



**Health Data Research (HDR UK)**  
**UK National Data Library: Technical Architecture**  
**White Paper**

**Licence**

This work © 2024 by HDR UK and is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)  

## Our Funders



## Introduction

[Health Data Research UK \(HDR UK\)](#) is the national institute for health data science that includes England, Scotland, Wales and Northern Ireland with a mission to accelerate trustworthy data use to enable discoveries that improve people's lives.

Supported by nine of the largest government and charity research funders in the UK our 20-year vision is for large-scale data to benefit every interaction with patients, every clinical trial and every biomedical discovery, and to transform public health.

We work in partnership with the NHS, industry, charities and universities to realise the potential of the UK's wealth of health data in life-changing research. Patients and the public are actively involved in shaping HDR UK's work and ensuring it delivers public benefit.

We are a multi-disciplinary, geographically distributed Institute involving researchers across more than 39 organisations and our activities and programmes across >150 organisations.

HDR UK has >70M of funding from 2023 to 2028. This significant core funding gives us the ability to work consistently over the long term towards accelerating the trustworthy use of data for public benefit and provides a strong base to attract further grants and funding (£131M, and 98M in-kind since 2017).

The UK's health data should be treated as critical national infrastructure that can underpin the health of the nation<sup>1</sup>. Access to de-identified, linkable health data is a requirement to answer many cross-sectoral research questions. A recent [workshop](#) run by [DARE UK](#) (a UKRI programme jointly led by HDR UK and [ADR UK](#)), looked into the opportunities for research using linked data from different domains and found that of the 52 use cases developed:

*“Use cases were cross-cutting, with 69% of use cases bringing together three or more different domains of data to answer their research question. Health data was used in 85% of use cases, but in broad ways and with improved health outcomes often as a secondary motivation: typically use cases focused on improving public health, wellbeing and productivity and only one use case targeted a specific clinical intervention.”*

We see the National Data Library (NDL) as a secure, distributed ecosystem of environments that should – in the context of health and health-relevant data – unite the UK's diverse health and social care datasets. Its core objectives as they pertain to health data should be to:

- **Accelerate health research:** Enable secure, equitable access to data for researchers, policymakers, and innovators.
- **Transform the NHS:** Support personalised medicine, operational efficiency, and health equity through data-driven insights.

---

<sup>1</sup> The [Sudlow Review](#)

- **Drive economic growth and attract global industry to the UK:** Strengthen the life sciences sector by fostering innovation in AI, digital health, and therapeutics.
- **Promote and enable industry:** through common, standardised collaborative communication between services, enabling the exchange of data, analysis workloads and other information across secured cloud networks, with accessible standards and with minimal restrictions on the implementing software.
- **Align with international frameworks:** such as the European Health Data Space (EHDS) to enable global collaboration and maintain the UK's leadership in health data innovation.
- **To understand the social/environmental determinants to health:** supporting the quest to pivot from "sickness to prevention".

This white paper focuses on principles and best practice the NDL should consider in assembling the necessary data architecture for future health data research. The associated [DARE UK white paper](#) provides an infrastructure-architectural view across all data domains.

To use the library analogy, **Health Data should be a highly visited and valuable floor of any NDL.**

## The Landscape

The [Sudlow Review](#) provides an overview of the health data landscape. In summary, the UK has rich, abundant sources of health data and we are unique worldwide thanks to the NHS. Our large, diverse population of 67 million people are all largely seen in the same health system with data that goes back decades. The HDR UK [COALESCE](#) study exemplified the power of this approach as the first to analyse health records of the entire UK population, revealing that under-vaccination against COVID-19 led to a significant number of preventable hospitalisations and deaths. The UK also has some of the world's leading research-collected consented cohorts of data, e.g., the UK longitudinal population studies (LPS) are a national asset based on the generous donation of detailed data by participants over decades. In addition to the flagship cohorts of [UK BioBank](#), [Our Future Health](#) and [Genomics England](#), there is an extensive collection of smaller, more focused population [cohorts](#) as well as disease-based clinical cohorts, audits and registries.

However, complexity and inefficiencies in the UK's data systems make the sharing of data both legally and technically challenging, impeding important research. Furthermore, health is devolved across the UK making it difficult to share health data between England, Scotland, Wales and Northern Ireland. It can take years to get access to health data before research can even begin. Even with the consent of participants to link their routinely collected health data to data within research cohorts, it is challenging to get data controllers to agree to share this data. **These are some of the factors impeding important research across diseases impacting the lives of millions of patients and their families in the UK.**

There is huge potential to do better. The health data landscape has evolved over recent years, moving from a model in England where copies of de-identifiable datasets were shared for researchers and companies to analyse on their own infrastructures (applying data sharing agreements and stringent data security requirements) to a model where the default is that data has to be accessed within Trusted Research

Environments ([TREs](#)), also referred to as Secure Data Environments (SDEs). Scotland and Wales have utilised this model for over a decade.

In the UK, TREs broadly adhere to the Office for National Statistics (ONS) '[Five Safes Framework](#)' Whilst TREs provide a step forward in secure data access, protecting patient confidentiality, their designs and builds have, unavoidably, created secure data silos. The NDL needs to provide the same level of controls and public assurance whilst at the same time supporting seamless data access across a distributed landscape of TREs.

## **HDR UK is a UK-wide ecosystem convener**

HDR UK is uniquely placed to map, track, surface and engage with stakeholders across the UK-wide health data ecosystem, including making visible the 'starting position' of multiple, small-scale and/or siloed data assets. Our mechanisms to achieve this include:

- The [UK Health Data Research Alliance](#) ("Alliance") and associated sub-groups and workstreams (e.g., the [Pan UK Data Governance Steering Group](#)).
- Strategic programmes and partnerships e.g., [COVID-19 Data & Connectivity National Core Study](#); [Digital Innovation Hub programme](#); [Dementia Trials Accelerator](#); COALESCE consortia; [OHDSI UK node](#) and pilot [Real World Evidence Network](#); Cancer Data Collaborative; [DARE UK](#).
- Our [Core Funders Committee](#), particularly the ability to engage with medical research charities, four-nation health funders and three UKRI research councils.
- Our programme of events and conferences.
- International profile and involvement, including recent visits from French Data Hub, Health-RI (Netherlands) and [memorandum of understanding with Singapore government](#), alongside our International Advisory Board.
- Facilitating the Sudlow Review and its implementation.
- Representation on other ecosystem groups and boards (Government Life Science Council, [NHS Data Enabled Research Advisory Group](#), [ELIXIR](#) etc.).

There is an urgent need to align and simplify across a traditionally competitive sector. HDR UK has identified current barriers and challenges and prioritised investment in data curation, service development, standardised governance and access requests. Our aim is to maintain a clear sense of focus and collaborative problem solving, as seen during the pandemic the opportunity is to assemble a UK-wide member-owned cooperative that subscribe to a portfolio of world-leading secure messaging and data curation services.

## **NDL design and implementation principles**

Some of the principles which the NDL should follow are:

- **Implement the [Sudlow Review](#) recommendations:** For the NDL's health data component.

- **Demonstrate trustworthiness to patients and the public:** The public need to understand and agree with the controls put in place by the NDL to ensure confidentiality of their data, whilst supporting research at scale for public benefit.
- **Build upon existing capabilities:** There are a range of national TREs and data repositories providing services and access to health data for research. These should be leveraged and “dock” into the NDL.
- **Streamline information governance:** Reduce approval times for data access through standardisation and addressing legal complexity.
- **Utilise open data and technical standards:** To streamline interoperability and support a multiplicity of technical solutions for governance and access processes.
- **Work with existing communities:** Leverage existing communities to share best practices, enhance interoperability, reduce complexities and streamline the research journey.
- **Balance the strengths of centralisation and federation:** There are efficiency benefits to certain degrees of centralisation which need to be balanced with the need for federation due to information governance and technical constraints.
- **User centred design:** The NDL must support users’ needs, including training machine learning (ML)/artificial intelligence (AI) models, utilise high performance computing (HPC) and graphics processing units (GPUs), install specialised software and run federated projects across different TREs.
- **Follow the FAIR principles:** Ensuring the NDL adheres with Findability, Accessibility, Interoperability, and Reuse ([FAIR](#)) is essential for transparent, collaborative, efficient and accelerated research.

We expand on each point below, with background, key NDL considerations, and relevant existing capabilities, infrastructures, and standards.

## Implement the Sudlow Review recommendations

### Background

The [Sudlow Review](#), *Uniting the UK’s Health Data: A Huge Opportunity for Society*, provides five recommendations for a pathway to establishing a secure and trusted health data system for the UK:

1. Major national public bodies with responsibility for or interest in health data should agree a coordinated joint strategy to make England’s health data a critical national infrastructure.
2. Leading government health and research bodies should establish a national health data service for England with accountable senior leadership.
3. The Department of Health and Social Care should oversee and commission a strategy for ongoing coordinated engagement with patients, public, health professionals, policymakers and politicians.
4. The health and social care departments in the four UK nations should set a UK-wide approach for data access processes, and proportionate data governance.

5. National organisations in the four UK nations should develop a UK-wide system for standards and accreditation of secure data environments (SDEs) holding data from the health and care system.

Key data priorities for the national service should be to:

6. Establish a national system for general practice data, enabling secure access to comprehensive, whole-population, structured, coded general practice data, linkable to other data sources and accessible for a wide range of beneficial uses.
7. Enhance and accelerate access to other major national and regional NHS data assets: hospital episodes, medicines data, lab data (including genomics), national audits and registries, screening data and unstructured clinical data (including imaging and free text).
8. Transform access to data from other sectors linked to health and care data at national scale.

## **NDL Considerations**

For health data to be successfully connected to the NDL, these recommendations should be implemented.

## **Demonstrate trustworthiness to patients and the public**

### **Background**

Public support should not be assumed. A lack of 'social licence' for data use severely impacted programmes designed to provide greater access to data for research, e.g. Care.Data, GP Data for Planning and Research (GP DPR). Key stakeholders – such as GPs – have been seen to shape public views on new data initiatives, in part due to their concerns regarding the legitimacy and risks associated with data sharing models.

### **NDL Considerations**

NDL must **involve** and engage the public and stakeholders from outset to earn trust and understanding. NDL must advocate for public transparency by implementing a UK-wide communications strategy to build public awareness and understanding of how data is accessed, processed and safeguarded. However, communication and engagement alone will not be sufficient. NDL should include public involvement in data use decision making to ensure proportionality and to drive continuous improvement.

### **Leverage**

Public Engagement in Data Research Initiative (PEDRI) is championing best practice in this area, co-developing to support meaningful public involvement and engagement and running practitioner events to tackle challenges and develop community wide solutions for data and good practice standards; and public events to involve public contributors, such as in the design of data access processes. This is yielding a robust framework for enhancing involvement and engagement practices.

PEDRI has engagement and support from across data domains (e.g. ADR UK, ONS, SDR UK, DARE UK). **The NDL should consider leveraging this existing capability, infrastructure and best practices in PPIE.**

Transparency in data use is key to building trust. The [Alliance](#) developed a [data-use register standard](#) with input from over 100 members across data domains. . The Health Data Research [Gateway](#) (the Gateway) now implements a [registry](#) enabling over 90 data custodians to share project information in a transparent, standardised way. **The data-use standard is not health data specific and could be adopted by the NDL.**

## Build upon existing capabilities

### Background

The UK has range of existing repositories critical to science and health research. These include (but are not limited to):

- [UK Biobank](#)
- [Clinical Practice Research Datalink \(CPRD\)](#)
- [ONS Secure Research Service \(SRS\)](#) and [Integrated Data Service \(IDS\)](#)
- [Our Future Health](#)
- [Genomics England](#)
- Centralised environments for consented (e.g. [UK Longitudinal Linkage Collaboration](#), [Dementias Platform](#))
- [NHS England Data Access Request Service](#)
- [NHS England Secure Data Environment](#) (including curation by [BHF Data Science Centre](#))
- [NHS England SDE Network](#)
- [NHS England OpenSAFELY Service](#)
- [NIHR BioResource](#)
- [SAIL Databank \(Secure Anonymised Information Linkage\)](#)
- [Scottish National Safe Haven](#) and [Network of Safe Havens](#)
- [Northern Ireland Honest Broker Service](#)
- Hubs including [DATAMIND](#), [PIONEER](#), [INSIGHT](#), [Alleviate](#), [DISCOVER-NOW](#) and [DATACAN](#).

There are also private companies that provide access to NHS and other public data.

### NDL Considerations

Each TRE and national data repository provide services and access to valuable health data. The NDL should leverage these infrastructures, enabling them to interoperate and link with other datasets.

Data cleaning and curation should be done once by trained domain experts, with transparent steps and data provenance available to researchers, reducing inefficiencies and errors. Leverage

There is no need to start from scratch. The NDL can work with TRE providers and key private-sector firms to build on existing networks, services and expertise.



## Streamline information governance

### Background

The current information governance landscape is complex: legal frameworks governing data use differ across health and non-health sectors; between operational, research and statistics purposes; and between statute and case law. For example:

- **Health data:** The governance and legal framework for NDL must address the distinct legal requirements for health and social care data, which vary across the UK and fall outside the scope of the Digital Economy Act ([DEA](#)).
- **Non-health data:** The DEA supports the research use of administrative data sharing, which can include health-relevant data but does not support the use of NHS data.
- **Operational vs. research use:** Operational uses often rely on direct care exemptions, while research use requires explicit ethical and legal approval, creating fragmentation.

Complex considerations exist where the Common Law Duty of Confidentiality applies, with differing legal routes to address this via informed consent or devolved legal and procedural frameworks:

- **In England and Wales,** [Section 251](#) of the Health and Social Care Act 2006 allows the Duty of Confidentiality to be lifted where consent is not practicable.
- **In Scotland,** data access is overseen by the [PBPP](#), which assesses requests to link and use health data for research or service improvement.
- **Northern Ireland** lacks a clear legislative gateway for lifting the Duty of Confidentiality, which can result in uncertain and potentially inconsistent processes.
- **DEA** applies UK wide, allowing the Duty of Confidentiality to be lifted for non-health data.

Resulting from these differing laws, associated codes of practice and their interpretation, data custodians have developed different data access policies and procedures. Data sharing and access agreements are often bespoke and on a per-project basis. This makes data access time consuming and complex.

### NDL Considerations

Addressing this legal complexity will be essential for the NDL to maximise the value of different types of data. We do not consider legislative change to be the primary solution here, rather the barrier often lies in confusion and inconsistent interpretation of existing legislation and regulations; and that some regulations (such as Mental Capacity Act) are not suited to all contexts. We also note the critical importance of context specific governance, such as the mechanisms supporting study-participant trust in longitudinal studies that span lifetimes.

The NDL will need to advocate for:

- **Alignment of legal gateways**, promoting legislative clarity and consistency across the four nations, ensuring equitable access while respecting devolved and domain specific structures and aligning between DEA and health-specific requirements (e.g., aligning TRE accreditation standards).
- **Strengthened governance mechanisms**, collaborating with existing bodies (e.g., HRA CAG, PBPP) to create standardised review processes that accommodate national differences, including harmonised data access processes, while respecting devolved regional governance.
- **Establishment** of a UK-wide governance board with representatives from patients, researchers, and policymakers.

## Leverage

The Alliance, representing data custodians and policymakers across the four nations, and the [Pan-UK Data Governance Steering Group](#) cross-data domain working group have advocated for alignment of TRE accreditation around the UK Statistics Authority mechanism and have published community-developed standards and templates aimed at streamlining information governance for research:

- A template [Data Access Request](#) form based on the Five Safes Framework, to standardise data access processes.
- A template [Data Access Agreement](#) for TREs to reduce complexity, standardize contracts, and clarify data protection obligations, speeding up the process.
- [Transparency Standards](#) establishing best practice for transparency of processes for both researchers and the public.
- Recommendations for a [Data Use Register](#) standard.

The [Gateway](#) provides a way for researchers to submit data access requests to data custodians using the data access request template questions, for those data custodians who have adopted the standard. The solution is currently being enhanced to support applications across multiple data custodians.

The NDL could utilise existing standards and contribute to the **Pan-UK Data Governance Steering Group** to support the adoption of community agreed standards to streamline information governance processes.

## Utilise data and technology standards

### Background

Data and technology standards underpin the success of any large-scale scientific infrastructure, and the NDL should be no different.

Operational and research standards often differ. [OMOP](#) is a widely used common data model (CDM) for health research data, while the [HL7 FHIR](#) data exchange standard is widely used by the NHS to support interoperability between clinical systems. Within healthcare coding standards such as [SNOMED CT](#) and [ICD-11](#) are key.

Metadata standards are also invaluable for data discovery. Examples include [DCAT](#), [BioSchema.org](#) and the [Health Data Research metadata schema](#).

## **NDL Considerations**

The NDL must invest in both data architecture and technical infrastructure, distinguishing between operational and research data. Where they diverge, it must invest in mapping them, a labour-intensive process requiring data and domain experts and extract-transform-load (ETL) tools. The data science rule-of-thumb – all projects are 80% data wrangling and 20% analysis – applies to developing the NDL.

## **Leverage**

We recommend building upon international standards already available e.g., from the Global Alliance for Genomics and Health ([GA4GH](#)). Originally designed to meet the needs of the genomic data community, these have much wider applicability. We endorse the GA4GH white paper submission “Establishing a Foundation for a UK National Data Library: A Federated Approach to Data”.

HDR UK are utilising different standards including the GA4GH standards (for details see the DARE UK associated white paper) and is looking to integrate the GA4GH [Beacons](#) standard to support data discovery and to federate metadata catalogues.

Standards for TREs are key. The Alliance developed ‘[Principles and Best Practice for Trusted Research Environments](#)’ and explored TRE federation. The 300+ member [TRE Community’s](#) Standard Architecture for Trusted Research Environments ([SATRE](#)) has been adopted by TREs including the NHS England SDE Network and European National TREs through the [ESOC-ENTRUST](#) programme.

It is challenging to answer questions when data are highly heterogeneous. Standardising datasets using CDMs such as the OMOP (managed by [OHDSI](#)) for health data can help support interoperability. HDR UK partners with the [OHDSI UK node](#) to promote adoption of OMOP across the UK and is piloting a [OMOP-based Real World Evidence network](#). This effort has been significantly enhanced by investments made through the European Health Data Evidence Network ([EHDEN](#)).

## **Work with existing communities**

Existing communities like Alliance and the Pan-UK Data Governance Steering Group, PEDRI, the DARE UK [TRE Community](#), [Research Data Scotland](#) and SDE Network are already working to standardize, collaborate, and reduce complexity. These should be utilised to the design and develop the NDL rather than creating new structures which could increase complexity.

## Balance the strengths of centralisation and federation

### Background

The volumes of data across the four nations UK from different data domains are vast. Health data is fragmented due to it being devolved and the large number of separate data controllers (Boards, Trusts, GPs and organisations hosting research collected data) each collecting data in different systems across the UK. The spectrum of data volume or size in health adds further challenges, e.g. imaging and genomic data compared with coded electronic health records.

A single, central environment for all data domains for research is not feasible technically, nor would such an approach gain public support. It could not include all the different data needed to answer all of the research questions posed, could not scale to support all the necessary data analysis needs, nor retain the necessary depth of expertise regarding the datasets it holds.

However, there are, as articulated in the Sudlow Review, efficiency benefits to certain degrees of centralisation. Some key health data is collected centrally in England, Scotland, Wales and NI. These datasets are primarily collected for administrative purposes however they are regularly used for research and service improvement purposes. Regional and research study data is often higher resolution, and essential to answer some research questions. In the quest for greater depth of data, it is key that the “higher-level” national scale data is not deprioritised, as these will be core to a wide range of researchers.

### NDL Considerations

The NDL should take a balanced approach to centralisation versus decentralisation with a blended approach, a “mesh network” or power stations on a grid, most likely to succeed. E.g., centralising and standardising key datasets in different data domains to be accessible through the NDL. Conversely, given the existing distributed nature of UK data and the benefits thereof a decentralised approach to the technical underpinnings of the NDL is the most feasible. These are just examples, there are several strategic axes to consider across a spectrum of centralised/decentralised approaches that must be considered in any design of the NDL.

### Leverage

From a technical perspective, the NDL is only feasible as a federation and must be underpinned by a technical architecture that supports this (see the associated [DARE UK white paper](#)).

## User centred design

### Background

The move from a data distribution towards a data access model through TREs is positive, since TREs offer enhanced security and reduce data privacy risks. However, they must be fit for diverse research needs in terms of functionality, scalability, performance, interoperability, federation and speed of access.

### NDL Considerations

As TREs mature, researchers will increasingly expect them to resemble their own preferred working environments, and offer the powerful tools available in other research settings e.g.:

- **Support for machine learning (ML)/artificial intelligence (AI) training** within TREs and safe export of trained algorithms out of TREs. There is much to be done before TREs routinely support ML/AI projects.
- **The availability of HPC and GPUs**, providing the key powerful compute capabilities to carry out research at scale.
- **Software and infrastructure to support analysis of unstructured datasets** such as imaging and genomic data.
- **The ability to install and configure highly specialised software** including software from code repositories. Groups develop software outside of TREs, utilising it on different datasets which then needs to safely be bought into the TRE environment. Many TREs currently provide a limited suite of software, impacting researcher efficiency.
- **Research services at scale.** Large scale infrastructure also requires efficient research services.
- **Support for projects requiring data from multiple TREs.** There are many research questions which can't be answered by data sourced from just a single TRE. The NDL cannot just be a data catalogue, where different datasets can be found, but must be a place where researchers work seamlessly with data linked from multiple sources.

Comprehensive training material for researchers in using these new, complex environments will also be vital. Increasing the standardisation of TREs will help significantly as will ensuring that solutions are cloud and vendor agnostic where possible.

### Leverage

DARE UK [GRAIMATTER white paper](#) recommends ways TREs can safely support AI/ML projects. The [SACRO project](#) followed this with foundational tools supporting automated approaches to the necessary risk assessments.

DARE UK [transformational programme](#) is working with TRE early adopters to test standards and reference implementations to enable users to log into one TRE and see approved data from multiple TREs as though it were in one place, and to develop further the ideas of automating risk assessments for ML models within TRE-private environments.

## Follow the FAIR principles

### Background

A core component of the [FAIR](#) (Findability, Accessibility, Interoperability, and Reuse) principles is provision of rich, standardised and publicly accessible metadata (the descriptions of the data assets). Empowering and requiring researchers and organisations, through both adoption of common data standards and development of technical solutions is key to efficiently catalogue, harmonise and share the wealth of underlying data resources at a national level.

The UK's data are diverse and complex, and the effort to harmonise data from across the UK to be comparable, searchable and accessible is considerable. It is also important that beyond each data asset being richly described that the full research output and data use is captured and displayed in connection to it. These issues with data and data analysis re-use and discovery are well publicised for health, e.g., [Goldacre Review](#).

Beyond data, the development of Reproducible Analytical Pipelines that adhere to FAIR principles is important to ensure consistency and reproducibility through automation, version control (through platforms like GitHub or GitLab), and the transparent and reproducible sharing and licencing of code repositories. This is in keeping with the Government Analyst Service ([GAS](#)) which provides resources to support such implementation.

### NDL Considerations

Ensuring that the NDL adheres to the principles of ([FAIR](#)) is essential for transparent, collaborative, efficient and accelerated research. Open APIs and open-source (where appropriate) will be key. The NDL should provide mechanisms for everything that is not potentially sensitive row level data within TREs to be exported, including metadata, data cleaning/transforming and analysis scripts and software developed.

### Leverage

HDR UK has several on-going efforts supporting the FAIR principles:

- Health Data Research [Gateway](#) provides a way for researchers to discover standardised metadata of >800 datasets and associated data analysis scripts, data uses and publications.
- [Open science and open code policies](#) and [development principles](#).

## Components of a Technical Architecture

DARE UK is a programme jointly led by HDR UK and ADR UK with the aim to develop a UK wide research infrastructure for analysis of linked sensitive data across different data domains. We endorse the **separate [DARE UK white paper providing more technical details of a potential federated architecture for the NDL](#)**.

Here we summarise **some** of the components of a federated approach which the NDL could leverage, with a focus on harmonisation of the data services layer (specifically health).

## Data Discovery and Access Services

Users of the NDL must have powerful, yet easy-to-use, ways to discover what data it holds. NDL data discovery services need to be much more than simple static catalogues. They must be able to respond to regular, rapid changes in the underlying data, and provide the means for users to explore the dimensions of data assets in detailed ways, all the while maintaining the security of potentially sensitive data.

The Gateway is an enterprise level solution, with scaling and future proofing baked into its core. Built upon the Google cloud, it utilises cutting-edge, large-scale frameworks, such as Next.js and Laravel. Many different health data groups use the Gateway as a way for the community to find their data. E.g., NHS SDE Research Network use the Gateway as their 'common front door'.

There are several ways that the NDL could use this open-source solution:

- Federate metadata catalogues
- Query information within the Gateway via the API
- Enhance the solution to support additional types of data datasets
- Co-develop the solution so that the Gateway becomes the front-door interface of the NDL

A component of data discovery is not just finding metadata about a dataset, but also understanding how many records match a specified criteria of a research question supports high level research feasibility questions or to potentially find recruits for clinical trials. The Gateway has embedded a Cohort Discovery tool.

## Indexing Services

Effective data linkage is critical for realising the full potential of the NDL. Linking datasets across individuals, places, and organisations enables comprehensive analyses that drive advancements in health research, public policy, and service delivery. Fragmented identifiers, inconsistent standards, and gaps in metadata significantly limit our ability to unlock insights from existing data assets. Different governance approaches, legal bases for data sharing and contractual arrangements for sub-licencing must also be addressed.

As a minimum, NDL indexing services should support linkage by:

- **Individuals:** Use secure identifiers, such as pseudonymised NHS numbers to enable linkage while preserving privacy and to undertake rigorous bias assessments to ensure linkage processes do not exacerbate inequalities.
- **Places:** Leverage the Unique Property Reference Number ([UPRN](#)), which government has mandated to standardise geospatial data and allows for efficient linkage of health and social outcomes to environmental, housing, and public service and generation of household indicators.

- **Organisations:** Maintain a register of organisational identifiers, such as NHS England’s [Data Search and Export Service](#) by Organisation Data Service, to streamline linkage across administrative datasets.

HDR UK ‘s [Social and Environmental Determinants of Health](#) driver programme is focused on linking health and environmental data utilising UPRN and the NDL may be able to leverage the lessons learnt.

A centralised indexing system or national data spine is essential for ensuring consistent, scalable data linkage. This spine would:

- Maintain a registry of key identifiers across datasets.
- Act as a secure mediator for linkage requests, reducing duplication and inconsistencies.
- Support federated queries to avoid moving sensitive data unnecessarily.
- Provide a sampling frame for instigating new randomised control trials and longitudinal studies.

One possible candidate is the [Reference Data Management Framework](#) (RDMF), that the ONS [Integrated Data Service](#) is building to improve the way in which data can be linked for analysis purposes.

## Researcher Registry/Researcher Identity Services

All data custodians agree that, for a particular project, the researchers and the organisations involved must be assessed to be “safe” (adhering to the Five-Safes Framework) and that the identity of the researcher can be authenticated. An HDR UK-led, DSIT/UKRI funded project is building the Safe Organisation and User Registry for Sensitive Data (SOURSD), with a minimal viable product planned for March 2025. It is a centralised system for all researchers and their host organisations to securely share information with TREs, to enable TREs to assess whether they are “safe”. This reduces manual steps and duplication, and supports standardisation.

A range of single sign-on (SSO) mechanisms can be associated with accounts – providing a way for users to log in to TREs for approved projects and for TREs working together on a federated project to all “trust” the identity of a user.

The Registry does not make decisions; those remains with data custodians. However, it can flag to TREs where a particular “rule” (from a customisable set) has been breached.

Funding allowing, we plan to integrate a range of automated feeds including [DEA accredited researchers](#) from ONS and ethical approvals from IRAS.

## Data Wrangling Services

The NDL will need experts who have deep understanding of data’s provenance, semantics etc. to clean, curate, transform and manage the data.



The BHF Data Science Centre is an example of this, acting as default delivery partner in the development of the NHS England SDE, through providing access to the English component for COVID-IMPACT collaboration. Delivery has included data engineering for research-ready data assets, including resources to support understanding data (e.g. Data notes and Dashboards) and Clinical/domain specific data science expertise to support efficient and effective use of the data. 87 active or completed studies have utilised the linked, nationally collated healthcare datasets.

More could be done to specify and productionise ‘core national health data products’, building on the work of the BHF Data Science Centre and leveraging wider expertise and partnerships including through the HDR UK Research Driver programmes: [Inflammation and Immunity](#), [Medicines in Acute and Chronic Care](#), [Social and Environmental Determinants of Health](#), [Molecules to Health Records](#), and [Big Data for Complex Diseases](#).

## Federation

There are three widely understood types of federated analysis: data pooling, meta-analysis, and federated analytics/learning. Each has advantages and disadvantages. The HDR UK [Federated Analytics Programme](#) and the DARE UK Transformational Programme are closely aligned, utilising global standards to provide a streamlined and secure way for federated solutions to be executed within and across TREs. They are NOT developing yet another federated solution as there are many commercial and open-source solutions already available. **Details are provided within the [DARE UK white paper submission](#).**

This approach combines the best attributes of each type of federated analysis, enabling a researcher to log in to a “Front door TRE” and see all of the de-identifiable data they have been approved to see across all of the TREs collaborating on a specific project. To facilitate this, all of the TREs must trust each other (their security levels must be equivalent) and they must be connected to an ephemeral, per-project, mesh network. Disclosure control can be carried out at the project network level (and individual TRE level if required by the project). The model supports different federated analytics solutions if they work to specific international standards, where queries can be “packaged” up to run within a secure bubble within each TRE.

The key here is that the “five safes” controls for each TRE remain the same. The definitions of “Safe Setting” and “Safe Outputs” are shifted to the level of a networked set of TREs rather than individual TREs. The Transformational Programme will work with early adopter TREs to test these assumptions with information governance teams and PIE groups, along with implementing and testing technical reference implementations of the standards.

This model aligns fully with the [DARE Federated Architecture Blueprint](#), which itself incorporates design work on [European data spaces](#) (including emerging [European Health Data Space](#) legislation).

## Conclusion

The NDL can deliver a cohesive set of entry points for researchers and analysts across academia, public sector, government **and industry** to find, access, and analyse public data at varying scales across the UK to answer questions that are in the public interest. We have outlined the principles that should underpin a health data ‘floor’ of the NDL; these principles could apply to all data types in scope for an NDL for research. The NDL must leverage and build on existing efforts. The various components and services of the NDL technical architecture already exist or are planned to (e.g. discovery, access, indexing, registry, wrangling, federation), the NDL should focus on bringing these components and services together as a coherent set of functionalities, whether delivered by the public or private sector We welcome further dialogue on the approach and design of the NDL. , The expertise across the HDR UK community – beyond just health – is committed to working in partnership to support a successful and high impact NDL for the public good.