UK Research and Innovation

ADR UK

HDR UK
Health Data Research UK

Data Research Infrastructure Landscape

A review of the UK data research infrastructure

UKRI Digital Research Infrastructure, DARE UK Programme – Phase 1

October 2021

Table of Contents

DARE UK 2021. DOI: https://doi.org/10.5281/zenodo.5584696.

Executive summary

UK Research and Innovation (UKRI) brings together the seven UK Research Councils, Innovate UK and Research England into a single organisation to create the best environment for research and innovation to flourish. As part of the overarching UKRI Digital Research Infrastructure Programme, the UK trusted and connected Data and Analytics Research Environments (DARE UK) programme has been launched to understand the needs of those using current research environments, including trusted research environments (TREs) – highly secure digital environments that provide access to sensitive data for approved researchers – to support the development of a coordinated vision for digital research infrastructure in the UK, with a particular focus in those managing sensitive data.

During August and September 2021, interviewees and workshop participants, including members of the public, were invited to discuss their unmet needs related to digital research infrastructure in the broad areas of creation, maintenance and access; especially in the context of TREs. The rationale for this review of the UK digital research infrastructure is to establish the foundation of an extensive and evolving dialogue with the UK research and innovation community as part of Phase 1 of the DARE UK programme.

The scope of this review was intentionally broad to ensure that the DARE UK programme establishes a fundamental understanding of the context and overarching challenges within the UK research and innovation ecosystem, to inform how best the programme should address those challenges that fall within the DARE UK remit. Six key themes of unmet needs were identified: 1) Data and discoverability, 2) Access and accreditation, 3) Digital research infrastructure, 4) Capability and capacity, 5) Demonstrating trustworthiness, 6) Funding and incentives.

There was a total of **60 interviews (79 individuals)** and **two workshops (c.50 participants in each)** with representation across the research and innovation spectrum as well as members of the public. There was widespread support for the DARE UK programme and its ambitions amongst interviewees and workshop participants, especially in the context of a more federated ecosystem that could address new cross-disciplinary research use cases. There is a clear need for a coordinated, cohesive effort supported by sustained funding to better enable the research and innovation ecosystem in the UK. DARE UK has an opportunity to play a key role in contributing to this effort, particularly around sensitive data.

Key findings:

- There was discussion around **standards for data and metadata**, and the development of UK-wide standards for **access and accreditation**. Accreditation of platforms as well as researchers could be improved to include the diversity of platforms in use now, including international researchers, while registries of approved environments and researchers could support safe use of data.
- A **proportionate approach to data risk** has the potential to gain the trust of the public, data custodians and commercial organisations.
- Many individuals reported that the technical opportunities which could be addressed in the development of a more federated system are, while by no means trivial, less challenging relative to the **governance challenges** involved in coordinating across digital research infrastructures.
- Many observe the challenges of **demonstrating trustworthiness**, with the public, other researchers, and commercial organisations. There is an opportunity to engage directly with these different groups through outreach activities, building on the success of or working in partnership with others in this space and demonstrating examples of best practice interdisciplinary working.

- A broad issue that, while perhaps not surprising, was raised often was the retention of **capability and capacity**, whereby teams struggle to retain data scientists and data engineers.
- Finally, **funding** was raised as an incentive that UKRI could use to facilitate agreement about sets of standards and collaboration. Sustained funding in maintaining and operating digital research infrastructures, clarity in working definitions and standards for accreditation will incentivise the development of work towards more federated environments.

Some of these observations, such as the need to address the recruitment and retention of data scientists and infrastructure engineers, will fall outside the scope of the DARE UK programme. They are nevertheless essential to the success of the DARE UK programme and should be considered as complementary to the work of the programme. This feedback has been retained within this report, and it will be critical for the DARE UK programme to align with those bodies addressing these areas. Investment in digital research infrastructure will not be effective without a broader holistic view on how to deliver this as a sustainable service to the research and innovation community.

There is an exciting opportunity to bring together data and make use of modern capabilities within the UK research and innovation ecosystem as never before. This is in the context of the Covid-19 pandemic, which has accelerated data sharing and demands for insight reaching across traditional disciplines. In that context, the question of how to address these unmet needs, and more specifically which needs should be considered within the scope of the DARE UK programme, is an exciting challenge that will be elaborated as Phase 1 of the DARE UK programme moves apace.

Introduction

Aims

UK Research and Innovation (UKRI) brings together the seven UK Research Councils, Innovate UK and Research England into a single organisation to create the best environment for research and innovation to flourish. The vision is to ensure the UK maintains its world-leading position in research and innovation.

The aim of this landscape review is to summarise the key themes of unmet needs and opportunities within the UK research and innovation ecosystem as part of a new UKRI Digital Research Infrastructure programme – the UK Trusted and Connected Data and Analytics Research Environments (DARE UK). Digital research infrastructure includes the collection of underlying physical infrastructure, data storage and handling, the underpinning software tools and services, flexible computing capacity and skills that - operating in concert - enable researchers to turn big data into scientific breakthroughs. It should be clearly stated that this document is an initial landscape review and identifies potential areas of opportunity for the DARE UK programme to add value to the research and innovation space within the UK, aligned with the DARE UK remit. However, these recommendations are by no means prescriptive and further work will be done throughout Phase 1 of the DARE UK programme to identify those recommendations most pertinent to the DARE UK remit and those which are better addressed through parallel initiatives. The Appendix provides an overview of the key investments in digital research infrastructure across UKRI, as well as key infrastructures used by researchers which are not directly funded by UKRI. This overview is non-exhaustive and relates specifically to those digital research infrastructure investments that formed part of the discussions within this landscape review.

Researchers funded by UKRI have a variety of needs including in certain instances access to sensitive data. Sensitive data, for the purposes of the DARE UK programme, includes personally identifiable information such as names and addresses or data which is commercially, legally or politically sensitive or sensitive from an intellectual property perspective. It could also be data which has been de-identified (has had all personal identifying information removed) but remains sensitive due to the potential for re-identification.

TREs are highly secure spaces for researchers to access this sensitive data, and offer additional security measures to protect people's privacy, whether the data includes personally identifiable information in its current form or not. UKRI funding has contributed towards the creation of TREs, also known as data safe havens. TREs represent a strategy to meet the needs of researchers, and a mechanism to build public and organisational trust – the Trusted Research Environments (TRE) Green Paper published by HDR UK provides further detailed insight.[1]

The Covid-19 pandemic has resulted in increased emphasis on sharing more data, including sensitive data, more regularly and with a greater degree of near real-time accuracy, for national population health management across the four nations. It has also resulted in a more flexible offer to researchers in some cases given the need for lockdowns and social distancing, for example increasing remote access.

[1] Trusted Research Environments (TRE) - A strategy to build public trust and meet changing health data science Needs: https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf

Scope

The scope of this initial landscape review is intentionally broad and includes all current digital infrastructures with a focus on research data. The scope of the DARE UK programme is more specific, including all research conducted by UKRI councils that uses, or anticipates use of, sensitive data from different research disciplines and from across different sectors, including, but not limited to, social, biomedical, and environmental sciences.

Approach

Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK), as the DARE UK Phase 1 delivery partners, commissioned CF (Carnall Farrar) to complete this landscape review of the existing data and digital research infrastructures for research based on desk research, interviews and workshops with both people building the infrastructure and those using the data for research (primarily academic as opposed to commercial research). CF is a management consulting and data science company.

60 interviews were held in August and September 2021, with stakeholders selected from across the spectrum of disciplines. Some interviewees brought colleagues; therefore 79 people were interviewed in total. As well as researchers, interviews were planned with individuals representing separate infrastructures. The distribution of interviews across councils is skewed towards those whose researchers use sensitive data, and those who invest in a wider diversity of infrastructures. The interviews aimed to be inclusive across the research councils, infrastructures and researcher communities. Recommendations to interview specific contacts were taken up where possible.

Two virtual workshops were held in mid-September 2021, aimed at researchers and technologists, and open to members of the public. Workshop details were shared with those invited to interviews (c. 100 individuals), those who had signed up to the DARE UK communication list (c. 600 individuals), and on the HDR UK and ADR UK websites, and publicised on social media by HDR UK. It should be noted that the workshops were focused towards researchers and technologists, while the workshops were open to public participation there will be future engagement activities within the DARE UK programme dedicated to the public perspective and input.

Approximately 50 individuals attended each of the two workshops. The topics for the interviews and workshops revolved around current infrastructures, definitions of federation, potential benefits of the DARE UK programme and opportunities for the DARE UK programme.

This report summarises the themes of unmet needs and opportunities for the DARE UK programme. Unmet needs relate to the issues faced by researchers and technologists today. Opportunities relate to the actions that could be taken, by DARE UK or other initiatives, to address the needs.

Acronyms

	Name
ADR UK	Administrative Data Research UK
ADRC-NI	Administrative Data Research Centre Northern Ireland
AHRC	Arts and Humanities Research Council
BBSRC	Biotechnology and Biological Sciences Research Council
CDRC	Consumer Data Research Centre
CEDA	UKRI's Centre for Environmental Data Analysis
CF	Carnall Farrar
DARE UK	UKRI Trusted and Connected Data and Analytics Research Environments
DCMS	Department for Digital, Culture, Media & Sport
EPCC	Edinburgh Parallel Computing Centre
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
HDR UK	Health Data Research UK
HPC	High performance computing
MRC	Medical Research Council
NERC	Natural Environment Research Council
ONS	Office for National Statistics
PHE	Public Health England
SAIL	Secure Anonymised Information Linkage
SeRP	Secure eResearch Platform
STFC	Science and Technology Facilities Council
TRE	Trusted Research Environment
UBDC	Urban Big Data Centre
UKRI	UK Research and Innovation

Overview of the existing landscape of research data infrastructure

The figure below is a snapshot of the relevant infrastructure bodies involved in handling research data across the research councils of UKRI. Individuals from these organisations were interviewed as part of this report. See the Appendix for more detail on each council and relevant asset.

Figure 1. Map of subject areas, key infrastructure investments and examples of dataset types, by UKRI funder

	BBSRC	ESRC	EPSRC	MRC	NERC	STFC	Other
Subjects	• Molecular biology • Farming • Industrial biotechnology • Pharmaceuticals	• Social e.g., digital • Administrative data • Finance • Census • Longitudinal	• Engineering • Materials • Physical sciences • Cities and infrastructure	• Human health and disease • Primary care • Hospitals • Cohort studies	• Environmental science • Atmospheric • Oceanography • Geoscience	• Particle physics • Nuclear physics • Space science • Astronomy	• Healthcare • Social care • Administrative data e.g. HMRC tax data
Examples of key digital research assets	• EMBL-EBI • CyVerse	• UK Data Service and Secure Lab TRE • CDRC (Consumer data) • UBDC (Urban data) • ONS Secure Research Service	• Alan Turing Institute • ARCHER • National Quantum Computing Centre	• Secure e-Research Platform and SAIL Databank • Health Data Research UK • Francis Crick Institute	• Environmental Information Data Centre • Centre for Environmental Data Analysis	• JASMIN • DAFNI • Hartree Centre and EPCC	• OpenSAFELY TRE • HMRC Datalab • NHS Digital TRE • HSC Northern Ireland Honest Broker Service
Example sensitive datasets	• Personal data – including genetic in some cases	• Personal data e.g., social media • Commercial data	• Commercial data from industrial partners	• Personal health data /genomics data	• Rarely, unless linked with clinical data (e.g. pollution epidemiology)	• Data of significance to national security (e.g., aerospace)	• Firm-level financial data • Personal data
Example non-sensitive datasets	• Biochemical or non-human organism data	• Economic indicators	• Materials measurements data	• Biochemical or non-human organism data	• Atmospheric measurements	• Physical experimental data	• Population-level data

Note – AHRC does not fund its own research data services. AHRC-funded researchers make use of a wide variety of data across disciplines, particularly historic data

Unmet needs and opportunities

Interviewees and workshop participants were invited to discuss their unmet needs. The needs of the researcher and technologist communities reflect problems in the creation and maintenance of digital research infrastructure, or access to it. In addition, individuals were asked to share their views on the opportunities and how, if at all, they viewed those opportunities in the context of the DARE UK programme. Members of the public were able to participate and contribute to the workshops.

A synthesis of these has grouped ideas into six themes:

1) Data and discoverability
2) Access and accreditation
3) Digital research infrastructure
4) Capability and capacity
5) Demonstrating trustworthiness
6) Funding and incentives

In the section below, each theme is expanded, alongside the variety of unmet needs reported and the variety of relevant opportunities to address these needs in the context of sensitive data.

The scope of these unmet needs and opportunities is intentionally broad to ensure that the DARE UK programme establishes a fundamental understanding of the context and overarching challenges within the UK research and innovation ecosystem, to inform how best the programme should address those challenges that fall within the DARE UK remit.

1) Data and discoverability

Use of technical standards can lead to interoperability of data. However, data from different sources is recorded in variable ways, using a variety of data standards and common data models. Data is also described in different ways using a variety of metadata standards.

Unmet needs

Data standards

There are currently separate sets of standards across the UK, each adopted by a limited number of parties. Data is therefore often not interoperable. Even if the same data standard has been used, other features of data can differ. It was pointed out that data standards in use now are already partially outdated.

Metadata and discoverability

Poor recording of features of data such as missingness limit the usefulness of data resources.

Data is not always discoverable, particularly to those looking for data from a new discipline. There have been attempts to bring together metadata from disparate sources into metadata catalogues. Metadata catalogues

are available in some cases e.g., the Innovation Gateway for health data. Funders see lots of projects using the same data source, simply because it is well-documented, meaning many datasets remain under-utilised. Availability and visibility of data can drive research questions. The right questions are therefore not necessarily being asked of existing data. It is not always clear who is using infrastructure or datasets, meaning collaborations are hard to foster.

Conversely, well-documented datasets can be available via multiple TREs. This can lead to duplication of the human and environmental cost, and confusion for researchers.

Opportunities

Data standards

There is an opportunity to assist in making recommendations for use of certain data standards, or convening groups to collaborate on developing, enhancing and/or adopting data standards. Committed collaborations of bodies (such as universities or hospitals) could be best placed in the implementation of standards, particularly those with similar interests in order to share their learnings. Although easiest to implement at the beginning of projects, conventions must be regulated throughout their usage. Data standards and metadata standards were of importance to interviewees, as well as API standards which were regarded by some as the key to federation. Quality measures would allow understanding of comparability. Related to this, standardised data capture forms would enable control at data input.

Metadata and discoverability

Making data discoverable, for example through the development and publication of user-friendly sets of metadata to describe datasets (or objects). This was highlighted by multiple individuals as a first step in the direction of federation. The creation or enhancement 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 not only within each research discipline but across disciplines as well. This could enable increased transparency, for example allowing an individual (e.g., a patient) to see where their data is being used, further demonstrating trustworthy use of data.

Participant quotes

“Efforts so far have been from the grassroots level, so you get separate sets of standards, each adopted by a few parties, but we need a single set of standards for all” Technologist, Research organisation

“Standards are like toothbrushes. Everyone’s got one. No one wants to use someone else’s. We’ve got plenty of standards. It just depends what you’re trying to do” Technologist, Commercial provider

“Sticking 101 TREs in won’t solve it if you can’t find what exists in a safe way” Technologist, University

“See lots of projects using the same data because it’s well-documented” Researcher, University

2) Access and accreditation

Beyond interoperable systems and interoperable data, there are governance, rules and frameworks for managing data access and enabling researchers to conduct analysis.

Unmet needs

Standards

Unclear, inconsistent or lack of standard, agreed processes means that many ways of working have evolved between different groups. Varied standard in accreditation ways of working can result in 'reinventing the wheel' for every new project. Researchers also experience slow timelines of accessing data, hindering the speed of innovation and the ambition of what people can do. Lack of standard processes and minimum levels of 'service' can result in large administrative burdens for accessing data.

Despite a wish for standardisation across organisations, European or international institutions do not want to feel locked into a UK-approach. Decision-making therefore needs to have an international perspective. Developing environments for rapid, short-term collaborations can be difficult, for example during the Covid-19 pandemic.

TRE standards, TRE accreditation and researcher accreditation

Many interviewees recognise federation as data in different physical places being used together through a common interface. It became apparent however that many have multiple definitions for federation or remain unclear.

Similarly, most individuals described TREs as allowing researchers going to the data, physically or remotely, without data moving from the safe environment. The Five Safes framework[2] was referred to as the basis for multiple TREs (safe data, safe projects, safe people, safe settings and safe outputs), however there is not one clear definition of a TRE. There are also several TREs across the UK, however there is no central register of best practice thus leading to a lack of clarity around what is and is not a TRE.

Researchers wishing to access TREs must complete training and gain accreditation for their use however training can be duplicative across platforms, and the accreditation process can take time as there are backlogs of people waiting for training in the Five Safes framework.

Risk governance

Researchers currently face inflexible data access processes via TREs, often justifiably so given the legal responsibilities that TREs, and the data custodians they provide services to, have. Despite huge variation in the risk and sensitivity of data, the requirements on researchers for data access can be similar and as such

[2] What is the Five Safes framework? — UK Data Service

researchers need a more flexible approach to data risk. The Alan Turing Institute has proposed such an approach – a policy and process framework that incorporates data security threat and risk profiles into five sensitivity tiers[3]. Such an approach has however not been widely adopted, it is unclear why this is the case and should be further investigated. Future categorisation of risk will need to assess the heightened risk of linked data, and what consent models were involved in the data collection.

Opportunities

Standards

There is an opportunity to work with the community and relevant stakeholders to develop and assist in the implementation of a best practice protocol or guidance on acceptable approaches, acting as an authority that groups could trust and rely on to move in a common direction. A pan-UKRI recommended approach could remove the responsibility from individual groups and therefore improve consistency and increase efficiency.

Some specific areas of opportunity for standards that some interviewees highlighted were in information security, platform specifications and service descriptions, as well as a centralised codified approach to data licensing. There is an opportunity to work with data custodians to agree to standards, and work with governance teams to navigate the interpretations of legal positions. The collective development and agreement of standards among data custodians could also then inform higher-level government bodies in policy-setting. Interviewees were strongly in support of pursuing the opportunity to ensure platforms are made more accessible to researchers across disciplines, and work for most of the research and innovation community.

TRE standards, TRE accreditation and researcher accreditation

An opportunity could be to convene a process with the research councils and key stakeholders to agree on definitions of TRE and related terms including federation and interoperability. For example, capturing the working definitions as discussed during the interviews, agreeing these with research and infrastructure communities, and continuing to evolve them as technology user needs change over time.

An approach for aligned standards for TREs could be further developed and delivered. As one example, TREs could commit to minimum standards in a service level agreement, to provide users with a minimum service in terms of staffing, standards of information relating to data holdings, compute and speed of disclosure control. Further supporting the work of the UK Statistics Authority (UKSA) in the standard accreditation governance for TREs, with regular independent quality testing as part of this.

There is an opportunity for further work to standardise and smooth the researcher accreditation process, 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 making the process feel as though the researcher were accessing data on their own machine. Training could therefore be made portable across TREs, through a 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 TREs.

[3] Design choices for productive, secure, data-intensive research at scale in the cloud (2019) https://arxiv.org/abs/1908.08737

Health data collected for an organisation's health functions is not however part of the DEA, although it should be noted that the majority of the DEA-accredited TREs also hold health data (for example eDRIS, ONS SRS and SAIL). International researchers would also need to be considered, as currently accredited researchers need a link to a UK institution. A surge in people using TREs would also need to be prepared for and staffed.

Successful TREs, based on the views of the interviewees, are those with teams of individuals to support researchers and data providers, including the ability to tell data providers what research their data is being used in.

Given that TREs operate in a global context, connectedness with global partners is essential, including those in low-resource settings.

Risk governance

Mechanisms to define and categorise risk of data, and environments to be reflective of this risk has already been explored, for example the Alan Turing Institute framework for categorising risk of data was mentioned by multiple interviewees.[4] Implementation of these tiers of risk could also make the use of de-identified data more appealing for researchers, although it should be noted that under the DEA, researchers can *only* access de-identified data, and this via a DEA-accredited TRE. However, use of a risk-based proportionate approach to data could therefore incentivise safe use of data and use of the least sensitive data to answer a research question. There is an opportunity to strengthen and consolidate existing data risk frameworks so that they can be pushed for adoption more broadly in order to establish standard data risk frameworks that are widely understood, accepted and adopted.

[4] Design choices for productive, secure, data-intensive research at scale in the cloud (2019) https://arxiv.org/abs/1908.08737

Participant quotes

"We probably don't want a single TRE environment for everyone, but instead the principals and guidance on approaches that are acceptable... Risk aversion and inconsistent or unclear standards are the main barrier... Publishing how other groups have navigated those requirements would be helpful for new teams to follow their example" Technologist, Research organisation

"We'd love for you to set standards for what an accredited TRE should be… At the moment it's a self-selective process… There's currently lots of interpretation around certification" Technologist, Commercial organisation

"When researchers realise the secure data requirement, they're trying to avoid it, i.e., people just change the variables they request access to. If DARE wants the TREs to work, there needs to be a level of flexibility… So many regulations and requirements make the use of data slow and difficult" Researcher, University

"We've been exploring mechanisms to define and categorise risk of data... This approach gives autonomy back to researchers, and makes requesting more anonymised data more appealing to researchers [i.e., lower risk, fewer controls]" Technologist, Commercial organisation

"[DARE UK should seek to be a] centralised broker for assisting interoperability of existing domain specific infrastructures" Workshop participant

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

3) Digital research infrastructure

Systems, including the physical and software infrastructure, vary widely depending on the types of data, the requirements of the users, the group who built the systems, as well as the subject areas. These systems have in many cases not been set up to be interoperable.

Unmet needs

Federation

The many physical and software infrastructures across the research landscape result in siloed working, particularly between research organisations and disciplines, within the UK. As a result, research communities can be unaware of other communities and data outside of their sphere. There is an increasing need for cross-disciplinary research to answer questions of importance, for example the impacts of climate change on infectious diseases. Despite an abundance of data, current environments do not facilitate cross-disciplinary working, and are in fact limiting the scope of research and the questions that can be asked and answered. Researchers wishing to access data from multiple environments face hurdles in terms of duplicative request

applications and delays. Researchers wishing to carry out cross-disciplinary research, for example incorporating environmental and health data, must work out from scratch how to access the data for each project, instead of creating legacy for the research community. Federation, where data in different physical places can be used together through a common interface, is a potential method of linking different sources of data. There are different perspectives of the best avenues to reach federation, and the best starting point for research infrastructures to begin to federate. De-identification of data and being able to link data across environments with common keys, whilst protecting sensitive information, is currently not solved.

Flexibility for researchers

Systems have not all been built with common requirements in mind. Irregular demand on compute power to drive high-power models results in system shortages and delays for researchers. TREs can be inflexible environments for researchers to work in if they are limited to a geographic space or limit the range of applications the analyst can employ. As a result, when researchers apply for secure data and realise the extent of requirements to access the data, some researchers avoid working in these environments by simply changing the variables they request access to.

Auditability

Increasing use of TREs is leading to a need for more review of outputs leaving TREs. This process currently varies by environment, and in some cases involves an individual manually 'eyeballing' extracted data. A manual process such as this cannot be audited easily, and limits reproducibility of analysis. There are however several good practices, for example the processes that TREs need to demonstrate are often in place before DEA accreditation cover this particularly around decisions being reproducible and replicable.

Opportunities

Federation

There is clear requirement and opportunity for creating environments that support linking of data from different disciplines. The federated approach has huge potential across UKRI-funded research, and there is an opportunity for DARE UK to help define federation at different levels. Federation of data and analysis could solve some of the unmet needs, particularly related to health data. Joining data across hospitals, linking primary and secondary care data, and linking health to crime, housing, education, environmental and consumer data were key examples. Federation could also fulfil specific non-healthcare-related use cases such as across the UK nations and cross-disciplinary research into the environment, human movement and economic opportunity.

Further, supporting efficiency by assessing the reuse of existing infrastructure and promoting best-practice examples of infrastructures such as JASMIN. TREs with widespread support across interviewees and potential for federation included ONS Secure Research Service, SAIL Databank, CO-CONNECT and OpenSAFELY. These are expanded on in the Appendix.

UK exemplar projects addressing specific aspects of federation could be supported as part of the DARE UK programme and the initiation of international dialogue to deepen understanding of international case examples of federated environments.

Enabling researchers to work flexibly and to be able to see data will support research involving advanced analytical capabilities. One researcher highlighted that text mining and natural language processing could be a central resource enabled across a federated network. There is also an opportunity to ensure access to data is as broad as possible, including remotely to international researchers. There is a fear that some TREs will remove remote access or individual data owners will remove permissions for their data to be accessed remotely post-Covid.

Cloud computing should be recognised for having better security and governance than on premise computing in some cases, and that not all cloud computing is commercial. Better understanding of the technical challenges to federation, such as the ability to use de-identification keys to link data whilst protecting sensitive information, require more detailed investigation.

Organisations wishing to make data available face a variety of platform options, and there was a call from interviewees to 'help the buyers buy'.

Flexibility for researchers

Technical challenges were highlighted by some individuals, particularly around managing irregular compute demands, and establishing the systems to allow compute resource to grow and shrink with demand, there is an opportunity to support the development of such capability. For example, combining resources such as bolting high performance computing (HPC) capability onto TREs is one theoretical approach.

Auditability

Another challenge that there is an opportunity to support is addressing the need for tools to be publicly auditable. One such example is the need for outputs from TREs to be audited. The audit process could be made smoother through automation of a peer review process, as opposed to 'eyeballing' of extracted data. Audit processes could be further enhanced to make the review process more efficient, fair, reproducible, objective, and be publicly auditable.

It was suggested that TRE providers could be motivated to publish source code for public benefit.

Many interviewees pointed out that the technical challenges though by no means trivial are however more straightforward relative to governance challenges.

Participant quotes

"In the field, I see a proliferation of amateurish TREs. More importantly, I see a problematic land grab of this area by commerce… I see a risk of 'lock-in' and disconnection from other environments" Researcher, University

"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

"If the data is going to used over and over again… You need legacy for the research community… The way you achieve a goal is through a change in behaviour, not just a change in infrastructure" Technologist, Research council

4) Capability and capacity

Research is underpinned by the people supporting infrastructure, using data, sharing data and in some cases, the data subjects. This section addresses training and staffing.

Unmet needs

Training

Institutions have skills shortages and the technical skillsets required of employees will continue to grow over time. Sensitivity of data is not understood by all researchers using data, and 'cloud skills' were referred to by participants as being particularly in demand; further discussion is required to elaborate what these skills are. Despite a move towards digital research, some fear that training must not become so digitally dominant that researchers lose the ability to work on physical sources of information.

Staffing

There is a need to support the career structures of individuals creating or engaging with digital research infrastructure. Institutions are broadly short of data scientists, statisticians, infrastructure, development and operations, and bioinformaticians, including those doing larger integration work. There may be a lack of capacity within organisations to adopt the recommendations coming from DARE UK.

Opportunities

Training

It was clear from the interviews that there are opportunities for upskilling researchers across disciplines, especially in the technical aspects of research using sensitive data. Examples of this are training researchers on how to code well for large scale analysis, or the fundamentals of good data management. There were comments that dedicated funding, structured career trajectories and the development of training programmes could be opportunities to address these challenges. In addition, there is an opportunity to raise the overall consciousness of security, governance and ethical issues, especially around sensitive data.

Staffing

Two approaches to support staffing were raised during interviews. Firstly, the creation of centralised capacity to help build and maintain TRE infrastructures, this needs to be investigated further in terms of its feasibility. Alternatively, increasing the use of secondments and collaborations, particularly between the private sector and public service.

Participant quotes

"We need to raise the consciousness of security issues, even if researchers don't feel the data is sensitive. Don't want it to become a tick-box exercise, but bring real benefit" Technologist, Research council

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

5) Demonstrating trustworthiness

Research is enabled by trust between data subjects (the public), data custodians (including commercial organisations), funders and researchers themselves. This section addresses trust and risk management associated with the use of sensitive data for research.

Unmet needs

Trust

Public concern exists around the risks associated with data sharing, particularly regarding commercial access to data. There is a need to demonstrate trustworthiness in order to gain the trust of individuals and data custodians. Gaining trust takes time and can be lost almost instantly.

Risk management

Data custodians have a duty of responsibility, often statutory in nature, to safely manage access to the data in their care; this in tandem with significant resource pressure due to staffing capacity relative to the volume of data access requests being received. Not surprisingly this has driven what can be perceived by the research community as an excessively risk averse approach to data access requests by data custodians, especially in environments where there is a greater degree of sensitivity linked to the nature of the data itself (for example personal health data). The lack of standardised and agreed risk management frameworks often leads to excessive risk aversion, for example through misunderstanding of data management methods, which leads to unnecessary delays to research. Data custodians are often understandably wary of losing control of the data if it is shared and there is also risk aversion on the part of the platforms distributing the information, which can lack mechanisms of prioritising time-sensitive projects.

Opportunities

Trust

All organisations working in research data, including the DARE UK programme, have an opportunity to address concerns regarding data collection and sharing, particularly around healthcare data. As an example, supporting researchers to work with communications experts to advertise security and governance measures, and successful examples of linked research. Examples of successful engagement exercises were brought up during interviews such as the OneLondon Citizens' Summit and by Understanding Patient Data.[5] By demonstrating compliance and safe use of information researchers have the opportunity to build trust, thereby enabling future researchers to do more.

Risk management

There is an opportunity to support data custodians in more efficiently and consistently managing their responsibilities around the risks to data under their purview by clarifying the legal position of different groups. As an example, a key opportunity would be to map the relevant legislation in each UK nation to understand what is and is not possible in each legislative geography - UK-wide support, including but not limited to the legislative landscape, was recognised as crucial in this regard. Interviewees acknowledged the challenges faced by data custodians in managing the volume of data access requests, however there was consistent feedback that there is an opportunity to improve the efficiency with which data access requests for research are processed. For example, gaining confidence of legal teams, contracts teams, and governance teams partially involved in data access around a UK-wide risk framework could support in addressing the trend of aversion to risk and subsequent effect on research outputs. Further, data custodian engagement and clarification of misunderstanding needs to happen, at all levels, including government sources. As a final point there is the opportunity to support platforms to manage and prioritise time-sensitive data applications, potentially with the support of automation as an example.

[5] OneLondon Citizens' Summit: https://onelondon.online/citizenssummit

Participant quotes

"This is not so much a technical challenge, but a 'hearts and minds' issue about trust and trustworthiness" Researcher and technologist, University

"Risk aversion amongst middle-management who don't necessarily understand novel ways to implement safe procedures, is leading to delays" Technologist, University

"[A challenge will be] bringing the public with you, upskilling and supporting them as advocates for data sharing" Workshop participant

6) Funding and incentives

The creation and maintenance of digital research infrastructures requires sustained funding. The execution of research using those infrastructures is also dependent on funding. When organisations work together in exchanging knowledge, responsibilities of the different organisations should be defined.

Unmet needs

Research culture

Research cultures do face challenges with incentivising collaboration, incentivising research cultures to improve and value more collaboration across institutions or across disciplines is extremely important for fostering cross-disciplinary research. For example, competition for funding can push researchers into institutional silos as opposed to collaboration. Standard funding timeframes can be short compared to data access processes and reviews, meaning researchers often fit their research questions to suit what data is available and accessible.

Funding and incentives

A challenge across the research and infrastructure communities is operating under limited and often time-limited funding, resulting in inefficient cycles of refresh along with funding cycles. This can drive organisations to trade off sustaining a resource and innovating. The DARE UK programme itself will have competitors and funding competitors. Researchers can experience funding issues in cross-disciplinary research when it is unclear to the researcher which research council is or should be responsible for funding. Engaging the public, which alone can be hard to define, is an important step in building trust, and working with communities requires long-term ongoing funding.

Responsibilities

Another challenge in data sharing is the division of rights and responsibilities between different groups. Rights and responsibilities for contributing organisations, are not always clear, for example there can be confusion as to which body is responsible for data quality.

Opportunities

Research culture

Research councils could encourage more productive cross-disciplinary relationships through funding that stipulated cross-disciplinary input. In addition to funding stipulations, showing the value of data sharing and demonstrating successful cross-discipline work and the methods by which it was achieved. There is a broad opportunity to improve the research and infrastructure cultures by bringing together for example key research environments stakeholders and major infrastructure providers in a friendly forum for exchange, to help innovation and communication, a process one interviewee said was successful in a recent large-scale TRE venture. Through any work, research that does not require use of TREs should not be penalised. The planning for such measures also needs to be with a long-term view, giving stakeholders enough time to align, so as not to feel alienated.

Funding and incentives

There is an opportunity to support the development of the funding and business model for connected or federated infrastructure, which is yet to be determined. Commercial organisations can be offered incentives in order to engage in data provision, and not necessarily financial incentives. To support the formation of connected organisations and build committed communities, clarifying the common objectives between groups. A collective approach can never be perfect for all groups, as such organisations must be prepared to compromise.

Responsibilities

There is an opportunity to clarify the responsibilities of different groups within a partnership. For example, those who maintain and curate data, or invest in structured data collection, could be given credit through the formation of guidelines that afford recognition through co-authorship on academic papers or acknowledgements. In addition, some datasets are published in peer review literature. Citation of these dataset-specific papers could be made a requirement of funding, and therefore be used to reflect the impact of a dataset.

Participant quotes

"There are a range of issues around interaction and sharing between universities… Stopped by systems to compete for research funds.. we are less inclined to work collectively on problems... There are deep-seated structural reasons why this is difficult… I see potentially risk in spending time designing the perfect system for interoperability etc, whereas that might not be the ideal solution to address the fundamental blocking issues" Researcher, Data research centre

"Grants always time-limited and short. Energy always gets diverted temporarily. Conditions attached to the grant are always different. A massive opportunity cost is created from small grants and diverting energy consistently. This is hugely disruptive locally" Researcher, University

"First of all, you need to spell out the common objectives between different groups" Researcher, Government body

"There need to be safeguards… for those that have invested [in data curation] e.g., co-authorship or acknowledgement in outputs" Researcher, University

Conclusions and recommendations

The purpose of DARE UK is to enable research, in particular cross-disciplinary research that maximises the trusted and secure use of sensitive data. There was widespread support for the DARE UK programme and its ambitions amongst interviewees and workshop participants. Some individuals are however cautious about how the programme will support the delivery of a coordinated vision for digital research infrastructures, and there was a degree of uncertainty about the remit of the programme.

Interviewees highlighted many use cases for a federated network of environments. Examples of cross-disciplinary use cases included linkage of environmental (e.g., climate change data) or social data to health data. The systems to link data across environments are not yet fully developed or widely adopted. Data across the UK has untapped potential and current systems limit the ambition of what could be asked of data, there is an opportunity for DARE UK to play a key role in a coordinated effort towards a better ecosystem for research and innovation. The six themes outlined in the section above encompass the broad variety of opportunities discussed during interviews or workshops.

There is a clear preference for specific measures. We repeatedly heard the twin challenges researchers face in 1) discovering data they could use, and 2) accessing this data. We understood a common picture of data being “locked” into bespoke environments, and fear that a proliferation of TREs could lead to the creation of more silos of disparate infrastructure. This is in the context of a gathering storm of opinion that de-identified data should not be released.

There is frustration with the current ecosystem’s efficiency and delays to research in TREs, there is an opportunity for DARE UK to bridge across environments, increase interoperability and pave the road to federation.

Recommendations:

- Regarding **data and discoverability**, continuing to explore the diversity of data standards used by subject area and maintain dialogue as to why these are not used consistently. For example, although favoured by certain TRE providers, the FIHR health data standard is not consistently used across healthcare datasets. Similarly, there is diversity in metadata standards, and distilling this could assist cross-disciplinary discovery of data. There is the opportunity to convene groups to set standards for recording of data lineage to support those running environments to audit the movement of data, and researchers to understand where data has come from.
- Support for the development of standards in **access processes and accreditation** is an opportunity to further support the work done by UK Statistics Authority (UKSA) under the Digital Economy Act (DEA); for example, by closely working with UKSA and TRE platforms to smooth out the access processes including setting standards to the staffing, timeframes, steps and individuals to be involved. In addition to the process, accreditation of TREs and researchers could be improved to include the diversity of platforms in use now, and international researchers. Registries of approved environments and researchers could support safe use of data and assist in consistent removal of unsafe users. **Accreditation of environments** would alleviate the burden on individuals (such as data custodians) to make their own assessments of TREs and allow for clarity in what is available. There is clear evidence that this is a sensible approach in the success of the work done by UKSA in the accreditation of TREs and the opening of access to more data via those TREs.

- In addition, a **proportionate approach related to data risk** has the potential to reduce barriers to accessing low-risk data for researchers, and gain trust of the public, commercial organisations and other data custodians.
- The concerns we heard about **digital research infrastructure** focused primarily around supporting irregular demand for compute, supporting advanced de-identification techniques, and automating the assessment of outputs extracted from TREs. Many individuals reported that the technical opportunities in the development of a more federated system are more straightforward, though certainly not trivial, relative to the governance challenges involved in coordinating across infrastructures. There were repeated calls for new digital infrastructure investments not to 'reinvent the wheel'.
- A bigger issue relative to digital research infrastructure was **capability and capacity**. Teams struggle to retain individuals, meaning there are often too few data scientists and engineers. Developing career pathways would build resilience within the UK research and innovation structures, secondments and partnerships with the private sector could also help support sustained resourcing.
- Many observe the challenges of **demonstrating trustworthiness**, with the public, other researchers, and commercial organisations. Engaging directly with members of the public (including patients), privacy activists, and those outside the 'usual' sphere, through outreach activities, building on the success of others in this space. In addition, demonstrating examples of best practice interdisciplinary working and publicising these properly could help to build trust.
- Finally, **funding and incentives** were raised by individuals, partially as an issue in itself but primarily as an obvious incentive that UKRI could use to facilitate agreement about sets of standards and collaboration. These standards could become a part of the license to operate similarly to other standards currently in place. Sustained funding, clarity in working definitions and standards for accreditation will incentivise the development of work towards more federated environments.

Although it is important that the approaches taken in future phases of DARE UK recognise previous attempts, including unsuccessful ones, DARE UK has an exciting opportunity to further investigate bringing together data and the use of modern capabilities as never before. This is in the context of the Covid-19 pandemic, which has accelerated data sharing and demands for insight reaching across traditional disciplines.

Appendix

Digital research infrastructure investments by UKRI funder – non-exhaustive overview of the interviewees' host institutions or investments affiliations

1. Arts and Humanities Research Council (AHRC)

AHRC funds outstanding original research across the whole range of the arts and humanities. This research provides economic, social and cultural benefits to the UK, and contributes to the culture and welfare of societies around the globe.

AHRC does not have its own data services and does not conduct audits of the data used by funded researchers. Researchers funded by AHRC use information, including data, held in collections. These collections include galleries, archives, museums, including those not necessarily funded by AHRC, and including those publicly available. Of all data arts and humanities researchers consult, only a small proportion is held in digital form.

AHRC is leading a programme of infrastructure work 'Towards a National Collection', a major five-year £18.9 million investment in the UK's world-renowned museums, archives, libraries and galleries. The programme will work towards creating a unified virtual 'national collection' linking metadata, to make access to information easier for researchers.[6] Funding is being provided through UKRI's Strategic Priorities Fund.[7]

2. Biotechnology and Biological Sciences Research Council (BBSRC)

BBSRC invests in world-class bioscience research and training. This research is helping society to meet major challenges, including food security, green energy and healthier, longer lives and underpinning important UK economic sectors, such as farming, food, industrial biotechnology and pharmaceuticals.

A recently published BBSRC review of data-intensive bioscience found that at least 50% of BBSRC's research grants now involve large-scale biological data and can be considered 'data-rich'.[8]

[6] Towards a National Collection: Opening UK Heritage to the World: https://ahrc.ukri.org/research/fundedthemesandprogrammes/tanc-opening-uk-heritage-to-the-world/

[7] UKRI - Strategic Priorities Fund: https://www.ukri.org/our-work/our-main-funds/strategic-priorities-fund/

[8] BBSRC Review of Data-Intensive Bioscience: https://www.ukri.org/wp-content/uploads/2020/11/BBSRC-201120-ReviewOfDataIntensiveBioscience.pdf

European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EMBL-EBI)

The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. EMBL-EBI is the largest data service provider , with many smaller organisations also generating datasets. Funders include BBSRC, Medical Research Council, the European Commission, the US National Institutes of Health, the Wellcome Trust, and industry partners.[9] EMBL is an international treaty organisation with partners across Germany, the UK, France, Italy and Spain.

The data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information (such as that from the Wellcome Sanger Institute), for which access depends on research consent agreements. All of the data and many of the software systems can be downloaded and installed locally.[10] EMBL-EBI has on site premise, local private cloud and also uses commercial cloud providers.

A £45 million boost to data and building infrastructure from UKRI's Strategic Priorities Fund is supporting academic and industrial demand for open access to biological data at one of the world's largest centres, EMBL-EBI. The investment will support intensifying growth in data resources driven by new technologies such as single cell sequencing and cryo-electron microscopy, and help ensure data are FAIR (Findable, Accessible, Interoperable and Reusable) for users globally.

EMBL-EBI is a node of the ELIXIR network - European life science infrastructure for biological information.[11]

Data standards: All resources are based on open and FAIR data. As a community leader, EMBL-EBI is concerned data has good metadata, so that the data is reused by others. EMBL-EBI uses infrastructure standards (BAM and CRAM etc.) but bespoke, specific and open standards for the metadata. EMBL-EBI is part of the Global Alliance for Genomics and Health, and uses published standards (e.g., for ontology, workflow execution standards, privacy standards) and toolkits that correspond with community standards. For molecular data, EBI uses a small number of highly used toolkits such as Galaxy and Bioconductor.

ELIXIR UK

The UK ELIXIR Node brings together 18 UK organisations to coordinate and provide training and services so that life sciences researchers can more easily discover, distribute, analyse, and store data, as well as exchange expertise and agree on standard approaches. It is a part of the broader ELIXIR distributed infrastructure for life sciences information, which aims to coordinate and develop vital bioinformatics resources such as databases and portals, toolkits, software, training materials and computing across Europe. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. It should be noted that the ELIXIR UK Node also receives funding through the Medical Research Council (MRC).

CyVerse

CyVerse is a project started in the United States, and the UK CyVerse node funded by the BBSRC, one node of the federated 'CyVerse' system, a way of sharing data across disciplines. UK CyVerse at the Earlham Institute in Norwich can potentially support a large proportion of the UK biological sciences community's data requirements, from genomics to phenomics.[12]

---

[9] EMBL-EBI – About us: https://www.ebi.ac.uk/about

[10] EMBL-EBI Tools & Data Resources: https://www.ebi.ac.uk/services

[11] ELIXIR – European life science infrastructure for biological information: https://bbsrc.ukri.org/research/international/engagement/research-infrastructures/elixir/

[12] CyVerse UK: https://cyverseuk.org/about/faqs/

It will enable UK researchers to access extensive data storage/back-up, local and global compute power, and structured, integrated analysis applications and workflows. It will also allow BBSRC-funded tools to become available globally and will help build a common international biological science platform supporting reuse of data, applications and resources, with consistent rules and formatting.

Data standards: The CyVerse Data Commons supports good data description through metadata templates (e.g., DataCite metadata template), bulk metadata upload and automatic collection of analysis parameters, inputs, and outputs.[13]

3. Economic and Social Research Council (ESRC)

ESRC is the UK's largest funder of research on the social and economic questions facing us today. This research shapes public policy and contributes to making the economy more competitive, as well as giving people a better understanding of 21st century society. ESRC funds organisations to create data assets or data collections (such as 'Understanding society', 'Census longitudinal studies'), and also provides access to data (e.g., UK census data, government funded surveys, longitudinal studies, international macrodata, qualitative data and business microdata) and support for researchers via data services.[14] The ESRC submits cross-UKRI proposals such as the 'Digital Footprints' proposal, which would involve input from multiple councils including NERC and BBSRC for example, and would be aided by more data sharing.

UK Data Service and UK Data Service Secure Lab TRE

ESRC's UK Data Service, which has existed since 1967, is one of the world's largest repositories of social sciences, economic, psychology and political data in the world, providing access via a large comprehensive metadata system.[15] The platform has thousands of users, most of these using open data. The UK Data Archive, at the University of Essex, provides researchers with training, support and data access as the lead partner of the UK Data Service.

The UK Data Service Secure Lab provides secure access to ONS data and non-ONS data that are too detailed, sensitive or confidential to be made available under the standard End User Licence or Special Licence (for example business surveys, and linked data from the Understanding Society survey). The Secure Lab was originally established in 2010. The UK Data Service Secure Lab TRE provides researchers access to more sensitive versions of the data – deidentified (e.g., names and addresses removed), but not anonymised. Linkage of datasets, for example the business surveys, is facilitated in the Secure Lab. Data accessed in this way cannot be downloaded. Once researchers and their projects are approved, they can analyse the data remotely from their organisational desktop, or by using the Safe Room.[16]

Data standards: The UK Data Service data adheres to the Open Archival Information System (ISO 14721:2012) as the bedrock of preservation and curation activities. ISO 15489 and ISO 16363 are also standards used to inform activities, and the UK Data Archive is accredited to ISO 27001 (for the provision of

[13] Data Management Overview: https://learning.cyverse.org/projects/foss-2020/en/latest/Data_management/overview.html
[14] ESRC Data Infrastructure Strategy Stakeholder Engagement: https://esrc.ukri.org/files/news-events-and-publications/publications/esrc-data-infrastructure-strategy-engagement-document/
[15] UK Data Service: https://esrc.ukri.org/research/our-research/uk-data-service/
[16] UK Data Service - Access levels and conditions: https://www.ukdataservice.ac.uk/use-data/secure-lab.aspx

the Secure Lab service). For metadata, DDI (Data Documentation Initiative) is used for the data collections, including controlled vocabularies and codes. UK Data Service does not harmonise variables on the basis of standards since this would a) impinge on the integrity of the data deposited and b) possibly create misleading codes. In addition, the UK Data Archive is accredited by the CoreTrustSeal as a trustworthy digital repository (TDR) which covers organisational infrastructure as well as digital object management.

Urban Big Data Centre (UBDC)

The UBDC is a research centre and national data service based at the University of Glasgow, promoting the use of big data and innovative research methods to improve social, economic and environmental well-being in cities. Themes include transport and mobility, education skills, the labour market, and spatial data from a range of domains such as administrative data, social media data, earth observation systems and CCTV sensor networks. UBDC has been jointly funded by the ESRC and the University of Glasgow since 2014.[17] Some data held by the UBDC is individual and disclosive, and some is commercially sensitive.

Consumer Data Research Centre (CDRC)

The CDRC was established in 2014 with funding from the ESRC and brings together world-class researchers from the University of Leeds, University College London, University of Liverpool and the University of Oxford.[18] The CDRC is the UK's leading source of consumer data, part of the ESRC's Big Data Programme, offering data under three tiers (and three services): Open, Safeguarded and Secure. Access to Safeguarded/Secure data is through a reviewed application process. Secure data (or 'controlled' data) is accessed through the labs in London and Liverpool, or remotely through UCL's Data Safe Haven. Topics include population and mobility, retail futures, transport and movement, finance and economy and digital.

Both the UBDC and CDRC apply their own standards and are working towards accreditation for Digital Economy Act approval, so they can deposit data in the ONS Secure Research Service.

Administrative Data Research UK (ADR UK)

ADR UK, a partnership of government, academic groups, and the ESRC team, was funded initially from July 2018 to March 2022, supported by £59 million drawn from the National Productivity Investment Fund (NPIF) via the ESRC. Each ADR UK partner (ADR England, ADR Northern Ireland, ADR Scotland, ADR Wales, ONS), including the Strategic Hub, is funded directly by ESRC with a portion of the total investment. A further £90 million funding extension was announced in 2021[19].

ADR UK works with UK government departments and devolved administrations to create linked research datasets from administrative sources, the facilitating safe and secure access for approved researchers to these newly joined-up and anonymised datasets via the ADR UK TRE network.[20] The ADR UK network is a federated research data infrastructure of TREs, and includes the ONS Secure Research Service, NISRA (ADR

[17] UBDC - Our Work: https://www.ubdc.ac.uk/about-ubdc/our-work/
[18] CDRC Data: https://data.cdrc.ac.uk/protecting-data
[19] ADR UK – Funding extension: https://www.ukri.org/news/data-research-initiative-secures-90m-funding-extension/
[20] About ADR UK: https://www.adruk.org/about-us/about-adr-uk/

Northern Ireland), Research Data Scotland (ADR Scotland) and SAIL Databank (ADR Wales).[21] See below for further information.

Office for National Statistics (ONS) Secure Research Service (SRS) TRE

The ONS is the UK's largest independent producer of official statistics and the recognised national statistical institute of the UK. The ONS receives a dedicated portion of the total investment in ADR UK (from the ESRC), initially from July 2018 to March 2022, to expand the SRS.[22] The SRS gives accredited or approved researchers secure access to de-identified, unpublished data (including the Census) for research projects for the public good.

The SRS is an accredited processor under the Digital Economy Act (DEA) and provides a safe setting (private cloud) as part of the Five Safes framework used to protect data confidentiality.[23] ONS procedures mean that analysis results don't disclose sensitive information, and the SRS operates within a legal framework with penalties for breaking these rules. Most datasets are available to access through remote access to the SRS. In some instances, the data can only be accessed from a SafePod, an approved Safe Setting or a secure connection to one. Safe Settings are located in London, Newport, Titchfield, Belfast and Glasgow.

The ONS is also leading the effort to bring an Integrated Data Platform (IDP) for government.[24] The programme will provide the opportunity to unlock the vast potential of linked data to enhance decision making for the public good and providing a quality evidence base. It will be a digital collaborative environment that enables cross-government teams and wider communities to deliver complex analytical outcomes by bringing together analysts, data, information governance and domain expertise in a safe, secure and trusted infrastructure.

The ONS has long advocated for sound data foundations and has contributed to a range of cross-government initiatives to shape data foundations and support the government in its data-driven decisions. As part of that effort the ONS will be developing and validating a set of data principles to be applied across the government.

Northern Ireland Statistics and Research Agency (NISRA) for administrative data TRE

NISRA's Research Support Unit (RSU) provides access to administrative datasets, allowing researchers safe access to project specific de-identified data in a secure environment to carry out secondary data analysis. NISRA holds data from across the government in Northern Ireland including data on travel, justice, housing, births and the census. NISRA also provides linkages between data from the Northern Ireland Longitudinal Study (NILS) and health and social care data. The NILS is a large-scale record linkage study of approximately 500,000 people (a representative c. 28% sample of the NI population).

NISRA operates a purely physical environment, with approximately 25 machines in one room for use. Discussions to make some of the data via remote access to the ONS SRS are currently underway. The Digital Economy Act does not allow incorporation of health data (as health data collected for an organisation's

[21] ADR UK- Trusted Research Environments: https://www.adruk.org/data-access/trusted-research-environments/
[22] ADR UK - Office for National Statistics: https://www.adruk.org/about-us/our-partnership/office-for-national-statistics/
[23] ONS Secure Research Service: https://web.www.healthdatagateway.org/collection/5493969548153421
[24] ONS - National Data Strategy – the ONS takes centre stage: https://www.ons.gov.uk/news/news/nationaldatastrategytheonstakescentrestage

health functions isn't part of the DEA), and NI does not have secondary use legislation, therefore adult health and social care information is not available via NISRA.

ADR NI brings together NISRA and ARDC NI (Queen's University Belfast and Ulster University). ADR NI is funded by the ESRC with a dedicated portion of the total investment in ADR UK, initially to March 2022.[25]

Social Data Science Lab

The Social Data Science Lab at Cardiff University is an ESRC Data Investment and forms part of the £64 million Big Data Network for the social sciences. An ESRC Capability Methods and Infrastructure Grant provides the Lab's core funding, and brings together crime, social, computer, and statistical scientists to study the empirical, methodological, theoretical and technical dimensions of new and emerging forms of data in social, policy and business contexts.[26]

HateLab

HateLab, based at Cardiff University, is a global hub for data and insight into hate speech and crime, it is funded by ESRC and the US Department of Justice.[27]

CLOSER

The CLOSER (Cohort and Longitudinal Studies Enhancement Resources) network was funded by the ESRC and MRC, with the initial five-year grant extended by the ESRRC from 2017 to 2022. The UCL Social Research Institute is the lead research institute.

CLOSER brings together world-leading longitudinal studies. The work maximises the use, value and impact of longitudinal studies to help improve understanding of social and biomedical challenges. CLOSER lead research to link data held by government to survey data collected by longitudinal studies across a range of areas, including health, geography and education. CLOSER's flagship resource, CLOSER Discovery, enables researchers to search and browse questionnaires and data from the UK's leading longitudinal studies to find out what data are available in unprecedented detail.

Population Research UK (PRUK)

PRUK is an initiative funded by ESRC, MRC and Wellcome, and currently being scoped by HDR UK.[28] PRUK will be a national data infrastructure, increasing the insights, innovations and research efficiency of the UK's wealth of social, economic and biomedical longitudinal population studies (LPS). It will focus on increasing access of data to new and potential users across academia, public bodies, charities and industry, and providing services and expertise to users that facilitates research. The programme will support current LPS through tackling challenges in data curation, linkage and analytics.

---

[25] ADR Northern Ireland: https://www.adruk.org/about-us/our-partnership/adr-northern-ireland/
[26] Social Data Science Lab: http://socialdatalab.net/
[27] HateLab: https://hatelab.net/data/
[28] Population Research UK: https://www.hdruk.ac.uk/population-research-uk/

4. Engineering and Physical Sciences Research Council (EPSRC)

EPSRC invests in world-leading research and postgraduate training across the engineering and physical sciences. This research builds the knowledge and skills base needed to address scientific and technological challenges and provides a platform for future UK prosperity by contributing to a healthy, connected, resilient, productive nation.

Alan Turing Institute

The EPSRC is the primary funder of the Turing Institute, a joint venture among 13 UK universities, and UK's national institute for data science and artificial intelligence. Work from the Turing Institute includes assessment of choices for secure, data-intensive research at scale in the cloud. One proposal includes a policy and process framework that incorporates data security threat and risk profiles into five sensitivity tiers, and, at each tier, specifying recommended policies for data classification, data ingress, software ingress, data egress, user access, user device control, and analysis environments.[29] With secure research environments for each project appropriate to their sensitivity classification, the Turing Institute hopes to maximise researcher productivity and minimise risk.

The Institute is at the heart of the recently announced business-led Prosperity Partnerships, in support of the government's ambitious new Innovation Strategy, and funded by EPSRC, businesses and universities.[30]

ARCHER

From 2013 until January 2021, ARCHER (Academic Research Computing High End Resource) was the UK Tier-1 National Supercomputing Service.[31] The ARCHER2 High Performance Computing (HPC) service, currently being developed, should be capable, on average, of over eleven times the science throughput of its predecessor, ARCHER.[32] ARCHER is managed by EPSRC as a joint investment with NERC.

The ARCHER2 Service is a world class advanced computing resource for UK researchers. ARCHER provides a capability resource to allow researchers to run simulations and calculations that require large numbers of processing cores working in a tightly coupled, parallel fashion. The major users of the system are materials scientists, climate scientists, physicists, engineers, and biosciences but ARCHER also supports others including medical research and industrial simulations.

ARCHER2 is provided by UKRI, EPCC, HPE Cray and the University of Edinburgh. ARCHER2 will be an HPE Cray EX supercomputing system with an estimated peak performance of 28 PFLOP/s. The machine will have 5,848 compute nodes, each with dual AMD EPYC Zen2 (Rome) 64 core CPUs at 2.2GHz, giving 748,544 cores in total. [33]

---

[29] Design choices for productive, secure, data-intensive research at scale in the cloud (2019) https://arxiv.org/abs/1908.08737

[30] The Turing announces new Prosperity Partnerships: https://www.turing.ac.uk/news/turing-announces-new-prosperity-partnerships-support-governments-innovation-strategy

[31] EPSRC – HPC facilities: https://epsrc.ukri.org/research/facilities/hpc/

[32] ARCHER2 Hardware & Software: https://www.archer2.ac.uk/about/hardware.html

[33] ARCHER2 Hardware & Software: https://www.archer2.ac.uk/about/hardware.html

National Quantum Computing Centre

The EPSRC and STFC are leading a programme to establish the National Quantum Computing Centre (NQCC) as part of phase 2 of the National Quantum Technologies Programme (NQTP). The NQCC represents a £93m investment over 5 years and will establish 4 key technology work streams. The new National Quantum Computing Centre (NQCC) is set to open in 2023.[34]

Other EPSRC initiatives of interest include:

UK Collaboratorium for Research on Infrastructure and Cities (**UKCRIC**) - an integrated research capability encouraging disparate areas of infrastructure to work collaboratively with each other.[35] UKCRIC will enable the UK to develop a world-class national infrastructure capability combining physical and social sciences. EPSRC is working with the 13 UKCRIC university partners in this Collaboratorium as the delivery partner.

The **Henry Royce Institute** - the UK's national institute for advanced materials research and innovation. The Royce is a consortium of leading institutions working on interoperability of data. Operating with its Hub at The University of Manchester, Royce is a Partnership of universities, National Nuclear Laboratory, and UK Atomic Energy Authority.[36]

The **Faraday Institution** - established in 2017 as an independent institute for electrochemical energy storage research, skills development, market analysis and early-stage commercialisation.

The **Rosalind Franklin Institute** - dedicated to transforming life science through interdisciplinary research and technology development.[37] The mission is to develop and apply disruptive new technologies in physical and engineering sciences that will change life science research, and in turn impact the UK pharmaceutical sector.

Scottish National Safe Haven TRE

The Scottish National Safe Haven is the responsibility of the electronic Data Research and Innovation Service (eDRIS) which is part of Public Health Scotland.[38] Over the last few years, the eDRIS service has expanded services to support non-health research. Formally the Scottish National Safe Haven is operated by EPCC (at the University of Edinburgh) under contract to eDRIS, on a private cloud. National Records of Scotland provide de-identification services and create joins between different data systems. A wide range of de-identified administrative datasets held in the National Safe Haven are made available for research through eDRIS, whilst Research Data Scotland (RDS) offers safe, secure and cost-effective access to data for research in response to Covid-19.

RDS offers 35 datasets including health activity, prescribing, GP data, vital events, census, schooling and children's social work. An RDS Data Catalogue in currently in development, and RDS's offer has been made

[34] National Quantum Computing Centre: https://www.nqcc.ac.uk/
[35] UKCRIC: https://www.ukcric.com/
[36] About Royce: https://www.royce.ac.uk/about-royce/
[37] Rosalind Franklin Institute: https://www.rfi.ac.uk/about/

[38] Use of the National Safe Haven: https://www.isdscotland.org/Products-and-Services/eDRIS/Use-of-the-National-Safe-Haven/

possible through ESRC's funding of ADR UK.[39] RDS is working with the ONS as part of the Integrated Data Programme (IDP).

eDRIS is offered under the Scottish Informatics and Linkage Collaboration (SILC). The initial technical infrastructure that supports SILC was funded by the MRC, Scottish Government and a collaboration of Universities and NHS National Services Scotland.[40]

Data standards: Preliminary work is being undertaken around applying OMOP to some of the datasets. NHS Scotland has some common data standards for some of their national datasets such as the Scottish Morbidity Records (SMR) which have a published data dictionary and standards and validation carried out at source.

5. Medical Research Council (MRC)

MRC is at the forefront of scientific discovery to improve human health. Its scientists and clinical professionals tackle the greatest health problems facing humanity in the 21st century, from the rising tide of chronic diseases associated with ageing to developing new medicines to treat rare genetic disorders.

Secure e-Research Platform (SeRP) TRE platform and SAIL Databank TRE

The Secure e-Research Platform (SeRP) was developed by the Population Data Science group at Swansea University, with support from the Farr Institute of Health Informatics Research funded by MRC.[41]

In 2011, SeRP was developed and implemented to provide a secure virtual environment and remote desktop protocol so that data could be accessed safely anywhere in the world. SeRP have exported their concept internationally including to Australia and Canada.

SeRP, a high-powered data management and sharing technology operating as a private research cloud, benefits from carefully designed Information Governance to ensure person-based data with high privacy risk is managed to the highest standards, accredited to the ISO270001 information security standard and to the UK Statistics Authority under the Digital Economy Act 2017.

SeRP UK is used by many research organisations across the UK to host research data within a secure research environment enabling collaborative research. SeRP powers the Adolescent Mental Health Data Platform (ADP), Dementias Platform UK (DPUK) and SAIL Databank.[42]

The SAIL Databank uses SeRP to provide controlled data access and High Performance Computing. SAIL Databank gives researchers secure remote access to datasets with billions of anonymised person-based population, health and social care data records. SAIL utilises the services of a Trusted Third Party (TTP) to facilitate data linkage via personal identifiers. SAIL Databank does not receive or handle identifiable data.

[39] Research Data Scotland: https://www.researchdata.scot/
[40] Charter for Safe Havens in Scotland: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/pages/4/
[41] SAIL Databank Overview: https://saildatabank.com/about-us/overview/
[42] SERP: https://serp.ac.uk/

It makes anonymised data available for research purposes where the rationale behind the research is for public benefit and interest.

SAIL Databank is also the primary TRE to a wide range of projects. One such project is BREATHE – the health data research hub for respiratory health. BREATHE is one of seven Health Data Research Hubs funded through UKRI's Industrial Strategy Challenge Fund, and coordinated by Health Data Research UK (see below). SAIL Databank provides the environment for data hosting, data analysis, and a governance framework for data access and analysis to the respiratory research community.[43]

Health Data Research UK (HDR UK)

HDR UK are an independent, registered charity supported by 10 funders, including the MRC, EPSRC and ESRC, working across 31 locations within the UK. HDR UK's mission is to unite the UK's health data to enable discoveries that improve people's lives.

HDR UK does not control any health data. They work with organisations that hold and manage datasets and support connections between these datasets to support access for research and innovation. HDR UK supports the FAIR principles of Findability, Accessibility, Interoperability, and Reusability in research.[44]

HDR UK is responsible for the Health Data Research Innovation Gateway, providing a common entry point to discover and request access to UK health datasets through a metadata catalogue. Users can search for health data tools, research projects, publications and collaborate via a community forum.[45]

Francis Crick Institute

The Francis Crick Institute is independent organisation, established to be a UK flagship for discovery research in biomedicine. The founding partners are the MRC, Cancer Research UK, Wellcome, UCL, Imperial College London and King's College London. The Crick was formed in 2015 and is now the biggest biomedical research facility under one roof in Europe.[46]

Genomics England Research Environment TRE

Funded by the Wellcome Trust, Cancer Research UK and the MRC, the Genomics England Research Environment has over 2,000 researchers onboarded to carry out analysis with a range of tools and a full high performance computing environment, with over 20Pb of genome data from the 100,000 genomes project.[47,48]

[43] BREATHE - Our Trusted Research Environment: https://www.ed.ac.uk/usher/breathe/who-we-are/our-trusted-research-environment
[44] HDR UK – About us: https://www.hdruk.ac.uk/about-us/what-we-do/
[45] Innovation Gateway: https://www.healthdatagateway.org/
[46] Francis Crick Institute: https://www.crick.ac.uk/about-us
[47] Genomics England Research Environment: https://www.genomicsengland.co.uk/about-genomics-england/research-environment/
[48] UK Health Data Green Paper on TREs: https://ukhealthdata.org/wp-content/uploads/2020/04/200430-TRE-Green-Paper-v1.pdf

CO-CONNECT

CO-CONNECT is a project funded by MRC and the Department of Health and Social Care (part of NIHR) as part of their response to the Covid-19 pandemic. CO-CONNECT is led by the University of Nottingham and has built a collaboration of over 40 leaders from across 20 organisations across the UK. The goals are to standardise antibody data collection across the UK, configure an infrastructure which enables trustworthy, fast, de-identified, secure analysis of data sets from across multiple sources, and answer key questions about immunity to Covid-19 and the implications for patient outcomes.[49] This is a key example of a completely federated infrastructure in operation within the UK, enabling distributed querying of data. Data is not linked by CO-CONNECT, but advanced de-identification methods are used to understand the degree of overlap between datasets whilst protecting each individual's identity. The CO-CONNECT team is working with support of the Turing Institute in developing their technology.

Other MRC initiatives of interest include:

UK Biobank - a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK cohort study participants.[50] The data – the largest and richest dataset of its kind – is anonymised and made widely accessible to researchers around the world. UK Biobank makes record-level data available to approved researchers and approved projects, and is enabling this via a new cloud-based Research Analysis Platform (RAP).[51] Funding for the platform has come from Wellcome, and development has been by DNAnexus in collaboration with Amazon Web Services (AWS).[52] The RAP is currently in a test phase, assessable to invited researchers. General availability is expected in Q3 2021.[53]

UK Dementia Research Institute – made up of seven centres hosted in universities across the UK, the Institute represents a joint £290 million investment into dementia research from the MRC and others. Dementias Platform UK (DPUK) was funded by the MRC. It offers access to detailed information for over 3 million individuals, from 47 cohort studies via the DPUK Data Portal.[54] The data remains on the DPUK servers - data files provided via the Portal are not physically removed, only outputs such as results files are permitted.[55]

6. Natural Environment Research Council (NERC)

NERC is the UK's leading investor in environmental science. Its world-class research, skills and infrastructure solve major global issues such as the climate crisis and plastic pollution, and bring benefits to the UK, such as affordable clean energy, sustainable agriculture, clean air, and resilience.

[49] CO-CONNECT: https://co-connect.ac.uk/
[50] UK Biobank Background: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us
[51] Accessing UK Biobank Data: https://biobank.ndph.ox.ac.uk/~bbdatan/Accessing_UKB_data_v2.3.pdf
[52] Biobank - Data Analysis Platform: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-creates-cloud-based-health-data-analysis-platform-to-unleash-the-imaginations-of-the-world-s-best-scientific-minds
[53] DNAnexus - Research Analysis Platform: https://www.dnanexus.com/partnerships/ukbiobank
[54] Dementias Platform UK (DPUK): Multi-modal Data Access in a Digital Age: https://ukdri.ac.uk/events/dementias-platform-uk-dpuk-multi-modal-data-access-in-a-digital-age
[55] DPUK - Welcome to the Data Portal: https://portal.dementiasplatform.uk/

NERC funds six research centres:

- National Centre for Earth Observation
- British Antarctic Survey
- Centre for Ecology and Hydrology
- National Oceanography Centre
- National Centre for Atmospheric Science
- British Geological Survey

NERC commissions the Environmental Data Service.[56] This supports five data centres covering a range of discipline areas:

- British Oceanographic Data Centre (BODC) (Marine)
- Centre for Environmental Data Analysis (CEDA) (Atmospheric, Earth Observation, and Solar and space physics)
- Environmental Information Data Centre (EIDC) (Terrestrial and freshwater)
- National Geoscience Data Centre (NGDC) (Geoscience)
- Polar Data Centre (Polar and cryosphere)

Some of the data centres are embedded in their respective research centres (e.g. EIDC, NGDC and the British Antarctic Survey) whereas BODC and CEDA act more as separate entities, giving them more freedom to act as individual 'honest brokers' for their specific communities. The data centres hold both NERC-generated data from NERC grants and long-term programmes by the research centres, and a considerable amount of non-NERC data where the NERC data centre acts as repository and a portal, e.g., large amounts of Met office data.

Although each of the data centres are domain-specific, there is an ongoing programme of work fully supported by the data centres and NERC to better integrate the infrastructure, policies and procedures to provide a single point of entry to NERC's data centres. Data centres coordinate this through two groups - the Information Strategy Group and the Data Operations Group. A significant amount of work is linked to JASMIN (hybrid of high performance computing and data server), and NERC is in discussion about how other research councils can benefit from JASMIN (see STFC section below). NERC has been asked to lead the cross-UKRI net-zero digital research infrastructure programme and is currently entering an 18 month scoping phase.

Environmental Information Data Centre (EIDC)

EIDC is part of NERC's Environmental Data Service and is hosted by the UK Centre for Ecology & Hydrology (UKCEH). The EIDC manages nationally-important datasets concerned with the terrestrial and freshwater sciences.[57]

UKCEH is an independent, not-for-profit research institute, and a strategic delivery partner for NERC. UKCEH's 500 scientists provide the data and insights that researchers, governments and businesses need to create a productive, resilient and healthy environment. Through the national capability programmes, funded by NERC, UKCEH enable the UK research community to stay at the forefront of environmental science

[56] NERC – research sites: https://nerc.ukri.org/research/sites/

[57] NERC The Environmental Information Data Centre: https://www.ceh.ac.uk/nerc-data-centre

globally, meeting national strategic needs, informing government and business decision-making on environmental issues.

The EIDC holds terrestrial, freshwater and atmospheric environmental data generated by researchers in the UK. Most of the data is available under Open Licence agreements and is freely accessible.[58]

Centre for Environmental Data Analysis (CEDA)

CEDA is run jointly with the STFC. The CEDA Archive is the national data centre for atmospheric and earth observation research. Sources include aircraft campaigns, satellites, automatic weather stations and climate models, amongst many more. The CEDA Archive hosts over 18 Petabytes of atmospheric and earth observation data from climate models, satellites, aircraft, met observations, and other sources.[59]

Data standards: A standard thesaurus was developed to address the issue of standard terms, based on the Climate and Forecast (CF) Metadata Convention and the specific needs of the CEDA and atmospheric scientists. Many groups have adopted netCDF (network Common Data Form) as a standard way to represent their scientific data.

JASMIN is deployed on behalf of NERC and operated by the STFC. It is jointly managed by CEDA and STFC's Scientific Computing department. JASMIN is a globally unique data intensive supercomputer, and private cloud, for environmental science. JASMIN provides access to different types of data resources including curated data in the CEDA Archive. See the STFC section for detail.

JASMIN

JASMIN is the UK's data analysis facility for environmental science, designed, built, and managed by STFC on behalf of NERC's Centre for Environmental Data Analysis (CEDA).[60] JASMIN is one of the key UK facilities with major digital infrastructure, a super-data-cluster, and provides NERC scientists with the ability to create, share and access cutting-edge computing and storage technology on a flexible, collaborative platform, in general for free at the point of use. It is part supercomputer and part data-centre, with far more storage than computing, and provides a globally unique computational environment. The JASMIN infrastructure provides a compute and storage cloud for researchers in the UK, linked together by a very high bandwidth network in a unique topology. With its significant compute power and a bandwidth greater than usual in data centres, the JASMIN network topology is more typically found in the largest global-scale data centres.

JASMIN serves the scientific community by providing a range of computing services (batch, interactive, community cloud), supporting a variety of data types in a scalable environment, as scientists bring their data to JASMIN.[61]

Data standards: With an archive of hundreds of millions of files, both CEDA and users rely on data standards to facilitate data management and exploitation. CEDA supports the Climate and Forecast (CF) Convention for file metadata, and works with partners and international networks to promote the use of data standards.[62]

---

[58] EIDC: https://eidc.ac.uk/
[59] The CEDA Archive: https://archive.ceda.ac.uk/
[60] JASMIN: https://www.jasmin.ac.uk/about/
[61] CEDA - JASMIN: https://www.ceda.ac.uk/services/jasmin/
[62] Perspective from the Centre for Environmental Data Analysis: http://cedadocs.ceda.ac.uk/1381/1/MartinJuckes_at_JAMSTEC_v4.pdf

JASMIN supports a large range of standards such as CF Standard Names (for common vocabularies) and netCDF[63] data format. Other examples include the Open Geospatial Consortium (OGC), a not-for-profit organisation which promotes the development of data model and API standards for exchange of geospatial data, and STAC[64], a specification for catalogue and search of geospatial data that has grown from the community using Earth Observation satellite data. JASMIN uses REST as the preferred architectural pattern for web services.

7. Science and Technology Facilities Council (STFC)

STFC is a world-leading multi-disciplinary science organisation. Its research seeks to understand the Universe from the largest astronomical scales to the tiniest constituents of matter, and creates impact on a very tangible, human scale.

DAFNI

DAFNI is a computational platform, purpose-built, hosted and managed by STFC in a partnership led by the University of Oxford, and funded for its development years by a grant from the UK Collaboratorium for Research on Infrastructure and Cities (UKCRIC).[65] The DAFNI platform offers UK researchers a place to share their work and collaborate to study rich scenarios where changes in one area affect other areas. This might be the impact of climate change on the flooding in cities, or how new railways might affect where people live and work.

DAFNI brings together disparate data sources, high performance computing, analytics and visualisations into a collaborative platform, allowing research to be carried out more quickly, with larger research scope than otherwise, for models developed by researchers to be built on by others, and enables online collaborations.

Hartree Centre and EPCC (formerly Edinburgh Parallel Computing Centre)

Since 1990, EPCC has gained an international reputation for leading edge capability in all aspects of high-performance computing (HPC), data analytics and novel computing. The EPCC supercomputer centre at the University of Edinburgh hosts and administers a number of national-level facilities (such as ARCHER and ARCHER2) for use by UK researchers. EEPCC currently runs four national services: ARCHER, the UK's primary academic research supercomputer, the DiRAC Extreme Scaling service, Cirrus, an EPSRC a Tier-2 HPC service, and the UK Research Data Facility. EPCC is responsible for developing and hosting the Edinburgh International Data Facility (EIDF).[66] PCC funding has come from combined funding including the ESRC and EPSRC.[67]

Funded by STFC, the Hartree Centre is transforming UK industry through high performance computing, data analytics and artificial intelligence (AI) technologies. Backed by over £170 million of government funding and

[63] NetCDF: https://www.unidata.ucar.edu/software/netcdf/
[64] STAC: https://stacspec.org
[65] DAFNI: https://dafni.ac.uk/boost-to-uk-infrastructure/
[66] EPCC – A history: https://www.epcc.ed.ac.uk/about/history
[67] EPCC – About us: https://www.epcc.ed.ac.uk/about

significant strategic partnerships with organisations such as IBM and Atos, the Hartree Centre is home to some of the most advanced computing, data and AI technologies in the UK.[68]

In March 2021, UKRI announced the UK joining a European network to advance high performance computing. Two of the UK's leading supercomputing facilities, Hartree Centre and EPCC, have combined to form a national supercomputing competence centre, a high performance computing, data analytics and AI research facility.[69]

IRIS

IRIS works with the UK's major digital research infrastructure providers to help create and deliver resources to support science.[70] IRIS is a cooperative community driven project helping develop and grow the digital research infrastructure that will allow STFC to continue to play a leading role in world class science.

Other research environments without funding from UKRI

UK-wide

HMRC Datalab TRE

The HMRC Datalab allows approved researchers to access de-identified HMRC data in a government accredited secure environment. [71] The aim of the Datalab is to produce high quality analysis that benefits both HMRC and the wider research community. There is currently a relatively small research community of London-based universities using the Datalab due to the requirement to be in the London office. Datalab projects are not commissioned by HMRC.

OpenSAFELY TRE

OpenSAFELY is a collaboration between academics and health record software companies to analyse NHS primary care records from more than 24 million patients to understand the impact of Covid-19 in the UK.[72]

It is a fully open source and highly secure analytics platform for NHS data created during the Covid-19 pandemic. It is executing code across an unprecedented scale of data: 58 million patients full raw GP records - 100 billion rows of information - linked onto various other sources including SGSS, SUS/HES, ECDS, ISARIC, ICNARC, ONS death, and more. All code for the platform, and for data management and analysis of each output, is shared under open licenses for review and re-use. It should be noted that OpenSAFELY has received funding from the Medical Research Council (MRC) for Covid-19 related work.

---

[68] Hartree Centre – About us: https://www.hartree.stfc.ac.uk/Pages/About-Us.aspx

[69] UKRI News - UK joins European network to advance high performance computing: https://www.ukri.org/news/uk-joins-european-network-to-advance-high-performance-computing/

[70] Iris – About Iris: https://www.iris.ac.uk/about-iris/

[71] HMRC Research: https://www.gov.uk/government/organisations/hm-revenue-customs/about/research#the-hmrc-datalab

[72] OpenSAFELY: https://www.opensafely.org/about/

England

NHS Digital TRE

NHS Digital's TRE service for England provides approved researchers with access to essential linked, de-identified health data to answer Covid-19 related research questions. The service is being delivered in partnership with HDR UK. The TRE is hosted on a public cloud.[73] Compute power isn't directly allocated per user; compute and memory capacity is managed across the service. Flexible cloud compute is available through Amazon Web Services.[74] As part of the TRE process, NHS Digital routinely seeks advice from the Independent Group Advising on the Release of Data (IGARD) to ensure that the highest standards of data stewardship and governance are upheld.[75] All projects requiring access to the TRE service first have to apply for their data through the Data Access Request Service (DARS).

NHS Digital / British Heart Foundation (BHF) CVD-COVID TRE - NHS Digital is working with the BHF Data Science Centre to develop its TRE, which is already being used to analyse the impact of Covid on cardiovascular diseases and the safety of vaccines. The BHF Data Science Centre CVD-COVID is in the NHS Digital TRE. The work has been undertaken by NHS Digital working with the CVD-COVID Consortium. The BHF Data Science Centre has been the first client, and working alongside HDR UK the TRE service has delivered tools and data to support analysis across a range of linked data sources. The BHF Data Science Centre is a partnership between HDR UK and the BHF, and sits within HDR UK.[76]

NHS Digital / DATA-CAN TRE - DATA-CAN is working with NHS Digital to provide access to cancer data in the NHS TRE. Two of DATA-CAN's founding partners, the University of Leeds and Leeds NHS Teaching Hospitals Trust, are leading on this work. Nationally collected NHS Hospital Episode Statistics (HES) and Covid-19 testing data is available to approved researchers via the HDR Innovation Gateway. The next phase will include national cancer datasets (Cancer Outcome and Services Data set, chemotherapy and radiotherapy datasets) and be available to researchers from October 2021.[77]

Foundry (Palantir) NHS COVID-19 Data Store TRE

The NHS COVID-19 Data Store sits on a Microsoft Azure platform under contract with NHS England and NHS Improvement.[78] Within that secure cloud processing environment, Palantir (acting under instruction from NHS England) manage their platform which is called Foundry. Data and code do not leave the Foundry platform.

---

[73] NHS Digital TRE: https://web.www.healthdatagateway.org/collection/9118411587595364

[74] Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource (2021) https://www.bmj.com/content/373/bmj.n826

[75] NHS Digital - Trusted Research Environment service for England: https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england#governance-and-transparency

[76] BHF Data Science Centre: https://www.hdruk.ac.uk/helping-with-health-data/bhf-data-science-centre/

[77] DATA-CAN Trusted Research Environment: https://www.data-can.org.uk/health-data/trusted-research-environment/

[78] NHS Covid-19 data store: https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/nhs-covid-19-data-store /

International COVID-19 Data Alliance (ICODA) workbench TRE

ICODA is convened by HDR UK. The ICODA 'Workbench' has been separately commissioned. It is provided by Aridhia Informatics and allows researchers to discover, access and analyse global multi-dimensional datasets while respecting confidentiality and privacy. The range of partners includes UK TRE providers such as SAIL Databank, UK SeRP and Genomics England.[79]

Data standards: Aridhia Informatics is a supporter of the FIHR data standard, and ensuring it is used as standard.

ORCHID (Oxford-Royal College of GPs Clinical Informatics Digital Hub) TRE

ORCHID is the secure data processing environment for the Oxford-Royal College of GPs Clinical Informatics Digital Hub, operating via a cloud platform.[80] Theme leads and clinicians (with expertise in the theme) use the SNOMED CT code tool developed by the Oxford-RCGP RSC team to curate variables/codes within each theme.[81]

Northern Ireland

HSC Northern Ireland (Health and Social Care) Honest Broker Service TRE

The HSC Honest Broker Service provides access to health and social care data. Anonymised patient level data is provided for research, with access only permitted in a controlled fashion via a safe research environment. The safe setting is either accessed via attendance at the Safe Haven in the Business Services Organisation headquarters or remotely via the Health Data Research Northern Ireland UK Secure e-Research Platform (Health Data Research Northern Ireland UK SeRP).[82] Identifiable data does not leave the Honest Broker Service-governed environment.[83] The Honest Broker Service enables the provision of anonymised, aggregated and in some cases de-identified health and social care data to the NI Departmental of Health, Health and Social Care organisations and for anonymised data for ethically approved health and social care related research.

Scotland

Regional hubs – Local safe havens - NHS TREs

Local Safe Havens operate in the regional hubs of Aberdeen, Dundee, Edinburgh and Glasgow; with a national Safe Haven at National Services Scotland.[84] Safe Havens in Scotland were established as part of a

[79] ICODA: https://icoda-research.org/
[80] ORCHID: https://web.www.healthdatagateway.org/collection/5626663352808625
[81] Using Oxford-RCGP RSC for observational studies: https://orchid.phc.ox.ac.uk/index.php/orchid-data/
[82] Honest Broker Service: https://hscbusiness.hscni.net/services/2454.htm
[83] Safe and secure remote access to Northern Ireland's Health and Social Care data: https://www.hdruk.ac.uk/news/new-initiative-supports-safe-and-secure-remote-access-to-northern-irelands-health-and-social-care-data-for-researchers/
[84] NHS Scotland - Data Safe Haven: https://www.nhsresearchscotland.org.uk/research-in-scotland/data/safe-havens

need for delivering research excellence and rapid access to high-quality health data for research. They were developed in line with the Scottish Health Informatics Programme (SHIP) blueprint which outlined a programme for a Scotland-wide research platform for the collation, management, dissemination and analysis of anonymised Electronic Patient Records (EPRs), including sensitive data. The agreed principles and standards to which the Safe Havens are required to operate are set out in the Safe Haven Charter. Data remains under the control of the NHS and complies with legislative and NHS policies.

The four regional TREs include:

The Grampian Data Safe Haven (DaSH), Aberdeen - opened in May 2012 by NHS Grampian and the University of Aberdeen. DaSH provides a secure setting for data linkage and data hosting projects accessed through a Virtual Private Network (VPN). DaSH has been accredited by the Scottish Government (November 2017) and meets the Information Security and Governance standards outlined in the Charter for Safe Havens in Scotland 2015. Additionally, DaSH is accredited to ISO27001:2013 Information Security Management.[85]

Health Informatics Centre (HIC), Dundee - a leader in health data linkage, and the first centre in Scotland to offer a Safe Haven, which is now Nationally Accredited, and ISO27001 certified. HIC maintains a clinical data repository of eHealth data covering approximately 20% of the Scottish population. The eHealth repository combines routine collected datasets for the Tayside and Fife population and Tayside, with local speciality research, and clinical datasets.[86]

DataLoch, Edinburgh - DataLoch has been entrusted by NHS Lothian with routine data collected as part of people's day-to-day interactions with health and social care services. DataLoch is currently accepting applications from academics and health and social care professionals within the South-East Scotland region.[87] Once projects and users are approved, the necessary data are supplied to researchers either within NHS Lothian to specified staff, or accessed through the secure Scottish National Safe Haven facility managed by the eDRIS team within Public Health Scotland, hosted by the EPCC at the University of Edinburgh.

Glasgow Safe Haven: The safe haven facilitates researchers access to de-identified health datasets, offers a secure ISO-accredited data analytics platform and delivers expert support to enable data-driven discovery with de-identified NHS data. The Safe Haven provides secure access to projects from the safe room at the University of Glasgow, or VPN access to the Safe Haven research environment.[88]

[85] Grampian DaSH: https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php
[86] HIC Trusted Research Environment (Dundee): https://www.dundee.ac.uk/hic/hic-trusted-research-environment/
[87] DataLoch: https://dataloch.org/
[88] GLASGOW Safe Haven Secure NHS data research: https://www.nhsggc.org.uk/media/266674/glasgow-safe-haven-user-guide.pdf

List of interviewee organisational affiliations

The organisations listed below are simply a snapshot of the initial landscape review activities that the DARE UK programme has executed to date. As described in the introduction this consisted of 60 interviews across 79 individuals and two virtual workshops, open to the public, with an attendance of approximately 50 individuals per workshop. This list will continue to develop as the landscape review matures with additional contributions, engagement events and input throughout Phase 1 of the DARE UK programme.

1. AIMES
2. Aridhia Informatics Ltd.
3. Arts and Humanities Research Council (AHRC)
4. Biotechnology and Biological Sciences Research Council (BBSRC)
5. British Heart Foundation Data Science Centre
6. Centre for Environmental Data Analysis (CEDA)
7. Centre for Radiation Chemical and Environmental Hazards (CRCE)
8. Consumer Data Research Centre (CDRC)
9. Economic and Social Research Council (ESRC)
10. Edinburgh Parallel Computing Centre (EPCC)
11. Electronic Data Research and Innovation Service (eDRIS)
12. Engineering and Physical Sciences Research Council (EPSRC)
13. European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI)
14. Exeter University
15. Genomics England
16. Health and Social Care Northern Ireland (HSC NI)
17. Health and Social Care Northern Ireland Honest Broker Service TRE
18. Health Data Research (HDR) UK BREATHE Hub
19. Health Informatics Centre
20. Imperial College Health Partners
21. Imperial College London
22. Institute of Metabolic Science
23. Jisc
24. King's College London
25. London Health Data Strategy*
26. London School of Hygiene & Tropical Medicine (LSHTM)
27. Medical Research Council (MRC) Epidemiology Unit
28. National Health Service (NHS) X*
29. National Institute for Health Research (NIHR)*
30. Natural Environment Research Council (NERC)
31. Northern Ireland Statistics and Research Agency (NISRA)
32. Office for National Statistics (ONS)
33. OpenSAFELY*
34. Public Health England
35. Public Health Scotland
36. Research Data Scotland
37. RISG Consulting
38. Rolls-Royce plc

39. SAIL Databank
40. Science and Technology Facilities Council (STFC)
41. Secure eResearch Platform (SeRP)
42. The Administrative Data Research Centre Northern Ireland (ADRC-NI)
43. The Alan Turing Institute
44. The Francis Crick Institute
45. UK Centre for Ecology & Hydrology (UKCEH)
46. UK Data Archive
47. UK Data Service (UKDS)
48. UK Research and Innovation (UKRI)
49. UK Statistics Authority
50. Understanding Society, Essex
51. University College London (UCL)
52. University of Birmingham
53. University of Cambridge
54. University of Edinburgh
55. University of Essex
56. University of Exeter
57. University of Exeter Medical School
58. University of Glasgow
59. University of Leicester
60. University of Liverpool
61. University of Manchester
62. University of Nottingham
63. University of Swansea
64. Urban Big Data Centre
65. Warwick Business School
66. Wellcome Sanger Institute
67. Wellcome Trust

**Discussions took place outside of the initial landscape review interviews due to time constraints – insights are nevertheless integrated within this document.*

Contact

enquiries@dareuk.org.uk

DOI: https://doi.org/10.5281/zenodo.5584696

Visit the DARE UK website: www.dareuk.org.uk