Towards 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

Draft v1, July 2022

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

Contents

Acknowledgements

DARE UK is funded by UK Research and Innovation (UKRI) as part of its Digital Research Infrastructure portfolio of investments[1]. Phase 1 of the programme – Design and Dialogue – is led by Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK).

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.

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

[1] UK Research and Innovation. Digital Research Infrastructure. Accessed 15.07.2022.

1. 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 joined-up, 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 custodians 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 national data research infrastructure that is joined-up, demonstrates trustworthiness and supports 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 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 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, industry, the public and more.

1.1. Summary recommendations

Based upon the findings of engagement with stakeholders 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. *Proactive* transparency should be consistently practiced by all those handling and using sensitive data for research, particularly data collectors, data custodians and data guardians.
2. A public information campaign should be conducted to raise general awareness of how and why sensitive data is made accessible for research.
3. Data use registers should be published and maintained by the custodians of all types of sensitive data.
4. A culture shift is needed to recognise the crucial importance of public involvement and engagement in data research.
5. A central, independent coordinating function for public involvement and engagement in data research should be set up, either as a new entity or as an off shoot of a relevant existing body.
6. Where feasible, processes enabling access to sensitive data for research should be standardised and centralised across the UK.
7. Researchers hoping to access sensitive data should be stringently and ongoingly vetted and monitored to ensure public benefit is the principal motivation for accessing sensitive data.

Access and accreditation of researchers

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 user 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 to request access to sensitive data from TREs 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 establish a nationally recognised trusted research environment (TRE) standard, accreditation, and audit framework.

Data and discovery

1. Enhance the data lifecycle for cross-domain sensitive data research.
2. Explore implications of new data types, models for sharing and velocity of delivery.
3. Develop guidelines on privacy enhancing technologies (PETs).
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 council resources at UKRI-wide and council levels.

Core federation services

1. Develop reference architectures for trusted research environments (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 to validate that they are fit for purpose.

Capability and capacity

1. Establish clear technical career pathways that can be adopted across the UKRI research domains
2. Improve recruitment pathways for technical roles.
3. Improve the availability of career development resources and training.
4. Use automation to reduce the dependency on shortage skills.

Funding and incentives

1. Develop a new type of grant tailored for addressing the costs for maintaining cross-domain, national sensitive data research infrastructure.
2. Determine the funding requirements to establish the first phase of federated infrastructure for sensitive data research, with a focus on enabling federation across existing national data infrastructure and complimenting existing investments.
3. Investigate, test and prototype the operational model(s) for a federated ecosystem of national sensitive data research infrastructure. 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 national, federated infrastructure for sensitive data research.
5. Investigate the scope and funding requirements for the integration of large-scale compute availability in a federated 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. Raise awareness amongst data guardians regarding the legal framework around the secondary use of data for research.
8. Dedicate greater resource to incentivising data guardians to routinely make their data accessible for research in the public benefit.

2. Introduction

2.1. 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 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 joined-up, impactful and trustworthy way, to support research at scale for public benefit. Further, in line with the 'Levelling Up' agenda[2], driving a sensitive data research ecosystem (and indeed the data ecosystem more broadly) towards greater cohesion and collaboration will provide more equitable 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 programme, 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[3]. A key theme within the Digital Research Infrastructure portfolio strategy is to provide secure and trusted data services for sensitive data and the appropriate tools that will enable researchers, innovators, and decision-makers to derive benefit from these 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 joined-up, 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 different research disciplines and from across different sectors. The DARE UK programme is being undertaken in an open and inclusive manner, with involvement from researchers, technologists, funders, and the public embedded throughout.

[2] UK Government 2022. Levelling Up the United Kingdom. Accessed 13.07.2022.
[3] UK Research and Innovation. Digital Research Infrastructure. Accessed 13.07.2022.

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 data research use cases and there are opportunities for adoption of existing best practice across different data research domains.
- There is a **cross-domain need** – there are strategic use cases that span research domains and require a common architecture. These use cases are high priority, fundable research that involve data from across 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, align metadata and API (application programming interface) standards, and agree shared virtualisation approaches.
- It is **ethically and legally feasible** and would be **trusted by the public** – there are existing best practice governance practices which can be adopted and scaled across research domains.
- It is feasible to do within the **funding and time envelope** set by UKRI – by leveraging existing established 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 Administrative Data Research UK (ADR 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 will 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 3.

2.2. Definitions

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 as a by-product of 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 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. 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 is able to 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. This is because the TREs would all follow the mutually agreed security and governance protocols, and the different systems used across them would be able to work together coherently. This can occur in two different ways: in the first approach, the analysis of multiple sets of data occurs in the TREs in which they are held, and the results are brought together centrally in one TRE for final review. This approach works well for analysis on comparable data, known as horizontally partitioned data, held in different TREs. The second approach is to 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 separate TREs together.

Interoperability

Interoperability means different trusted research environments (TREs) – often managed by different organisations – can work together so that users 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 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 they were collected, the coverage of the data, and licencing arrangements. Metadata can include such information as publication date, description and search keywords.[4]

Metadata can be a 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).

[4] Office for National Statistics. Metadata policy. Accessed 12.07.2022.

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 to be 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.

3. Process and summary of input

Input from stakeholders and the community that forms the basis of this report has been gathered through a mixed method of activities throughout the programme's lifecycle. It should be highlighted at the start that the DARE UK Delivery Team, alongside the input described below, 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.

Initial landscape review, August-September 2021

The first engagement activity undertaken in DARE UK Phase 1 was the initial landscape review, the purpose of which was to engage broadly across the UK research and innovation ecosystem to establish a *"fundamental understanding of the context and overarching challenges within the UK research and innovation ecosystem"*.[5] 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 expedient to include multiple stakeholders in a single sitting. In total, 79 people (researchers, technologists, funders) were interviewed.

The second part to the review was two workshops aimed at researchers and technologists respectively of 1.5 hours each that focused on reviewing the synthesis of the interviews. The aim was to receive g feedback 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. The workshops were attended by approximately 50 participants per workshop and provided valuable additional input, feedback and constructive questions for the Delivery Team to consider further. 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.

Sprint Exemplar Projects, January-August 2022

Additional insights that have been incorporated within these recommendations are those from the portfolio of nine Sprint Exemplar Projects that were initiated and competitively selected at the end of Q4 2021. These projects were selected by an independent panel and focused on executing exploratory work that would inform the future directions of DARE UK. The projects started their work in January 2022 with delivery of their outputs scheduled for the end of August 2022. The level of engagement from the Sprint teams has been significant and has made a substantial contribution to shaping the views of the programme and the recommendations set out in this report. Through regular monthly connects with the projects, sharing of early findings and thinking and discussions at the virtual and in-person Sprint events, the DARE UK thinking has benefited greatly from the insights the project teams have shared. The DARE UK delivery team has incorporated the interim insights from the portfolio into this set of recommendations where relevant.

[5] DARE UK 2022. Landscape Review.

Public dialogue, January-February 2022

In January and February 2022, a total of 44 members of the public from a diversity of backgrounds and identities 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.

Two initial workshops were held online over two days across Thursday 13 January and Friday 14 January 2022. A single follow-up workshop was then held online on Tuesday 22 February with a cross-section of 10 participants from the two initial workshops, to check 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 data research community to address them. You can find out more about the methodology and findings in the full Public Dialogue report.[6]

Wider stakeholder workshops, March 2022

In Q1 2022, the first drafting of early thinking around the potential directions of travel for the DARE UK recommendations were completed, structured around the six broad thematic areas of focus as described in the initial landscape review. These early directions of travel were then shared and discussed throughout the month of March 2022 via six virtual workshops of two hours each, with each workshop focused on one thematic area of focus. The workshops were structured to provide participants from the community 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, attendance was just short of 300 participants across the six workshops, with some sessions attracting more participants than others. The feedback received was compiled and summaries published via the DARE UK website capturing the key messages from each workshops' feedback, with an open request to participants to correct any inaccuracies or misunderstandings[7]. This feedback has been incorporated into this set of recommendations.

Initial user personae definitions, April 2022

In April 2022, the Delivery Team convened an in-person, half-day design thinking workshop hosted at the Science and Technology Facilities Councils' (STFC) Hartree Centre on the Sci-Tech Daresbury campus, this design thinking workshop was facilitated and led by the STFC team with the DARE UK Delivery Team providing input around the scope of the workshops' focus. The workshop was attended by 30 participants (incl. the Delivery Team) from across different research domains and approximately half of the participants coming from variety industry sectors, participants were guided through the design thinking approach to begin to identify the unique user personae of a national sensitive data research infrastructure that is joined-up, demonstrates trustworthiness and supports research at scale for public good. This work is only the start of what will need to be an iterative discussion with the research community and industry to validate a set of representative user personae, and subsequently map out the kinds of user journeys these personae could take through the infrastructure that can

[6] DARE UK 2022. Building a trustworthy national data research infrastructure: A UK-wide public dialogue.
[7] DARE UK. We want to hear your views: join us for a series of workshops this March. Accessed 15.07.2022.

then guide design choices and decisions. The outputs from this work can be found in the user personae in Appendix 1, though as stated this is only the first step and will need to be iterated and validated in future phases of the programme.

3.1. Structure of the report

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 Phase 1:

Chapter 4: Demonstrating trustworthiness
What is needed in order to demonstrate trustworthiness in the handling and use of sensitive data for research, to maintain the confidence of the public and other stakeholders.

Chapter 5: Access and accreditation of researchers
Governance, rules, and frameworks for enabling data access and for researchers to conduct analysis on sensitive data for public benefit.

Chapter 6: Accreditation of research environments
Standards and frameworks for accrediting trusted research environments (TREs) which store, process, and manage sensitive data for analysis.

Chapter 7: 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 8: Core federation services
System requirements to begin enabling interoperability across a network of trusted research environments (TREs).

Chapter 9: Capability and capacity
Staffing, training, and improved career structures needed to support an appropriately skilled workforce that underpins data research.

Chapter 10: 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 these will have a variety of timeframes for delivery and a variety of different approaches to doing so, which will be premised on the priorities and resources available now and in the future. While some of the recommendations set out are more mature and could be actioned in the shorter-term, others will require further scoping – with involvement from across the research community – in future phases of the DARE UK programme.

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 move the landscape forward.

4. Demonstrating trustworthiness

4.1. Context

DARE UK's focus is on sensitive data. Sensitive data is often data about people, 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 (Waind 2020). This is reflected in the findings of the DARE UK public dialogue carried out in early 2022, which explored views towards current sensitive data research practices and where improvements might be needed to increase public confidence (Harkness et al., 2022). 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."*[8]

Various recent papers have expressed the importance of a focus on demonstrating trustworthiness in the context of research using sensitive data (Aitken 2016a; Sheehan 2021), while others have explored *how* trustworthiness can be demonstrated in the context of specific research topics (for example, Milne et al. 2021). 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 that *"most research finds dissatisfaction with the current ways in which data is used and managed, and a desire for this to change"* (Kennedy et al. 2020b). Specifically, the authors found that the public want more: *"honesty, transparency and genuine dialogue with the public; regulation, enforcing compliance, the existence of safeguards and accountability, and the right to redress; and personal control."* Change is needed to better demonstrate trustworthiness in the use of sensitive data.

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, why and by whom.
2. **Meaningful and inclusive public involvement and engagement**, in which members of the public are involved in decision-making processes.
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.

[8] O'Neill O., 2013. *What we don't understand about trust*. TEDxHousesOfParliament.

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

4.2. 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.[9] [10] More proactive efforts to reach out to the public with information about what is being done with their sensitive data and why are needed. This requires going beyond putting information on websites for those who seek it out and actively going 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'"* (CDEI 2020). A public attitudes tracker survey carried out by the CDEI in December 2021 found that *"people report feelings of uncertainty about current data practices and fairly limited knowledge regarding how data about them is used and collected in their day-to-day lives… This uncertainty, alongside perceived risks around data security, data control and data accountability are barriers that must be overcome to build confidence in data use"* (CDEI 2022).

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. Initially, participants of DARE UK's public dialogue had low understanding of the ways in which sensitive data is used for research (Harkness et al., 2022.). 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 (Cameron et al. 2014; Aitken et al. 2016a; Aitken et al. 2016b; Davies et al. 2018). One participant of the DARE UK dialogue commented: *"There's a gap between the great work being done and what people actually know. People are badly informed."* 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 (Harkness et al. 2022.).

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 public need to know, in simple terms, what data is being collected and how it will be used to feel secure"* (Maxwell M., et al. 2021.). A 2021 dialogue commissioned by the National Data Guardian 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 communicating widely about how*

[9] Triggle, N. 2013. Care.data: How did it go so wrong?. BBC News.
[10] Which? 2021. Around 20 million people unaware of plan to share GP medical records with NHS database, Which? finds. Which? Press Office.

data is used for public benefit" (Hopkins Van Mil 2021). 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 (Davies et al. 2018; Stockdale et al. 2019; Aitken et al. 2016a).

In terms of how to achieve 'proactive transparency', participants of the DARE UK dialogue suggested that honest, consistent and accessible public communications are key.

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, including data collectors/owners, data custodians and researchers; and
2. **increased general awareness of data research practices**, i.e. through a large-scale, 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 public information campaign should focus on the use of all types of sensitive data for research – including data about education, welfare, health, justice, 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, people who don't have much interaction with or trust of public services, and those who are geographically isolated"* (Harkness et al. 2022.).

A 2020 paper by H. Kennedy et al. stresses that *"social inequalities play a role in informing perceptions of data practices"* (Kennedy et al. 2020a) 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[11]) – to challenge *"the assumption that understanding is the main pre-requisite to developing views about data practices,"* and suggest that "*feelings offer a way to engage with some people about datafication."* This further emphasises the need for a tailored approach to public communications.

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 (Harkness et al. 2022.).

[11] Southerton, C., 2020. *Datafication*. Springer Link Encyclopedia of Big Data. Accessed 20.05.2022.

The National Data Guardian dialogue found that participants *"felt the communications should be widely distributed and displayed on and offline in public spaces used every day by everyone such as GP surgeries and websites, libraries, local authority websites and newsletters and community venues"* (Hopkins Van Mil 2021).

Regarding the messaging of public communications, participants of the DARE UK dialogue felt this 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. Again, 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 (Harkness et al. 2022).

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.

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 (such as de-identification), and how researchers access the data, right up to being informed of the intended findings and societal implications of the research*" (Harkness et al. 2022). 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 (in plain English) detailing who has accessed and uses the data (and why), the impact of the research undertaken, and distribution of any financial benefits to the NHS"* (OneLondon 2020).

One solution to easily provide 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 responsible for the collection, storage and sharing of data for the purpose of research, innovation and service evaluation should publish and actively promote a public record (data use register) of approved research studies, projects and other data uses"* (Karrar et al. 2022). 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 (published in beta version at the time of writing), 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.[12] A Power BI report is *"a multi-perspective view into a dataset, with visuals that represent different findings and insights from that dataset"*.[13]

[12] NHS Digital. Data Uses Register. Accessed 20.05.2022.
[13] Microsoft Build. Reports in Power BI. Accessed 20.05.2022.

Stakeholders engaged with during DARE UK Phase 1 showed widespread support for the adoption of data use registers by the custodians of all types of sensitive data – including data relating to education, justice, welfare, health and so on – as a key aspect of maintaining transparency in the use of this data for research. There was a clear view that these registers should be standardised, accessible in language and format and regularly reviewed, with their existence proactively communicated to the public.

Public involvement and engagement

Meaningfully involving 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.[14] However, we have heard from the researchers, technologists, members of the public and others that 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 and more – for which public involvement and engagement is currently far less routine than for research using health data. A **culture shift** is needed in which all those handing and using sensitive data for research – data collectors and custodians, technologists, funders, researchers and others – fully acknowledge the necessity of public involvement and engagement and dedicate appropriate resources to enable it to happen in a meaningful way.

Participants of the DARE UK public dialogue felt it was *"important that members of the public represent the public voice on panels for decision-making"* (Harkness et al. 2022). The National Data Guardian dialogue found that *"public benefit is undermined if authentic public engagement is not integrated into data assessment"* (Hopkins Van Mil 2021). 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 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"* (OneLondon 2020).

In addition, it is important that the researchers, technologists and others 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 these groups are to develop effective skills in this area.[15] 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.[16] A similar portal for public involvement in data research

[14] Why does public engagement matter? National Co-ordinating Centre for Public Engagement. Accessed 19.05.2022.
[15] DARE UK 2022. Early thinking: Demonstrating trustworthiness workshop and feedback.
[16] NIHR Learning for Involvement. Training and resources for public involvement in research. Accessed 20.05.2022.

relating to all types of data 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"* (Harkness et al. 2022). 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 that 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 for building trust.

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[17], led by researchers at Cardiff University and Diabetes UK, is primarily focussed on exploring the best ways to inform, engage and involve young people with health conditions, their families and the wider public – particularly from under-represented groups – in the use of their sensitive data for research. 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; a separate evaluation of public involvement and engagement in the DARE UK Sprints was underway at the time this report was published.

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 (Harkness et al. 2022). 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. They were concerned that too much 'red tape' may have an impact on the public benefits of research.

> *"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**

For storage of and access to sensitive data for research to be more trustworthy, 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 standardisation and transparent auditing of TRE structures and

[17] DARE UK. STEADAST: Education outcomes in young people with diabetes – innovative involvement and governance to support public trust. Accessed 06.07.2022.

processes – including standardisation of what qualifies as a TRE – to ensure best practice is met and security processes remain fit for purpose across the entire ecosystem. Participants of the dialogue also wanted this standardisation to occur on a UK-wide level, with involvement from each of the four nations in agreeing these standards. Another of the dialogue's recommendations is: *'The processes and systems supporting data research across the UK should be unified in their approaches'*. Participants 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.

More detail regarding DARE UK Phase 1 findings and recommendations regarding the standardisation and centralisation of data storage and access processes can be found in Chapters 5, 6 and 8 of this report.

Research for public good

Public conversations over the last decade and more have consistently found that people want public good (or 'public benefit') to be the principal motivation of any research using sensitive data (Waind 2020). 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 (Harkness et al. 2022).

Although 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 (Harkness et al. 2022). They were comfortable with commercial access to sensitive 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. Similarly, 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"* (Maxwell et al. 2021).

Assessing public benefit as the principal motivation for data access, however – and how this relates to commercial benefit – is not straightforward. 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 expressed similar views; they 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 (Hopkins Van Mil 2021).

For the public to be confident that public benefit is the principal motivation of research using sensitive data, DARE UK Phase 1 finds this to predominantly rely on two key things:

1. stringent and ongoing **vetting and monitoring** of researchers accessing sensitive data to assure their motivations; and
2. as discussed above, involvement from members of public on **decision-making panels** to help assess the potential public benefit of proposed research using sensitive data on a case-by-case basis.

There is also further work needed to explore the concept of public good/public benefit in more depth, and what it might mean for different groups in society in the context of data research. At the time of writing, a public dialogue led by ADR UK (Administrative Data Research UK) and the Office for Statistics Regulation was underway to explore views towards the public good of data and statistics.

4.3. Recommendations

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

1. ***Proactive* transparency should be consistently practiced by all those handling and using sensitive data for research, particularly data collectors, data custodians and data guardians.**
    - Although proactive transparency should be practiced by all those handling sensitive data for research, **data collectors, data custodians and data guardians** in particular should hold collective responsibility for actively reaching out and telling people their data is being made accessible for research, how and why.
    - Proactive transparency should span the **entire data journey**, from data collection to research impacts.
    - Messaging should be **honest, accessible and consistent**, but with particular focus on the **public benefits** of the research and the **security processes** in place to protect data. It should not shy away from communicating complex topics.
    - Public communications should be tailored to be accessible and relatable for different communities and groups to ensure a **diverse and inclusive** public is reached. Designing tailored messaging should include involvement from members of those communities and groups.
    - Sufficient **resources** should be dedicated to carrying out proactive transparency in an ongoing fashion, and **support** should be provided by host organisations for data researchers to carry out communications activities related to their projects.
2. **A public information campaign should be conducted to raise general awareness of how and why sensitive data is made accessible for research.**
    - This should be **funded and centrally coordinated by UKRI** – in consultation with relevant organisations working within the sector and with involvement from the public – and dedicate the **necessary resources** to raise general awareness.
    - It should focus particularly on the **security processes** in place to protect data and the **public benefits** of data research.
    - It should include the development of **honest, consistent and accessible** messaging and definitions to be adopted across the sector.
    - It should bring data research into the **mainstream** via channels such as newspapers, television and social media.
    - It should involve reaching out to different groups with **tailored messaging** which is accessible and relatable to them, to ensure an inclusive and diverse public is reached. Designing tailored messaging should include involvement from members of those communities and groups.
3. **Data use registers should be published and maintained by the custodians of all types of sensitive data.**
    - These should include information about **who** has accessed **what data**, **when** and for **what purpose**, and should cross-reference any research outputs and impacts as they emerge.
    - They should follow a **standard, clear and accessible format** across different data custodians – which could be **mandated by UKRI** for funded initiatives – and should be regularly updated as and when data access is granted.

- Data use registers could be **centrally collated by UKRI** – or directed to from a central place – alongside information describing them, so that people can find and access them easily.
- They should be **widely promoted** 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. A culture shift is needed to recognise the crucial importance of public involvement and engagement in data research.

- **Public involvement and engagement should be embedded** as a central component across the entire data research lifecycle, by both data custodians and researchers.
- Members of the public should be **involved in decision-making processes** relating to the use of sensitive data in research, particularly in relation to assessment of 'public benefit'.
- Public involvement and engagement should be **meaningful and inclusive**, involving people from across different groups, communities, identities and backgrounds and giving them the time and resources needed to fully understand and respond to the issues being discussed.
- The **necessary resources should be dedicated** to enable public involvement and engagement to happen in a meaningful way. This should include dedicated funding as part of research grant applications; dedicated members of staff within data research organisational structures; and incentives for researchers.
- UKRI could embed **mandatory requirements** for public involvement and engagement within funding applications, and data custodians and guardians for data access requests, and these activities could be included in monitoring and reporting processes.
- Researchers, technologists and others should have access to **training and resources** to support them to embed public involvement and engagement as a crucial element throughout the research lifecycle.

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

5. A central, independent coordinating function for public involvement and engagement in data research should be set up, either as a new entity or as an off shoot of a relevant existing body.

This function should cover all types of sensitive data, be appropriately and sustainably resourced and be responsible for:

- Leading the creation of a **more standardised approach** to public involvement and engagement across the whole UK data research sector. Including: refining and promoting definitions, principles and best practice standards whilst maintaining flexibility to ensure inclusivity; and driving their adoption across the sector.
- Leading on **better understanding and documenting public attitudes** towards the use of sensitive data for research. This should involve a full and ongoing audit of existing knowledge of public attitudes, highlighting variations in attitudes, where the gaps lie and where views may have changed over time; and leading on public dialogue work – in collaboration with relevant organisations – to fill gaps in knowledge.
- Providing 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.

- Developing and providing **public involvement and engagement training and resources** for researchers, technologists and others working within the data research sector. In addition, maintaining a central store of case studies and examples of public involvement and engagement best practice.

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). It could also 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.

In the next phase of the DARE UK programme, this function will be fully scoped out via engagement with relevant organisations across the community, to decipher what is needed (in terms of both remit and resources), which organisations would be responsible for leading it, and how.

6. **Where feasible, processes enabling access to sensitive data for research should be standardised and centralised across the UK.**

   This recommendation is detailed in greater depth in Chapter 5.

7. **Researchers hoping to access sensitive data should be stringently and ongoingly vetted and monitored to ensure public benefit is the principal motivation for accessing sensitive data.**

   Further detail about our recommendations for researcher accreditation can be found in Chapter 5.

5. Access and accreditation of researchers

5.1. Context

Safe and timely access to sensitive data is crucial to enable research and innovation for the public good at scale. However, one of the biggest challenges for researchers from both academia and industry is the time that it takes to apply 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 because of many data guardians having different processes and collecting 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 different custodians in a slightly different format.

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

- User accreditation process
- Unifying user authentication capabilities to access TRE services
- Data access request standardisation

Data Access Committee processes, policies, and standards are outside the scope of this report, as they are often subjective and subject to the combinatorial requirements of the data guardian, data custodian (e.g., TRE provider), and research project. This chapter therefore focuses on the underlying baseline standards and support required to enable such complex governance procedures to take place by streamlining how users are accredited within the network and the processes for subsequently accessing sensitive data for analysis.

A consistent challenge that has been identified across all stakeholders during DARE UK Phase 1 has been a lack or unclear and inconsistent application of user 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 environment. A simple and clear example of this is user authentication – every TRE needs to provide some form of user identification and authentication for researchers to access services within their environment. Despite the availability of many identity federations (for example, UKFed[18], Geant[19]) and industry standards (for example, OIDC[20], OAUTH2[21]), 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 user accreditation (information governance and sensitive data handling training) requirements which place a large administrative burden on TREs to verify and validate each user separately for access to each TRE. User authentication and 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.

[18] Jisc. The UK Access Management Federation for Education and Research. Accessed 13.07.2022.
[19] GÉANT. Trust and Identity Services. Accessed 13.07.2022.
[20] OpenID Connect Federation 1.0 – draft 20. Accessed 13.07.2022.
[21] OAuth 2.0. Accessed 13.07.2022.

5.2. Existing challenges and opportunities

Alongside the input received as part of DARE UK Phase 1 as outlined in Chapter 3, this chapter summarises additional international input from the Global Alliance for Genomics and Health (GA4GH), Australian Research Data Commons (ARDC), Towards European Health Data Space (TEHDAS), World Health Organization (WHO), and National Institutes of Health (NIH). 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 4 above.

Federated identity and user authentication standards

There is a need for authority that could provide a user authentication protocol or guidance on acceptable approaches that infrastructure providers and users could trust and rely on to move in a common direction. Conceivably, this would need to be UKRI itself, as it has the appropriate remit 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 that this is a prerequisite for all forms of federation to occur and will aide 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[22], which will need to either be expanded or linked to other federations being created, such as NHS Care Identity Service 2 (CIS2)[23] or GovRoam[24]. The existence of modern industry and community standards of user authentication (for example, SAML[25], OIDC[26], OAUTH2[27], GA4GH Passports[28]) were also highlighted by stakeholders and hence we recommend leveraging existing standards 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 blueprint, our respondents also highlighted the need to support different forms of identity verification and have logging and auditing embedded across the system.

User accreditation

A key requirement highlighted by stakeholders has been the need for a streamlined approach to user accreditation. While there are a number of existing training modules for sensitive data handling (for example, those provided by ONS and the MRC), many of these trainings are duplicative without allowing for equivalence between modules. Respondents 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

[22] Jisc. The UK Access Management Federation for Education and Research. Accessed 13.07.2022.
[23] NHS Digital. NHS Care Identity Service 2. Accessed 13.07.2022.
[24] Jisc. Govroam. Accessed 13.07.2022.
[25] OneLogin. SAML Explained in Plain English. Accessed 14.07.2022.
[26] Trust and Identity Services. Accessed 13.07.2022.
[26] OpenID Connect Federation 1.0 – draft 20. Accessed 13.07.2022.
[27] OAuth 2.0. Accessed 13.07.2022.
[28] Global Alliance for Genomics and Health. GA4GH Passports and the Authorization and Authentication Infrastructure. Accessed 13.07.2022.

standard foundation for all accreditation courses with the possibility of 'extended' modules in specific cases as needed.

Stakeholders affirmed that work to standardise and smooth the researcher accreditation process was sorely needed, along with reciprocal or unilateral recognition of accreditation. Providers should be aiming to provide a consistent researcher user 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 therefore be made portable across TREs through standard accreditation for researchers acting as a TRE passport. The Digital Economy Act (DEA) already works as a passport in some respects, with shared accreditation across certain TREs. However, health data collected for an organisation's health functions is not covered by the DEA.

International and industrial researchers would also need to be considered, as currently accredited researchers need a link to a UK institution. Stakeholders considered it best practice for TREs to maintain teams of individuals to support researchers and data providers, including the ability to tell data owners and members of the public what research their data is being used in and by whom – a surge in people using TREs would need to be prepared for and staffed, as discussed in Chapter 9. Given that TREs operate in a global context, connectedness with global partners is essential, including those in low-resource settings. A key recommendation is therefore to provide online training modules that can delivered at a national and international scale with on-site drop-ins to scale the delivery and regular maintenance of user accreditation.

Data access request standardisation

Standards in information security, platform specifications and service descriptions were highlighted by respondents, as well as a centralised, codified 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 in policy-setting would address the desire from respondents, 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[29] or HRA's Integrated Research Application System (IRAS)[30], as well as from consortiums such as the BHF Data Science Centre and 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 4).

[29] HDR UK. Data Access Request Process Overview. 15.07.2022
[30] Integrated Research Application System. Accessed 13.07.2022.

5.3. Recommendations

The following key recommendations would support access and accreditation of users 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. **Provide a unified user authentication capability to enable researchers to access services more easily across the entire sensitive data research ecosystem (see also Chapter 8).**
    - 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 user 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 user accreditation.
    - Develop consistent guidance for stakeholders to undertake user accreditation.
    - Develop user 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 user accreditation.

3. **Develop a standardised and streamlined – yet extensible – process to request access to sensitive data from TREs 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 a centralised service.
    - Align Data Access Request forms using the Five Safes – Safe People, Safe Project, Safe Data, Safe Setting and Safe Outputs.
    - Interlink external identifiers of resources – datasets, tools, funders/sponsors, people, and project/grant identifier 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.

6. Accreditation of research environments

6.1. Context

The focus of this chapter is around the accreditation of trusted research environments (TREs). Standard and transparent accreditation of research environments 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 8). 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 2017 Digital Economy Act (DEA)[31] 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 in-compatible 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 7 and 8 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 4) – 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.

6.2. 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, SAIL Databank, the Scottish National Data Safe Havens (facilitated through the electronic Data Research

[31] UK Government. 2017 Digital Economy Act. Accessed 14.07.2022.

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[32]. Although a simplified definition of a TRE has been provided in Chapter 2 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. 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 – i.e. 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 important considerations from stakeholders during DARE UK Phase 1 have been around:

- Any TRE standard and accreditation framework should look to consolidate the range of certifications that already exist (for example, ISO27001[33]), as a basis for developing 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.
- 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 public dialogue (see Chapter 4) 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.

[32] Desai, T. et al. Five Safes: Designing Data. University of the West of England.
[33] ISO - ISO/IEC 27001 — Information security management

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 (e.g., specific use cases or 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 commercial 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.

TRE audit

Alongside any accreditation process is the 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 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 public and data guardian confidence in TREs handling sensitive data.

As above, under the DEA there already exists an established process for the accreditation and audit of processors (including the enabling technical environments) of sensitive data with the authority assigned to the UK Statistics Authority. It is important to acknowledge that this 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 or as a point of reference for extending the existing 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) thereof should reflect this. Equally it must be acknowledged that there will be many differences as well and appropriate, independent audit thereof, including work on extending the existing framework, is critically important.

6.3. 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 establish a nationally recognised trusted research environment (TRE) standard, accreditation, and audit framework.**
    - Led by the UK Statistics Authority (UKSA) – as the independent authority – and with involvement from stakeholders across the UK research community, develop a working definition of a TRE and iterate to get to 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). 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.
    - 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.

7. Data and discovery

7.1. Context

Data and metadata lifecycle management underpins every trusted research environment (TRE) and every project within a TRE. {Meta}Data lifecycle management 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 setting(s). The Findable, Accessible, Interoperable and Reusable (FAIR) principles (see Figure 4 below) ensure efficient data sharing practices are the[34].

	Level 1 Initial	Level 2 Repeatable	Level 3 Defined	Level 4 Managed	Level 5 Optimised
Findable					
F1. (meta)data are assigned a globally unique and eternally persistent identifier. F2. data are described with rich metadata. F3. (meta)data are registered or indexed in a searchable resource. F4. metadata specify the data identifier.	No URI or PID and no documentation	PID without metadata or documentation	PID with limited metadata, just enough to understand the data	PID with standardised metadata registered or indexed in a trusted data repository	Extensive metadata and rich additional documentation available and searchable in a trusted data repository
Accessible					
A1 (meta)data are retrievable by their identifier using a standardized communications protocol. A1.1 the protocol is open, free, and universally implementable. A1.2 the protocol allows for an authentication and authorization procedure, where necessary. A2 metadata are accessible, even when the data are no longer available.	No user licence / unclear conditions of reuse / metadata nor data are accessible	No metadata and user Access restrictions apply with only bespoke access	Appropriately licensed and limited (meta)data retrievable using standardised protocols	Public access (after registration) With (meta)data accessible (even when data is no longer available)	Open Access (unrestricted)
Interoperable					
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles. I3. (meta)data include qualified references to other (meta)data.	Proprietary, non-open format data	Proprietary format accepted by certified and trusted data repository	Non-proprietary, open format (archival format)	Data additionally harmonised/standardised using a standard vocabulary	Data is additionally linked to other data to provide context
Re-usable					
R1. meta(data) have a plurality of accurate and relevant attributes. R1.1. (meta)data are released with a clear and accessible data usage license. R1.2. (meta)data are associated with their provenance. R1.3. (meta)data meet domain-relevant community standards.	No clear provenance of data (to facilitate replication and reuse)	Explication of how data was or can be used is available with user access restrictions	Data automatically usable by machines and (meta)data meet domain-relevant community standards	Data stored in a trusted data repository	Data is reliable and tested against gold standard (reference data)

Figure 4: FAIR Capability Maturity Model

- **Findable** – data should include metadata and persistent identifier to make it discoverable.
- **Accessible** – metadata should be freely accessible with documented routes to request access to sensitive data.

[34] Go Fair. FAIR Principles. Accessed 14.07.2022.

- **Interoperable** – metadata should use controlled vocabularies, be machine-readable and include references to other metadata. Where possible, data should be made available in open formats and standards.
- **Re-usable** – metadata and data should conform to standards for greatest reusability.

Making metadata FAIR is critically important, benefitting all participants in the research ecosystem; but it requires consistent effort and should continue to be supported both within and across research domains. Increased visibility and documented routes for the sensitive data available within TREs allows more research to be conducted, but importantly improves the efficiency with which this increasing scale of research takes place (see Chapter 5). Interoperable metadata allows discovery of innovative research through novel linkage of datasets, not only within research domains themselves but increasingly between different research domains. Reusable {meta}data ensures the outputs of innovative research can be reused and built upon by others in the research ecosystem, though this is a real challenge when considered from a cross-domain perspective.

At least a decade ago, funders and charities improved policies that require the inclusion of research data management statements from research projects and their investments. This has significantly improved the findability and accessibility of the data through metadata deposited in various deposition databases however this has also resulted in further challenges in interoperability and reusability due to their subjectivity in research. It must be acknowledged, as we heard from stakeholders throughout, that 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 data 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 thus 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 which leads to the decreased utility of the data. Simple metadata attributes such as missingness, spatial and temporal coverage can lead to a significant improvement in the utility of the data. Lack of such consistent metadata coverage 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. Initiatives such as metadata catalogues are available in some domains, but not all, and further do not interoperate with each other both within and across research domains. Lack of such catalogues also prevent datasets from being assigned persistent identifiers[35] that support to 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.

"Standards are like toothbrushes, everyone's got one, but no one wants to use someone else's."
Commercial technologist

[35] https://en.wikipedia.org/wiki/Persistent_identifier

7.2. Existing challenges and opportunities

This chapter summarises the input from a wide range of stakeholders and previous work from research data lifecycle management efforts from funders, universities, and research organisations. The engagements throughout this first phase of DARE UK have helped to identify several challenges that need to be addressed as the infrastructure and the ecosystem evolves 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, ethics best practices, and improve efficiency of data access so that research is not unduly delayed by differing data management processes[36].

Data management lifecycle

There was overwhelming feedback for more streamlined data management lifecycle approaches across data custodians and TREs, with a recurring theme from many stakeholders being the need for more data stewardship capacity to help manage data collection, curation, harmonisation, and linkage (see Chapter 9). 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 and archiving capability that supports use cases within research council domains and cross-domain use cases as well, as currently required for ESRC funding with disposition of artefacts with the UK Data Service[37] (see Chapter 10, recommendation 6).

Many respondents also highlighted the lack of cross-TRE approaches for data provisioning especially when linkage is required, the administrative burden of coordinating the provisioning of data across TREs is orders of magnitude greater than provisioning the data sets individually. This covered all stages of the data management pipeline from access requests through multiple data custodians, coordination data preparations and linkage, and then through to provisioning into the TREs. Similarly, respondents also highlighted significant challenges in obtaining approvals from multiple data custodians for access. Trusted third parties to support data provisioning and linkage have been proposed in the past, however concerns have been noted on its fairness, security and costs.

Data standards

The challenges of dealing with the volume and variety of data was highlighted in our engagements with a particular emphasis on streaming, wearable, and near real-time data. Beyond these specific use cases, the respondents also highlighted the need to consider the storage of new data types, for example imaging will consume more storage than is practical with some data just needing to be processed at the point of productions and the raw data discarded. These emerging data requirements are already impacting the volume, velocity, and variety of data within the sensitive data research ecosystem and of course within the research and innovation ecosystem more broadly. The emerging data velocity – how quickly data is generated and how quickly it moves – requirements pose a particular challenge in the sensitive data research ecosystem in how to ensure the governance and privacy protections in place fit the tempo of the data velocity itself to ensure trustworthy and ethical use of the data.

Data standards, metadata standards, and interoperability standards were of importance to interviewees though there was clear feedback that introducing additional standards would not serve to alleviate the challenges, as well as API standards which were regarded by some as the key to federation.

[36] DARE UK. Building a trustworthy national data research infrastructure: A UK-wide public dialogue. Accessed 14.07.2022
[37] UK Data Service. Deposit Data. Accessed 14.07.2022.

Data governance and privacy

Following the development of a data lifecycle management approach, our stakeholders highlighted the need for a streamlined approach to data governance and improvements in privacy risk management. Traditionally data custodians have often been risk averse as a consistent approach, supported by appropriate tools, to supporting risk assessment were not available to them. While this may have been well-suited to limited data in the past, the advent of new techniques to assessing privacy risk and importantly introducing proportionate privacy risk would support improved management of data governance processes at scale. Alongside this, new techniques to minimise privacy risk such as privacy enhancing technologies (PETs), where there is already work ongoing for example with the ICO, UN, Royal Society/Alan Turing Institute on the deployment of PETs alongside TRE, provide new opportunities to develop approaches that do not only minimise privacy risk but also support data custodians and guardians in their decision-making around data governance, thus improving the efficiency and consistency of those decisions. This then allows risk management to be more proportionate.

Sensitive data taxonomy

Throughout discussions with stakeholders and members of the DARE UK programme board and Scientific and Technical Advisory Group it has been clear that there is no simple definition of sensitive data, nor a taxonomy that could be used to describe such data. The development of such a taxonomy is important to support work on privacy, linkage and the approaches that might be applied to different data for federation, for example, whether the data is typically horizontally or vertically portioned[38]. The taxonomy will also need to encompass not just existing structured and unstructured data but also emerging types such as data from wearables, and that delivery of these data may not be through conventional datasets but potentially through approaches such as publish/subscribe distribution. This work would also assist in the development of cross-domain synthetic data to support a variety of use cases such as early development of models and training and development of analysts to work on cross-domain sensitive data research.

It is therefore proposed that the next stage of the DARE UK programme with work with the community to develop a common cross-domain taxonomy for sensitive data.

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. Respondents 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. Committed collaborations of bodies (such as universities) could be best placed in the implementation of standards, particularly those with similar interests in order 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, to define a minimally acceptable metadata standard across UKRI councils with opportunities of each council to extend the minimum standard as their recommended standard for capturing their domain-specific metadata.

[38] Towards Data Science. Database Terminologies: Partitioning. Accessed 14.07.2022.

UKRI could also support the creation of infrastructure that allows for sharing of metadata, browsing services for different types of data, and pointing towards places or groups that could provide good feedback on the data. Understanding quality, missingness, and how a dataset was generated requires collaboration. This report recommends the development or selection of a reference implementation of Metadata Catalogue that can support one or more minimally acceptable metadata standards and the use of digital object identifiers (DOIs)[39].

Researchers need to have user-friendly sets of metadata to describe datasets (or objects). UKRI could support with more tools or capabilities for data-holders to register their datasets and their terminologies. Increased transparency enabled an individual (for example, a patient) to see where their data is being used, this could even help to demonstrate trustworthy use of data.

To facilitate cross-council discovery and reuse, we also recommend the creation of a federated registry to hold a list of available catalogues, standards, vocabularies, and terminologies.

The respondents also highlighted the challenge of identifiers of data being made available across the fragmented landscape. Each release of a dataset and revisions needs a unique DOI that can be cross linked with other resources. There are existing services such as NERC already doing this for archival data, so we recommend reviewing existing resources being made available throughout UKRI-councils and mandate each council to have a DOI for any resource post-investment.

The respondents also highlighted that this needs to be a production services and not just an academic-only service and should be rich enough to provide all the technical details necessary for discovery, cross-linkage and distribution. The DOI service will also need continuous investment for maintenance. Hence, we recommend investing in the creation and maintenance of a DOI.

7.3. Recommendations

The following key recommendations are made for investment in future phases of DARE UK to support data and discovery as part of UKRI's wider 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 initiative from both the UK and internationally.

1. Enhance the data lifecycle for cross-domain sensitive data research.

- Develop a common cross-domain taxonomy 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.

2. Explore implications of new data types, models for sharing and velocity of delivery.

- Review emerging data requirements for new data types (for example, use of wearables), delivery models beyond datasets such as streaming data, and requirements for near real time access.
- Develop a data management lifecycle model to address the requirements of new, emerging data modalities.

[39] https://www.doi.org/

- To facilitate the development and support of streaming data modalities, develop a Minimum Viable Product (MVP) service to support near real time flow of IoT data such as wearables using streaming technology.

3. Develop guidelines on privacy enhancing technologies (PETs).

- In collaboration with existing initiatives (for example, the Information Commissioner's Office, the United Nations, the Royal Society, the Turing Institute) 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 in DARE UK Phase 1[40].
- Execute a focused 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 selected 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.
- Develop or select a reference implementation of a metadata catalogue that that can support the metadata specifications and use of digital object identifiers (DOIs).
- Define a federated registry to hold a list of available catalogues, standards, vocabularies/terminologies.

5. Leverage existing digital object identifier (DOI) minting services to provide persistent identifiers for all UKRI Council resources at UKRI-wide and council levels.

- Provide a central UKRI-council level service and guidance of UKRI data custodians.
- Federate and syndicate large data custodians that already have DOIs assigned.
- Review option to mandate that all UKRI-councils have all resource metadata post-investment registered either at the central UKRI level or at the UKRI-council level.

[40] DARE UK. PRiAM: Privacy Risk Assessment Methodology. Accessed 14.07.2022.

8. Core federation services

8.1. 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 joined-up, 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, 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. Without carefully guided, strategic investment, this is unlikely to provide the optimal approach for a UK-wide capability to support cross-domain research on sensitive data. There is a real risk that the current evolution will result in an even more fragmented landscape without a trustworthy model for accreditation nor effective interoperability. DARE UK needs to address this by providing leadership to ensure that an open, 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.

8.2. Existing challenges and opportunities

As outlined above, this analysis has benefited from a wide range of input from the data research community. This 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 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 wishing to access data from multiple environments face hurdles in terms of duplicative request applications and delays, whilst researchers wishing to carry out cross-disciplinary research must work out from scratch how to access the data for each project, instead of creating legacy data linkages for the wider research community. 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 critical that investment is made to provide best-of-breed analytics capability that can be used across a federated network of research environments. The research community also 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, and for this to be provisioned with the UK

in a cost-effective way, it needs to integrate with a federated infrastructure. 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 around supporting **'safe return'**, for example i.e., the ability to provide a service that would securely reidentify records so that identifiable information can be returned to a data custodian in exceptional circumstances; for example, if a researcher were to identify someone as being at risk of an adverse health outcome. Providing such as a service is technically complex, requiring approaches to data linkage and de-identification that can support reidentification whilst still ensuring security; the governance processes also need further definition. This isn't proposed as a package of work for DARE UK Phase 2, but could be considered for Phase 3 based on the deployment of a third-party linkage service.

- There was strong support in DARE UK Phase 1 stakeholder discussions and workshops for the development of an open but formally governed – perhaps using an existing framework such as from the Apache Software Foundation – community led project to build a **reference architecture for TREs** that could be deployed using cloud native technologies[41]. This would then support integration with the 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 the work by the Alan Turing Institute[42], Microsoft[43] as well as output from the DARE UK Phase 1 Sprint Exemplar Projects. 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 implementation 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, currently, 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 the infrastructures could be recovered with a move to a more software-defined infrastructure model. In addition, some felt that the data custodian holds responsibility for re-provisioning the primary data and the researchers for re-provisioning their research artefacts, and any investments would therefore likely be disproportionate to risk. 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 that this could be through a centralised service. There was also more consensus that further study was needed and that there

[41] https://www.infoworld.com/article/3281046/what-is-cloud-native-the-modern-way-to-develop-software.html
[42] https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud
[43] https://github.com/microsoft/AzureTRE

should be a strategy to cover business continuity and disaster recovery requirements for a federated network of TREs.

- 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 respondents 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 potential 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 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.
- The public input – including from the recent DARE UK public dialogue (see Chapter 4) has been overwhelmingly in favour of research on sensitive data, provided that it is undertaken securely, transparently and with clear 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 will require 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.

Participants of the DARE UK Phase 1 initial Landscape Review said:

> *"In the longer term, I'm interested in how TREs work together - transferring data from one to another… we're keen DARE can delivery productive facilities."*
> **Technologist, research council**
>
> *"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**
>
> *"When we remove data from a TRE, the data review should be on basis on the script that created it, not eyeballing the extracted data. i.e., auditing the computer programme."*
> **Technologist, university**
>
> *"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**

8.3. A federated network of cross-domain trusted research environments

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, for example, providing capability to securely link health, crime, housing, education, environmental, consumer and retails data. Federation could also fulfil specific use cases 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 collaborate 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 industry and the public sector, with public involvement and engagement embedded throughout, 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 point of control. To be successful, the deliverables must enhance democratisation of access to data and infrastructure, addressing 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.

A federated infrastructure should not be delivered through a net-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 around capacity building and a carbon neutral future for research and research infrastructure. This work should also align closely with other major investment – such as the NHS Federated Data Platform – to ensure interoperability.

Phase 2 of the DARE UK programme needs to design and deliver proof-of-concept deployment of a scalable set of 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 that 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 to support federated analysis across the research domain. Briefly outlined below are the key areas to be addressed.

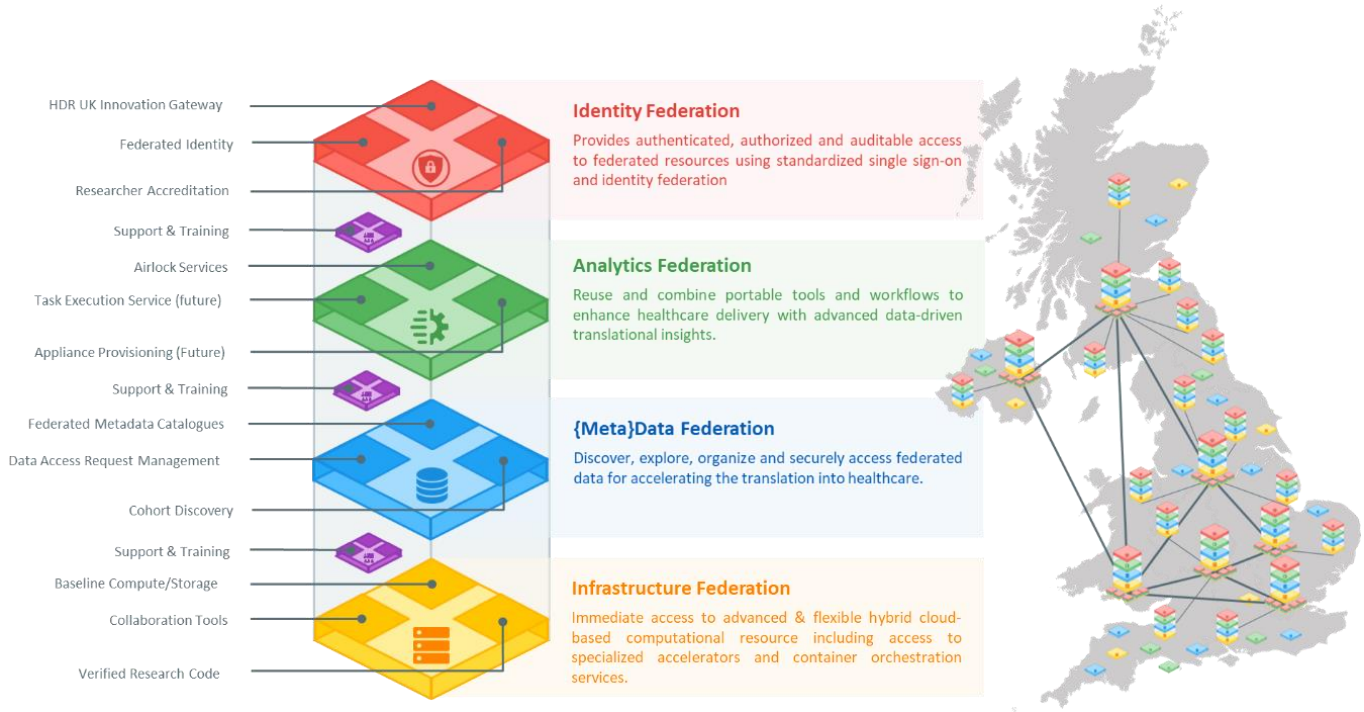

Figure 1: Overview of Core Federation Services

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 5.

There are a wide range of existing initiatives in this space, such as: the UKRI-funded JISC National AAAI Framework for Researchers[44] ; open-source projects including the widely adopted Keycloak platform[45] ; and commercially available offerings such as mvine[46], 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.

Analytics federation

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 for example Docker[47] and Kubernetes[48]). 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

[44] https://www.jisc.ac.uk/rd/projects/national-aaai-framework-for-researchers
[45] https://www.keycloak.org/
[46] https://www.mvine.com/identity-access-management.html
[47] https://www.docker.com
[48] https://kubernetes.io/

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 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[49], 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 programme should not, however, look to deliver the AI research or algorithms themselves, but focus on enabling capabilities.

Metadata federation

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 as appropriate to their requirements.

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[50]. The programme should also look at emerging open-source metadata distribution projects such as the Linux Foundation Egeria Project[51], and novel discovery and visualisation, an example of which is the Linked Data Explorer from Agrimetrics[52].

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 implementations (blueprints)

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 practice on implementation, operation, and integration. It is proposed that DARE UK Phase 2 should develop a standard blueprint (or reference architecture) that can be used for the development of these new environments that is already fully integrated with the core federation services discussed in this chapter. This reference architecture should be supplemented with an open-source, cloud-native reference implementation building on learnings from the DARE UK Phase 1 Sprint Exemplar Projects, and work by partners and industry. Any reference implementation 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 providers.

[49] https://www.ga4gh.org/how-we-work/2020-2021-roadmap/2020-2021-roadmap-part-ii/cloud-2020-2021-roadmap/
[50] https://www.healthdatagateway.org/
[51] https://egeria-project.org/
[52] https://app.agrimetrics.co.uk/linked-data

The reference architecture should include both technical capability and an integrated governance framework. This activity should be built on the existing work underway in the Alan Turing Institute, 'Data safe havens in the cloud' project[53], the work by Microsoft to provide an open-source TRE on Azure[54] and other initiatives such as the DARE UK Phase 1 TREEHOUSE Sprint Exemplar Project[55]. The architecture should fully support the ONS Five Safes model with appropriate controls. The implementation should be containerised to allow for deployment within existing infrastructures as well as to be deployed onto public cloud.

This TRE blueprint should then be used as the basis for developing a shared TRE capability (pop-up TREs). This will cover use 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.

These blueprints will provide assets to support the scale-out of federation and ensure that the barrier to participation is low for both new and existing infrastructures. This work should be undertaken in collaboration with the equivalent activities proposed by HDR UK and other work across the community to avoid duplicated 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 that covers the whole of the data pipeline from provisioning from data custodians through to deployment within the TRE network and on 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' data[56] and near real-time data, 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 as well as integrating into more recent technologies such as event streaming. This will require new methods to integrate traditional approaches to data management with those that follow a publish/subscribe pattern[57], for example over Kafka[58] or MQTT[59] infrastructure.

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. The successful development of these capabilities will be central to enabling cross-domain research use cases. This service could also be the basis for later support for Safe Return.

[53] https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud
[54] https://github.com/microsoft/AzureTRE
[55] https://dareuk.org.uk/sprint-exemplar-project-treehoose/
[56] https://en.wikipedia.org/wiki/Internet_of_things
[57] https://en.wikipedia.org/wiki/Publish%E2%80%93subscribe_pattern
[58] https://kafka.apache.org/
[59] https://mqtt.org/

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 researcher community to understand this area and in particular the linkage requirements for cross-domain use cases.

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 climate 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 in the on the importance of investment in this area, with some respondents 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) and Recovery Point Objective (RPO) expectations[60].

Some environments have already implemented High Availability (HA) support, and this is adequate where the loss of an environment that 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 a 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 elsewhere in this

[60] https://www.rubrik.com/blog/technology/19/5/rpo-rto-disaster-recovery

chapter. 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 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 10. 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 operation 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 during DARE UK Phase 1 – from NERC in particular – was intermittent access to large-scale compute. This has also subsequently been raised by the artificial intelligence and machine learning community. 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.

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 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[61]. Flexible access should extend not just to HPC/HTC capability but to providing efficient and cost-effective access to specialist compute such as GPUs[62] and IPUs[63].

[61] https://www.ukri.org/what-we-offer/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/ukri-position-on-next-phase-of-large-scale-compute-investments/
[62] https://www.nvidia.com/en-gb/data-center/
[63] https://www.graphcore.ai/ipus-in-the-cloud

Next generation statistical disclosure control

There are already significant issues with staffing resources to support statistical disclosure control (Safe Outputs). This staffing issue is already acting as a barrier to scaling up the use of TREs for research, and work is therefore be 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 address this skills shortage are covered in Chapter 9. This section briefly outlines the technology and supporting governance-led approaches that could be used to supplement staff resources.

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 Phase 1 Project, PRiAM[64] 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 ensure skilled personnel are used with areas of most significant risk. This should explore the extensive existing research and work that has been undertaken in the area of developing tools for automations, such as those developed by PRiAM and Eurostat[65].

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 GitHub[66].

This is an urgent area of need and should be prioritised for DARE UK Phase 2 investment.

This work will require detailed public scrutiny to ensure the approaches are technically rigorous, 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 of the scope of DARE UK.

[64] https://dareuk.org.uk/sprint-exemplar-project-priam/
[65] https://op.europa.eu/en/publication-detail/-/publication/2e1b3f08-2fbb-11ec-bd8e-01aa75ed71a1/language-en
[66] https://github.com/sdcTools

8.4. Preparing for production deployment

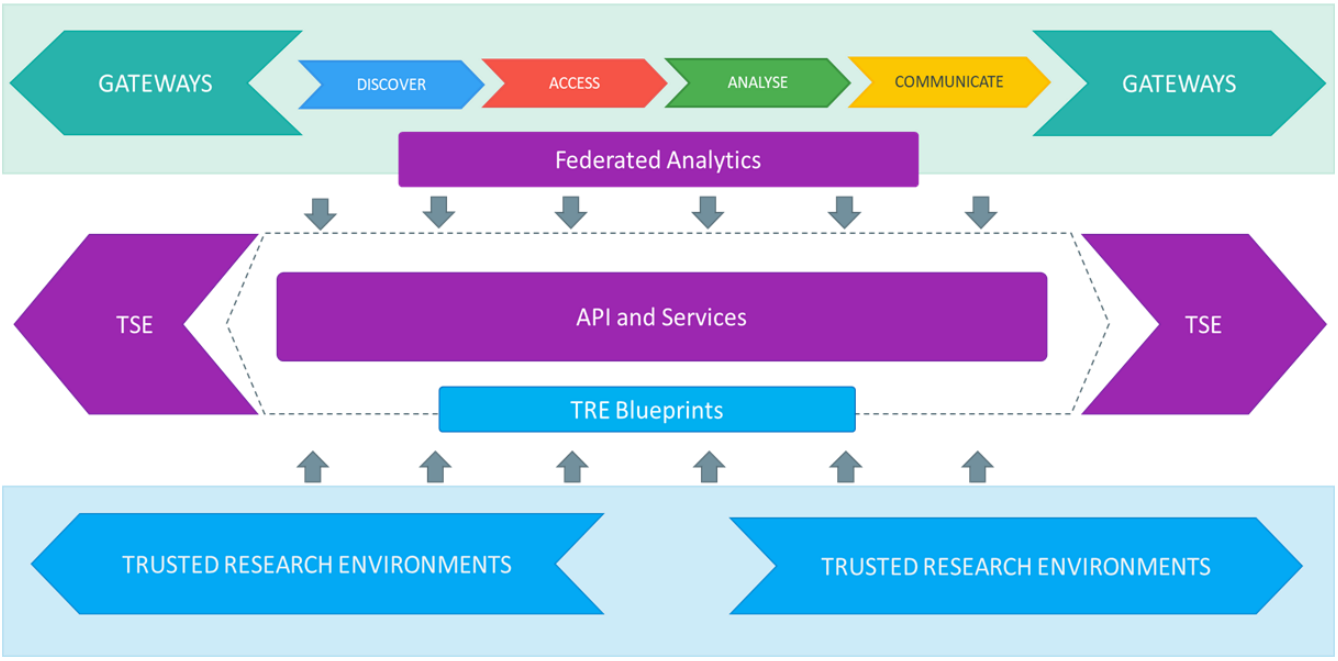

Figure 2: High level environment for deployment

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.

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.

One of the risks of assembling these core low level services – and the high-level services discussed above – from existing capability, will be the lack of consistency. Therefore, investment will be needed to ensure a consistency of design across the library of services and APIs, including documents and samples. This work will also need to ensure that 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.

8.5. Driver projects

It will be essential for the programme to ensure that its work is based on design thinking principles and guided by the requirements of the research communities working with cross-domain sensitive data. This will need to include

researchers from across the UKRI councils' domains, as well as academia, the public sector and industry to ensure that new capabilities can be generalised across a wider range of data and use cases. All proposed work should be associated with driver projects that can actively participate in the co-design of the capability and validate its usefulness in support research.

In addition to the pilot and proof-of-concept work in DARE UK Phase 2, there therefore needs to be a programme of driver projects that tests and validates the new infrastructure capability. These could be delivered as a competitive call with projects starting later in Phase 2 and continuing in Phase 3. It will not 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.

8.6. 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.

Key partnerships

The following key 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 practices and capabilities.

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

Data custodians – 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[67], including those related to data infrastructure; large-scale computing; skills and career pathways; and foundational tools, techniques and practices.

[67] https://www.ukri.org/what-we-offer/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/

FAIRness and Levelling Up the UK

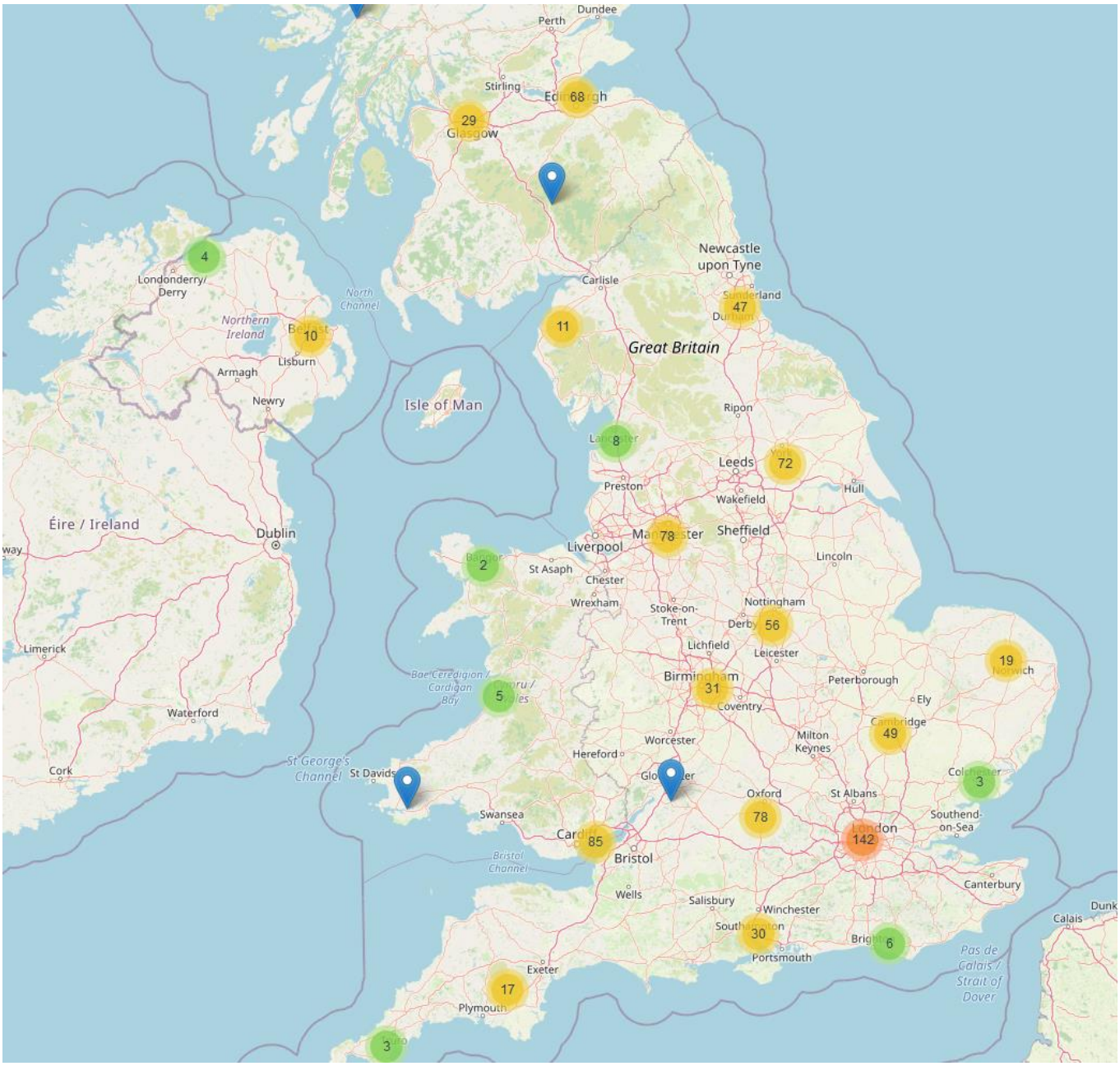

Figure 3: Distribution of UKRI Infrastructure[68]

Not only should the Core Federation Services deliver against the DARE UK aims outlined at the start of this chapter, they should also unpin 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 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.

[68] https://www.infraportal.org.uk/Searchmap

Borrowing from the well-established FAIR principles for data and metadata[69], these can also be applied with little modification to infrastructure:

Findable – building from the UKRI Infrastructure Portal[70], 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.

8.7. 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 initiative from both the UK and Internationally..

1. Develop reference architectures for TREs

- Develop a reference architecture and open implementation for a 'TRE in a Box' using open-source technologies suitable for deployment on-premise or on public cloud, and that would 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 reference 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.

2. Assemble an API library to support Core Federation Services

- Design and assemble an open reference API library to support core federations 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.

[69] https://www.go-fair.org/fair-principles/
[70] https://www.infraportal.org.uk/

3. Run a competitive call for driver projects to utilise the new infrastructure services and to 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.

9. Capability and capacity

9.1. Context

Data research is underpinned by the staff providing data preparation, curation, linkage and analysis, and by those developing and supporting the digital infrastructure.

This chapter addresses the challenges of the academic and third-sector communities in the recruitment and retention of staff. 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 this as it was a key area of concern from many stakeholders. Some of the potential solutions may be a mix of recruitment, training, and technology and are therefore in scope for the next Phase of the programme. This is particularly important in the current climate where there is a significant shortage of skills and recruitment in the public and third sectors.

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 infrastructure operational staff, especially with skills in modern cloud computing technologies.
- 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.
- Information governance specialists to support data access management and ethics approvals.
- 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.

It was also 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 learning development. There was also strong support for use of internships, secondment between organisations (including between the public sector and industry) to share skills and best practices, and expansion of apprenticeships, especially at Level 6 and Level 7[71].

"I see availability of staff throttling the work that can be done."
Technologist, data research centre

[71] Level 6 is an apprenticeship at Bachelor's degree level, and Level 7 is at Master's degree level.

9.2. 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 to develop and support world-class data research infrastructure for the UK. The public-funded research sector is in competition with industry, which has the advantage of more established career pathways, higher salaries and better job security (fixed terms contacts were identified as a major risk factor), and often also 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 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 and third sectors, 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 scheme, flexible working and a perceived less intense environment 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 industry, 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 industry. 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 here.

- 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[72]. 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[73], 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 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 often **roles are 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 to 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 on the need to improve the **flow of skills between industry and academia** more generally. More consistent role descriptions and career paths were seen has 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[74] and AstraZeneca[75]. 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

[72] https://www.10000blackinterns.com/
[73] (https://eu.detroitnews.com/story/business/2022/02/21/employers-try-skipping-resumes-improve-diversity-hiring/6879317001/
[74] https://discover-now.co.uk/
[75] https://www.astrazeneca.co.uk/

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 current 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 focussed on digital requirements and is not holistic.

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 retrain staff returning from other roles or career breaks. The key requirements identified were around cyber security, public involvement and engagement and output checking. However, many different skills areas were also raised during discussions.

There are several initiatives already in place which are beginning to address these needs, including the Hartree National Centre for Digital Innovation (HNCDI) Explain programme[76] and the HDR UK Futures platform[77]. 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 commercial sector has already engaged in specific training support, for example around development for GPUs, and further opportunity to engage industry 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[78], which is a joint initiative between the BCS, (British Computer Society) the Royal Statistical Society, the Alan Turing Institute and the National Physical Laboratory. However, the HDR UK Alliance was not mentioned in any of the stakeholder discussions, so clearly has limited traction at this point but could be a valuable initiative with UKRI support.

[76] https://www.hartree.stfc.ac.uk/Pages/Explain.aspx
[77] https://hdruklearning.csod.com
[78] https://alliancefordatascienceprofessionals.com/

Short term (6-8 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 high skilled staff to focus on the critical or highest risk areas. 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[79].

Recruitment and retention

A central issue for many institutions over recent years has been staff retention. Indeed, many stakeholders have expressed a view that retention is more of a challenge than even recruitment. It is inevitable that some staff – both those directly involved with data science and those engaged in the development and operations of infrastructure – will seek more financially rewarding opportunities in the commercial sector, so those roles within the public sector need to be made more attractive. Key to achieving this is the need to establish clear career pathways that value these highly skilled technical roles and give long-term career pathways with equivalence to research and managerial roles. To address retention, it is important that the public research domain focuses on differentiation from industry through clear, standardised career pathways, excellence in training, establishing a diverse and inclusive workforce and communicating the opportunities to engage in work that make 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 commercial sector. It is the norm within the commercial 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 careers rather than feel the need to move out to managerial or executive roles. Many organisations in industry have implemented dual career paths that are porous to allow movement between the technical and managerial pathways at several points[80]. 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.

[79] https://n8cir.org.uk/events/digital-research-infrastructure-retreat/
[80] https://medium.com/@edrisdzulkifli/encouraging-innovation-dual-ladders-self-empowerment-59f139d5f9c3

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.

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 the 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 (RSE)[81], the Software Sustainability Institute[82] and the Technicians Commitment[83]. 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 Digital Research Infrastructure initiative in UKRI 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 MI Talent[84].

9.3. 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 there is a recommendation that should be delivered by UKRI or in whole or part through the DARE UK programme, this is noted.

1. Establish clear technical career pathways 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

- 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 industry 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).

[81] https://society-rse.org/
[82] https://www.software.ac.uk/
[83] https://www.technicians.org.uk/technician-commitment
[84] https://www.mitalent.ac.uk/theTALENTcommission

- 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 career development resources and training

- Establish a pan-UKRI virtual learning environment for high quality modular training delivery, possibly via extending the HDR UK Futures Platform[85], 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 reduce the dependency on shortage skills

- 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 journey (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).

[85] Health Data Research UK. HDR UK Futures. Accessed 15.07.2022.

10. Funding and incentives

10.1. Context

The traditional investigator-based grant model that is the current structure through which research funding is granted 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 funding allocation cycles and purpose-built grant structures that are cognisant of the complexity of requirements inherent within the digital infrastructure ecosystem itself and designed with this complexity in mind. In order to nurture a UK ecosystem that balances collaboration and competition, new funding and incentive structures for providing the infrastructure and related services within the ecosystem 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 science research for public benefit.

This chapter will address 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 data research infrastructure, through the lens of research on 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 publics.

Funding and incentives for data research infrastructure, specifically for research on sensitive or potentially sensitive data, is a key focus area for the overall UKRI Digital Research Infrastructure programme's strategic vision[86]. It is crucial to consider that the nature of data research infrastructure is highly interwoven and, 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.

10.2. 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:

- There exists **limited dedicated, tailored funding allocations** for the operation, maintenance, and refresh of digital 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 research infrastructure, which are sporadic and often awarded from within specific Research Council remits rather than with a cross-domain, holistic view of the landscape in mind. There are four interdependent areas of funding to consider in this context:
  1) Hardware environments – these are the base layer ('bare metal') components (for example, CPUs, GPUs, network cables, RAM, 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) through public cloud providers and how this is categorised from a funding point of view. For the

[86] UK Research and Innovation. Digital Research Infrastructure. Accessed 13.07.2022.

purposes of this report, we will not dive 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 as explored in Chapter 9.
4) 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.

- When funding is made available, it is often not executed in a way that acknowledges 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 9). Additionally, it makes it difficult to understand holistically what the funding footprint for data 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.[87]
  - 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 – as the adage goes, you cannot steer what you cannot measure.
  - 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 8.
  - 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 data research infrastructure is funded could begin to address these challenges. However, establishing transparency across the UKRI spectrum for both where

[87] UK Research and Innovation. Moving towards net zero. Accessed 15.07.2022.

and how funding is currently allocated to data 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 – 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 available rather than fostering the 'right' questions (see Chapter 7).
  - 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.

- 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 will the providers of federated infrastructure and related services 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 not yet been established how to cohesively **integrate industry-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 inconsistency 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.

- 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?
- A lack of sustained, dedicated **funding allocation for public engagement and involvement activities** – particularly those working 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 with sensitive data research (see Chapter 4).
- 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 8. 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 where a growing number of strategic, cross-domain, high priority research areas require a means of linking sensitive data securely without cost and risk prohibitive data considerations.

It is important that this work be 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 kick start the move towards a federated infrastructure for sensitive data research, 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 8. 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 the discussion around the qualifying criteria for 'national, critical digital research infrastructure' and the need to clearly define these alongside those infrastructure that would meet these criteria.

Access to large-scale compute

There is an increasing demand from the data research community for affordable access to large-scale compute capacity, driven through several growing knowledge domains such as artificial intelligence and machine learning, as an example. 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 or potentially 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 that need 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 the absolute costs thereof combined with the 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 the community 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 custodians around making sensitive data accessible for secondary purposes, including research. This wariness is driven by their – often statutory – obligations to protect the data which they hold in their care and to ensure any secondary use of that data is ethical and in line with the legislative frameworks that govern it. 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 the increased awareness of the

existing legal framework around the secondary use of data for research (2017 Digital Economy Act) 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 custodians. 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 custodians. Ultimately, as a bare minimum standard, covering the resource costs (through license fees and so on) realised by data collectors and custodians in making their data available for research should be encouraged and supported through funding of research applications.

With reference to Chapter 7, a key concern from stakeholders within the data lifecycle is the need for a more consistent approach to data archiving and archiving capability that supports use cases within research council domains and across research domains as well. While UKRI councils all have individual approaches to data lifecycle management, stakeholders put forward that a more consistent approach across councils would be beneficial though they were clear that this should be built off existing best practice (for example the 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 key national data research infrastructure and related services 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 as the costs for these underlying capabilities are built into research grant application budgets, effectively meaning that these underlying capabilities are funded indirectly through new research grant awards. However, this is no longer an efficient approach due to a myriad of reasons that revolve primarily around the 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 data threats develop alongside technological advancements.

It should be acknowledged that there are certain research domains with a greater requirement 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 the kind of capability that can only be answered 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 a reasonable balance between establishing transparency as an ongoing activity and 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. Thus, it is rather convening 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 the taxpayer pounds spent.

10.3. 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 for 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.
    - Investigate the current advantages and disadvantages of funding flowing through 'host' organisations (for example, higher education institutions) and define how to dovetail with those existing funding streams if applicable.
    - 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 8.

2. Determine the funding requirements to establish the first phase of federated infrastructure for sensitive data research, with a focus on enabling federation across existing national data infrastructure and complimenting existing investments (with reference to Chapter 8).

- 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
  - Federated analytics
  - Metadata federation
  - Infrastructure (compute, transfer, and storage) federation

3. Investigate, test and prototype the operational model(s) for a federated ecosystem of national sensitive data research infrastructure. 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 above, 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 for sensitive data research 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 national, federated infrastructure for sensitive data research.

- In tandem with Chapter 8 recommendation 4, investigate and determine the financial feasibility of business continuity and disaster recovery scenarios and responses.
- Determine the funding requirements to pilot business continuity and disaster recovery scenarios between selected digital research infrastructure sites.
- In tandem with Chapter 8 recommendation 4, investigate and determine the financial feasibility of business continuity and disaster recovery within a federated infrastructure.
- 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 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 for sensitive data research.

- 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 7).**

- Understand and analyse current approaches across UKRI research councils, leveraging examples of best practice (for example the ESRC funded UK Data Service) 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. **Raise awareness amongst data guardians regarding the legal framework around the secondary use of data for research.**

- Develop a toolkit for data collectors and guardians regarding the legal gateways in place for making sensitive data accessible for secondary research purposes.
- Through mixed methods (e.g., information campaigns, conferences), proactively raise awareness around the provisions and operations of the legal gateways in place for making sensitive data accessible for secondary research purposes, to drive a more consistent understanding across the landscape.

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

- Improved baseline funding to support the development of sustainable, research-ready data resources from across domains and sectors – especially 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.

11. References

Aitken, M., et al. 2016a. *Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme*. Science and Public Policy, 43:5.

Aitken, M., et al. 2016b. *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.

Cameron, D., et al. 2014. Dialogue on Data: Exploring the public's views on using administrative data for research purposes. IPSOS Mori.

Centre for Data Ethics and Innovation (CDEI) 2020. Addressing trust in public sector data use.

Centre for Data Ethics and Innovation (CDEI) 2022. Public Attitudes to Data and AI Tracker Survey – Wave 1 (December 2021).

Davies, M., et al. 2018. Public attitudes to data linkage. NatCen Social Research.

Harkness, F., et al. 2022. Building a trustworthy national data research infrastructure: A UK-wide public dialogue. DARE UK.

Hopkin Van Mils 2021. Putting Good into Practice: A public dialogue on making public benefit assessments when using health and care data. National Data Guardian.

Karrar, N., et al. 2022. Improving transparency in the use of health data for research: Recommendations for a data use register standard. UK Health Data Research Alliance.

Kennedy, H., et al. 2020a. *Approaching public perceptions of datafication through the lens of inequality: a case study in public service media*. Information, Communication and Society, 24:12.

Kennedy, H., et al. 2020b. Public understanding and perceptions of data practices: a review of existing research. Living With Data.

Maxwell, M., et al. 2021. Public dialogue on location ethics: engagement report. Geospatial Commission.

Milne, R., et al. 2021. *Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries*. Genome Medicine, 13:92.

OneLondon 2020. Public deliberation in the use of health and care data.

Sheehan, M., et al. 2021. *Trust, trustworthiness and sharing patient data for research*. Journal of Medical Ethics, 47:26.

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.

Waind 2020. Trust, Security and Public Interest: Striking the Balance. ADR UK.

Appendices

Appendix 1: User personae profiles

Sharon Wakefield

44 • Domain researcher • career driven

Motivations

PUBLIC BENEFIT	* * * *
RECOGNITION	* * * *
EASE	* * * * *

Goals

- to do more impactful research by accessing & 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 science skills
- gaining access to restricted data
- poor data quality
- lack of interoperability between disparate and disjointed data
- slow workflow

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

Bio

Sandra 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.

DARE UK

Sarah Greenshaw

47 • Budget holder • building value

Motivations

SUSTAINABILITY	* * * *
GROWTH	* * * *
RECOGNITION	* * * * *

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

"I am constantly spinning plates and I have no thinking time."

Bio

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.

DARE UK

"I need to know how my data is being used."

Bio

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.

Grace Opedemi

27 • Member of the public • security focussed

Motivations

PUBLIC BENEFIT	* * * * *
DATA SECURITY	* * * * *
DATA ACCURACY	* * * *

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

DARE UK

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

Bio

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.

Pritesh Navdra

32 • Data scientist • technical

Motivations

IMPROVING SOCIETY	* * * *
CAREER DEVELOPMENT	* * * * *
REDUCED WORKLOAD	* * *

Goals

- to stay up to date with 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

DARE UK

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

Bio

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

Peter Shaw

53 • Data Custodian • process driven

Motivations

DATA SECURITY	* * * * *
RECOGNITION	* * *
MAXIMISING DATA VALUE	* * * *

Goals

- to share data with others easily and securely
- to connect with other datasets
- 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

DARE UK

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

Bio

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

Jeremy Foster

59 • Business collaborator • building value

Motivations

MAKING PROFIT	* * * *
RECOGNITION	* * * *
EASE	* * * * *

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 science skills
- gaining access to restricted data
- poor data quality
- lack of public trust in private companies
- accessing and building a relevant data community

DARE UK