every team are subtly different. Each researcher has different expertise, access to different local digital infrastructure and different ethics and governance arrangements which need to be adhered to.

This project, led by researchers at the Francis Crick Institute, has documented use cases covering the main barriers faced by sensitive data research collaborations, and demonstrated how they can be met using cloud-based data technologies (technologies accessible via a secure internet connection rather than locally).

The project team's vision for the next generation of trusted research environments (TREs) is a service which works for the majority of researchers, the majority of the time. The project demonstrates a move away from previous concepts of independent, inflexible research environments towards an infrastructure in which existing TREs can work better together.

up' TREs, it will be important to undertake structured [threat modelling](#) to understand the new vulnerabilities, impacts and appropriate mitigations. The need for this was emphasised in feedback received on the draft recommendations. This work will need the active support of subject matter experts from across different disciplines and would benefit from the active engagement of the [National Cyber Security Centre](#).

It was also highlighted by some stakeholders that there is a significant difference between data security in theory and in production, with the requirements for handling data securely within TREs being very different to those required when data is published. Further work is required to understand how TREs and other privacy enhancing technologies (PETs) – such as secure multiparty computation, homomorphic encryption, secure enclaves and synthetic data – can be effective combined in production and at scale as part of a holistic approach to addressing potential security threats. This is in addition to providing a secure approach to implementing the Five Safes across a network of TREs.

**Provision of a centralised sandpit environment(s)**

In response to input on the provision of a sandpit environment where researchers could explore potential cross-domain use cases using synthetic and open data, we recommend an exploratory project that uses existing, open environmental datasets with synthetic health datasets. This

would allow linkage on, for example, UPRN to allow the utility of such environments.

Longer term consideration could be given to a centrally operated environment with community donation of open and synthetic datasets with light touch access control to support research, and even potentially public and citizen scientists.

**Business continuity and disaster recovery**

A perceived shortfall in business continuity and disaster recovery strategy for some infrastructures was raised by some stakeholders. However, there were significantly divergent views on the importance of investment in this area, with some viewing this as one of the most urgent and critical needs, whilst others felt it would be wasted investment based on the level of risk and other options for mitigation. This indicates that further study and the development of a strategy is required before major investment is undertaken in this area.

As the dependency of UK research increasingly moves to rely on TREs and the continuous availability of a federated network of capabilities supporting those TREs, this needs to be addressed through a clear business continuity and disaster recovery strategy. It is therefore recommended to include pilot projects in the DARE UK Phase 2 programme to help determine the production deployment models for Phase 3. It is also likely that a sustainable approach to business

continuity and disaster recovery will be at least partially dependent on a move to make greater use of public cloud capability and a change in the investment model to focus on the need for such capability.

It will be important to establish proportionate expectations for TREs and these will differ across use cases and research communities. A starting point will be to have clear service level agreements (SLAs) and metrics based on recovery time objective (RTO – the maximum time under which a failed workload must be recovered) and recovery point objective (RPO – the maximum amount of data that an organisation can afford to lose)<sup>34</sup> expectations.

Some environments have already implemented high availability (HA) support (so systems can operate continuously at a high level without intervention<sup>35</sup>), and this is adequate where the loss of an environment is not critical – for example, where reprovisioning would recover the capability through failover. However, for large-scale, national TREs, there needs to be consideration of how to recover from a site level failure or critical loss – for example, through a ransomware attack. This needs to include processes to support business continuity; RPO/RTO/SLA targets; technical implementation; and testing strategy.

The technical support for disaster recovery will differ depending on the environment and the criticality of the

use. It may also be appropriate to design HA into the TRE reference architectures described above. As these architectures will be based on open cloud technologies, these will extend easily to provide for appropriate HA capability for less critical environments.

The move to a federated network of TREs should provide an infrastructure that would allow for the replication and failover of capability between sites. This will, however, require collaboration around processes and governance. Use of a grid approach is likely to be more cost effective than implementing standby capability for each of the key environments.

It is recommended that in DARE UK Phase 2, two pilot projects are undertaken. The first should be a study into the risk scenarios and responses required for disaster planning, and the second a pilot to model this through replication between TRE sites.

### Sustainable investment model

Detailed discussion on moving to a sustainable investment model for infrastructure is covered in Chapter 9: Funding and incentives. However, it is worth noting here that many of the recommendations in this chapter are only viable with a shift to a more sustained model that is not dependent on the top slicing of grant funding supplemented by sporadic capital grants. The current approaches will not sustain

a progressive move to public cloud deployment through operational expenses, where multi-year contracts deliver very significant discounts, nor a strategic approach, for example, to business continuity and disaster recovery.

### Flexible access to large-scale compute

One key need identified by stakeholders during DARE UK Phase 1 is intermittent access to large-scale compute. The current provisioning results in delays to data access and, subsequently, delays to research. This has been identified as sufficiently severe that it causes some researchers to avoid areas of work where this could be problematic. This requirement needs further clarification, but is likely to include access to clusters of both conventional CPUs (central processing units) and GPUs (graphics processing units) / IPU (intelligence processing units) on a shared basis but only perhaps for short periods. Such capability does exist within the UK, for example with access to the [ARCHER](#) and [ARCHER2](#) national supercomputing services, so this may be a funding rather than technical requirement.

There are clearly several options to consider for solving this need; however, it may be appropriate to have some of this capability provisioned using cloud resources via an on-demand model where the need is not for traditional large-scale compute which is likely to remain on-premises in the

<sup>34</sup> Guglielmi P. 2019. [Understanding RPO and RTO](#). Rubrik. Accessed 15.08.2022.

<sup>35</sup> Cisco. [What Is High Availability?](#) Accessed 15.08.2022.

near to medium term. Most of the major cloud providers support a ‘spot-market’ for compute instances, and for use cases requiring short usage (under 24 hours) of large-scale compute, this would therefore be an effective solution. For those modelling use cases requiring repeated intermittent access over a longer period, collaboration with the facilities provided through EPSRC and STFC might be more appropriate and should be investigated. Both approaches should be evaluated further in DARE UK Phase 2.

Any work in this area should be delivered such that it can be accessed flexibly from across the network of TREs and not require duplicative investment, and should be delivered in collaboration with [UKRI programmes focused on future large-scale compute](#). Flexible access should extend not just to HPC/HTC capability but to providing efficient and cost-effective access to specialist compute, such as GPUs and IPU.

### Next generation statistical disclosure control

There are already significant issues with staffing resources to support statistical disclosure control (safe outputs). This staffing issue is acting as a barrier to scaling up the use of TREs for research, and work is therefore needed to address this key area alongside complimentary activity by HDR UK, NHS Digital, ESRC and ONS. This area also needs to align with international activity and learnings from other sectors, including the finance and banking communities. The

recruitment and training actions that could help to address this skills shortage are covered in Chapter 8: Capability and capacity. This section briefly outlines the technology and supporting governance-led approaches that could be used to supplement staff resources.

Feedback was received that the disclosure of trained machine learning/artificial intelligence models from TREs is particularly challenging, especially as this is often using multi-modal data such as imaging. The DARE UK [GRAIMatter Sprint Exemplar Project](#) (see box to the right) has provided a base for further developments in this area.

It is recommended that a three-stage approach is taken to establishing a statistical disclosure control framework:

**Stage 1:** Establish a proportionate risk model for reviewing disclosure. Not all projects require the same level of review. The DARE UK [PRiAM Sprint Exemplar Project](#) (see box on page 44) is already developing a model based on past research and this will be made available as an open-source project which could form the basis for this activity.

**Stage 2:** Where possible, automate the review of outputs to support and focus the use of skilled personnel on the areas of most significant risk. This should explore the extensive existing work that has been undertaken in the area of developing tools for automations, including those developed by PRiAM and [Eurostat](#). This would need to include checking

### Sprint Exemplar Project



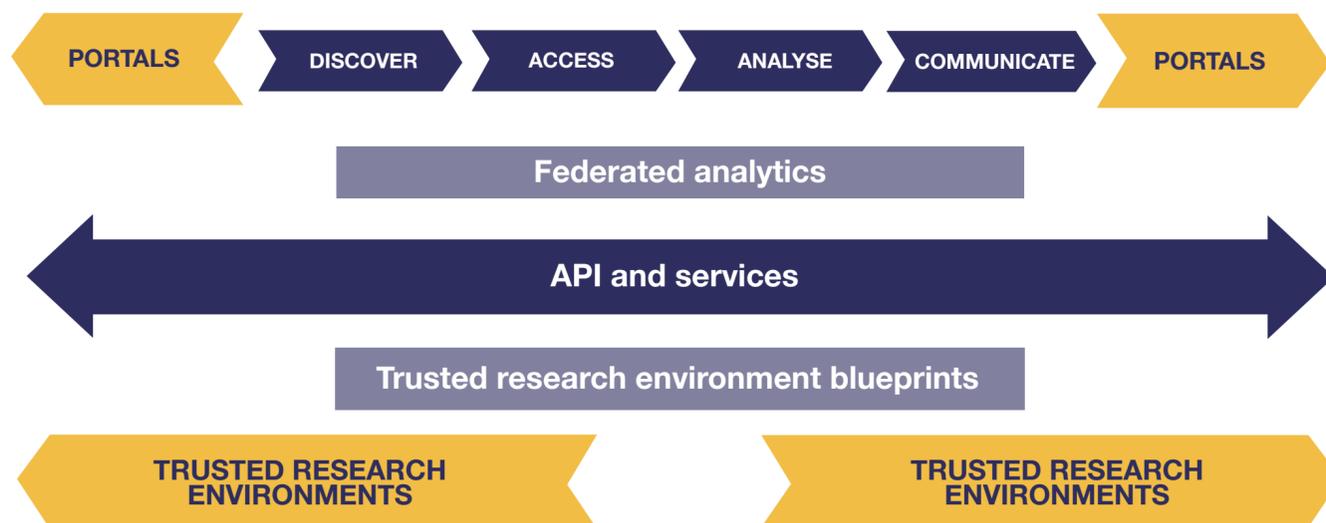
## GRAIMatter: Guidelines and Resources for Artificial Intelligence Model Access from Trusted Research Environments

Trusted research environments (TREs) have historically supported only traditional statistical data analysis, and there is an increasing need to also facilitate the training of artificial intelligence (AI) models. AI models have many valuable applications, such as spotting human errors, helping with repetitive tasks and supporting clinical decision making. The trained models then need to be exported from TREs for use.

The size and complexity of AI models presents significant challenges for the TRE output checking process. Models may be susceptible to external hacking, with more potential to lead to re-identification than conventional statistical methods.

Led by researchers at the University of Dundee and with input from public representatives, GRAIMatter has assessed a range of tools and methods to support TREs to assess output from AI methods for potentially identifiable information, investigating the legal and ethical implications and controls and producing a set of guidelines to support TREs with export controls of AI algorithms.

Figure 1: High level environment for deployment



of more conventional outputs as well as generated machine learning/artificial intelligence models. Development in the area of automation should draw on the expertise of those working in this field, such as the [Safe Data Access Professionals group](#).

**Stage 3:** Extend automation to allow the coordination of statistical disclosure control across a federated network to allow different organisations to collaborate on controlling the disclosure of outputs. There are examples of existing tools [on the GitHub platform](#).

This is an urgent area of need and should be prioritised for DARE UK Phase 2 investment. The work will require detailed public scrutiny to ensure the approaches are technically rigorous and proportionate, and meet the expectations of

the public. Any technical approach in this area will need to be delivered alongside appropriate information governance approaches, which may be outside the scope of DARE UK.

### Preparing for production deployment

Many of the high-level services needed to establish a federated network of TREs will have dependencies on a set of services to support orchestration, secure data transfer, high availability and network access to compute and storage (see Figure 1). The DARE UK Phase 2 work programme should assemble these services in collaboration with the wider technology community in the UK and internationally, progressively deploying them as consistent, open API libraries and reusable containerised services. Whenever possible, these services should be assembled from

existing open-source projects, of which there are numerous examples. Following the principles already established by, amongst many others, the [Health Data Research Innovation Gateway](#) and the [OpenSafely](#) project, all deliverables should also be made available as open-source, with permissive licencing through an established framework.

One of the risks of assembling these core low level services – and the high-level services discussed above – from existing capability, will be a lack of consistency. Therefore, investment will be needed to ensure consistency of design across the library of services and APIs, including documentation and samples. This work will also need to ensure there is a licencing model that is self-consistent across the services and supports the envisaged use cases and enables reuse.

Feedback from stakeholders during DARE UK Phase 1 was that there would be a need to have a funded environment for these services with a supporting operational structure, including a helpdesk. In DARE UK Phase 2, these services would then need to be progressively deployed into an operational cloud environment to provide a proof-of-concept minimal viable product that can be transitioned to being a production environment in Phase 3. During Phase 2, a proposal with business case should be developed for the establishment of a sustainable operating environment.

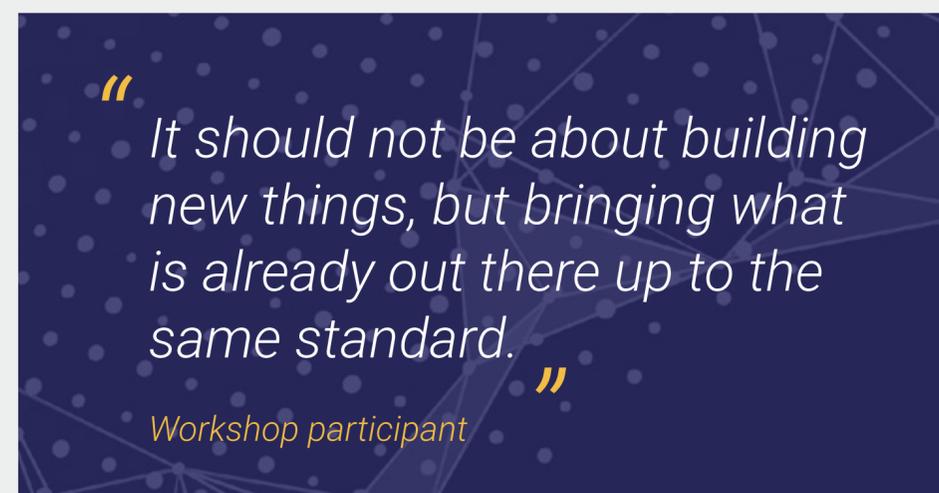
The successful delivery of the core federation services outlined in this chapter will require significant planning, programme management and collaboration across the UK. The challenges of this delivery should not be underestimated and therefore prioritisation will be important to deliver on those areas of highest value, rather than attempting to deliver all recommendations in parallel. A long-term and sustained funding approach will be needed to deliver, enhance and maintain the services described in this chapter.

## Driver projects

It will be essential for future phases of the DARE UK programme to ensure that its work is based on [design thinking principles](#) (empathise, define, ideate, prototype and test) and guided by the requirements of the research communities working with cross-domain sensitive data. This will need to include researchers from across all UKRI council domains, as well as academia, the public sector and the private sector to ensure that new capabilities can be generalised across a wider range of data and use cases. In addition to the pilot and proof-of-concept work in DARE UK Phase 2, there needs to be a programme of driver projects that can actively participate in the co-design of the capability and validate its usefulness in support of research. These could be delivered as a competitive call with projects starting later in Phase 2 and continuing in Phase 3. These

projects could only commence once sufficient progress of the technical development outlined in this chapter has been delivered. It will not therefore be viable to concurrently deliver complex technical proof-of-concepts and driver projects in the proposed Phase 2 timeframe, so this needs to be scoped to provide continuity across Phase 2 and 3, which will require funding agreements that span later phases of the programme. If this does not prove possible, any call for driver projects will need to be delayed until phase 3 of the programme.

It should also be anticipated that the driver projects would identify further non-technical aspects of successful federation that would need to be addressed by the DARE UK programme and more broadly in the research community.



## Partnerships and collaboration

The requirements covered in this chapter overlap with many other programmes and initiatives. It is important therefore that all elements are delivered in collaboration with the organisations (both within the UK and internationally) delivering these initiatives, reusing existing technologies wherever possible and integrating with existing infrastructure.

The following partnerships will be key to the delivery of future Phases of the DARE UK programme:

**UK infrastructure providers** – to develop and operate a next generation TRE configurable to their requirements and compliant to national and international standards, best practice and capabilities.

**The private sector** – who will utilise the TRE network to develop tools and services and access high-value datasets to develop high-impact research that delivers public benefit. Also, to ensure that private sector know-how and open assets can be used to assemble core services.

**Data collectors and data guardians** – to ensure that data is available for research and can be linked and provisioned into the infrastructures, maintaining transparency and trustworthiness.

**International** – to demonstrate alignment and commitment to international standards, policies, processes, tools, frameworks, and infrastructure services, which will allow participation in national and international programmes. Researchers – from within academia and the public, third and private sectors. To prioritise requirements and run driver projects to validate service capability.

**The public** – to ensure all approaches to implementation meet public expectations and enhance trustworthiness, and that they can be communicated effectively to a non-specialist audience.

**Complimentary UKRI Digital Research Infrastructure projects** – to ensure consistency between the DARE UK programme and the [complimentary UKRI projects](#), including those related to data infrastructure; large-scale computing; skills and career pathways; and foundational tools, techniques and practices.

## FAIRness and levelling up the UK

Not only should core federation services deliver against the DARE UK aims outlined at the start of this chapter, they should also underpin the unique opportunity to provide more equitable access to data, storage and compute to enable research across the whole of the UK. The current investment in infrastructure in the UK, as shown in Figure 2 – which

maps the number of infrastructures per area as recorded on the [UKRI Infrastructure Portal](#) – is geographically uneven and siloed; this must be addressed if the UK is to maximise its cross-domain research capability. This work should recognise the twin needs to level-up investment in the UK and lay the foundations of moving to a net zero future.

Borrowing from the well-established [FAIR Principles](#) for data and metadata (see also Chapter 6: Data and discovery), these can also be applied with little modification to infrastructure:

**Findable** – building from the UKRI Infrastructure Portal, there should be capability to understand the infrastructure across the UK and the availability of data for research within those infrastructures.

**Accessible** – there should be transparent processes for access to infrastructure and for the use of DARE UK provisioned service across the federation of TREs, supported by common identity management services and accreditation processes.

**Interoperable** – TREs should agree a common framework for core federation services which will be delivered through community-led, open-source projects. These should be implemented to provide a federated network of TREs at both national and sub-national level.

**Reusable** – specialist services, such as access to on-demand, large-scale compute, should be available across the network, providing reuse to support more efficient and cost-effective provisioning.



Figure 2: Distribution of UKRI Infrastructure (source: [UKRI Infrastructure Portal](#))

## Recommendations

The following key recommendations are made for investment in DARE UK Phase 2 to support core federation services with delivery in collaboration with the wider community and existing initiatives from both the UK and internationally:

### 1 Develop reference architectures for TREs.

Develop a **reference architecture and open implementations for a 'TRE in a box'** using open-source technologies suitable for deployment on-premise or on public cloud, to accelerate the ability of existing infrastructure to move to a hybrid cloud model.

Develop a **reference architecture and open implementation for a 'Pop-Up TRE'** that can be deployed within existing TRE environments and alongside the TRE in a box architecture to support secure transient analysis of data from multiple TREs.

Investigate approaches to integrate the TRE network with future **large-scale compute provisioning**.

Investigate options to provide an early proof-of-concept for a **'sandpit' environment** for open and synthetic data.

Develop a **threat model** to address changes anticipated through the DARE UK programme.

### 2 Assemble an API library to support core federation services.

Design and assemble an **open reference API library** to support core federation services, building on existing open-source projects.

Deploy the federation API library as a proof-of-concept with a **driver use case** across three TREs from different research domains.

Develop a proof-of-concept for a cloud-native implementation of a **portable analytics workspace**.

Conduct a study to identify the requirements for a **cross-domain data management and linkage service**.

Develop a proposal with business case for the establishment of a **sustainable operating environment** for these services and APIs.

### 3 Run a competitive call for driver projects to utilise the new infrastructure services and validate that they are fit for purpose.

Pilot **cross-council use cases** to validate the capabilities delivered in core federation services Recommendation 3.

Identify use cases to act as **driver projects** to validate the progressive rollout of production deployment in DARE UK Phase 3.

### 4 Establish an approach to business continuity and disaster recovery.

Undertake a **study** to establish the business continuity and disaster recovery requirements for a production network of TREs.

Pilot a **network failover capability** to support disaster recovery requirements.

# 8 / Capability and capacity

## Context

Data research is underpinned by those providing data preparation, curation, linkage and analysis, and by those developing and supporting digital research infrastructure. This chapter addresses the challenges in the recruitment and retention of staff, particularly in the public sector. It also looks at areas where there may be significant opportunity for change that would support a more efficient use of staff and skills, such as output checking.

Whilst DARE UK Phase 1 has not focused extensively on the area of capability and capacity, as this is subject to other areas of investment by the UKRI Digital Research Infrastructure programme, it is important to consider as it was a key area of concern from many stakeholders. Some of the potential solutions may also be a mix of recruitment, training and technology and are therefore in scope for the next phase of the of the UKRI Digital Research Infrastructure programme. This is particularly important in the current climate in which there is a significant shortage of skills and recruitment in the public and third sectors.

There was strong feedback that the focus on funding is often on hardware and technical capabilities, without strong enough consideration for the training and support of infrastructure development and management staff and on building governance systems around infrastructure to support, maintain, improve and share it more effectively. During DARE UK Phase 1, several key areas of skills shortage were identified, including:

- **Data scientists and/or analysts** to support research projects throughout the research project lifecycle. This was highlighted by some groups as their most serious current exposure.
- **Digital research infrastructure operational staff**, especially with skills in modern cloud computing technologies.
- **Information governance specialists** to support data access management and ethics approvals, particularly where expertise is needed to provide guidance and policy development for cross-domain research.

- **Cybersecurity specialists** to support the development of infrastructure as well as to advise appropriate security engineering and privacy enhancing technologies.
- **Data scientists and/or engineers** who support projects with TREs by providing, for example, data preparation, curation, linkage and metadata management.
- **Output checkers** to provide skilled analysis of research results to ensure that safe output requirements are met.

The challenges for each skill area appear to be common; how to recruit, train and retain staff. In addition, given the scale of the challenges, it is appropriate to consider whether there are technology approaches that allow us to reduce the requirements on staffing in some areas (for example, accreditation, data provisioning and output checking) to ensure we make the best use of staff and ensure they are involved in the highest skilled and most rewarding work with support from validated automation tooling.

It was felt by many stakeholders that retention was at least, if not more, of a challenge than recruitment. Particular barriers identified included poorly defined career pathways with a lack of opportunity to progress; technical roles being under-valued; and a reluctance from some organisations to invest in training and development. There was also strong support for the use of internships, secondment between organisations (including between the public and private sectors) to share skills and best practices, and expansion of apprenticeships, especially at Level 6 and Level 7<sup>36</sup>.

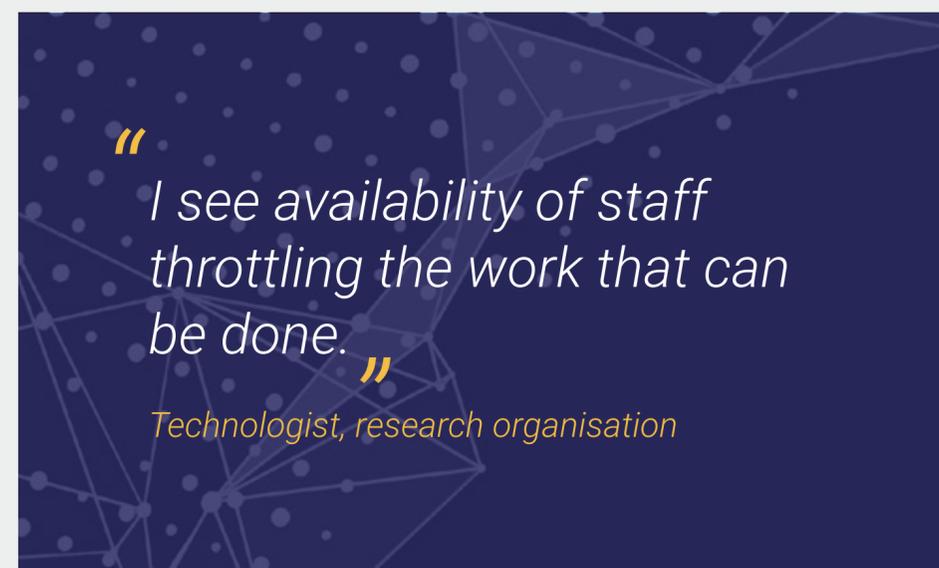
## Existing challenges and opportunities

DARE UK Phase 1 has identified significant capability and capacity challenges, and it should be anticipated that the skills requirements will evolve over the 2022-2026 period. This will include advances in AI that will require reskilling of researchers and data scientists, as well as technology advances in many areas – for example, in quantum computing, novel approaches to privacy engineering, federation/virtualisation, and the enhanced use of process automation.

There are several key challenges for building a sustainable pool of skilled staff to provide data science capacity

and develop and support world class digital research infrastructure for the UK. The public-funded research sector is in competition with the private sector, which has the advantage of more established career pathways, higher salaries and better job security (fixed terms contracts were identified as a major risk factor), and often also benefits from a greater general awareness of the roles available. This can, however, be countered with a focus on:

- improving the **visibility of roles** and the breadth and impact of the work undertaken;
- a focus on **excellence in training and retraining**, especially to attract a diverse and inclusive workforce and not only at an early career point;



- **clearer career pathways** that recognise and reward professional and technical skills; and
- a more **inclusive culture** that values technical roles alongside traditional academic roles.

The question of how to recruit successfully was raised in several discussions with stakeholders during DARE UK Phase 1. Central recruitment, use of secondments and approaches to making roles more widely known and attractive were all raised. Activity specific to recruitment is likely to be out of scope for the DARE UK programme, though critical to its success. It is clear we are in an exceptionally challenging recruitment market for technical roles in the public sector, and it is important to consider all three factors of pay, purpose and culture. The most difficult area is pay, but even here improvements are possible, and the areas of purpose and culture can be significantly addressed with focused activity.

However, as a contribution to the wider work in this area, a few key areas of feedback are outlined below.

**Public sector salaries** were unsurprisingly seen as a major challenge. The view was that in the past the additional benefits of public sector pension schemes, flexible

<sup>36</sup> Level 6 is an apprenticeship at Bachelor’s degree level; Level 7 is at Master’s degree level.

working and perceived less intense environments are no longer significant and so no longer offset the salary gap. Several stakeholders also expressed the view that the use of fixed term contracts further detracted candidates from considering public sector roles, particularly for roles where permanent positions are the norm in the private sector, such as in software development. It is also clear that some organisations, such as [OpenSafely](#), have shown greater commitment to competitive pay and that this is possible.

Several stakeholders shared success stories about the **recruitment of mid and later career staff from other sectors**. This included those returning from career breaks, as well as staff transitioning from successful careers in the private sector. Both groups have the potential to bring outstanding skills and life experiences, but will need support to retrain and are unlikely to be motivated to do so through formal postgraduate courses. Professional development approaches will be critical to success.

There was also concern that some organisations are reluctant to recruit more **junior members of staff** as there is significant pressure on staff, and organisations are reluctant to invest in skills development. This also results in a related concern of succession planning, as there is no established internal pipeline for progress into more senior roles.

Both HDR UK and ADR UK, as well as their partners institutions, have had a strong focus on **improving the diversity of the workforce** across researchers, data scientists and infrastructure engineering roles. This has been particularly successful with the [10,000 Black Interns](#) internship programme. It is recommended that this approach is further enhanced alongside the DARE UK programme and the later phases of the programme. This will bring benefits of enhanced recruitment and more diverse role models and will help focus on ensuring aspects such as diversity of data are kept at the forefront of the research agenda. We should also explore novel recruitment approaches, such as CV-less sifting as [piloted by HDR UK](#), which has shown to result in more equitable recruitment and therefore a more diverse workforce.

Stakeholder views on **internships** raised some interesting perspectives. Some organisations have concerns about the impact of supervision on already pressured staff, though others had the opposite view and had experienced positive benefits for the active use of interns. One novel proposal was whether it might be possible for interns to be funded (by other organisations) to work in key infrastructure groups, to then be recruited to the funding organisation to expend skills and share best practice following graduation. There was also interest in a coordinated approach to expanding the availability of **Level 6 / Level 7**

**apprenticeships** with more of a focus on data science. The view was that this needed central coordination and some level of seed funding.

There was a strong view that **roles are often poorly marketed**, with over-specific role descriptions and skills requirements, inconsistent role naming and excessive qualification levels (for example, requiring a PhD for output checkers) to meet organisational banding requirements. These practices significantly hinder recruitment, especially from outside of academia. The current culture appears oriented on recruiting against very defined immediate skills, rather than a more flexible approach that could be based on aptitude and investing in development and upskilling staff. Related to this was a view that much more could be done to support a coordinated approach to schools outreach to highlight the ‘backroom’ roles in research. Central coordination would help with the creation of programmes and the development of materials which could then be deployed locally.

There was strong consensus among stakeholders of the need to improve the **flow of skills between academia and the private sector** more generally. More consistent role descriptions and career paths were seen as important, but there was also a widely expressed view that secondments could be beneficial to allow the exchange of staff, skills and best practices between sectors, provided the financial

arrangements could be structured appropriately and the cultural differences addressed. This has been demonstrated in work between the [DiscoverNOW Health Data Hub](#) and [AstraZeneca](#). The option of sharing internship programmes was also raised, perhaps providing interns with the opportunity to derive the benefits of work across the two different worlds.

One area explored in DARE UK Phase 1 discussions was whether a **centralised pool of resources for key shortage skills** would be helpful. This attracted conflicting opinions. There was interest in a centralised approach to some critical skills, such as cybersecurity and output checking, and possibly as a ‘bank’ of resources to bridge the gap during local recruitment. However, there was concern amongst some participants regarding how a centralised pool would work to ensure fair access and avoid draining local skills. Overall, this looks to be an area worth exploring, but perhaps with a targeted approach around specific, highly sought-after skills.

### Training and skills

Excellence in training and staff development can provide a key tool to attract and retain staff in research support roles. There is a need to support career development by providing a rich set of training opportunities for all roles. This training offering will need to continuously evolve to meet the

demands of data science and infrastructure development, including cloud technology and AI methodologies. It is also clear that for some research domains, there is an ongoing need for training that still supports work on physical sources of information, and a risk that training is too focused on digital requirements and is not holistic. There is also a need for better access to training in quantitative skills for those in traditionally qualitative research domains – such as the social sciences – particularly to support and encourage cross-domain research using linked data.

Formal training at undergraduate and postgraduate level is widely available, but continuous professional development opportunities less so. Learning development needs to focus on bite-sized training to upskill existing staff, and retraining for staff transitioning from other roles or career breaks. The key requirements identified were around cybersecurity, public involvement and engagement and output checking, though other skills areas were also raised.

There are several initiatives already in place which are beginning to address these and other training needs across the sector, including the [Hartree National Centre for Digital Innovation \(HNCDI\) Explain programme](#), the [HDR UK Futures platform](#) and the [Nuffield Foundation Q-Step Programme](#). There are also other similar initiatives and there is risk of these initiatives operating as silos. UKRI should look to

coordinate these efforts to provide a rich UKRI resource with incremental funding to deliver across all UKRI councils. This could include training to support the development and sustainability of infrastructure skills which are currently seen to be poorly addressed. Whilst TRE providers have ultimate responsibility for their staff development, a more centralised approach to training platforms and resources is likely to be more efficient. The private sector has already engaged in specific training support, for example around development for GPUs (graphics processing units), and further opportunity to engage the private sector could be productive. Successfully addressing this requirement will be critical to ensuring there is capacity within organisations to adopt the recommendations coming from the broader DARE UK programme, other UKRI Digital Research Infrastructure initiatives and even to sustaining current investments.

Discussions during DARE UK Phase 1 have identified opportunities for upskilling researchers across disciplines, especially in the technical aspects of research using cross-domain sensitive data. Examples of this are training researchers on how to code well for large-scale analysis and the fundamentals of good data management. In addition, there is an opportunity to raise overall understanding of security, governance and ethics associated with research using sensitive data. Output checking was also identified as an area of specific concerns for skills.

Some stakeholders also raised whether there was an opportunity here for a recognised professional qualification to improve recognition. It is possible that this will be addressed by the [Alliance for Data Science Professionals](#), which is a joint initiative between the BCS (British Computer Society), the Royal Statistical Society, the Alan Turing Institute and the National Physical Laboratory.

Short term (six to eight week) secondments were also raised as a potential way for organisations to share skills and best practice. This would need a structured programme to avoid high administrative overhead for these short engagements, which could be trialled in DARE UK Phase 2. In addition, the development of high fidelity linked synthetic data deployed within a federated TRE network would enhance training opportunities significantly. Such datasets would enable focused training for data scientists and upskilling researchers moving into cross-domain projects, linked in with a UKRI training platform.

There are opportunities for technology to augment work in several areas that would address some of the impacts from the skills gap. DARE UK Phase 2 would provide an opportunity to look at, for example, partial automation of output checking and policy driven access request management. There was wide input on the need to use automation and AI to augment or indeed replace some

of these manual activities entirely. This would then allow highly skilled staff to focus on the critical or highest risk areas. It is important to note that partial automation to support these processes will require more than just technical development. Some of the current processes and policy implementation will need adjustment to reflect the needs for cross-discipline research. It will also take time to develop, validate and establish trust in automated approaches, however this will be necessary to scale research to its full potential.

The DARE UK Phase 1 Sprint Exemplar Projects have already shown opportunities around governance and risk analysis that could support approaches to automation. This should be explored further in Phase 2 with a view to significant investment in Phase 3.

Developing clearly defined and valued career pathways would build resilience within the UK research and innovation structures. UKRI should also consider further investment in conference style events that bring together the technical community across the councils, building on the excellent experiences from events such as the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) [Digital Research Infrastructure Retreat](#).

## Recruitment and retention

A central issue for many institutions over recent years has been staff retention. It is inevitable that some staff – both those directly involved with data research and those engaged in the development and operations of infrastructure – will seek more financially rewarding opportunities in the private sector, so those roles within the public sector need to be made more attractive. Key to achieving this is the need to establish clear, long-term career pathways that value these highly skilled technical roles and are equivalent to research and managerial roles, building on the work of the [Society of Research Software Engineering](#) and others. To address retention, it is important that the public research domain focuses on differentiation from the private sector through clear, standardised career pathways and excellence in training, establishing a diverse and inclusive workforce and communicating the opportunities to engage in work that makes a real difference to the public good.

Currently, there is a stark contrast between how the academic research community promotes technical careers, be they infrastructure or data science related, and the equivalence in the private sector. It is the norm within the private sector to have professional and technical pathways that support progression to a very senior level, and this is a core aspect of ensuring highly skilled technical staff are both retained and motivated to continue to pursue technical

“  
Currently, net flow direction  
is from academia to industry  
– it would be good to make it  
as normal for people to move  
from industry to academia as  
it is to move within industry or  
within academia.”

Workshop participant

careers rather than feel the need to move out to managerial or executive roles. Many organisations in the private sector have implemented dual career paths that are porous to allow movement between the technical and managerial pathways at several points<sup>37</sup>. There is no reason this type of approach could not be adopted within the public sector research community and thus provide a more structured approach that values technical excellence and breaks from these roles being seen as little more than administrative support.

Those organisations that are actively recruiting at more junior level have also expressed concern around their ability to retain staff, with a common issue of attrition after about five years at the point at which staff are starting to take on leadership activity. This was seen as particularly acute for data scientists and in software development.

Another issue that has been identified as affecting retention is poor role definition, with highly skilled staffing being used to assist with activities which would be better addressed by additional project management capacity. Clear career paths would be helpful in clarifying the roles and help to maintain the motivation of highly skilled staff.

The good news is that there is already a lot of excellent work underway to address capability and capacity across the UK's data research infrastructure. It will therefore be important for future phases of the DARE UK programme to positively support the work in the wider UKRI Digital Research Infrastructure programme around career development and training, and to engage with initiatives such as the Society of Research Software Engineering, the [Software Sustainability Institute](#) and the [Technicians Commitment](#). The RSE has proven very effective in improving the visibility of technical staff, whereas professional bodies such as the British Computer Society were seen as having little or no impact. Few of those engaged with during DARE

UK Phase 1 were aware of the Technicians Commitment, but there was broad support for its objectives. Collaboration with these initiatives was seen as critical to the wider UKRI Digital Research Infrastructure initiative and to encouraging research organisations to actively engage with them and embed these initiatives into their own learning and development pathways. It will also be important to ensure that any work initiated under the DARE UK programme aligns with existing UKRI initiatives, such as the [TALENT Commission](#).

<sup>37</sup> Dzulkifli E. 2018. [Encouraging Innovation: Dual Ladders & Self-Empowerment](#). Medium. Accessed 10.08.2022.

## Recommendations

It is expected that most of the following recommendations will be delivered outside of the DARE UK programme, though this should happen in parallel to DARE UK Phases 2 and 3 to ensure there is sufficient skilled capacity for the core DARE UK and the other UKRI Digital Research Infrastructure programmes to be successfully deployed. Where appropriate, the recommendation is flagged below as delivered through ‘**UKRI**’ – probably by another component of the overall DRI programme – or as ‘**DARE UK**’ where it should be considered for delivery directly by the programme.

### 1 Establish clear technical career pathways in data research infrastructure that can be adopted across the UKRI research domains.

Work with the Society of Research Software Engineering and Technicians Commitment initiatives to establish **agreed career pathways** across the UKRI research domains (UKRI).

Investigate and report on other **barriers** that exist for those pursuing careers in support of data research (UKRI).

### 2 Improve recruitment pathways for technical roles in data research infrastructure.

Establish a **recruitment taskforce** to explore effective recruitment options, including alignment with diversity and inclusivity work already underway across HDR UK, ADR UK and elsewhere. This taskforce could also examine approaches to providing exemplary approaches to attract those making career changes (UKRI).

**Pilot secondments and exchanges** with the private sector to bring in shortage skills. This could be used to supplement the DARE UK Phase 2 Delivery Team (DARE UK).

Embed participation in the **Black Internship Programme** in future activity, including interns as members of the future DARE UK Delivery Team (DARE UK and across UKRI).

Consider investment options for funding a **central pool of high skills resources** with potential to pilot for one of cybersecurity or output checking in DARE UK Phase 3 (DARE UK).

Fund centralised development of a **schools outreach programme** and supporting material. This would not be for centralised delivery, rather to establish a reusable structure and material for wider adoption (DARE UK).

### 3 Improve the availability of resources and training for career development in data research infrastructure.

Establish a **pan-UKRI virtual learning environment** for high quality modular training delivery, possibly via extending the [HDR UK Futures Platform](#), the focus of which should be to support the development of the skills identified above (DARE UK).

Develop a rich set of high fidelity **synthetic linked datasets** to support training in cross-disciplinary data research (UKRI).

Establish an **annual UKRI Technical Retreat** using the approach and learnings from the N8 CIR Digital Research Infrastructure Retreat (UKRI).

4

**Use automation to ensure data research infrastructure services are reliably secure, auditable and reproducible.**

Use the outputs from the DARE UK Phase 1 Sprint Exemplar Projects to create **open-source projects** on risk assessment and information governance processes to support progressive automation of research user journeys (DARE UK).

Pilot the delivery of **automation** to augment **output checking** (DARE UK).

Pilot the delivery of **automation** to support **policy-driven access request management** (DARE UK).



# 9 / Funding and incentives

## Context

The traditional investigator-based grant model that is the current structure through which research is funded is not efficient for supporting the increasing requirement for infrastructure and related services that have become essential to a large proportion (arguably almost all) of research work today.

The infrastructure and related services – be it hardware, software, or human resource – that enable data research require stable, continuous funding allocation cycles and purpose-built grant structures that are designed with the complex requirements inherent within the digital infrastructure ecosystem in mind. To nurture a UK ecosystem that balances collaboration and competition, new funding and incentive structures for providing the infrastructure and related services need to be tested and designed. This is necessary not only as a fundamental cornerstone of modern research, but also to recognise and reward the contribution of these services as a foundational part of the delivery of data research for public benefit.

This chapter addresses the challenges for data research infrastructure funding linked to the current structures and cycles of funding. It will also look at opportunities for change that would support a more efficient and sustainable infrastructure for research using sensitive data. In addition, it will address the challenges around incentivising collaboration, particularly in a federated context, while counterbalancing that with the need to incentivise innovation through competition and broader engagement with the public.

Funding and incentives for data research infrastructure are a key focus area for the overall [UKRI Digital Research Infrastructure programme](#)'s strategic vision. It is crucial to consider that the nature of data research infrastructure is highly interwoven and is in fact a system of systems – as a result, there is no single solution to funding or incentives that will prove a solution to all challenges. Ultimately, a cohesive, considered blend of new funding structures, adjustments to existing funding structures and additional incentives will provide the necessary fiscal support needed to enable a shift to a more sustainable but, importantly, more productive ecosystem that will enable era-defining research on sensitive data.

## Existing challenges and opportunities

Based on the input received from the community to date during DARE UK Phase 1, there are several key challenges around funding and incentives that should be addressed.

First, there exist **limited dedicated, tailored funding allocations** for the operation, maintenance, and refresh of digital research infrastructure and related services – including the legal, contractual, and service level (if applicable) frameworks that would underpin such funding allocations. This includes capital grants for digital research infrastructure, which are sporadic and often awarded from within specific UKRI research council remits rather than with a cross-domain, holistic view of the landscape in mind. There are five interdependent areas of funding to consider in this context:

- 1) **Hardware environments** – these are the base layer ('bare metal') components – for example, CPUs (central processing units), GPUs (graphics processing units), network cables, RAM (random access memory), persistent storage,

servers, operating systems and so on – needed to operate a computer system. There is also a discussion to be had around the provision of infrastructure as a service (IaaS) – in which infrastructure is provided on an on-demand basis through public cloud providers – and how this is categorised from a funding point of view. For the purposes of this report, we will not delve deeply into this but acknowledge that it is a nuance that needs to be addressed.

2) **Software environments** – these encompass the layers of a computer environment running on the hardware environments and the digital tools which enable data research (for example, middleware, web servers, runtimes, applications and so on).

3) **Human resource** (software engineering) – the skills necessary to effectively operationalise the hardware and software environments; certainly a clear challenge within the UK research ecosystem.

4) **Human resource** (data curation) – the skills necessary to create, organise and maintain research-ready datasets that are then available at pace to answer research questions.

5) **Human resource** (information governance) – the skills needed to ensure the work being done in the hardware and software environments is in line with the relevant legislative and ethical oversights and, critically, to ensure the research is in the public benefit.

The final three areas are explored further in Chapter 8: Capability and capacity.

Second, the rules associated with the funding of digital research infrastructure often do not acknowledge the realities of maintaining digital infrastructure, leading to **large amounts of operational overhead** to effectively keep the lights on:

- Organisations are forced to ‘top slice’ (include specific budget lines within multiple grant applications) off existing new research grants to fund their operational expenditures. This distorts the grant approach, inhibits mid to long-term planning and is a factor in limiting the acquisition and retention of talent in this critical area (see Chapter 8: Capability and capacity). Additionally, it makes it difficult to understand holistically what the funding footprint for digital research infrastructure is at a level of detail that would enable more efficient and effective funding decisions.
- From a technology perspective, organisations are forced to maintain aging hardware components that are inefficient both in terms of modern hardware standards of performance but also from an energy usage perspective; in light of the strategic UKRI net zero aspirations, this will have to change.
- Organisations make trade-off decisions between sustaining resource (be that the hardware assets, software assets or

the human resource maintaining it), maintaining and further improving the quality of that resource, business continuity and disaster recovery, and innovation.

- It is not clear from a UK-wide perspective how much compute capacity is required both today and into the future, nor what the full spectrum of benefits are that can be derived from expanding or optimising that capacity.
- Business continuity and disaster recovery planning are not prioritised for national infrastructure that is critical for the UK data research landscape. As the criticality of data for driving policy and decisions that can improve people’s lives increases, so does the necessity to put in place prudent measures for protecting against failure risk. This need is covered in further detail in Chapter 7: Core federation services.
- There is an ever-increasing software maintenance burden linked to the constant creation of new, standalone methods and tools that is driven by the incentives linked to publishing compared to those incentives for maintenance post-publishing, especially for more foundational methods and tools with wide applicability. Highly specialised domain-specific software should not fall under this category.
- Structured collection, maintenance and curation of data is not always sufficiently incentivised nor formally recognised as having a critical impact on research outputs.

A tailored, fit-for-purpose and consistent approach to how digital research infrastructure is funded could begin to address these challenges – critically, any funding approach must be driven by a long-term strategy including an appropriate performance measurement framework for eligible (or not) digital research infrastructures. However, establishing transparency across the UKRI spectrum for both where and how funding is currently allocated to digital research infrastructure is a critical first step to inform the best approach to making the necessary adjustments.

Current average time horizons for funding – approximately around the 12-month range based on input received – also do not always provide the **stability and long-term perspective** required to effectively enable sensitive data research:

- Standard funding timeframes for data research are short compared to data access processes and approvals, often leading to research questions aligned to what data resources are more readily available, rather than fostering the ‘right’ questions (see Chapter 6: Data and discovery).
- Effectively operating, maintaining and refreshing foundational hardware and software environments (including retaining the human capital with the right skills) requires stable planning horizons that in most cases extend well beyond a single year time horizon.



- Current funding timeframes heavily favour those applications with existing data access agreements and do not consider the challenges of making an effective application for funding without a degree of confidence regarding both whether a data access approval will be successful, and when that feedback would be received.
- While there are examples of digital research infrastructure programmes and projects that have managed to maintain continuity beyond a single funding cycle (for example, [ARCHER2](#), the [Joint Academic Data Science Endeavour \(JADE\)](#) and the [Distributed Research Utilising Advanced Computing \(DiRAC\)](#) consortium) this has been in spite of rather than due to current UK funding models.

In a federated environment, where conceivably infrastructure and related services are provided by the research ecosystem for the research ecosystem, the **UK-wide operating model(s) for such a federated ecosystem has not been developed and established**. Some components to consider here are:

- **Financing** – how to structure and allocate the funding that will provide the initial and subsequent investment to drive the establishment of such a model(s).
- **Cost recovery** – at the appropriate stage of maturity, how the providers of federated infrastructure and related services will recover (or cover) costs in a sustainable way.
- **Service levels** – what minimum levels of service to the data research community are required from the providers of federated infrastructure and related services.

It has also not yet been established how to cohesively **integrate private-led cloud compute capability** more seamlessly into the fabric of the sensitive data research infrastructure landscape, and the associated costs are not widely understood:

- Cloud compute capability most often requires multi-year contracting to secure favourable pricing. However, this does not always fit with grant timelines and the research itself (often time-limited and project based).

- There is no comprehensive, consistent overview of what cloud providers can offer, the constraints inherent in that offering, and clear guidance or frameworks to support decisions around when the cloud model is best utilised and when it is not required.
- There is no consistent definition of cloud compute within grant applications, particularly around the classification of such costs under operational or capital expenditure – this does not serve the researchers themselves nor does it lead to efficient spend of UKRI funding in many cases.
- There is a misconception that public cloud technology stacks alone can address many of the valid privacy and security concerns. Appropriately skilled people, procedures and processes in combination with the technology itself are essential to manage privacy protecting, secure and trustworthy research environments.

Additionally, there is a challenge in meeting the irregular demand from the sensitive data research community for **large-scale compute capacity** (high performance compute or high throughput compute) that needs to be addressed:

- Particularly in the domains of linked sensitive data, the challenge is how to leverage large-scale compute capacity while maintaining the security of the data itself, which has not traditionally been a consideration for large-scale compute environments.

- There is a need to establish how funding that will provide the initial and subsequent investment to drive the integration of large-scale compute can be structured and allocated in a way that is appropriate for research on sensitive data.
- At the appropriate stage of maturity, there is a question of how the providers of large-scale compute for research on sensitive data will recover (or cover) costs in a sustainable way. Alternatively, is the improved utilisation (assuming in principle this would be the case) of large-scale compute infrastructure considered an adequate return?
- In addition, what minimum levels of service to the sensitive data research community are required from the providers of large-scale compute?

Furthermore, a lack of sustained, dedicated **funding allocation for public engagement and involvement** – particularly for research on sensitive data about people – and coordinated guidance on how best to utilise those funds often results in inefficient spend and a lack of meaningful public involvement and engagement. This is covered in greater depth in Chapter 3: Demonstrating trustworthiness.

Finally, competition for funding can push researchers into **institutional silos** as opposed to the kind of cross-disciplinary collaboration that is critical in cross-domain research.

### Novel, tailored funding allocations

Addressing these challenges requires more than new methods of funding, but also novel approaches that optimise the utilisation of underlying digital infrastructure, build resilience within the data research ecosystem within the UK and enable cutting-edge, cross-domain research at scale and pace.

Core to the DARE UK programme is the concept of federation and the capability through federation to interoperate securely across a diverse landscape of data research infrastructures; details of our findings and recommendations related to federation can be found within Chapter 7: Core federation services. However, this effort requires an injection of seed funding to accelerate the evolution of the data research landscape towards a more federated model. This is particularly timely within the sensitive data research landscape in which a growing number of strategic, cross-domain, high priority research areas require a means of linking sensitive data securely without cost prohibitive data considerations.

It is important that this work is undertaken by those within the landscape with the necessary expertise, understanding and experience of the challenges that federation will present in the context of sensitive data research. The focus at this stage should be on enabling existing infrastructure providers

in sensitive data research with a proven track record to test and deliver the first federation elements across a selected number of such environments.

While this would address the need for seeding the technical innovation to kickstart the move towards a federated network of trusted research environments (TREs), there needs to be development of operating model(s) that provide a basis for determining the sustainability of such a federated network in tandem with this work. In theory, federation should deliver efficiency, resilience, and novel approaches to answering research questions. However, the implications for the UK balance sheet need to be understood and deemed worth the return.

### Business continuity and disaster recovery

There are mixed views on the criticality (and urgency) of the need for business continuity and disaster recovery as outlined in Chapter 7: Core federation services. Based on the outcomes of a risk scenarios and responses study, as well as a pilot model to replicate this between TRE sites, investigation into the cost implications of possible approaches extrapolated at UK scale is needed to determine the financial feasibility of such a model(s).

Ultimately, a trade-off between the risks of a site level failure or a critical loss and the resource requirements to guard

against such scenarios is needed to provide a foundation for decision-making by UKRI around the appropriate degree of resilience that should and could be implemented. Within this context was also discussion around the need to clearly define national, critical digital research infrastructure.

### Access to large-scale compute

There is an increasing demand for affordable access to large-scale compute capacity, driven through several growing knowledge domains, such as artificial intelligence and machine learning. While cloud-providers can certainly provide this effectively – at varying levels of scale as required and increasingly with the levels of security needed for research on sensitive data – this model becomes cost prohibitive in instances where longer-term access to large-scale compute is required.

Often, attractive pricing options for cloud compute capacity are coupled with long-term contractual commitments that in most cases do not align with the existing funding cycles for research grants. As such, there are two primary challenges to be addressed: there is an increasing need for short-term, on-demand, large-scale compute capacity; and there is a need to address those instances where long-term access to large-scale compute is needed but not possible due to absolute costs combined with existing research grant time boxing.

Regarding short-term access, most cloud providers support a ‘spot-market’ for compute instances that may prove sufficient in addressing this requirement though, this needs to be investigated in more detail. Considering longer-term access to large-scale compute may require investigation and collaboration into utilising the existing national facilities provided by the EPSRC and STFC, as this may prove more effective not only from a cost for research perspective, but also ensuring optimal return on investment through high utilisation of those facilities. The investment implications for both approaches need further investigation.

Feedback from stakeholders engaged with during DARE UK Phase 1 was clear, however, that the starting point for defining these requirements should be driven out of strong use cases around sensitive data that require such a level of compute capability.



### Incentives for data collectors and data guardians

There is, justifiably, wariness amongst data collectors and guardians around making sensitive data accessible for secondary research purposes. This is driven by their – often statutory – obligations to protect the data which they hold in their care and ensure any secondary use of that data is ethical and in line with legislative frameworks. Greater efforts – and there has been excellent work already in this regard by the likes of the UK Statistics Authority and ADR UK – are needed to support increased awareness of the existing legal framework around the secondary use of data for research (the [Digital Economy Act, 2017](#)) and how this compliments other legal frameworks providing guidance on the use of data (for example, the [2018 Health and Social Care Act](#)).

It must be acknowledged that producing high quality, research-ready data resources is not normally part of the core functions of data collectors or guardians. There are a few factors contributing to this, but fundamentally the purpose of collecting and safely storing the data in the first place and the resources allocated to that purpose are the primary driver for data collectors and guardians. Ultimately, as a bare minimum standard, covering the resource costs (through license fees and so on) realised by data collectors and guardians in making their data available for research should be encouraged and supported through funding of research applications.

Further, there is a need to investigate the funding requirements to enable greater support for data guardians – provisioned through sensitive data research infrastructure teams – with data curation (for example, data quality improvement) and analysis of linked data.

With reference to Chapter 6: Data and discovery, a key concern from stakeholders is the need for a more consistent approach to data archiving that supports research both within and across different UKRI research domains. While all research councils have individual approaches to data lifecycle management, stakeholders emphasised that a more consistent approach would be beneficial. They were clear this should be built off existing best practice – for example, ESRC have standard clauses in their grant awards for making data assets discoverable and, importantly, provide the infrastructure through the [UK Data Service](#) for doing so.

### Transparency, coordination and collaboration

The notion of stable, predictable funding for a national data research infrastructure is widely supported. The question is how to adjust the established legacy structures that exist, for good reason, around the awarding of grant funding to address the clear need for longer-term funding horizons, purpose-built grant awards and effective coordination across the UK.

The traditional research funding structures heavily favour new, novel work that captures the imagination of what can be discovered through research. As such, grant awards are largely aimed at funding new research applications rather than the underlying infrastructure and related services. Certainly, this has largely been successful and to some extent unnoticed to date if the costs for these underlying capabilities are built into research grant application budgets. This effectively means that these underlying capabilities are funded indirectly through new research grant awards.

However, this is no longer an efficient approach, primarily due to increasing absolute costs for the infrastructure and related services; inefficiency in disjointed funding; practicalities in managing the costs associated with the increasing size and scale of the data itself; and the need to manage the overall environmental ‘bill’ that this incurs. And finally, in the context of sensitive data research, there is a critical requirement to protect the privacy and security of sensitive data, both today and looking forward as new threats to the security of data develop alongside technological advancements.

It should be acknowledged that there are certain research domains with a greater need for the kind of capability that requires intensive capital investments. However, there is an increasing demand from all research domains, and certainly

in cross-domain research, for improved access to capability that can only be enabled through prioritised capital investments.

This speaks to a need for greater transparency, coordination and collaboration across the sensitive data research community to jointly steer and manage the national sensitive data (and beyond) footprint while extracting maximum value for each taxpayer pound spent:

- **Transparency** is critical as a starting point, especially in understanding the as-is picture which will provide the context for the directions of travel that will need to be taken to incrementally pivot towards the evolving to-be picture. It should be noted that there is a risk of analysis paralysis in this regard, and establishing transparency as an ongoing activity in tandem with working towards an evolving to-be picture is required.
- **Coordination:** due to the complexity and breadth of the sensitive data research landscape, coordination is crucial. Effective coordination across the landscape enables sensible agility in response to this complex, fast-paced environment. This is especially true considering the broad scope of activities, with a need to execute in an agile mode rather than via more traditional waterfall approaches.

- **Collaboration:** as such, an undertaking cannot be achieved in isolation, nor will a ‘top-down’ approach be effective in sustainably addressing the challenge. Further, it is evident that the existing landscape has both the legacy infrastructures and expertise in place to address future challenges. It is therefore necessary to convene these players around the goal of interoperability while providing

the necessary resources for them to define and deliver this interoperability ‘glue’ for the ecosystem – in a way that is open and competitive – as a driving force for innovation.

These three characteristics are especially important in the UK research and innovation ecosystem, where resources are limited and there is an increasing need for optimising efforts to deliver the most return on investment for taxpayer money.



## Recommendations

Based on the above, DARE UK Phase 1 recommends the following in the context of funding and incentivising a coordinated national data research infrastructure:

1

### Develop a new type of grant tailored to addressing the costs for maintaining cross-domain, national sensitive data research infrastructure.

Establish a comprehensive, rolling, periodically refreshed **overview of the sensitive data research infrastructure landscape** and related services across the UKRI research domains – this should form a subset of a broader view of the digital research infrastructure landscape.

Establish a comprehensive, rolling, periodically refreshed overview of the actual and projected UKRI funding – be it full or partial – of **operational costs** for national sensitive data (and beyond) research infrastructure and related services across the UKRI councils.

Establish a comprehensive, rolling, periodically refreshed overview of the active and projected UKRI funding – be it full or partial – of **capital investments** for national sensitive data (and beyond) research infrastructure and related services across the UKRI councils.

Design, develop, and implement a **new criteria of grant award** tailored for the funding of operational and capital expenses for sensitive data research infrastructure and related services in a federated ecosystem.

- Consider carefully how full economic costing applies and how the variety of sensitive data research infrastructures will impact funding parameters (for example, appropriate timeframes may differ).
- Define a matrix of complimentary funding requirements across both functional (for example, data management, reuse of software assets) and structural (for example, human resources, hardware resources) requirements – including a clear strategy for associating funding to the use of sensitive data research infrastructures.

- Define, in collaboration with the different research communities, the definition of ‘national sensitive data research infrastructure’ and ‘core, national research-ready datasets’ in the context of this new criteria of grant award – carefully considering factors such as whether there should be a tiered approach to such a definition; the balance between breadth and depth of data available for research within such a definition; and ensuring such a definition encourages healthy innovation while maintaining stability and consistency in the long-term.
- Develop an independent, competitive process(es) to allocate funding in such a grant category.
- Develop the legal, procurement and contractual frameworks that would be required to execute on such a grant category.
- Investigate and define minimum service levels for providers of national sensitive data research infrastructure that receive baseline operational funding with clear provision for different types and maturities of sensitive data research environments.

Consideration must be given to how those minimum service levels integrate with the core federation elements as outlined in Chapter 7: Core federation services. Including whether support for the curation of core, national research-ready datasets provisioned through sensitive data research infrastructures should form part of such minimum service levels to support data collectors and guardians to make their data available for research.

**2 Determine the funding requirements to establish the first phase of a federated national infrastructure for sensitive data research, with a focus on enabling federation across existing infrastructure and complimenting existing investments (with reference to Chapter 7: Core federation services).**

Make funding available to **investigate, test and evaluate approaches** to core federation services required across the ecosystem and that could be scaled in the mid- to long-term, namely:

- Federated identity management
- Enablement services to support federated analytics
- Metadata federation
- Infrastructure (compute, transfer, and storage) federation

**3 Investigate, test and prototype the operational model(s) for a federated national infrastructure for sensitive data research. Critically, ensure federation lessons and insights from those outside of the sensitive data space are considered.**

In tandem with the development of the core federation services outlined in Chapter 7, develop an **operating model(s)** around these services that could be considered for scaling in the mid-to-long term.

Based on the operating model(s) developed, determine the feasibility and comparative options for **cost recovery** across a federated infrastructure that would be sustainable over the long-term.

Determine whether federation (or components thereof) will deliver **additional value** for the data research ecosystem – either through cost efficiencies, additional capacity, or both – as a decision criterion for further scaling in the mid-to-long term.

**4 Investigate the cost implications for appropriate business continuity and disaster recovery requirements for a federated national infrastructure for sensitive data research (with reference to Chapter 7: Core federation services, Recommendation 4).**

Investigate and determine the **financial feasibility** of business continuity and disaster recovery scenarios and responses within a federated infrastructure.

Determine the funding requirements to **pilot** business continuity and disaster recovery scenarios between selected digital research infrastructure sites.

Based on the points above, develop an **options appraisal** of different approaches to addressing – if there is agreement on need – the business continuity and disaster recovery requirements.

**5 Investigate the scope and funding requirements for the integration of large-scale compute availability in a federated national infrastructure for sensitive data research.**

Investigate and define the **use cases** for large-scale compute requirements for sensitive data research (for example, short-term versus long-term access requirements).

Based on the use cases identified and prioritised, validate the **feasibility and initial investment(s)** needed to integrate large-scale compute capabilities into a federated infrastructure.

Investigate, together with the EPSRC and STFC, a **long-term access model** for large-scale compute capacity for sensitive data research in a federated ecosystem and how the costs should be considered within existing or new research grant structures.

Investigate the **on-demand model** for cloud compute capacity for sensitive data research in a federated ecosystem from a cost perspective and how these costs should be considered within existing or new research grant structures.

**6 Building upon existing best practice, improve the availability of all data produced through publicly funded grants for reuse and investigate the funding requirements for provisioning such archival capability (with reference to Chapter 6: Data and discovery).**

Understand and analyse current approaches across UKRI research councils, leveraging examples of best practice to develop a more **standardised approach** to how data assets are made discoverable, not only in each research council domain itself but also how this could be federated to support cross-domain discovery as well.

Investigate the funding requirements for provisioning an **archival capability** both within and across UKRI research council domains, as well as the business case that would underpin this.

**7 Support raising awareness amongst data collectors and data guardians regarding the legal framework around the use of data for research.**

Develop a **toolkit for data collectors and data guardians** regarding the legal gateways in place for making sensitive data accessible for research purposes.

Through mixed methods (for example, information campaigns, conferences), **proactively raise awareness** around the provisions and operations of the **legal gateways** in place for making sensitive data accessible for research purposes, to drive a more consistent understanding.

**8 Dedicate greater resource to supporting data collectors and data guardians to routinely make their data accessible for research in the public benefit.**

With reference to Recommendation 1, consider improving **baseline funding** to support the development of sustainable, research-ready national datasets from across domains and sectors through provision of skills and capacity from national sensitive data research infrastructures – building on the existing support provided by UKRI research councils to grant holders.

**Raise awareness** regarding the security processes in place to protect data from harm (particularly the Five Safes framework); evidence of public support for data research; and the policy benefits associated with making data accessible for linkage and research.

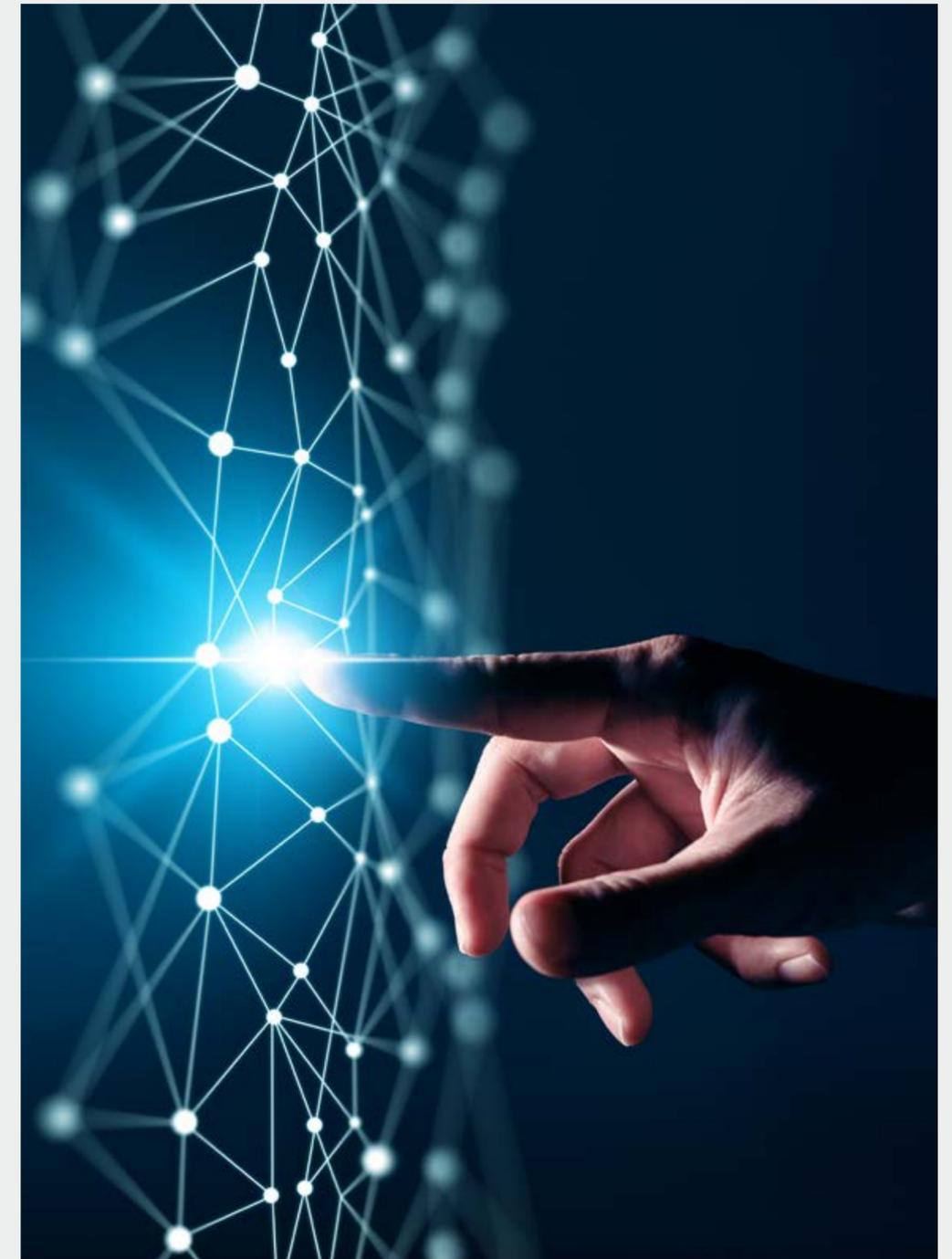
# 10 / Next steps

UKRI has confirmed further funding for the DARE UK programme with a total of £4.5 million from September 2022 to October 2023 as part of the [UKRI Digital Research Infrastructure \(DRI\)](#) programme. This additional funding comes as an extension of DARE UK Phase 1, to continue to be led by Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK).

UKRI embarked on the first phase of developing a national digital research infrastructure in the 2021-2022 financial year, with £17 million invested in a portfolio of interventions and pilot projects – one of which is the DARE UK programme. The DRI Programme has been allocated £129 million in the recent spending review, with an increasing profile over three years of £17 million, £42 million and £70 million. Due to the rising profile of funds, the DRI Programme will continue in a pilot phase for financial year 2022-2023. A more substantial portfolio of projects will be developed to commence in the financial year 2023-2024 – ‘Phase 2’ – following the establishment of the Digital Infrastructure Advisory Committee (DIAC).

This further funding into DARE UK Phase 1 will enable the DARE UK Delivery Team to begin to take forward some of the more immediate recommendations outlined in this report, as well as dedicate time to further scope and refine longer-term recommendations in continued collaboration with stakeholders and communities from across the sensitive data research landscape. A roadmap for how to take the DARE UK programme forward into the extension of this first phase will be shared and discussed with the community in due course.

**To keep updated about the work and progress of the DARE UK programme, please [sign up to our mailing list](#).**



# 11 / Appendices

## Appendix 1: DARE UK Phase 1 governance

### Programme Board members

**Professor Patrick Chinnery**, University of Cambridge (Chair)

**Professor Felix Ritchie**, University of the West of England  
(Deputy Chair)

**Dr Mike Ball**, Biotechnology and Biological Sciences  
Research Council

**Dr Catherine Bromley**, Economic and Social  
Research Council

**Dr Angela Coulter**, DARE UK Public Contributor

**Professor David Ford**, Swansea University (Chair of the  
DARE UK Scientific and Technical Advisory Group)

**John Marsh**, DARE UK Public Contributor

**Dr Justin O'Byrne/Dr Richard Gunn**, UK Research  
and Innovation

### Scientific and Technical Advisory Group members

**Professor David Ford**, Swansea University (Chair)

**Dr Claire Bloomfield**, Centre for Improving Data  
Collaboration, NHS England

**Professor David De Roure**, University of Oxford

**Professor Ben Goldacre**, Bennett Institute for  
Applied Data Science

**Professor Søren Holm**, University of Manchester

**Alison Kennedy**, Science and Technology Facilities Council

**Phil Kershaw**, Centre for Environmental Data Analysis

**Maisie McKenzie**, DARE UK Public Contributor

**Chris Monk**, DARE UK Public Contributor

**Professor Máire O'Neil**, Queen's University Belfast

**Professor Mark Parsons**, University of Edinburgh

**Professor Tom Rodden**, UK Department for Digital,  
Culture, Media, and Sport

**Professor Elena Simperl**, King's College London

**Peter Stokes**, Office for National Statistics

### Oversight Group members

**Dr Michael Ball**, Medical Research Council

**Dr Ekaterini Blaveri**, Medical Research Council

**Dr Paul Colville-Nash**, Medical Research Council

**Kirsten Dutton**, Economic and Social Research Council

**Rosie French**, ADR UK

**Dr Emma Gordon**, ADR UK

**Richard Welpton**, Economic and Social Research Council

**Yan Yip**, Medical Research Council

### Public Contributors

**Dr Angela Coulter**, member of the Programme Board

**Joyce Fox**, DARE UK Phase 1 Delivery Team Advisor

**John Marsh**, member of the Programme Board

**Maisie McKenzie**, member of the Scientific and Technical  
Advisory Group

**Chris Monk**, member of the Scientific and Technical  
Advisory Group

## Appendix 2: User personas

### Pritesh Navdra

32 • Data Scientist • technical

Pritesh completed a degree in Computer Science followed by a Masters in Data Science. He has been working as a data scientist in the private sector since he graduated in 2012, but has recently transitioned into the not for profit sector as he wants to make a difference. He is highly skilled in working with data in general but his skills aren't specific to a particular domain.

#### Goals

- To stay up to date with the latest technology
- To make a difference to society
- to discover data easily

#### Pain points

- Poor data quality, wrangling/cleaning required
- Understanding jargon/domain specific language barriers
- Lack of interoperability between disparate and disjointed data
- Gaining access to restricted data
- Cost of accessing lots of data
- Visualising large quantities of disparate data
- Considerably lower income in public sector
- Limited tech in public sector



#### Motivations

- Improving society • • • •
- Career development • • • • •
- Reduced workload • • •

“  
It's frustrating that  
the latest tech is out  
of reach. ”

### Peter Shaw

53 • Data Custodian • process driven

Peter is a highly experienced data custodian from Manchester. He has been working with environment data for over 15 years.

#### Goals

- To share data with others easily and securely
- To connect with other datasheets
- Raising the profile of my organisation
- Keeping data safe

#### Pain points

- anonymising sensitive data
- Understanding policies and regulations
- Duplication of data
- Lack of interoperability between disparate and disjointed data
- Not receiving any credit when data is used by others
- Understanding domain specific jargon



#### Motivations

- Data security • • • • •
- Recognition • • •
- Maximising data value • • • •

“  
I want to combine  
my data safely with  
other data to get  
more value for my  
community. ”

## Appendix 2: User personas

### Grace Opedemi

27 • Member of the public • security focussed

Grace is an accountant who lives in London. Recently, her mum was informed that her health data had been breached and this has made Grace keen to find out more about how personal data is stored and used in the UK.



#### Goals

- To ensure the public purse is yielding good value for money
- To ensure data security practices are being followed
- To help achieve the greater good

#### Pain points

- Missing technical and data skills
- Knowing about and finding relevant data
- Understanding jargon
- Understanding policy and regulations
- Data inaccuracies

#### Motivations

Public benefit	• • • • •
Data security	• • • • •
Data accuracy	• • • •

“  
I need to know how  
my data is being  
used.”

### Jeremy Foster

59 • Business collaborator • building value

Jeremy is an ambitious product manager who has been working for a leading Edtech company for the past 5 years. He like to draw on research to inform product development. However, gaining access to such data is difficult and time consuming.



#### Goals

- Generating business value/ROI through accessing and sharing data
- To discover new insights
- To make an impact on society
- To access and benefit from data skills I don't have

#### Pain points

- Missing technical and data skills
- Gaining access to restricted data
- Poor data quality
- Lack of public trust in private companies
- Accessing and building a relevant data community

#### Motivations

Making profit	• • • •
Recognition	• • • •
Ease	• • • • •

“  
I want to make use  
of existing datasets  
to help drive product  
development in my  
company.”

## Appendix 2: User personas

### Sarah Greenshaw

47 • Budget holder • building value

Sarah has been working in research for over 25 years and is an established leader in the public health domain. She leads a university based research centre and is well connected with UKRI.



#### Goals

- Build commercial opportunities and protect IP
- National and international recognition
- Talent retention
- Maintain and grow funding

#### Pain points

- Competition
- Exploitation of research
- Lack of access to non-academic expertise
- Unable to retain talent due to funding insecurity and low salaries
- Accessing and building a relevant data community

#### Motivations

Sustainability	• • • •
Growth	• • • •
Recognition	• • • • •

“  
I am constantly spinning plates and I have no thinking time.”

### Sharon Wakefield

44 • Domain researcher • career driven

Sharon is a mid career researcher who has been working in the field of research for almost 10 years. She is an expert in the agricultural domain but will be moving into the public health space.



#### Goals

- To do more impactful research by accessing 7 linking multiple data sets
- To access and benefit from data skills I don't have
- To raise the profile of myself and my organisation
- To make an impact on society
- To diversify my skillset
- To speed up my workflow

#### Pain points

- Missing technical and data skills
- Gaining access to restricted data
- Poor data quality
- Lack of interoperability between disparate and disjointed data
- Slow workflow

#### Motivations

Public benefit	• • • •
Recognition	• • • •
Ease	• • • • •

“  
I'm terrified by my own lack of understanding in the new domain I'll be working in.”

# DARE UK

## Get in touch

✉ [enquiries@dareuk.org.uk](mailto:enquiries@dareuk.org.uk)

🌐 [www.dareuk.org.uk](http://www.dareuk.org.uk)

🐦 [@DARE\\_UK1](https://twitter.com/DARE_UK1)

