DARE UK

A Federated Architecture for a National Data Library

M. Amugi, R. Baxter, C. Cole, C. Goble, P. Harrison, W. Igbo, E. Jefferson, F. McDonald, A. Morris, P. Quinlan, D. Seymour, C. Smith, J. Smith, B. Stewart, S. Thompson, J. Wood

November 2024

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

Contents

Words: 5,999 (*MS Word* count).

1. Introduction

The call for a National Data Library (NDL) is timely. Granting authorised researchers controlled access to sensitive public data at scale will be key to enabling more efficient services and research within the UK.

In 2021 the DARE UK programme[1] was set up by UKRI with a mission to put the UK at the forefront of sensitive data research and innovation by assembling the tools, technologies and standards needed to streamline secure data linkage and use.[2] Since then, the programme has conducted public and professional consultations, landscape reviews and programmes of research, development and proofs-of-concept to assemble most of the key elements needed for a federated approach to the NDL [1][3][4][16][17][18].

Our focus is **access to sensitive data for approved research purposes**, and this white paper describes a comprehensive and self-consistent way to achieve this. In defining the overall architecture we have assimilated not only the pioneering results from the DARE UK programme to date, but also current thinking and practice from large-scale scientific infrastructures, data spaces and digital public infrastructures from both home and further afield. The associated HDR UK white paper provides a health data perspective.

We are guided by a handful of key principles:

- Public sector data used for research should always be anonymised where possible, or if not de-identified (pseudonymised), following the principles of the UK GDPR.
- Research with public sector data should be conducted inside specially secured computing environments known as trusted research environment, or TREs.[3]
- Secure computing environments must implement appropriate governance and processes for all aspects of their operation.
- The use of data for research should be transparent to the public. Public trust and support are key.
- The processes and systems supporting data research across the UK should be unified in their approaches where possible.
- Where feasible, processes enabling access to sensitive data for research should be standardised and centralised.

[1] https://dareuk.org.uk/how-we-work.

[2] Sensitive data are data and information, the loss, misuse, or unauthorized access to or modification of, that could adversely affect commercial, the national interest or the conduct of research and innovation programs, or the privacy to which individuals are entitled to under GPDR.

[3] TREs come in various shapes and sizes, with no currently settled definition. For the purposes of this paper we use TRE to mean a class of systems which include Secure Data Environments (SDEs) in the National Health Service in England, Safe Havens in Scotland, processing environments as defined in the Digital Economy Act 2017 (DEA) and secure processing environments as defined in European Health Data Space legislation. These systems are typically operated according to information governance practices and processes modelled on the Five Safes approach developed by the Office for National Statistics (ONS) [2]. Their general purpose is to enable researchers to access sensitive datasets across administrative boundaries whilst maintaining governance control of the data. The UK Health Data Research Alliance & NHSX 2021 paper, *Building Trusted Research Environments –Principles and Best Practices* [12] and the SATRE specification [4] take important steps towards defining what a TRE actually is.

- Services can be provided by public and/or private sector partners.

Outputs from Phase 1 of the DARE UK programme are already being adopted across the UK and Europe. The NHS England SDE Network, for example, has adopted the community developed Standardised Architecture for Trusted Research Environments (SATRE) specification for TREs [4][5] and the EU-funded EOSC-ENTRUST project[4] has adopted both the DARE UK Federated Architecture Blueprint [1] and the SATRE specification as foundations for an EU-wide network of national TREs.

Community buy in and engagement is illustrated by the 300 member, DARE UK-funded UK-TRE community, whose membership spans data provider and TRE organisations alike, including The Alan Turing Institute, The Francis Crick Institute, Health Data Research UK, UK SeRP, ONS, the NHS England SDE Network, SAIL Databank, the Scottish Safe Haven Network, Genomics England, UK Biobank, Our Future Health, industry partners and the UK Data Service.

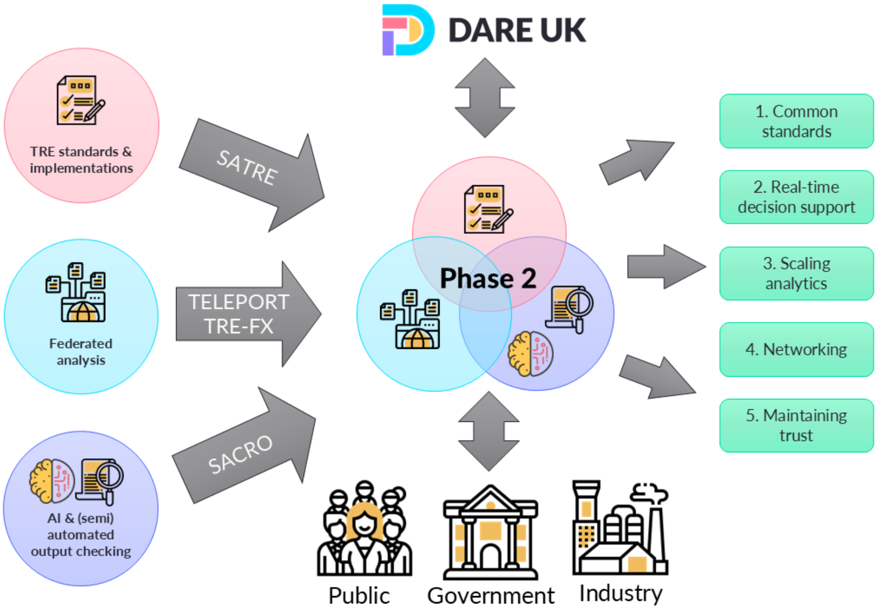

Figure 1. DARE UK Phase 2 builds firmly on the foundations of Phase 1.

The £18.2 million Phase 2 of the DARE UK programme commenced in August 2024 and is perfectly placed to rise to the challenge of the NDL. The consortium is working to integrate outputs from recent years into a comprehensive reference implementation of a federated network of TREs (Figure 1).

In addition to its integrating transformational programme addressing federation, interoperability, semi-automated disclosure control and machine learning, DARE UK is in the process of collaborating with key national TREs to build a prototype network of TREs to test a range of technical solutions and new capabilities, the outputs of which could feed into the NDL. Beyond these core developments, the programme will be seeking ideas for new innovations and scientific exemplar lighthouse projects in the next few years (cf. section 6).

[4] https://eosc-entrust.eu/

Although the federated architecture for TREs has been designed for research access, the same architecture could readily support operational intergovernmental access to data. The architecture is data agnostic and API-driven, making it widely applicable. Its key features are:

- **for service operators**, a common trust domain, including certifications and required levels of compliance for all participating services;
- **for data providers**, an infrastructure that is data-agnostic, placing no restrictions on the kinds of data it could support;
- **for information governance**, standardised per-project security architecture based on overlay networks between approved collaborators, creating distributed "safe settings" for each and every project;
- **for researchers**, the ability to log on to single "front door" TRE and to be able to see and analyse their approved data from multiple TREs as though they were all in one environment;
- **for industry**, common, standardised collaborative communication between participating services, enabling the exchange of data, analysis workloads and other information across secured networks, with minimal restrictions on the implementing software;
- **for the public**, greater assurance that sensitive data in research are handled according to national best practice, and greater transparency of the whole process;
- **for the future**, an infrastructure that is expandable and extensible, placing no restrictions on the number of participants it can accommodate.

For all, its key benefits will be:

- faster, fit-for-purpose federated analysis of cross-domain data by researchers;
- automation for faster, cheaper more efficient research;
- new research enabled by safer use of AI models in TREs;
- reduced friction for researchers to move between TREs;
- greater visibility for research using public data and its benefits;
- reduction in the costs of TREs from standardisation, yielding more funds for research.

The National Data Library has the opportunity to leverage a wealth of open work, and work in partnership with a large, established and diverse cross-sectoral community convened by the DARE UK consortium.

2. The need

In an increasingly complex and uncertain world, the need for good quality data has never been more pressing. The 2020 National Data Strategy [6] called upon the UK to embrace the exponential growth of data and the resulting 'tangible opportunities to improve our society and grow our economy.' The value of data as part of the UK's evidence base, and the urgency of the need to improve how different data sources are accessed and linked is clearly set out in *Transforming the UK's Evidence Base*, the 2024 report from the Public Administration and Constitutional Affairs Committee [7].

Research with public sector data from sources across public bodies, government departments and institutions already happens in the UK, in pockets of good practice connected by ad hoc technical processes. This data is most often directly or indirectly connected to the lives of individuals, communities and/or populations, thus demanding due consideration to maintaining the social contract with the public about how this data is used alongside secure, privacy-preserving management thereof. In other words, this data exists on a spectrum of sensitivity and must be treated accordingly.

There are inherent vulnerabilities and challenges when sensitive data is stored, analysed or shared at scale. Multiple bodies including The Royal Society [8], has called for new technological and governance approaches that will underpin trustworthy and responsible data use by enhancing data protection and collaborative data analysis. Reducing the barriers to entry for researchers not yet familiar with the use of sensitive data is essential to unlocking the UK's potential to deliver groundbreaking interdisciplinary research and to attract additional inward investment drawn by frictionless approaches to accessing large-scale datasets across the UK. The demand for data is growing fast [9].

If we are to realise UK Government ambitions on housing, living standards, safer streets, energy, children and the NHS, we must address the strategic need for a coordinated but flexible approach to increasing trustworthy access to public sector data. The NDL will address this, promoting responsible innovation whilst maintaining public trust in revolutionary technologies.

A preferential shift from a "research-by-download" model to a "research-by-access" model through the adoption of TREs was a key recommendation of the 2022 Goldacre Review [10]. The increasing complexity of the landscape and the need to apply a much more joined-up approach to safe research with public data – health data in this case – was a key recommendation of the 2024 Sudlow Review [11]. The review speaks to public data as "critical national infrastructure" and the need for a "UK-wide system for standards and accreditation of SDEs", these recommendations, amongst many others, point towards the need for an NDL ***for research***.

The increasing adoption of TREs for research with sensitive public data is a shift away from a "lending library" model and towards a "reference library" model, where researchers must go to the data to work with them *in situ* – in a "reading room", as it were – and cannot take them away. As highlighted in the Sudlow Review, we must now enable research with data linked from multiple "safe data libraries". These safe data libraries already exist in the UK and have done for many years, supporting research on sensitive public data across a range of research areas, albeit siloed around a specific dataset or set of research themes. This fragmented landscape suffers from attendant frictions and bottlenecks in data sharing, though: a significant drag on researcher productivity.

While the numbers and locations of data sources and services within this landscape will ebb and flow, for reasons of data governance or volume, or the divorcing of data from knowledgeable custodians, there is no likely future scenario which brings all data together in one location.

Consequently we must develop a secure and trustworthy method of "inter-library loans" between reference libraries, whereby the safe data libraries can share data securely behind the scenes, providing researchers with seamless access – within a suitable TRE reading room – to the datasets they have approvals to use. These interoperating services need technical and organisational measures in place to provide appropriate levels of trust between all of them.

To address this need, the NDL – certainly the section of the library that deals in ***sensitive*** public data – must be a designed as a federation.

3. Motivating use cases

A federated architecture which can connect data providers ("safe libraries"), TREs providing analytics environments ("reading rooms") and other services (e.g. discovery, software, indexing) must be standardised but must also be as ***minimally disruptive as possible to the good practice already in use***. It must support two underpinning use-cases: moving datasets into a single location for analysis ("**data pooling**"), and moving analysis workloads to distributed datasets ("**remote execution**").

The data pooling use-case occurs more often in current TRE use, especially when data have different attributes (e.g., health and education data for the same population) (Figure 2). Here datasets may need to be linked together using a common "master index", created by a trusted third-party index service in a way that ensures that the resulting linked dataset is only ever created within the hosting TRE.

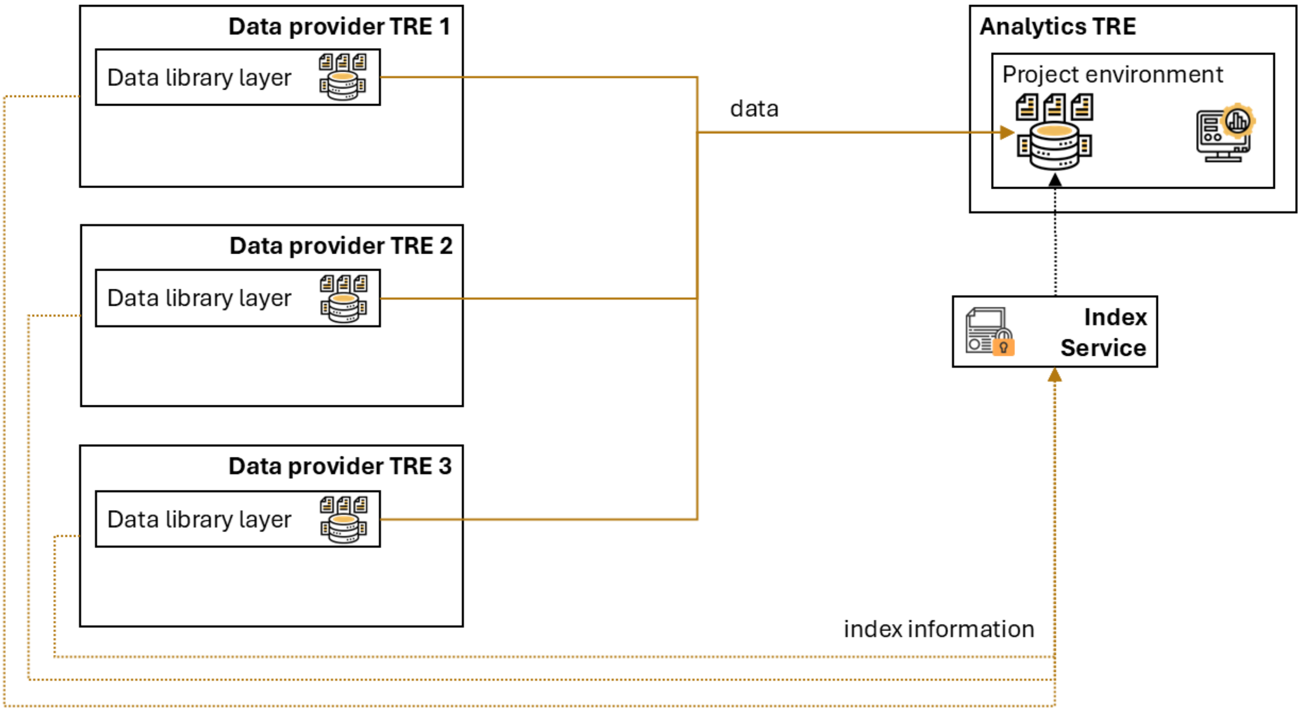

Figure 2. The data pooling use-case. Research-ready data for an approved project are combined in one single location and linked using an index provided by a trusted third-party index service.

The remote execution pattern (Figure 3) works very well when data are uniform in attributes and only split across populations (e.g., census data divided by region). It can be made to work when data have different attributes although it is technically more challenging to include the additional index service needed to make the join between the remotely calculated query results.

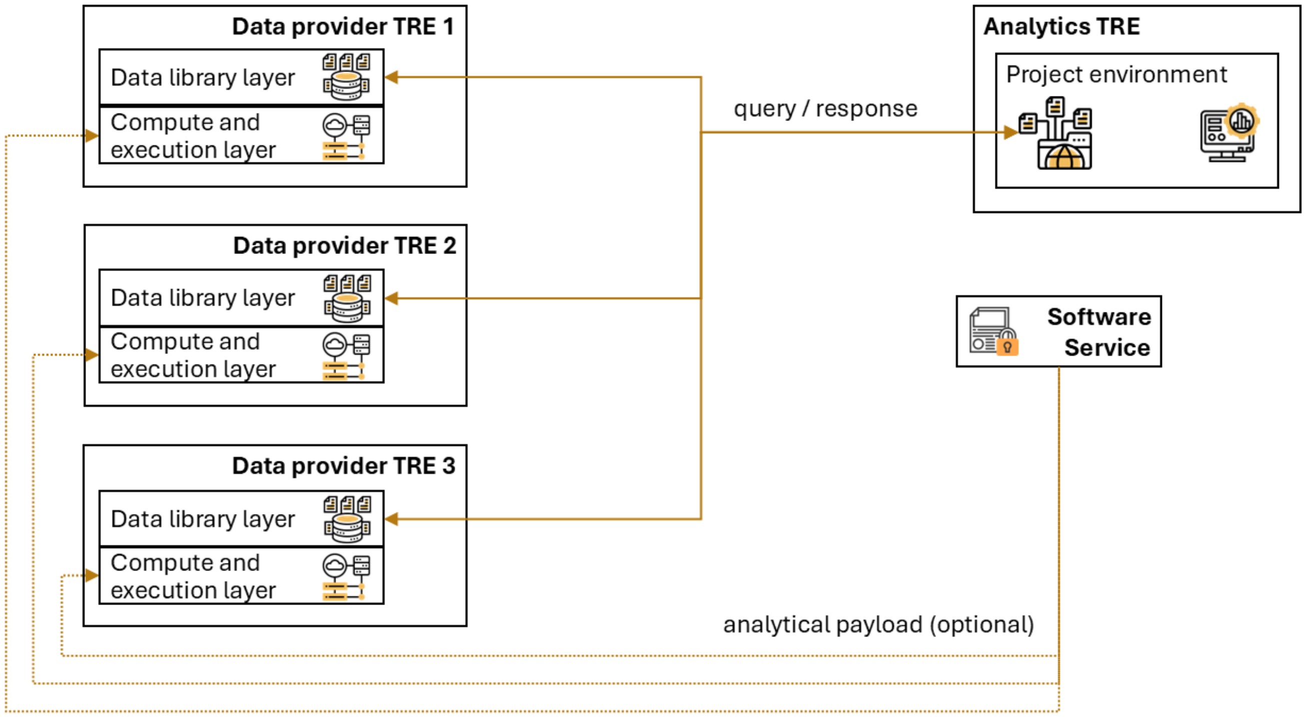

Figure 3. The remote execution use case. Research-ready data for an approved project remain in situ. Instead, each hosting Data provider TRE is able to run analytical workloads against their data asset, either as a result of a direct query from the Analytics TRE hosting the project, or by first retrieving (and approving) a "payload" from a known software service.

Remote execution may also involve a third-party software service from where the actual analytical payload to be run must be retrieved by each TRE wanting to execute it.

Federated **machine learning algorithms** can involve elements of both: partially learned models moving between TREs can be considered as data to be pooled, or as query responses to be processed. Which choice is appropriate for a given project will depend on the project's governance arrangements.

3.1. The research appetite

In early 2024, DARE UK and the PSC[5] convened a workshop of UK researchers and posed the question: "if there were no obstacles to data linkage, what research questions would you ask?" In a few hours the workshop developed 52 research use-cases involving the linkage of population scale data across 10 broad data domains, from health and social care to climate and environment [13].

The top three use-cases, as chosen by the workshop, illustrate the appetite and ambition: transforming the food economy for long-term health and prosperity; addressing domestic abuse; reducing NHS bottlenecks. In addition to the societal benefits that could arise from enabling research like this, an economic analysis of the top 10 use-cases suggests **potential benefits to the UK of £319.11 billion (+/- £79.14 billion) to 2050** [13].

[5] The Public Service Consultants, https://thepsc.co.uk/

4. Key elements of a federated National Data Library

To address these needs we propose **a federated architecture for the National Data Library**. It defines a backbone for secure information exchange between all participants, with strong guarantees of confidentiality, integrity and availability, connecting data providers, TREs and other service providers together in a high-assurance network with common trust and strong governance oversight.

To realise a federated NDL, we need two principal elements, one we might call the "foundation element" and one the "access element".

The **foundation element** is a standard, trustworthy way of connecting different NDL resources – safe data libraries, TRE "reading rooms" and other services – together to support secure movement of both data and analysis workloads. This provides the trustworthy basis of the NDL (cf. [1]).

The **access element** is the ability to configure specific sets of resources into projects, for federated research access, data movement or other purposes. This provides the dynamic access mechanisms for the NDL.

The architecture we describe below is **loosely coupled**, reflecting the distributed nature of the NDL, and focuses on **standard connections** between resources rather than specific software stacks. This lends itself to a future NDL of **independent interoperable nodes**, each built from a variety of implementations, whether commercial, open source or some combination.

4.1. Library foundations

Figure 4 illustrates the key foundation elements of the federated NDL.

In these diagrams we use "TRE" to represent a class of "safe settings" which may host sensitive research-ready data in data libraries (what we term "Data Provider TREs"), may provide analytic environments to researchers through which they can access sensitive data, either directly or indirectly ("Analytics TREs") – or may do both. We define these different specialisations of TRE below.

Data Provider TREs focus on curating and managing sensitive data in a "**safe data library**" and providing research-ready data for safe analysis. Any and all use of this data library will be overseen by a strong regime of **agreements, contracts and information governance**. Data Provider TREs may have sophisticated **compute and execution** environments alongside their data library, or they may not. In the latter case they may instead rely on other TREs to provide researchers with secure access to their data library.

Analytics TREs focus on providing secure access to research-ready data. They provide the "reading rooms" to researchers, rich **project environments** supporting a range of tools for modelling, analysis, data wrangling, provenance capture and semi-automated disclosure control. As with Data Provider TREs, any and all Analytics TREs will be overseen by a strong regime of **agreements, contracts and information governance.**

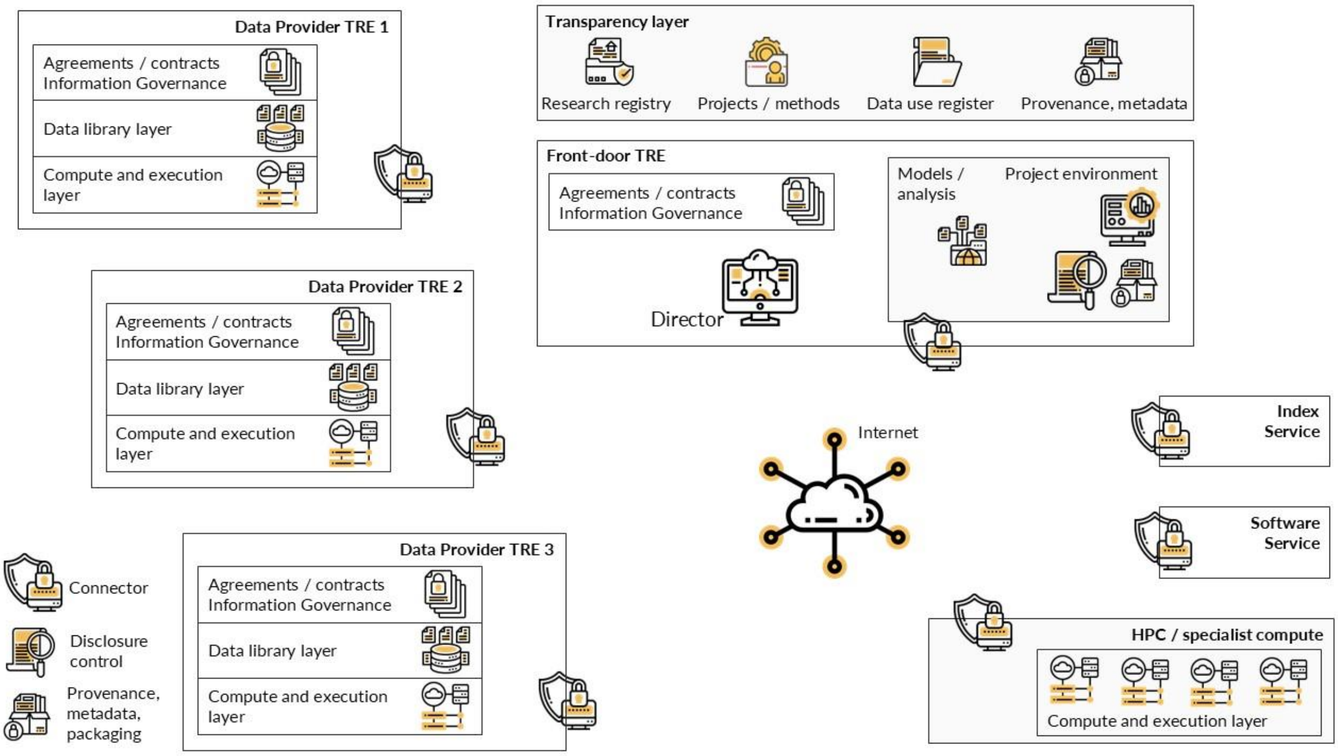

Figure 4. Architecture of a federated NDL, where nodes of different types provide data, computation or other capabilities into a federated network through standardised "Connector" components.

Front-door TREs or "host" TREs are specialised Analytics TREs with additional capabilities. They are critical components in enabling federated library access. We do not expect every TRE to be either willing or operationally and technically capable of undertaking this role. A key additional capability offered by a Front-door TRE is the operation of a **Director** component (cf. section 4.2).

HPC/specialist compute services provide high performance compute capabilities, whether CPU, GPU or other specialised hardware. A specialist compute service may be connected to a TRE or may be a standalone instance that has been selected to process the data for a particular access project.

Transparency Layer. This provides metadata information about elements of the NDL to users outside the federation boundary. Services in this layer record evidence of safe data use, captured as FAIR metadata and provenance, to demonstrate trustworthiness to the public. The Transparency Layer includes registries of approved researchers, approved research projects, methods and data use. It supports transparency of operations (often called "observability") across the network based on common, detailed metadata about its composition and operations.

Index Services create linkage spines for different access projects. How a given service does this will depend first and foremost on the principal index key in question. For personal data, for example, an Index Service might create depersonalised linkage spines by converting between "bare" personal identifiers and project-specific linkage keys.

Software Services provide access to approved sources of software from outside the federation. In this role, a Software Service may:

- act as a direct network proxy for Internet-based third-party software services (e.g., CRAN);
- act as an independently curated, high-assurance mirror service for popular software packages (e.g., Anaconda Python Enterprise);
- act as a proxy for defined and approved user accounts on a public open-source software repository (e.g., GitHub);
- act as a proxy for researcher workflows or analytical scripts stored in external repositories (e.g., WorkflowHub) to be used as payloads for federated analysis queries.

All of the nodes described above share a common element called a **Connector**. A Connector (known as the "Security Server" in [1]) is a set of software components that manage all aspects of the interactions between nodes.

Connectors are the gateways by which each TRE, specialist compute or other node becomes part of a federated network, and installation and configuration of a Connector is the minimum requirement for a node to participate in the NDL. Connectors receive configurations from a Director for each project they participate in, and for all data movement, analytical jobs and connections to the broader network.

4.2. Library access

The principal goal of the NDL is not to provide secure storage for sensitive public data but to provide safe access to these data for approved research by accredited researchers. Secure library access mechanisms are thus crucial to its success.

Figure 5 below extends Figure 4 to illustrate how the foundation elements described in section 4.1 interact to enable safe research projects using NDL resources.

The **project** is a key concept in the use of the federated NDL. A project defines an authorisation context for an approved research activity – a "safe project" – including the researchers involved, information about the data they are authorised to use, the Front-door TRE that hosts it, its duration and so on. Projects provide key information for overall federation governance. Metadata about projects is recorded in the Transparency Layer's **Research registry**.

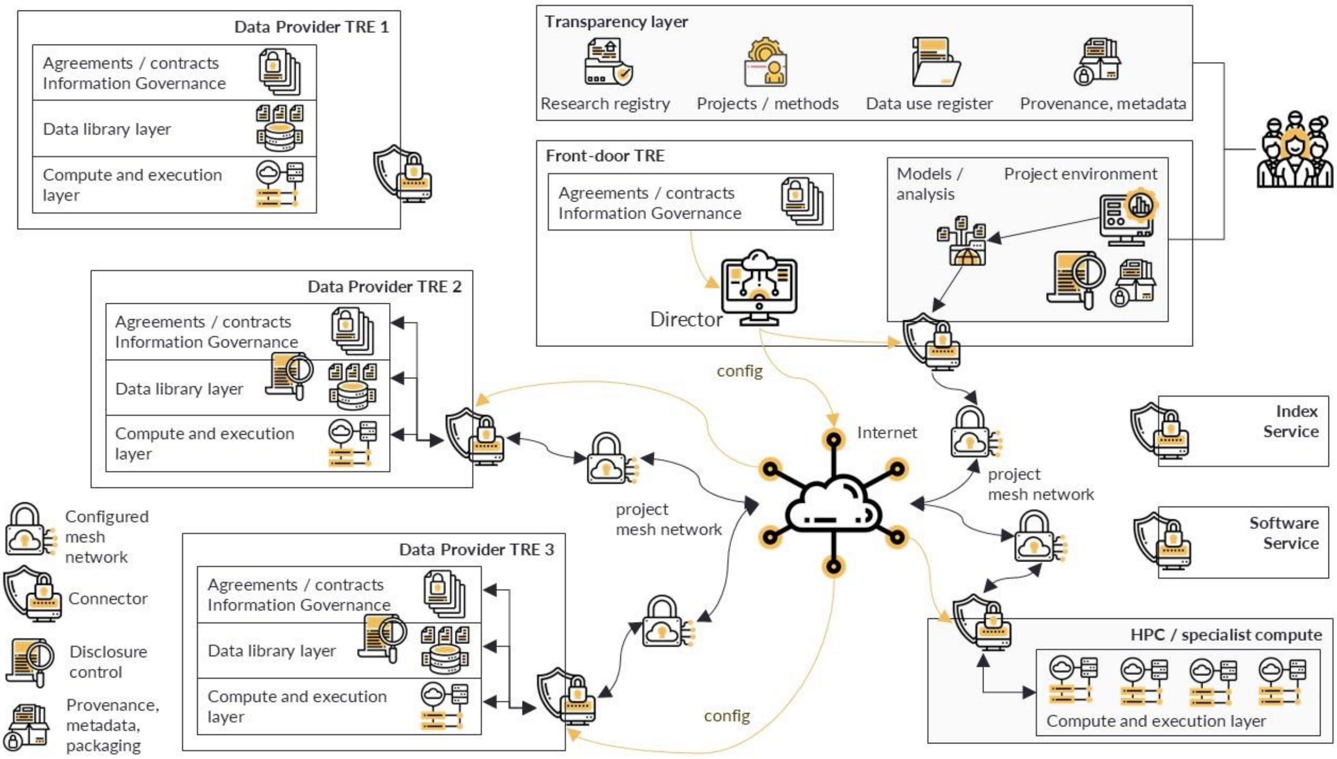

Figure 5. Orchestration of nodes in a federated NDL to create a project-specific safe setting spanning multiple distributed data and compute resources. In the illustration, a lead TRE, the "Front-door", uses its Director component to define and configure an overlay mesh network, unique to the project, connecting itself to two data providers and a specialist compute service.

Federated projects are initiated in **Front-door TREs**. For any given project, the Front-door TRE has the trust of the participating TREs and accepts the overall risk for the project. Within the Front-door TRE, the **Director** component is responsible for configuring the encompassing "safe setting" for the project – configuring the environment that the researcher will be given to perform their analyses, orchestrating the network of Data Providers and other nodes to connect the required resources, and configuring the mesh network between the participating nodes' **Connectors**.

The project's overlay **mesh network** dynamically and securely connects all the required resources. The mesh network creates a transient, project-specific, secure network boundary such that all the project's components are within one overarching safe setting.

In this fashion, every approved data access project is able to redefine its safe setting to encompass distributed resources with no loss of security or increase in disclosure risk. The effect is to the move the disclosure control boundary for project outputs from around individual TREs to around the entire federated safe setting.

Management and control of disclosure risk in federated projects requires enhanced approaches beyond those currently used in single-TRE projects. Accordingly, we introduce the concept of a **risk aggregator** running in the Front-door TRE. This uses semantic reasoning to create a global risk appetite for the project, and an approval process constructed from the individual TREs' profiles and risk appetites.

In the ‘data pooling’ use-case the risk aggregator automatically orchestrates reasoning over both combined and per-TRE level data, to create global and local risk assessments.

For the ‘remote execution’ use-case, outgoing queries can be passed to federated project nodes to be run by the risk assessment engine on their subsets of the data. The query and local risk assessment outputs are returned to the Front-door TRE where the risk aggregator combines the information to produce a global query result and global and local risk assessments.

In both cases all necessary information can be assembled within a dashboard with a view for each participating node, from where the necessary approvals can be collated to inform an egress decision.

These two models of disclosure risk assessment can be combined, via a stateful workflow running in the Front-door TRE, to support **federated machine learning** (ML). In ML, training is an iterative process requiring the distribution of ‘partially trained’ ML models to participating TREs, as remote-execution queries that return the updates for ‘their’ data, followed by the collation of results to create the next version of the model. Once the model is fully trained, egress from the federated project environment can be requested, triggering a distributed risk assessment process as described above. Again, the risk aggregator combines these into a visualisation dashboard for information governance to make a final decision.

5. Additional elements of a federated National Data Library

To complete the picture, there are a number of elements that complement this technical architecture.

5.1. Harmonised data landscape

This architecture is entirely data agnostic. Work still needs to happen in parallel to prepare public sector data for sharing across the federated NDL. However, the advantages to our approach are that the foundational infrastructure we describe can be assembled quickly and independently of slower efforts to harmonise data standards. Then, the existence of a secure, federated infrastructure can make the practice of data harmonisation between sites safer and more streamlined: data custodians and other providers can have the confidence to exchange data and test ideas for indexing and linkage over an infrastructure designed for just such a purpose.

5.2. Streamlined data governance

Creating a trustworthy infrastructure is a necessary part of a research-focused NDL but is insufficient in itself. Multiple data custodians operate multiple governance regimes with multiple risk appetites. Initiating a federated project between multiple data providers requires a certain degree of harmonisation of data governance policies. Currently, such projects can take months or even years to begin.

In the health and health-relevant data domains this is an approach championed by the UK Health Data Research Alliance. With over 100 members from across the UK, the Alliance is working towards standards in transparency, metadata, data access and data use.[6] Information governance standards will be a big driver of the implementation details of the federated NDL.

5.3. Public information and engagement

A strong, transparent and honest programme of information for, and engagement with, different sectors of the public will be vital to maintain the trust and social licence by which a federated NDL for research will operate. Drawing on the success of groups like PEDRI[7] will be essential.

5.4. Operational model

The functionality required of a federated NDL already implies a certain logical organisational structure. However, there is flexibility in how that logical structure could be realised in practice, depending on how the community of potential participants might agree on its setup and operation.

To meet public need for a more standardised, more trustworthy environment, the federated NDL needs to be real, in the sense of some kind of **membership organisation** with rules and standards. This organisation will oversee five key elements:

- Rules and policies – the underpinning agreements about joining the NDL federation, and the machine-actionable implementations of them.
- Trust, identity and certification – agents and methods for vouching for digital identities.
- Registry (participating nodes) –records of which services are members of the federation.
- Catalogue (data) – queryable records of what data assets could be accessible within the Federation, and how to go about applying for access.
- Transparency – records of datasets exchanged by participants, in which Project contexts, with what authorisation.

5.5. Economic model

A reasonable economic model for a federated NDL can be found in research over the past decade and a half on the costs of digital preservation and archiving. In the UK, Beagrie et al [14] remains one of the most comprehensive studies, including archives of both sensitive and non-sensitive data. A good rule of thumb for a data library is that 55% of the lifetime cost arises in acquiring and rendering data “research ready”, 30% arises from data preparation for access and project use, with the remainder split between long-term archive costs (c. 5%), administration (c. 6%) and software and networking (< 5%).

The software and networking required by the architecture proposed here is unlikely to impose proportionate costs significantly different from the c. 5% figure. The vast bulk of costs for the NDL arise in the preparation and curation of the data themselves. The value of doing this well was illustrated in section 3.1.

[6] UK Health Data Research Alliance, https://ukhealthdata.org/alliance-outputs/

[7] Public Engagement in Data Research Initiative (PEDRI), https://www.pedri.org.uk/

In a federated system each of the TREs will still require appropriate funding to cover their involvement within any research project. One of the challenges will be around how to work with a mixed model where some TREs work on cost recovery, but others receive core funding.

6. Implementation and delivery of a federated National Data Library

The architecture described in the preceding sections is not purely conceptual – it exists. Over the past 24 months the DARE UK programme has developed specifications, designs and proofs of concept of many of the required elements. Phase 2 of the programme, begun August 2024, has defined plans to mature these approaches into a federated network of TREs and data providers – exactly what is needed to underpin the NDL.

Below are a series of starting points and evolutions **already planned or in progress** towards delivery.

6.1. A willing coalition

Central to timely delivery is a core group of committed organisations supported by a broad-based collaborating community. Over the last 24 months DARE UK has convened just such a coalition, with three key dimensions.

Public trust and support. DARE UK has always required a public information and engagement component in everything it funds. It is also delighted to support and fund the work of PEDRI.

Legal and governance steer. DARE UK's parent organisations HDR UK and ADR UK are conveners of the Pan-UK Data Governance Steering Group,[8] the principal forum for harmonisation of data governance standards for a federated future.

Delivery and adoption partners. DARE UK's core partners include some of the leading TREs in the UK, and our wider community encompasses every major data provider and service operator from across the devolved governments of the UK, and major commercial organisations such as Google, AWS and others from the Confidential Computing Coalition.[9]

6.2. Reference implementations

By decoupling data providers and other services using this federated approach, infrastructure based on this architecture can be delivered incrementally. Further, once the foundations are in place, test and development versions of application services – new analytics tools for federated machine learning, for instance – can be deployed in safety and tested in environments that match their ultimate deployment.

6.2.1. Specifications and designs

Starting point:

- SATRE v1.0 [4], DARE UK Federated Architecture Blueprint v2.2 [1].

[8] https://ukhealthdata.org/projects/data-access-and-governance/

[9] https://confidentialcomputing.io/

- The Task Execution Service (TES) and Workflow Execution Service (WES) specifications from GA4GH.[10]

Evolution:

- Extend SATRE to v2.0 to include federation.
- Extend Federated Architecture Blueprint to v3.0, building on Phase 2 developments.

6.2.2. Common metadata standards

Starting point:

- Fife Safes RO-Crate community standard [15]; W3C Community Data Privacy Vocabulary;[11] W3C Recommendation Open Digital Rights Language;[12] sensitive community data standards from the UK HDR Alliance[6] and the Global Alliance for Genomics and Health (GA4GH).[13]

Evolution:

- Extend Five Safes RO-Crate to match SATRE 2.0.
- Extend SACRO [16] metadata profiles for federated risk appetites.

6.2.3. Library foundations

Starting point:

- Proofs-of-concept from TELEPORT [17] and TRE-FX [18] and earlier DARE UK Sprint Exemplars.[14]
- Existing Transparency Layer services such as the Health Data Research Gateway[15] and the ADR UK Data Catalogue.[16]
- Emerging Index Services such as the Integrated Data Service Reference Data Management Framework.[17]
- Existing open-source digital public infrastructure software such as NIIS's X-Road.[18]

Evolution:

- Develop a Kubernetes-based TRE reference implementation.
- Develop a Director network and project environment orchestration tool.
- Identify Connector software, selected from existing solutions already proven in operation (technology readiness level 8 or 9 in standard industry terms).

[10] https://www.ga4gh.org/

[11] https://w3id.org/dpv/

[12] https://www.w3.org/TR/odrl-model/

[13] https://www.ga4gh.org/our-products/

[14] https://dareuk.org.uk/how-we-work/previous-activities/dare-uk-phase-1-sprint-exemplar-projects/

[15] https://healthdatagateway.org/en

[16] https://www.adruk.org/data-access/data-catalogue/

[17] https://integrateddataservice.gov.uk/news/introducing-the-reference-data-management-framework-rdmf

[18] https://x-road.global/architecture

6.2.4. Library access

Starting point:

- Proofs-of-concept from TELEPORT & TRE-FX projects.
- The GA4GH Workflow Execution Service backend software [19].

Evolution:

- Integration of Phase 1 components into end-to-end federated network.

6.2.5. Disclosure control and risk aggregation for federated analytics and ML

Starting point:

- SACRO TRE-local assessment tools.

Evolution:

- Extend SACRO to support federated analysis and federated ML.

6.3. Delivery, adoption and rollout

Our loosely coupled architecture lends itself an incremental delivery approach: introducing a common low-level foundation while aiming for minimal disruption to existing services and supporting maximum innovation at application level. Services wanting to join the NDL federation can do so at their own pace, giving them time to meet whatever rules and policies the membership organisation has agreed.

Securing the foundation layer allows for greater innovation at the interface and application level without increasing risk. The core library access services that run on top of the foundation can be drawn from a wider ecosystem. Experimentation between NDL federation participants is possible at this level without undermining the security of data exchange.

The introduction of a member-owned cooperative providing secure standards and messaging services. with oversight of on-boarding provides an incrementally scalable model, placing no restrictions on the growth of the future NDL.

7. Conclusion

The key elements described here are necessary and sufficient to create the foundational technical infrastructure for a federated National Data Library. From the perspective of a public sector data provider the architecture offers three key capabilities:

- it creates a network of trustworthy services which have agreed to operate with a shared security posture, answering the question “is the IT safe enough for me to share data with it?”
- it creates a community of operators around a single set of technologies at the infrastructure level, reducing the complexities of data sharing and increasing the pool of expertise available, answering the question “who can help me do this more easily?”
- it provides an authoritative point of truth for services, data assets, users and projects, answering the question “how can I tell whether my data are being used for the purposes I’ve approved, and by whom?”

To realise the value of sensitive public data inside Britain's existing "safe data libraries" and secure "reading rooms", "do nothing" is not an option. To build everything from scratch would be expensive and risky, and result in an immature product in an environment requiring mature security. The work of DARE UK – led by our partners, collaborators, contributors and communities - has delivered and is delivering many of the components required for a federated NDL for research.

8. References

[1] DARE UK (2024). *DARE UK Federated Architecture Blueprint* (2.2). Zenodo. https://doi.org/10.5281/zenodo.14192786

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

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

[4] C. Cole et al (2023); *SATRE Standard Architecture for TREs*; Zenodo; https://doi.org/10.5281/zenodo.10055345

[5] Standard Architecture for Trusted Research Environments (SATRE); specification v1.0.0; https://satre-specification.readthedocs.io/en/stable/

[6] UK Government (2020); *National Data Strategy*; https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy#the-data-opportunity

[7] House of Commons Public Administration and Constitutional Affairs Committee (2024); *Transforming the UK's Evidence Base*; https://committees.parliament.uk/publications/44964/documents/223187/default/

[8] The Royal Society (2023); *From privacy to partnership*; https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf

[9] The Economist (2023); *AI is setting off a great scramble for data*. https://www.economist.com/business/2023/08/13/ai-is-setting-off-a-great-scramble-for-data.

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

[11] C. Sudlow (2024); *Uniting the UK's Health Data: A Huge Opportunity for Society*; Zenodo, Nov. 08, 2024. https://doi.org/10.5281/zenodo.13353747

[12] UK Health Data Research Alliance, & NHSX. (2021). *Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems* (1.0). Zenodo. https://doi.org/10.5281/zenodo.5767586

[13] DARE UK (2024). *Scientific use cases for cross-domain sensitive data research in the UK*. Zenodo. https://doi.org/10.5281/zenodo.14025303

[14] N. Beagrie, B. Lavoie and M. Woollard (2010); *Keeping Research Data Safe 2: Final Report*; Joint Information Systems Committee (JISC); https://repository.essex.ac.uk/2147/

[15] S. Soiland-Reyes, S. Wheater, T. Giles, P. Quinlan, and C. Goble, (2024); *Five Safes RO-Crate: FAIR Digital Objects for Trusted Research Environments*, in *International FAIR Digital Objects Implementation Summit 2024*, TIB Open Publishing, March 2024.

[16] J. Smith et al. (2024), *SACRO-ML*. Zenodo. https://doi.org/10.5281/zenodo.7080279

[17] C. Orton et al. (2023), *TELEPORT: Connecting researchers to big data at light speed*, Zenodo, Oct. 2023. https://doi.org/10.5281/zenodo.10055358

[18] D. T. Giles et al. (2023), *TRE-FX: Delivering a federated network of trusted research environments to enable safe data analytics*, Zenodo, Jan. 2023. https://doi.org/10.5281/zenodo.10529669

[19] José María Fernández, Laura Rodríguez-Navas, Adrián Muñoz-Cívico, Paula Iborra, Daniel Lea (2024): *WfExS-backend*. Zenodo, https://doi.org/10.5281/zenodo.6567591