UK Sensitive Data Research Infrastructure: a Landscape Review

DARE UK Delivery Team

October 2023

UK Research and Innovation
HDRUK Health Data Research UK
ADRUK Data-driven change

Contents

Executive Summary

Between November 2022 and June 2023, as part of Phase 1 of the DARE UK programme, the DARE UK Delivery Team conducted a programme of surveys, interviews and ad hoc consultations with organisations providing infrastructure and other services in support of research using sensitive data with a particular focus on trusted research environments (TREs).

Overall, we found a lively but fragmented landscape of multiple services and data providers, exhibiting islands of excellence but lacking nationwide coherence across scientific disciplines. Changes in research patterns away from data distribution and towards data access via secure services are evident in the significant growth over the last two decades in the numbers of trusted, secure digital research environments – based on survey responses received. This growth in service numbers is accompanied by the growth and increasing maturity of a community of practice, and a steady convergence of ideas and technology solutions around what makes a trusted research environment (TRE) fit for purpose.

The review highlights four key conclusions around the current state of the infrastructure landscape fit for supporting research using sensitive data:

1. The shift away from the data dissemination model towards the data access model gives rise to more infrastructures and more complex datasets. As such the "enterprise architecture" for a future secure digital research infrastructure landscape must be distributed, it must be trustworthy, and it must be flexible enough to accommodate services of very different capabilities. The landscape is vibrant, with an increase in both overall number and capability of sensitive data research infrastructures. However, it is difficult to navigate, especially across scientific disciplines where cross-domain, inter-disciplinary research remains challenging.
2. Inter-operability will, increasingly, be critical. Encouraging and supporting the development of standards for inter-operation between digital research environment service providers, of which trusted research environment (TRE) providers are a key subclass, and data providers in a necessarily distributed landscape will underpin UK plc's national capability to address inter-sectional societal challenges at pace.
3. Trustworthiness, especially within the infrastructure landscape, remains fundamental. Continuing to embrace trustworthiness, and seizing the opportunity to build public-facing information systems which collect and surface research activities making use of public data from across the landscape.
4. The challenge of delivering better inter-disciplinary research using sensitive data is a global challenge. The UK is well positioned to lead in this space with several islands of excellence around the UK, the opportunity is ensuring the whole is greater than the sum of the parts. It is important to maintain a watching brief on global developments, particularly those in Europe, leading where possible and adopting where appropriate.

The landscape of UK digital infrastructure fit for the purpose of supporting research with sensitive data has grown steadily over the years and is poised, off the back of cloud-first technology approaches and the response to inter-sectional societal challenges such as the COVID-19 pandemic, for a period of evolution and significant growth. As it stands, the landscape is fragmented and lacking national coherence across traditional research silos – though as proven through the response to the COVID-19 pandemic this national coherence is more than possible, the challenge is transforming this into business-as-usual rather than as an emergency response. Contemporaneous with this review has been the emergence of working groups across the UK, driven in no small part by the UK Research Software Engineering community, sharing intelligence and ideas. Community-driven innovation will be

key but equally community consensus on adoption to turn innovation into business-as-usual practice that will drive an uplift in the overall quality standard across the UK.

Enabling sensitive data assets to be linked and analysed at scale and at far greater pace than is currently possible has the potential for very broad scopes of inquiry spanning all manner of UK sensitive data research, with both national and international impact. As it stands, valuable UK data assets have significant potential to deliver public benefit, including to support industrial science through trustworthy collaboration with industry, but remain underutilised. Inter-operability is key, ensuring the UK landscape is greater than the sum of its parts requires a level of national coherence and coordination that demands a foundational level of join-up across the landscape, both organisational and technical. The convergence on common technology strategies across the landscape, particularly cloud-first principles and approaches, alongside the growing communities of practice within the landscape make the ambition of inter-operability far more feasible than even only a decade ago.

From a DARE UK perspective this review should not be considered final, ongoing review through a regular cadence and consistent mechanism(s) should be conducted to monitor this evolving landscape. Nevertheless, this snapshot does point to several components that are needed to deliver on the DARE UK vision for the digital research infrastructure landscape fit for supporting sensitive data research:

- Leadership: address the fragmentation of the landscape by providing coherence, thought leadership, convenorship, and an enterprise vision for a joined-up landscape.
- Technical and operational governance: providing strategic, transparent, and collective decision-making structures to oversee and ensure an inter-operable infrastructure landscape delivers value for the public, research, and UK plc.
- Standards: address inter-operability by facilitating the collective development of information governance, process, data, and technology standards to enable a foundational level of inter-operability across the landscape, ensuring the whole is greater than the sum of the parts.
- Trustworthiness: ensuring the landscape continues to demonstrate trustworthiness to the public and responsibly uphold the mandate from the public to deliver public benefit through research using sensitive data.

The prize is a world-leading UK capability that would stand to vastly outperform current practices, allowing for broad cross-domain, cross-jurisdictional analysis, accelerating basic and applied research as well as informing fast-moving national policy priorities.

The Review

This review combined information from two sources:

- A survey of 88 research infrastructures (as on 13 June 2023), of which 45 answered "yes" to the question "does your research infrastructure store or process sensitive data".
- Ad-hoc engagements with a further 23 infrastructures from which we were able to construct model answers qualitatively (though not quantitatively) equivalent to those in the survey sample.

In general, this report focuses on these 68 combined research infrastructures. Where the distinction between these two sample sets is important, we distinguish this throughout the report. Around two-fifths of the 68 organisations providing these infrastructures were universities, another two-fifths public sector bodies with the balance made up of private firms and charities.

Based on our analysis of the samples, augmented with desk research as required, we have classified each infrastructure's primary role in the landscape: as purely supplying data (data provider); as providing value-added data products or services using raw data inputs (data service providers); as providing analytical computing and data storage services for sensitive data research (trusted research environment [TRE] service providers); and as providing platforms on which TRE services can be built (TRE platform providers). Note that for the purposes of this review we have classified digital research environments providing services around sensitive data as trusted research environment (TRE) service providers. Two-thirds of the organisations were best classified as TRE service providers, 16% as TRE platform providers, 13% as data providers and the balance as data service providers.

Funding for these organisations comes from a wide variety of sources through both core funding – underpinning block funding from a single research council, for instance – and cost-recovery mechanisms. All nine UK Research Councils are involved in funding the research infrastructures themselves and the projects which use them. Distribution is fairly even, although projects associated with the MRC stand out at over three times the number of any other. UKRI as a body funds over a quarter of the organisations directly, and 69% of the organisations report funding from non-Research Council sources. These "other" sources are evenly spread across universities, government, charities and private firms, with a peak coming from health services.

Use of health data dominates the landscape. Some 85% of the research infrastructures work with health-related data, 50% work with government administrative data and around 30% with commercially sensitive data. A small number (7%) work with defence or national-security data.

Among the TRE service providers, two-thirds are operational, and one-third are planning to come on-stream over the next two to three years. The numbers of TRE services operational or planned rises steadily from a handful in 2005 to more than 40 by the end of 2025. Among currently operational TRE services, two-thirds reported on their annual usage in terms of numbers of researchers and projects. Most serve relatively small communities with a spread of 50 to 1,300 users and a median of 165; while the spread of annual active projects is 5 to 400 with a median of 50.

Around half the TRE services hold some form of formal accreditation, with two-fifths certified under ISO27001 and a handful under the UK Statistics Authority Digital Economy Act framework. Formally certified TRE services form the core set of Trusted Research Environments (TREs) across the UK.

In technology terms, most TRE services have adopted a cloud-first approach, building on virtualisation and container software services typically of public cloud services. Over half the TRE services are deployed on public cloud, a quarter on-premises and around 10% using a mix of both. A relatively small number (16%) use platform-

as-a-service environments from organisations offering specialised, usually TRE-like, infrastructure. The vast majority of TRE services planned for over the next two to three years use public cloud infrastructure.

The maturity of the research infrastructure landscape for sensitive data varies across the devolved nations of the UK. SAIL Databank in Wales is the most mature resource, combining health and demographic data for the whole Welsh population in a rich TRE service. Scotland has a mature network of four regional and one national TRE, all now under the umbrella of a new coordinating body, Research Data Scotland. Northern Ireland leverages the Secure eResearch platform developed by Swansea University to support SAIL but now operating as a TRE platform for UK and international tenants.

In England, administrative data is well-served by the ONS Secure Research Service and the newer Integrated Data Service, while infrastructure for health data is much more fragmented. The British Heart Foundation Data Science Centre enables access to national data and is an exemplar for the relatively new NHS England National Secure Data Environment (SDE). Development of a complementary network of eleven NHS regional SDEs is underway. It is currently unclear how these might integrate with non-health data services.

Finally, we observe that the UK is not alone in investing in secure, digital research infrastructure for work with sensitive data, but it is in a position of leadership. Work in continental Europe, especially among the Nordic nations, is impressive (and to be tracked carefully), but nothing to date has been attempted at the scale of the UK population. This presents a significant opportunity for UK plc to build world-leading research capability to match its world-leading stocks of public data.

1. Introduction

The UK has vast data assets of a sensitive nature that through appropriately managed access, analysis and linkage stand to vastly improve the lives of people and to inform better, deeper, and timelier governmental policy decisions. There exist to date several 'islands of excellence' and 'flagship' investments that enable sensitive data research and that have provided remarkable insights. Examples include, in the health data space, Genomics England [1] and UK Biobank [2], and in the administrative data space, vital analyses conducted by the Office for National Statistics (ONS) [3]. However, only a fraction of all UK sensitive data assets are accessible, and researchers are increasingly impatient to see successful extraction of value from this data. The fragmentation of the sensitive data assets and the access challenge, especially the difficulty of linking data from different sources, presents a risk to UK competitiveness, limits research productivity, and lowers the return on investment of public research funding.

This report is a snapshot of the UK's provision of digital infrastructure to support public research with sensitive data. It follows on from the 2021 DARE UK report "Data Research Infrastructure Landscape" [4] which conducted a broad-based survey of digital research infrastructure and helped form the basis for DARE UK's Phase 1a recommendations [5]. We also acknowledge timely publication of the much broader review of the UK's research sector by Sir Paul Nurse [6] which offers a comprehensive framing of this (more limited) landscape review.

The aim of this report is to drill further into the "sensitive data" part of the data research infrastructure landscape using a mix of surveys, interviews and desk research (see *Methods* below). The primary focus of the report has been on infrastructure supporting publicly funded, rather than private, research use, often qualified further (especially with regards to the use of individual-level public data) as "research in the public benefit". Large, privately operated research infrastructures (those internal to pharmaceutical firms, for instance) are thus out of scope for this report.

The raw results from this work have already fed directly into the first version of the DARE UK "Federated Architecture Blueprint" [7].

1.1. Why This Matters

"Sensitive data" is a broad class of data for which appropriate precautions must be taken when making available for research. These types of datasets might contain confidential information about living natural persons (e.g., data about citizens such as various government-held records or behavioural data through internet or social media interactions) or legal persons (copyright, business financial or intellectual property data), or about sensitive locations (sites of national infrastructure or of endangered biodiversity); whatever the reason for their sensitivity they need to be handled in ways that maintain the trust of the data owners in question.

Much sensitive data about UK citizens is recorded by government but this is by no means the only potential source of data for research. Take research into dementia and other neurodegenerative diseases. A 2015 report from the OECD [8] highlighted the potential for using "big data" – data from social media, supermarket loyalty cards, mobile phone location data – linked with health data as a powerful tool to detect subtle behavioural changes that could help to predict the onset of dementia. For any given individual, the linkage of such data would be incredibly sensitive but might result in an early diagnosis of – and potentially life-changing treatment for – one of the developed world's great challenges. Not only are individual sensitivities involved: social media, supermarket and phone data are all potentially sensitive commercial properties of a variety of firms. The solution has the potential to be revolutionary, but the challenge is significant.

Access to sensitive data is a powerful enabler of research with vast potential for public benefit, and access to sensitive datasets that are linked together provides an even more powerful research tool. As an example, a retrospective cohort study [9], linking police domestic abuse data and routinely collected health data, indicated that vulnerable individuals are detectable in multiple data sets before and after involvement of police, demonstrating the possibility of targeted interventions that might temper the escalation of domestic abuse and related health outcomes. This is an example of the excellent instances of richly linked population data within the UK, in this case the SAIL Databank at Swansea University [10], covering the whole Welsh population. There is currently no equivalent of this at UK scale. Data linkage for other regions or populations can be done case-by-case on a per-project basis but this *ad hoc* approach is inefficient and often repetitive.

With increased linkage of individual-level sensitive data comes a greater responsibility on researchers and supporting research infrastructures to be as transparent as possible. The DARE UK public dialogue [11] showed that there is widespread public support for data research and the public are generally reassured by processes currently in place to protect their data. Moreover, they wanted all types of researchers – including commercial organisations – to have access to their data when the proposed research is in the public benefit. Mechanisms to increase the visibility of sensitive research and maintain these levels of public trust will become ever-more important as the landscape develops.

Research with sensitive data is shifting from a data distribution model, where researchers download de-identified data to their local systems, to a data access model, where researchers access data remotely within a secure computing environment – commonly referred to as a trusted research environment (TRE) or secure data environment (SDE). This shift is highlighted – and strongly encouraged – by the growing consensus (including from the DHSC Data Saves Lives policy paper [12], Phase 1 DARE UK recommendations, "*Better, broader, safer*" the review of health data use by Professor Ben Goldacre [13], and the UK Health Data Research Alliance TRE Principles and Best practice paper [14]) that all sensitive data should only ever be accessed and analysed by researchers within a TRE. The data access model increases the security around sensitive data: TREs are designed with data safety in mind, disabling download, managing available software tools and requiring researchers to complete appropriate training[1]. There is however a tension that needs to be managed in ensuring that the security and privacy measures that a TRE enables are maintained while avoiding creating silos of data that raise new barriers to research with the potential to deliver significant public benefit, especially as regards to data linkage.

The landscape itself is changing rapidly. The COVID-19 pandemic triggered shifts in the sharing for research purposes of both health and administrative data, resulting in new initiatives, new infrastructure and new ways of thinking about the longer term. TREs across the UK, brought together under the HDR-UK Data and Connectivity programme [15] worked together in a new way; new TREs emerged (OpenSAFELY [16]; the ISARIC4C Outbreak Data Analysis Platform [17]); and the UK Government published the "Data Saves Lives" strategy for healthcare data in England. The latter has triggered the National Health Service in England to develop plans for a network of "sub-national" secure data environments (SDEs; synonymous with TREs), a work in progress and a significant change to the sensitive data research landscape [18]. The UK Government has recently instigated a review into flows of health-relevant data across the UK by Professor Cathie Sudlow, HDR-UK's Chief Scientist [19].

A consequence of this rapidly evolving landscape is that this report, alongside other such efforts to understand TRE capabilities [20], may become outdated relatively quickly. Nevertheless, we believe it offers a valuable picture

[1] In this report we use the term "TRE" alongside digital research environment. At the research infrastructure level, in our context of supporting research with sensitive data, there is no appreciable difference between a TRE and a digital research environment and we occasionally use the terms interchangeably. However, TREs should be regarded not just as research infrastructures but as accredited services operating within a broad safety environment governed by the Five Safes approach. Where the distinction is important, we highlight it.

of the UK's digital sensitive data infrastructure landscape at the beginning of what could be a transformational period for research with sensitive data.

1.2. Methods

This review combines information from two sources: a self-reporting survey of UK research infrastructures conducted by the DARE UK programme from November 2022 to May 2023; and a series of ad hoc engagements and conversations with additional organisations and individuals over the first half of 2023. Both sources were augmented by desk research as required.

1.2.1. The survey

Between November 2022 and May 2023 we ran a survey of the UK's digital research infrastructures. This was conducted online (using Qualtrics) and combined with a series of clarifying follow-up interviews. The survey was open to research infrastructures in general, although our particular focus in this report is infrastructures handling "sensitive data" for research.

At a cutoff date of June 13, 2023, we had 88 responses. Of these, 13 were discounted as duplicates, spoiled or invalid. Of those remaining, 45 respondents answered "yes" to the question "Does your research infrastructure store or process sensitive data or does it have the intention to enable this in the future?"

These 45 respondents are our "**survey sample**".

1.2.2. Additional engagements

At the survey cutoff date a number of organisations we regarded as important to this review had not been able to respond to the survey. Therefore, through a series of ad hoc follow-up engagements and research of publicly available information we created model answers for a further 23 organisations, qualitatively equivalent to many – though not all – of the survey questions. We have not attempted to create model answers for quantitative survey questions; these remain blank.

These 23 organisations are our "**additional sample**".

1.2.3. Combining results

In the following report, where answers from the survey sample and additional sample are qualitatively equivalent we have combined the two into an "**extended sample**" of 68. Where they are not, we note this, and report only on the survey sample.

2. Organisations and their Roles

The sensitive data research landscape is a landscape of capabilities, of digital infrastructures, services and data. These capabilities are provided by different kinds of organisation – meaning, broadly, different kinds of legal entity. In classifying the types of organisations active in this landscape we have chosen to align as closely as possible with the types in the Nurse Review [6].

2.1. Types of Organisations

Results in this section draw from our "extended sample".

2.1.1. Universities

As noted in the Nurse Review universities remain the backbone of research activity within the UK, and this is true of research in the sensitive data space too. Universities are also very active in the provision of infrastructure-level services to support research in this area; around 40% of our extended sample are universities.

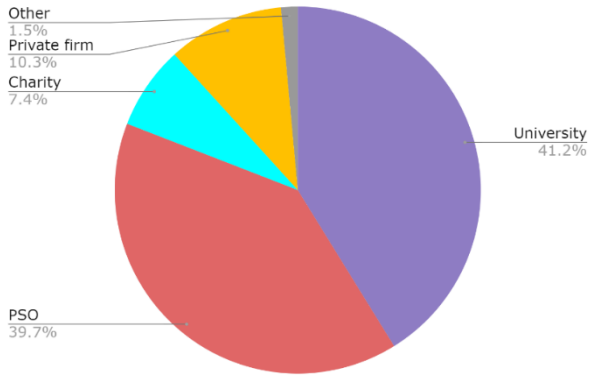

Figure i: Types of organisations. (PSO = public sector organisation.)

2.1.2. Public sector organisations (PSOs)

Public sector organisations cover a broad range of correspondents in our samples, from hospital trusts to national laboratories and government agencies. We have chosen not to break them down further this way but rather to focus on the roles they play in the wider landscape. Altogether, another 40% of our extended sample are PSOs.

2.1.3. Private firms

Private firms represent around 11% of our extended sample. In this review we have deliberately restricted ourselves to firms who provide infrastructure-level platforms or services for TREs or other organisations and have not surveyed firms who provide tools or application-level software. We are well aware that there is a much broader set of private sector actors involved in software tools and services relevant in the wider "sensitive data ecosystem" but not covered by this review.

2.1.4. Charities

The Nurse Review uses an organisational category of "Institute or Independent Research Organisation" to include charities. We have chosen to call out charities explicitly in our review. Around 7.5% of our extended sample are charities (excluding universities). Charity funding in this space, particularly around health-related research, is significant with big investments from large charities such as Wellcome, the British Heart Foundation, Cancer Research UK and the UK's big Alzheimer's research charities.

2.1.5. Others

"Others" is a catch-all category covering organisations that do not fall into one of the above. We use it to include large, multi-year but finite entities such as projects which do not have their own legal identity. Only around 1.5% of our extended sample fall into this category.

2.2. Roles in the Landscape

The focus of this review is less on the nature of the organisations that make up the landscape and much more on the roles they play in providing data or services to enable research to happen. To enable us to take this view we have classified infrastructures into a small number of types.

Because there was no specific survey question asking infrastructures to classify themselves we have interpreted free-text answers and public service descriptions to choose a principal role taken by each infrastructure. In most cases this was straightforward (e.g. from statements such as “we do not provide access to data”). In some cases the distinction between the role of service provider versus platform provider, for example, was nuanced. We are comfortable that the final classifications are robust enough for our purposes in this report.

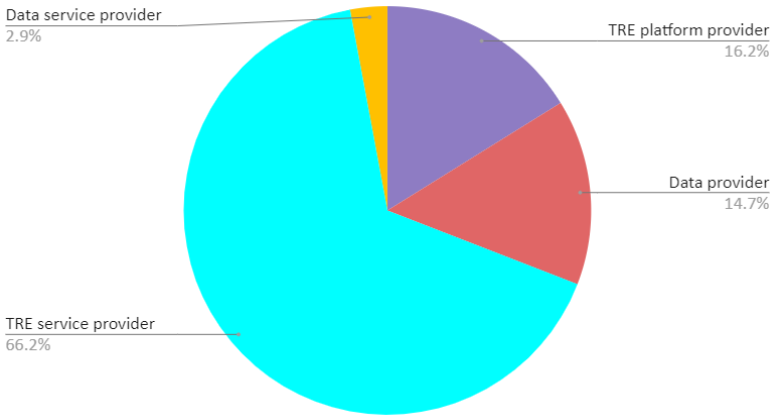

Figure ii: Primary roles within the landscape.

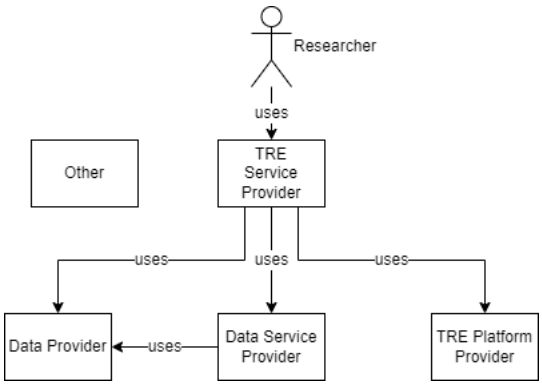

Figure iii: Relationships and dependencies between roles.

Figure ii (left) shows the proportional breakdown of the assigned primary role across our extended sample. Figure iii (right) sketches the relationships and dependencies between these roles. Briefly, these roles are:

- Trusted research environment (TRE) service provider, meaning providing a full analytical computing and data storage service directly to researchers; offering a “safe setting” in the Five Safes sense of a secure environment suitable for working with sensitive data and wraps the infrastructure as a fully managed service for researchers;
- Data provider, meaning providing sensitive datasets to researchers in various ways, but not providing analytic services alongside. A data provider may also be a data curator, preserving a dataset for the long-term;
- Data service provider, meaning providing a value-added service that makes use of data from data providers to create derived data products of additional use (e.g., a data indexing or linkage service, a cohort discovery service, a data aggregation service, and so on);
- TRE platform provider, meaning providing an infrastructure on which several independently governed or operated data services or TREs could be built (e.g., a cloud service provider). TRE platforms provide the “safe setting” aspect for a TRE service, for example, but do not offer the additional information governance capabilities needed for a full service;
- Other, meaning not one of the above.

As it turns out, for our extended sample none of them fall into the "other" category, suggesting our high-level picture of the landscape and roles in it is reasonably complete. An analysis of the landscape in terms of these roles forms the bulk of the rest of this report.

2.3. Patterns of Funding

Results in this section draw from our "extended sample".

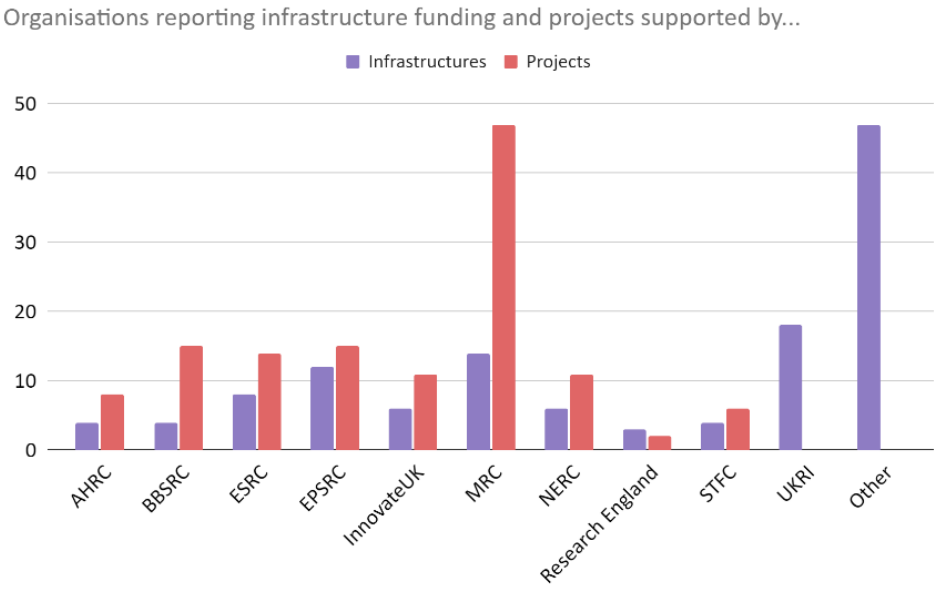

Figure iv: Infrastructure and project funding sources.

The following two charts show the numbers in our extended sample reporting funding from different sources, divided into UK research councils and "Other". The chart below (Figure iv) counts how many of our infrastructure or data-providing organisations receive funding from UK research councils (using their standard abbreviations). "UKRI" indicates funding from UK Research and Innovation rather than from a specific research council, either provided directly or channelled through a major investment programme such as the Digital Research Infrastructure programme[2]. We graph "core" funding, in response to the question "which bodies fund the infrastructure", in purple (series 1) and "project" funding, in response to the question "which research councils are your projects primarily aligned with", in red (series 2).

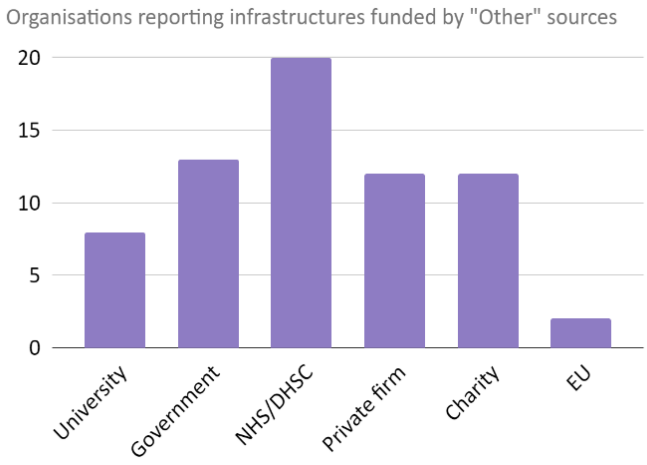

Figure v: Infrastructures funded by "Other" sources.

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

Some 69% report funding from "Other" non-research council sources; these are broken down further in the chart on the right. In this chart NHS/DHSC indicates funding from the National Health Service (NHS) or direct from the UK Government Department of Health and Social Care (DHSC) (or their Northern Ireland equivalents). Government means non-health government funding, and including both UK and devolved governments.

While the majority of our extended sample (37) report receiving funding from a single source a significant minority (30) rely on two or more (see chart right). Ten organisations rely on five or more sources.

It is possible that some infrastructures in our survey sample reported "core funding" – underpinning block funding from a single council, for instance – while others reported project funding – operating on a top-slice cost-recovery model from individual project grants, for instance. We would need to ask further questions to understand this.

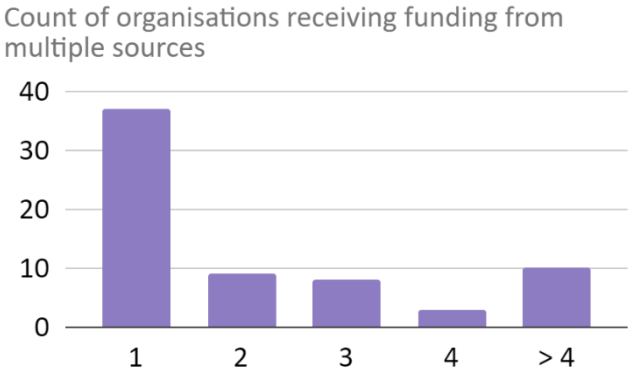

Figure vi: Organisations receiving funding from multiple sources.

The heatmap (Table 1) on the following page looks in a little more depth at infrastructure versus projects through a research council lens. The map should be read as "infrastructure with funding support from *X* hosts projects aligned with research council *Y*".

We do not have enough resolution in the data to drawn stronger conclusions; neither do we have enough information to determine whether some organisations use a cost-recovery model based on grant top slicing to cover part of their core costs. Nevertheless, the heatmap does give us more insight into the "UKRI" and "Other" infrastructure funding bars.

These charts illustrate a fairly mixed economy with all research councils playing a funding role alongside significant additional support from central government and the charity and private sector. Universities, too, support individual pieces of digital research infrastructure.

Table 1. Cross-comparison heatmap of organisations' infrastructure funding versus research council alignment of projects hosted.

	Project types supported by									
Infrastructures supported by	AHRC	BBSRC	ESRC	EPSRC	Innovate UK	MRC	NERC	RE	STFC	UKRI
AHRC	4	3	2	2	2	2	3	1	2	0
BBSRC	4	4	2	3	3	3	4	1	3	0
ESRC	4	3	6	4	2	6	4	1	2	0
EPSRC	5	5	2	11	5	5	5	2	5	0
Innovate UK	2	4	1	3	4	2	2	2	2	0
MRC	4	5	5	5	3	14	4	1	2	0
NERC	4	4	2	3	3	3	6	1	3	0
Research England (RE)	2	2	1	2	2	1	2	2	1	0
STFC	4	4	2	3	3	3	4	1	3	0
UKRI	5	6	6	7	5	12	5	2	4	0
University	0	2	3	0	2	7	0	0	0	0
Government	2	3	7	4	3	7	3	0	2	0
NHS/DHSC	0	2	2	1	3	19	0	0	0	0
Private firm	1	3	2	0	1	4	1	0	0	0
Charity	3	4	3	3	1	10	1	0	1	0
EU	0	1	0	0	0	2	1	0	0	0

3. Sensitive Data

Results in this section draw from our "extended sample".

In classifying the types of sensitive data prevalent in the landscape we chose broad categories, each with a small number of subdivisions.

It is worth highlighting that the provision of sensitive data for research invariably means the provision of "safe" data according to the Five Safes principle of the same name [21]. Sensitive data undergo de-identification, minimisation, perturbation and/or other confidentiality-enhancing techniques before being made available to researchers – and increasingly, being made available to researchers only within a TRE or equivalently secure research space.

Our sensitive data categories are:

- Health data, divided into
    - Primary care (secondary use of routine GP or community health data, typically unconsented but with opt-out);
    - Secondary care (secondary use of routine hospital data, again typically unconsented but with opt-out);
    - Genomics (primary capture from detailed lab analyses of individual genomes, typically provided by consent); and,
    - Clinical trials (health outcomes in the context of new treatments, typically provided by consent).
- Administrative data, divided into
    - Employment, welfare, social care and deprivation (secondary use of routine government data and derived products, typically unconsented but with opt-out);
    - Education (secondary use of data from schools, again typically unconsented but with opt-out);
    - Financial (secondary use of personal financial data, more often that of "legal persons" [firms] than "natural persons"); and,
    - Surveys (detailed "microdata" from individual-level surveys, typically provided by consent).
- Commercial data, divided into
    - Intellectual property (provided for particular research purposes, perhaps under non-disclosure agreements or other contractual means); and,
    - Sales and retail (again, provided for particular research purposes, perhaps under non-disclosure agreements or other contractual means).
- Other data, divided into
    - Defence or national security (potentially data government-classified as SECRET or higher); and
    - Other (anything that doesn't fit in one of the preceding categories).

The overall breakdown of these data types stored or processed by research infrastructures is shown in Table 2.

Table 2. Numbers of research infrastructures storing or processing data of certain "sensitive types".

	N = 68	
Any health data	58	85.3%
Primary care	41	60.3%
Secondary care	47	69.1%
Genomics	28	41.2%
Clinical trials	22	32.4%
Any administrative data	34	50.0%
Deprivation | Welfare | Social care | Employment	26	38.2%
Education	17	25.0%
Financial	11	16.2%
Surveys	20	29.4%
Any commercial data	20	29.4%
Intellectual Property	17	25.0%
Sales or Retail	6	8.8%
Any other data	5	7.4%
Defence or National Security related	5	7.4%
Other	5	7.4%

While fewer than a third of our extended sample handle commercial data, and slightly under a half handle administrative data, a substantial 85% work with health data of some kind. While there is a bias in the additional sample (the newly announced NHS secure data environments make up fully one half of our additional sample, and hence one sixth of the extended sample total) it does point to the importance of health data research as a driver of infrastructure and services in this landscape.

4. TRE Service Providers

Results in this section draw from both our “survey sample” and “extended sample”. We make clear which in the commentary below.

TRE service providers form the majority of our extended sample, some 65% overall. This group provide digital research services directly to researchers, typically computational capacity and project data storage. Some also provide long-term data hosting and curation, adopting the role of data provider as well.

Two-thirds of these TRE service providers are fully operational; the others are in development with plans to come online over the next few years. Cumulatively (Figure vii) we can see a steady increase in the number brought into service since 2005 (years 2023 onwards are expected dates). Note that the reports of TRE service providers currently in operation came exclusively through the survey, which we take to mean they are still operating. Some infrastructures are well into their second decade with even more are planned over the next few years.

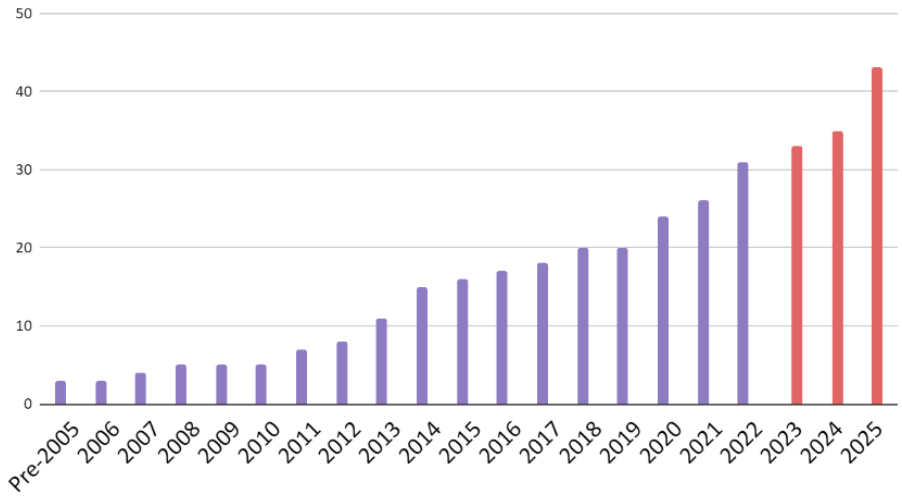

Figure vii: Cumulative count of TRE service providers by year of instantiation.

Of the TRE service providers in operation and who reported in our survey sample, their user bases range in size from 50 to 1,300 per year, with a median of 165 (see Figure viii). The typical numbers of projects supported annually ranges from 5 to 400 with a median of 50 (see Figure ix).

Almost all (95%) of TRE service providers from our survey sample reported applying information governance procedures to the research projects they support, from research accreditation panels, ethics panels, delegated authority arrangements or some combination of these. The balance of 5% can be attributed to providers who are still in development.

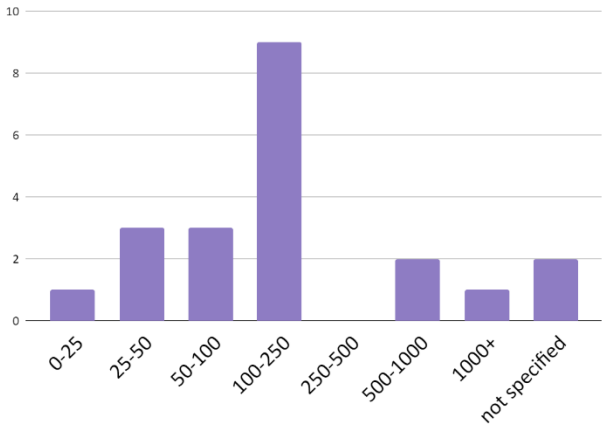

Figure ix: Counts of TRE service providers by number of users.

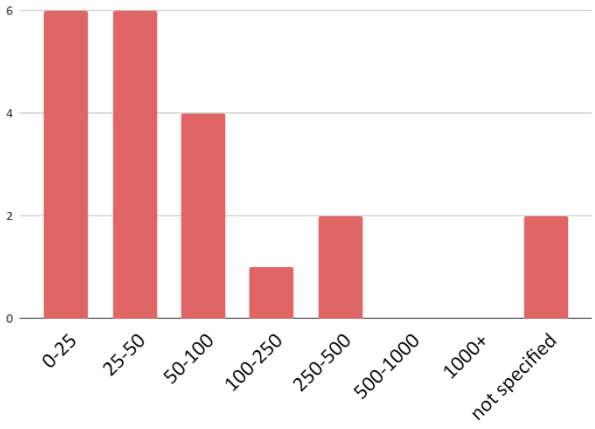

Figure viii: Counts of TRE service providers by number of projects.

A little over half of the providers from our extended sample report some sort of formal accreditation of the TRE services they offer. Nearly two-fifths of TRE service providers overall hold accreditation under ISO27001, the international standard framework for information security management. Around one-fifth each hold CyberEssentials+ or NHS Data Security and Protection Toolkit accreditation, and a handful of providers (13%) have obtained accreditation from the UK Statistics Authority under the Digital Economy Act[3].

In implementation terms, around half of the TRE service providers in our extended sample either use or plan to use public cloud as a platform, with around a quarter providing an on-premises solution and a handful taking a mixed approach (see Figure x). While many of these approaches adopt increasingly standardised software templates (otherwise known as infrastructure as code) for TREs they all make direct use of infrastructure-as-a-service (whether public cloud or on-prem). A relatively small number (16%) of providers make use of platforms-as-a-service, tailored and managed environments designed to host (usually) TREs in a multi-tenant fashion. Among the 15 TRE services that are either planned or are in current development, 13 make use of public cloud service providers and one plans a hybrid on-premises/public cloud approach (one correspondent was unspecified).

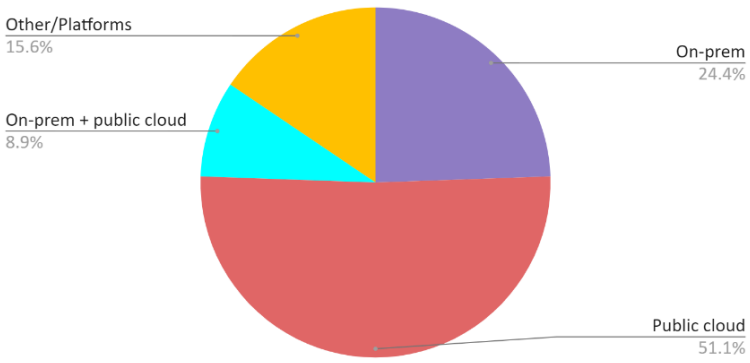

Figure x: Share of TRE service providers implementation approach

Our survey did not look in-depth at detailed capabilities but did ask a few high-level questions on support for linking to external data, support for bespoke or customisable software environments for researchers and provision of "advanced" hardware such as high-performance computing or GPU capability (often a prerequisite for AI and machine learning support). Two-thirds of TRE service providers in our survey sample support some form of data linkage, either project-by-project or as a curated data resource more akin to the data provider model. The other capabilities were fairly evenly distributed between yes and no, or not yet.

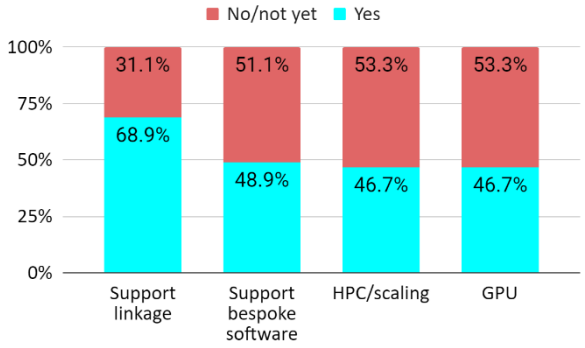

Figure xi: High level capabilities summary

We have no equivalent data to assess capabilities from our additional sample.

In the overview of data types in Chapter 0, provision and use of health data featured strongly. This picture is even more pronounced among TRE service providers: over 90% of our extended sample handle, or plan to handle, health data. A little under a half of TRE service providers support research with administrative data and a fifth with commercial data of some form. A handful work with other kinds of sensitive data – defence, biodiversity, natural history.

[3] See https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/list-of-digital-economy-act-accredited-processing-environments/

5. TRE Platform Providers

Results in this section draw from our "extended sample".

TRE platforms provide the underpinning computing, data storage and software infrastructure on which services can be built. This category includes both providers of infrastructure-as-a-service – generic public cloud with build-your-own-TRE software, for example – and providers of tailored and managed TREs. Our cohort size is small – 11 correspondents from across our extended sample fall into this category – but they divide evenly into managed platforms versus the un-managed "do-it-yourself" type of platform.

The split between commercial platform providers and university or public sector providers is around 40/60. While we caution against over-interpretation of low sample numbers, it is worth remarking that the overall population of TRE platform providers is itself not large. The ability to offer general multi-tenancy services comes only from significant compute and storage capacity and this necessarily limits the field. We are comfortable that our data include a significant number – perhaps even a majority – of UK organisations with these levels of capacity (cf., Appendix 0).

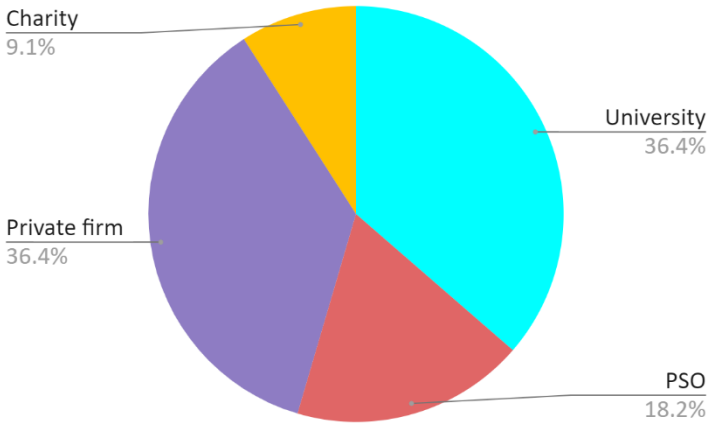

Figure xii: Types of organisations classified as TRE platform providers.

We do observe the beginnings of a maturing of the platform landscape, particularly among providers who support TREs rather than more general services. There is broad community interest in increasing the degree of standardisation for "do-it-yourself" TREs on cloud platforms[4] which in turn is shaping some of the platform offerings from the big public cloud providers[5]. Specialist managed TRE providers continue to innovate. It will be interesting to track the development of the platform space over the next 12 months.

6. Data Providers

Results in this section draw from our "survey sample".

Our "data provider" category captures "pure" data providers, organisations which supply data onward for research projects within TRE services but do not provide TRE services themselves. Note that a number of TRE service providers also act as persistent hosts and suppliers of curated data resources to their research users; for our purposes we classify them principally as TRE service providers and do not count them here.

Nine correspondents in our survey sample fit our definition of data provider. This is a low number in the context of all possible sensitive datasets across the UK; it is probable that our targeting of the original survey at research infrastructures has not sampled the broad pool of data providers very well.

[4] See, for example, the community around TRE standardisation that has emerged from the broader research software engineering community (https://rse-tre-community.readthedocs.io/en/latest/).
[5] Amazon Webservices, Microsoft Azure, *private communications*.

Table 3: Spread of data types across correspondents classified as data providers.

	N = 9			N = 9	
Any health data	5	55.6%	Any administrative data	5	55.6%
Primary care	2	22.2%	Deprivation | Welfare | Social care | Employment	4	44.4%
Secondary care	2	22.2%	Education	1	11.1%
Genomics	4	44.4%	Financial	2	22.2%
Clinical trials	1	11.1%	Surveys	4	44.4%
Any commercial data	7	77.8%	Any other data	3	33.3%
Intellectual Property	6	66.7%	Defence or National Security	1	11.1%
Sales or Retail	3	33.3%	Other	2	22.2%

The spread of data types from our nine correspondents is shown in Table 3. Bearing in mind again small sample size, the most common kind of sensitivity we see is commercial, particularly intellectual property. We might conjecture that the provision and sharing of commercial intellectual property data follows a different pattern to the sharing of individual-level health or administrative data: among correspondents handling health data (for instance), research is being carried out increasingly in TREs, whereas commercial data is still shared "off-site" under licence or data sharing agreement.

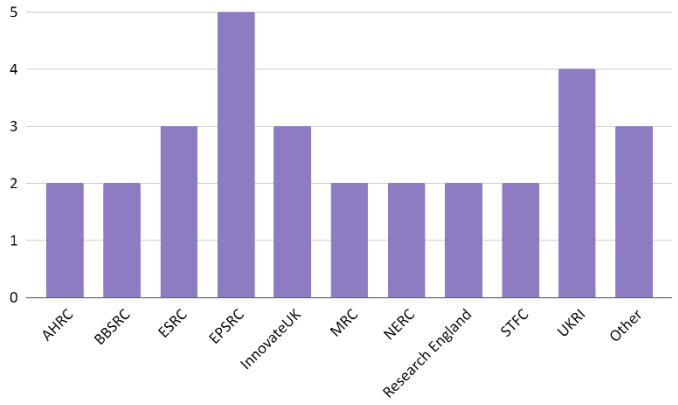

Figure xiii: Reported sources of funding for correspondents classified as data providers.

Figure xiii shows the reported sources of funding, by UK research council or other, across our data providers. "Other" typically means a large charity in our sample; the other labels use the standard UK research council abbreviations.

This shows a relatively even distribution across discipline areas (using research council as a proxy), in contrast to the dominance of health data in TRE service providers.

7. Data Service Providers

Results in this section draw from our "survey sample".

Our definition of data service provider encompasses organisations who consume data from some of our other sources – data providers or TRE service providers acting as data providers – and create value-added services or data products for onward use within the overall research space. This is a broad category and comes closest to the class of "tools and applications" that were not in scope for our infrastructure survey. Data service providers are, however, distinct from application providers in that they sit between data providers and TRE service providers rather than offering software tools to be used within a TRE service.

This is our smallest category in this analysis, with only three correspondents. Two of these are private firms offering innovative data services in the health data space; the third is the Office for National Statistics in its role as a principal trusted third-party data indexing service for many administrative data research projects.

The innovative nature of the services offered by the two private firms suggests that this type of role may see future growth. As with the TRE platform providers it will be interesting to track the development of the data service space over the next 12 months.

8. The UK Nations

Results in this section draw from our "extended sample".

When viewed through the lens of sensitive data research infrastructure the four nations of the UK are at quite different stages of development. Given the significant influence that health data research has on the overall landscape (passim) this may in part reflect the different structures of the UK's National Health Service (the main provider of health data for research) across the nations.

Our correspondents come from all four nations (see Figure xiv). Our picture of correspondents' nationalities matches the relative populations of the four nations reasonably well.

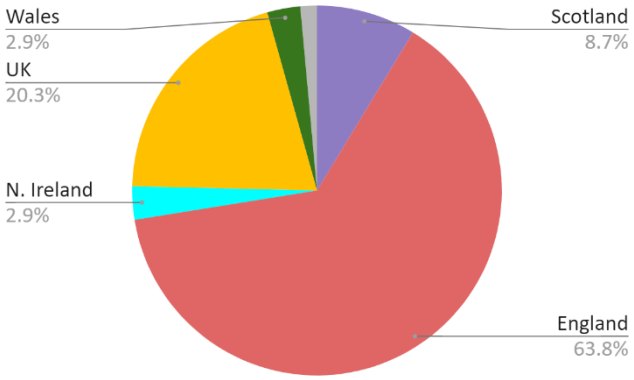

Figure xiv: Correspondents distribution across the four nations and UK-wide.

For some correspondents, classifying them by nation was straightforward – the Scottish National Safe Haven is hard to mis-place, for instance – but for others it was slightly more ambiguous. For entities such as UK national labs or firms we have chosen to designate them "UK" rather than pigeonhole them by where they happen to be sited. Where "nationality" becomes important is in the provision of individual-level public data for research. The slight variations in legal systems across the UK, the devolved nature of the administrations and the different governance arrangements in place for access to public data must all currently be navigated in different ways. One of the principal challenges of creating a federated network of secure research resources for sensitive data research across the UK is this varied governance component.

Reflecting this, the following "pen portrait" sections of the four nations focus most on TRE provision.

8.1. Wales

Supporting research with sensitive population data in Wales has been the purview of the Secure Anonymised Information Linkage (SAIL) Databank since 2007 [7]. Growing from the Health Informatics Research Unit at Swansea University SAIL Databank was one of the first UK sites to bring health and administrative data for a whole-nation population together in one place, under the auspices of the Farr Institute Wales (one of the forerunners to HDR UK) and the Administrative Data Research Centre Wales. That was in 2015, and SAIL Databank is still regarded as pioneering good practice in linked sensitive data research not only UK-wide but internationally.

SAIL Databank is the one-stop shop for population-level data in Wales. Since inception, SAIL Databank has worked with trusted third-party indexing agents at Digital Health and Care Wales to connect data for all 3.3 million Welsh citizens. As well as curating this significant linked data resource SAIL Databank provides a rich on-prem TRE and is

one of the few sites in the UK to have achieved accreditation from the UK Statistics Authority under the DEA as a processor of statistical data.

Research access to SAIL Databank is approved at national level by an independent Information Governance Review Panel comprising representatives of the Welsh Government, NHS and public health services in Wales, academic institutions and members of the public.

The development in 2011 of software to support off-site access to SAIL Databank led to the "spinout" of the software environment as the Secure e-Research Platform (SeRP [22]). Now in three flavours – SeRP UK, SeRP Australia and SeRP Canada – with a mixed on-prem/public cloud infrastructure, SeRP provides multi-tenant hosting for independent TREs worldwide. Three of our cohort of correspondents make use of SeRP UK as a TRE platform provider; altogether SeRP hosts over two dozen tenancies.

8.2. Scotland

Since the early 2010s Scotland has run a network of five "safe havens" (TREs) to support research with public data. Four of these are termed "regional safe havens" and are rooted in Scottish NHS regions: West of Scotland (Greater Glasgow and Clyde); Lothians, including Edinburgh; Fife and Tayside; and Grampian. These regional safe havens provide combinations of TRE-based research access and localised data and governance knowledge. The fifth is the "National Safe Haven" which provides TRE-based access to national population data for Scotland's 5.5 million citizens. Gao *et al* (2022) provides a good, contemporary summary of the nature and history of the safe haven network [23].

Safe haven operation in Scotland has since 2015 been governed by the guidelines in the Scottish Government Charter for Safe Havens [24]. This lays out sets of principles around separation of safe haven operations from information governance from research roles which still drives TRE development in Scotland today[6]. Each of the five safe havens runs as a partnership between local health authorities and four of Scotland's biggest research universities (Edinburgh, Glasgow, Dundee and Aberdeen).

The National Safe Haven is also the host TRE for the Scottish Medical Imaging Archive, a research-available copy of all medical images routinely collected in Scotland between 2010 and 2018 [25]. Scotland has ambitious plans to bring this archive up-to-date and "reconnect" with a live feed from Scotland's national PACS system. Such a research resource of routinely collected image data (currently around 1.5 petabytes, or 1.5 million gigabytes) is unprecedented in the UK, Europe and perhaps beyond.

Sensitive administrative data is provided for research under ADR UK Scotland, sourced from Scottish Government, National Records of Scotland (NRS) and other public agencies. NRS also provide trusted third-party indexing and linkage services for the safe haven network, often using Scotland's unique "community health index" number (CHI).

The National Safe Haven is formally owned as a service by Public Health Scotland, a Scottish NHS body with "sponsorship" from Scottish Government and Scottish local authorities. Research projects are approved by the Public Benefit and Privacy Panel for Health and Social Care. Management of research access to national datasets has recently become the preserve of Research Data Scotland[7], a relatively new body tasked with coordinating research using public data across Scotland.

[6] This author was, until the beginning of 2023, director of National Safe Haven development at TRE operators EPCC at the University of Edinburgh.
[7] See https://www.researchdata.scot/

8.3. Northern Ireland

Northern Ireland is unique in the four nations of the UK for having a combined health and social care system. Health and Social Care Northern Ireland (HSCNI) is technically separate from the UK National Health Service; its built-in connections with social care provision make provision of linked data for research a more streamlined affair than it can be in the rest of the Union.

The Honest Broker Service (HBS), set up in 2014, is Northern Ireland's principal TRE. The HBS allows data from across Northern Ireland's five HSC trusts to be joined together to gain a fuller understanding of the health of the population. The HBS is currently run as a tenancy of the UK SeRP.

In 2022 HSCNI published an ambitious digital strategy to 2030 [26]. The importance the strategy attaches to improving the use of data for research is perhaps best summed-up by these quotes from the *Data Strategy*:

> "We will publish metadata which clearly describes all the information we collect.
>
> "We will share public data in a safe and transparent way.
>
> "We will ensure that our people and our partners are able to safely and securely access the data they need… for research…
>
> "We will champion research as a fundamental part of health and care delivery."

8.4. England

The two central pillars of the sensitive data research landscape in England are the National Health Service and the Office for National Statistics (ONS). Both of these organisations take dual roles as data providers and TRE service providers. Both are looking ahead to a period of significant innovation in the way they enable data-driven research.

England's size relative to the other nations has meant it has no single whole-population data resource like SAIL Databank and no single coordinating body like Research Data Scotland. It does have plans, though, driven partly by the Data Saves Lives policy paper and partly by the successes of joint actions between the NHS, ONS, public health bodies and many research teams across the country during the covid-19 pandemic.

The picture of research using administrative data is in fact better than that for health. The ONS have operated a TRE – the Secure Research Service, SRS [27] – since 2004. With support from both ONS and the Administrative Data Research UK programme the SRS provides secure, cloud-based TRE access to more than 120 national datasets. In 2022 the ONS announced the creation of "SRS 2", the Integrated Data Service, IDS [28]. Targeted initially at harmonising access to national public data for government analysts, the longer-term plan is to broaden access to IDS to approved and accredited researchers from the wider community.

Almost as a side benefit, at the heart of the IDS system will sit an index spine that would provide mechanisms to link public datasets together in a far more efficient and effective way than happens today. ONS do have plans to make such a service available for research use, again, subject to approvals and accessed from within one or another TRE.

NHS England's first TRE was assembled rapidly at the onset of the covid-19 pandemic, driven by necessity in adversity but nevertheless a significant step along the road of the NHS's policy of moving from a data release model to a data access model[8]. This "version 1.0" TRE, deployed in the public cloud, is now developing into the

[8] As an indication of the shifting nature of the landscape, the covid-19 era TRE was assembled by NHS Digital. In early 2023 NHS Digital merged with NHSX and the "old" NHS England to create a "new" NHS England.

National Secure Data Environment (SDE) [29]. "Beta users" from the British Heart Foundation Data Science Centre [30], major early users of "TRE 1.0", have been working alongside NHS England to co-create the national SDE.

Alongside the NHS national SDE the period from the fourth quarter of 2022 has seen the announcement of two waves of regional or "sub-national" SDEs to support research with NHS data. Eleven of these have been approved for development (at the time of writing), each covering a region of England of around 5 million people. The first four ("Wave 1") SDEs have already piloted services built on public cloud using designs developed by HDR UK, DARE UK and the Alan Turing Institute[9]; the overall SDE network is very much a work in progress but will have a profound effect on the nature and mechanisms for the sharing of health data for research in England in the years ahead.

8.5. UK-wide

The differences in the research landscape discussed above arise principally from differences in the governance and legal frameworks in different parts of the UK for citizens' individual-level data. Looking beyond individual level data the landscape broadens to include many more potential infrastructure and service providers who operate across the UK as a whole.

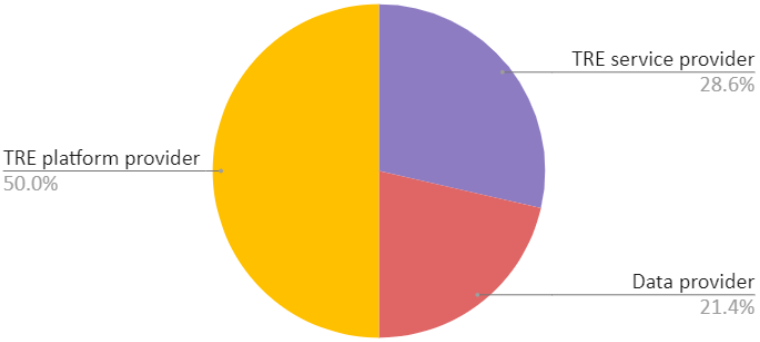

Figure xv: Distribution of UK-wide correspondent's primary role.

Among our correspondents we designated 14 as "UK wide" in scope. Of these four provide TRE services for sensitive but non-individual data, and three provide datasets of the same nature; the remaining 50% are platform providers.

8.6. The UK in the wider world

In planning digital research infrastructure, particularly to handle internationally relevant material such as health data, the UK must at least be aware of developments elsewhere. In fact, given the existing strength of UK-European connections in the genomics field through initiatives like ELIXIR [31] and GA4GH [32], it is fairer to say that the UK must ensure that its digital research infrastructure is compatible with these existing and emerging technologies.

One of the most significant international infrastructure developments the UK should track is the European Smart Middle Platform (variously SiMPl or SMP) [33]. SiMPl is designed to create an open standards-based approach to cloud interoperability and provisioning ("cloud-to-edge federation") and to underpin the European Data Strategy [34] and the development of "data spaces". The published timetable for SiMPl suggests a minimal viable product should be released "at the beginning of 2024".

Parallels between SiMPl and two further European "federating technologies", GAIA-X [35] and X-Road [36], are highlighted in the DARE UK Federated Architecture Blueprint [7].

[9] See, for example, the 2022 DARE UK portfolio of sprint exemplar projects, https://dareuk.org.uk/our-work/sprint-exemplar-projects/.

9. Conclusions

The value of cross-domain linkage and analysis of sensitive data is widely recognised, with broad applicability in academic research, industrial innovation, service delivery in governmental departments and the National Health Service, and to inform public policy. The UK's rapid, collaborative response to the COVID-19 pandemic is an exemplar of the kind of value and impact inter-disciplinary research can have, using sensitive data at scale appropriately and securely. There is clear and present opportunity for the digital research infrastructure ecosystem to capitalise upon the lessons learnt from this emergency response – collaboration at scale, policy impact, a joined up distributed national digital infrastructure – and translate this into a coherent, sustainable business-as-usual approach to enabling highly impactful research using sensitive data at national scale that will deliver significant impacts for public good, government policy and UK plc.

This report offers a snapshot of the UK's provision of digital infrastructure to support public research with sensitive data and reiterates several broad challenges across the landscape that are worth restating here for emphasis but may not come as a surprise to many of those operating within the sensitive data research ecosystem today. This review is focused primarily on the digital infrastructures within the sensitive data research ecosystem and the methods in undertaking this review reflect that. This specific review did not investigate, beyond anecdotal evidence during interviews, the data and information governance aspects that play a major role in the sensitive data research ecosystem. Future reviews could build on this work by investigating specific areas or gaps, for example detailed capability mapping not only today but also emerging (such as driven by AI) or more nuanced understanding of different cost recovery models across the landscape.

This review indicates the overall landscape of research infrastructures fit for the purpose of supporting research with sensitive data is fragmented, and increasingly so (cf. Chapter 4). There are both pros and cons to such fragmentation. On the one hand the move away from the data distribution model and towards data access through TREs is to be welcomed; TREs can be made much more secure than can multiple researcher laptops with the natural consequence being an increased proliferation of TREs of varying levels of capability and maturity. On the other hand, a major risk is that TREs become new silos, reducing our ability to link datasets together to increase their utility and impact for research that addresses major inter-sectional societal challenges and informs policy. From a DARE UK programme perspective, embracing the shift towards the data access model and the reality of a rich ecosystem of TREs and research services around them is important not least as a vibrant foundation for innovation in this space. Accepting that the future digital infrastructure landscape supporting sensitive data research will be distributed requires managing the risks of fragmentation, siloed infrastructures that raise barriers for inter-disciplinary research, and a lack of coherence across the landscape. This presents an opportunity to collectively develop an enterprise-level vision or target picture for the landscape that must be trustworthy, flexible enough to accommodate a spectrum of very different capabilities and maturities, provides strategic orientation for the ecosystem to coalesce around, and inform strategic investments in digital research infrastructure in the future. There is a window of opportunity to do this, the landscape is changing particularly quickly at the moment with the number of TREs scheduled to increase by a third over the next three years; the majority of these will be health data-focused and driven predominantly by the NHS in England. Alongside this are several UKRI initiatives that will require TRE services such as ESRC led Smart Data Research UK (formerly Digital Footprints) [37], BBSRC led BioFAIR [38], NERC led Digital Solutions Programme [39], the STFC led Centre of Excellence for Resilient Infrastructure Analysis [40], or the Smart Manufacturing Data Hub [41], to name a few. This flux provides an opportunity to create and maintain a coherent strategy for the landscape, without over-constraining future developments. While challenging, this is a significant opportunity.

In technology terms, the review strongly indicates that cloud-first is the dominant strategy, not necessarily meaning use of public cloud services but meaning adopting the same layers of software technology common across cloud – virtualisation, containerisation, webservice-orientation and microservice architectures. This is an encouraging development, a standardisation of technology approaches which supports wider adoption and deployment. Alongside this, use of public cloud as a preferred deployment environment is increasingly popular and is especially dominant in planned developments. Despite the convergence on common technology strategy, many TREs and related services (and their data governance arrangements) continue to specialise to meet particular needs. While many sensitive datasets, particularly from the public sphere, can be presented as simple (if sometimes large) flat files, many newer ones are more complex, requiring specialised tooling and staff to support their use. Increasing interest in research access to large medical imaging datasets, for instance – apparent anecdotally throughout this review – suggests that the TRE landscape may remain fragmented for reasons beyond governance caution or simple inertia. Large, complex datasets require large, complex analysis environments and enabling linkage of these data with others will mean that the "other" data will increasingly need to move to join the less mobile, specialist dataset. From a DARE UK perspective, this reinforces the need to plan for a landscape not only of distributed data but of distributed analytical capabilities.

Another theme apparent throughout the review is the appetite both for sharing and making use of sensitive data – safely and securely – in support of better research in areas of public policy. The UK's stock of public data is both large and of tremendous quality internationally; Professor Cathie Sudlow's ongoing review of health data flows [19] is clear recognition of this and will provide deep insight into this key aspect of the landscape. Add to this the clear trend towards use of TREs and the UK has a tremendous opportunity to take a global lead in this area of research.

The landscape of UK digital infrastructure fit for the purpose of supporting research with sensitive data has grown steadily over the years and is poised, off the back of cloud-first technology approaches and the response to inter-sectional societal challenges such as the COVID-19 pandemic, for a period of evolution and significant growth. Allowing sensitive data assets to be linked and analysed at scale and at far greater pace than is currently possible has the potential for very broad scopes of inquiry spanning all manner of UK sensitive data research, with both national and international impact. The prize is a world-leading UK capability that would stand to vastly outperform current practices, allowing for broad cross-domain, cross-jurisdictional analysis, accelerating basic and applied research as well as informing fast-moving national policy priorities.

10. References

[1] Genomics England; https://www.genomicsengland.co.uk/ (accessed 02/10/2023).

[2] UK Biobank; https://www.ukbiobank.ac.uk/ (accessed 02/10/2023).

[3] Office for National Statistics; https://www.ons.gov.uk/ (accessed 02/10/2023).

[4] DARE UK; *Data Research Infrastructure Landscape: A review of the UK data research infrastructure*; October 2021; DOI:10.5281/zenodo.5584696; https://zenodo.org/record/5584696 (accessed 12/06/2023).

[5] DARE UK; *Paving the way for a coordinated national infrastructure for sensitive data research*; August 2022; DOI:10.5281/zenodo.7022440; https://zenodo.org/record/7022440 (accessed 12/06/2023).

[6] The Nurse Review; *Independent Review of the UK's Research, Development and Innovation Organisational Landscape: Final Report and Recommendations*; March 2023; https://www.gov.uk/government/publications/research-development-and-innovation-organisational-landscape-an-independent-review (accessed 06/06/2023).

[7] DARE UK; *Federated Architecture Blueprint, initial version 1.0*; April 2022; https://dareuk.org.uk/wp-content/uploads/2023/04/DARE-UK-Federated-Architecture-1-Initial.pdf (accessed 12/06/2023).

[8] U. Deetjen, E. T. Meyer and R. Schroeder (2015); *Big Data for Advancing Dementia Research: An Evaluation of Data Sharing Practices in Research on Age-related Neurodegenerative Diseases*; OECD Digital Economy Papers, No. 246, OECD Publishing; http://dx.doi.org/10.1787/5js4sbddf7jk-en (accessed 02/08/2023).

[9] N. Kennedy, T.L. Win, A. Bandyopadhyay, J. Kennedy, B. Rowe, C. McNerney, et al.; *Insights from linking police domestic abuse data and health data in South Wales, UK: a linked routine data analysis using decision tree classification;* The Lancet Public Health, Vol 8, Issue 8, E629-E638, August 2023; https://doi.org/10.1016/S2468-2667(23)00126-3 (accessed 03/08/2023).

[10] SAIL Databank; https://saildatabank.com/ (accessed 15/06/2023).

[11] F. Harkness, J. Blodgett, C. Rijneveld, E. Waind, M. Amugi, & F. McDonald (2022); *Building a trustworthy national data research infrastructure: A UK-wide public dialogue* (1.0.0); Zenodo; https://doi.org/10.5281/zenodo.6451935 (accessed 27/06/2023).

[12] UK Government; *Data saves lives: reshaping health and social care with data*; policy paper; June 2022; https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data (accessed 12/06/2023).

[13] B. Goldacre et al; *Better, broader, safer: using health data for research and analysis*; 7 April 2022; https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis (accessed 02/08/2023).

[14] UK Health Data Research Alliance; *Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems;* December 2021; https://zenodo.org/record/5767586 (accessed on 03/08/2023).

[15] Health Data Research UK; *Covid-19 Data and Connectivity*; https://www.hdruk.ac.uk/covid-19-data-and-connectivity/ (accessed 12/06/2023).

[16] OpenSAFELY; *The OpenSAFELY Secure Analytics Platform*; https://www.opensafely.org/ (accessed 23/03/2023).

[17] Outbreak Data Analysis Platform; *Fast, FAIR and SAFE clinical data science in outbreaks*; https://odap.ac.uk/ (accessed 12/06/2023).

[18] NHS England Transformation Directorate; *Data for Research & Development Programme*; https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/accessing-data-for-research-and-analysis/work-in-progress/ (accessed 12/06/2023).

[19] The Sudlow Review; *Unifying Health Data in the UK*; June 2023, in progress; https://www.hdruk.ac.uk/helping-with-health-data/the-sudlow-review/ (accessed 12/06/2023).

[20] Kavianpour S, Sutherland J, Mansouri-Benssassi E, Coull N, Jefferson E.; *Next-Generation Capabilities in Trusted Research Environments: Interview Study;* https://www.jmir.org/2022/9/e33720 (accessed 03/08/2023).

[21] F. Ritchie (2016); *Five Safes: designing data access for research*; 10.13140/RG.2.1.3661.1604.

[22] Secure e-Research Platform; https://serp.ac.uk/ (accessed 15/06/2023).

[23] Gao C, McGilchrist M, Mumtaz S, Hall C, Anderson LA, Zurowski J, Gordon S, Lumsden J, Munro V, Wozniak A, Sibley M, Banks C, Duncan C, Linksted P, Hume A, Stables CL, Mayor C, Caldwell J, Wilde K, Cole C, Jefferson E; *A National Network of Safe Havens: Scottish Perspective*; J Med Internet Res. 2022 Mar 9;24(3):e31684. doi: 10.2196/31684. PMID: 35262495; PMCID: PMC8943560.

[24] Scottish Government; *Charter for safe havens in Scotland: handling unconsented data from national health service patient records to support research and statistics*; 2015; ISBN 9781785444968; https://bit.ly/3CwBJug (accessed 15/06/2023).

[25] Nind T, *et al*, *An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population*, GigaScience, Volume 9, Issue 10, October 2020, giaa095, https://doi.org/10.1093/gigascience/giaa095

[26] Health and Social Care Northern Ireland; *Digital Strategy 2022-2030*; https://www.health-ni.gov.uk/publications/digital-strategy-health-and-social-care-northern-ireland-2022-2030 (accessed 15/06/2023).

[27] Office for National Statistics Secure Research Service; https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice (accessed 16/06/2023).

[28] Office for National Statistics Integrated Data Service; https://integrateddataservice.gov.uk/ (accessed 16/06/2023).

[29] NHS England National Secure Data Environment; https://digital.nhs.uk/services/secure-data-environment-service (accessed 16/06/2023).

[30] BHF Data Science Centre; https://www.hdruk.ac.uk/helping-with-health-data/bhf-data-science-centre/ (accessed 19/06/2023).

[31] ELIXIR; *A distributed infrastructure for life science information*; https://elixir-europe.org/ (accessed 09/03/2023).

[32] Global Alliance for Genomics and Health; https://www.ga4gh.org/ (accessed 21/06/2023).

[33] European Commission; *Simpl: cloud-to-edge federations and data spaces made simple*; news article, 24/02/2023; https://digital-strategy.ec.europa.eu/en/news/simpl-cloud-edge-federations-and-data-spaces-made-simple (accessed 02/03/2023).

[34] European Commission; *A European Strategy for data*; policy paper; https://digital-strategy.ec.europa.eu/en/policies/strategy-data (accessed 09/03/2023).

[35] GAIA-X; *A Federated Secure Data Infrastructure*; https://gaia-x.eu/ (accessed 09/03/2023).

[36] NIIS; *X-Road Architecture*; https://x-road.global/architecture (accessed 02/03/2023).

[37] Smart Data Research UK; https://www.ukri.org/what-we-do/our-main-funds-and-areas-of-support/browse-our-areas-of-investment-and-support/smart-data-research-uk/ (accessed 02/03/2023).

[38] BioFAIR; https://biofair.uk/ (accessed 02/03/2023).

[39] NERC Digital Solutions Programme; https://www.digital-solutions.uk/ (accessed 02/03/2023).

[40] Centre of Excellence for Resilient Infrastructure Analysis; https://www.ukri.org/news/new-centre-of-excellence-for-resilient-infrastructure-analysis/ (accessed 02/03/2023).

[41] Smart Manufacturing Data Hub; https://smdh.uk/ (accessed 02/03/2023).

A. List of Correspondents

We would like to thank all correspondents for their time in helping us compile this review, in particular those organisations working with sensitive data in research, for whom we had more questions.

A.1 Correspondents working with sensitive data

AIMES TRE

Akrivia Health Clinical Research Interactive Search (CRIS)

Alan Turing Institute Data Safe Haven

Aridhia DRE

AWS Service Workbench

Barts Health Precision Medicine Platform

BHF Data Science Centre instance of NHS England TRE/SDE

Big Data and Analytical Unit Secure Environment (BDAU SE), Imperial College

British Ocean Sediment Core Research Facility

Centre for Macaques, Medical Research Council

Centre for Rapid Online Analysis of Reactions (ROAR)

CLARIN

Clinoverse

Consumer Data Research Centre (Leeds)

DAFNI - Data and Analytics Facility for National Infrastructure

DataLoch

Edinburgh International Data Facility

Electron beam lithography facilities, University of Cambridge

EPND (European Platfrom for Neurodegenerative Diseases)

FAIRDOM

FAIRDOM-SEEK

Genomics England RE

GG&C Safe Haven

Grampian Data Safe Haven, University of Aberdeen & NHS Grampian

Health Informatics Centre, University of Dundee

InterConnect and MRC Epidemiology Unit in-reach system

JASMIN

Leeds Analytic Secure Environment for Research (LASER)

Lifebit Federated Trusted Research Environment

Microsoft AzureTRE

National Survey of Sexual Attitudes and Lifestyles (Natsal)

Natural History Museum

NDORMS

NERC Digital Solutions

NHS England SN SDE Network

NI Honest Broker Service

NIHR BioResource

NURTuRE

ONS Integrated Data Service

ONS Secure Research Service

OpenSAFELY in OpenSAFELY-TPP and OpenSAFELY-EMIS

OurFutureHealth TRE

Personalised Medicine Centre, Ulster University

Royal Botanic Gardens Kew

SAIL Databank

Scottish National Safe Haven

Secure eResearch Platform (Serp)

Sir Peter Mansfield Imaging Centre

Software Sustainability Institute

STFC Scientific Computing Department

The Francis Crick Institute

The GW4 Isambard Tier-2 HPC service

UK Data Service

UK Health Security Agency (UKHSA)

UK Longitudinal Linkage Collaboration

UKAEA

UKAEA Materials Research Facility

UKRI - Medical Research Council - Mary Lyon Centre at MRC Harwell

United Kingdom Multiple Sclerosis Register

University of Liverpool

University of Portsmouth

University of Sheffield Sensitive Data Service

A.2 Other correspondents

Advanced Bioimaging RTP, University of Warwick

Bede, N8 CIR / EPSRC Tier-2 HPC service

BGS Space Geodesy Facility, Herstmonceux

British Ocean Sediment Core Research Facility

MRC Centre for Virus Research, University of Glasgow

National Oceanography Centre Discovery Collections

EPSRC National Dark Fibre Facility (NDFF)

EPSRC Quantum Communications Hub

Liverpool Hope University Science Facilities

NERC Geophysical Equipment Facility

North Wyke Farm Platform

Ocean Bottom Instrumentation Facility

Plymouth Marine Laboratory

Roland von Glasow Air-Sea-Ice Chamber

SAF oxford

STFC-IRIS (e-Infrastructure for Research In STFC)

The National Archives

The UK High-Field Solid-State NMR Facility

UK-EPMA

University of Hertfordshire High-Performance Computing facility

B. Survey Questions

Start of Block: Personal Information

Intro Responses and/or summary outputs of the responses from this survey, including the names of institutions, may be made publicly available. The names of individual respondents will not be published. By completing this survey, you agree to a record of the names and institutions of respondents being processed by the DARE UK Delivery Team. You can view the DARE UK Privacy Policy on our website. You may contact DARE UK at any time to have this information removed.

The results of this survey will be used to review the capabilities of sensitive data research infrastructures in the UK. This is the first step in a deeper landscape review. As such, the DARE UK team may be in touch with you to discuss your responses further.

○ I confirm I am happy for my responses to be used in this way. (1)

Display This Question:

If Intro = I confirm I am happy for my responses to be used in this way.

Q1 What is your name?*

Q2 What is your email address?* (please give a professional email address)

Q3 What is the name of the research infrastructure that you are responding on behalf of?*

Q3 - a If applicable, can you provide a reference link to a website for the research infrastructure?

End of Block: Personal Information

Start of Block: Sensitive Data

Q4 Based on this definition of sensitive data:

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

Does your research infrastructure store or process sensitive data or does it have the intention to enable this in the future?*

- ○ Yes (1)
- ○ No (2)

End of Block: Sensitive Data

Start of Block: Data

Q5- a What type(s) of sensitive data does your research infrastructure store or process? Please select all that apply.*

- ☐ Genomics (5)
- ☐ Biometrics (6)
- ☐ Primary care (7)
- ☐ Secondary care (8)
- ☐ Clinical trials (9)
- ☐ Wearables (10)
- ☐ Social care (11)
- ☐ Identification (12)
- ☐ Financial (14)
- ☐ Wearables (15)
- ☐ Geolocation (16)

- [ ] Judicial (17)
- [ ] Deprivation (18)
- [ ] Education (19)
- [ ] Intellectual Property (20)
- [ ] Sales or Retail (21)
- [ ] Defence or National Security related (22)
- [ ] Surveys (24)
- [ ] Employment (25)
- [ ] Welfare (26)
- [ ] Crime (27)
- [ ] Other (please specify) (23)

Q5 - b What domain does the sensitive data that your research infrastructure stores, or processes, primarily fall into?*

- [ ] Personal Health Data (1)
- [ ] Personal Administrative Data (2)
- [ ] Commercial and Industrial Data (3)
- [ ] Other (please specify) (8)

Q6 Does your research infrastructure allow projects to import external data?*

- Yes (1)
- No (2)

End of Block: Data

Start of Block: Infrastructure

Q7 What stage is your research infrastructure in?*

- Operational/serving users (1)
- Under development (2)
- Planned but not yet in development (3)
- In the process of closing operations (4)
- Other (please specify) (5)

Display This Question:

If Q7 = Operational/serving users

Or Q7 = In the process of closing operations

Q7 - a When was your research infrastructure established?*

- 2022 (1)
- 2021 (19)
- 2020 (2)
- 2019 (3)

- ○ 2018 (4)
- ○ 2017 (5)
- ○ 2016 (6)
- ○ 2015 (7)
- ○ 2014 (8)
- ○ 2013 (9)
- ○ 2012 (10)
- ○ 2011 (11)
- ○ 2010 (12)
- ○ 2009 (13)
- ○ 2008 (14)
- ○ 2007 (15)
- ○ 2006 (16)
- ○ 2005 (17)
- ○ Pre-2005 (18)

Display This Question:

If Q7 = Planned but not yet in development

Or Q7 = Under development

Q7 - b When is your research infrastructure planned to be established?*

- ○ 2022 (1)
- ○ 2023 (2)
- ○ 2024 (3)
- ○ 2025 (4)
- ○ 2026 (5)
- ○ Post-2026 (please specify) (6)

Q8 Please provide a brief summary of your research infrastructure's technical capabilities (either current or planned) using the prompts below to guide your response: *

- On-premise, public cloud, or hybrid cloud
- Support for the import and linking to external data
- Support for the import of custom code
- Library of tools and packages
- Bespoke OS/software available on request
- Access to specialist and high-performance infrastructure (e.g., HPC and GPU clusters

---

Q9 Which sectors does your research infrastructure support or hope to support in the future?*

- ☐ Academic (1)
- ☐ Commercial (2)
- ☐ Public sector (3)
- ☐ Third sector including charitable organisations (4)
- ☐ Other (please specify) (5) _____

Q10 Which research council(s) are your projects primarily aligned with?* (current or planned)

- [ ] Arts and Humanities Research Council (1)
- [ ] Biotechnology and Biological Sciences Research Council (4)
- [ ] Economic and Social Research Council (3)
- [ ] Engineering and Physical Sciences Research Council (6)
- [ ] Innovate UK (7)
- [ ] Medical Research Council (2)
- [ ] Natural Environment Research Council (8)
- [ ] Research England (9)
- [ ] Science and Technology Facilities Council (5)
- [ ] n/a (12)

Display This Question:

If Q7 = Operational/serving users

Or Q7 = In the process of closing operations

Q11 Approximately how many active projects do you host per year?*

Display This Question:

If Q7 = Operational/serving users

Or Q7 = In the process of closing operations

Q12 Approximately how many active users do you host per year?*

End of Block: Infrastructure

Start of Block: Governance and Impact

Q13 Do you currently have public representation in your formal governance structure - for example, members of the public on your boards or steering committees - or plan to do so?*

- ☐ Yes (1)
- ☐ No (2)

Q14 Do you require all researchers accessing your infrastructure to have public involvement and engagement embedded in their projects?*

- ○ Yes (1)
- ○ No (2)
- ○ n/a- researchers do not access our infrastructure (3)

Q15 Do you publish and maintain a data use register for public viewing?*

- ○ Yes (1)
- ○ No - but this is something we are currently considering (2)
- ○ No - this is not something we are currently considering (3)

○ Not applicable - we do not enable access to data (4)

Q16 Please provide a brief summary of the data access processes you utilise in coordination with your research infrastructure.

☐ Delegated authority arrangements, please elaborate (or provide links to the information if available): (4)

☐ Research project accreditation panels, please elaborate (or provide links to the information if available): (5)

☐ Ethics approval panels, please elaborate (or provide links to the information if available): (6)

☐ Other, please elaborate (or provide links to the information if available): (7)

Q17 Which body(ies) funds the research infrastructure you are responding on behalf of?*

☐ Arts and Humanities Research Council (1)

☐ Biotechnical and Biological Sciences Research Council (4)

☐ Economic and Social Research Council (3)

☐ Engineering Physical Sciences Research Council (6)

☐ Innovate UK (7)

☐ Medical Research Council (2)

☐ Natural Environment Research Council (8)

- [ ] Research England (9)
- [ ] Science and Technology Facilities Council (5)
- [ ] UK Research and Innovation (11)
- [ ] Other(s) (please specify) (10)

Q18 What accreditation standard(s) do you conform to? Please select all that apply.*

- [ ] ISO27001 (1)
- [ ] ISO22301 (4)
- [ ] DSPT (2)
- [ ] Cyber Essentials+ (3)
- [ ] Other (please specify) (5)
- [ ] n/a (6)

Q19 Are you Digital Economy Act accredited?*

- ○ Yes (1)
- ○ No (3)

Q20 Are you able to share approximately how much funding and/or investment your research infrastructure has been awarded (inclusive of previous and current phases)?*

- ○ Yes, please indicate the approximate amount below in m£ (e.g. 5 m£): (1)
- ○ I am unable to share this information. (2)
- ○ If you would like to give more information on your funding/investment such as the costs to maintain/run your infrastructure and/or the source of your funding, please do so below: (3)

Q21 What are the notable impacts that have been/will be achieved through the use of your research infrastructure?*

Q22 What do you see as your biggest challenges over the next 3-5 years?*

End of Block: Governance and Impact

Start of Block: Close

Display This Question:

If Q4 = No

This survey is solely for UK research infrastructures who store or process sensitive data or intend to in the future. As you responded "No" to this question, you have reached the end of the survey.

If you feel there are other research infrastructures that handle sensitive data and could contribute to this survey, please send this survey link on to the appropriate contact: DARE UK sensitive data research infrastructure survey

Please note that this survey is the first step in a review of the UK's sensitive data research infrastructure landscape. If you have any other comments or feedback on this survey, please elaborate below. Otherwise, please click the arrow below to complete this survey.

Display This Question:

If Q4 = Yes

If you feel there are other research infrastructures that handle sensitive data and could contribute to this survey, please send this survey link on to the appropriate contact: DARE UK sensitive data research infrastructure survey

Please note that this survey is the first step in a review of the UK's sensitive data research infrastructure landscape. Where necessary, the DARE UK Delivery Team will be in touch with respondents to build upon these responses in more detail.

If you have any other comments or feedback on this survey, please elaborate below. Otherwise, please click the arrow below to complete this survey.

End of Block: Close