DARE UK

Federated Architecture Blueprint - Part 1 (Executive Summary)

DARE UK Delivery Team

Version 2.2 final

November 2024

UK Research and Innovation

HDR UK Health Data Research UK

ADR UK Data-driven change

Licence

FOR CONSULTATION & COMMENT

Document control

Version	Date	Authors/Reviewers	Notes
0.6	22/03/2023	Rob Baxter	First complete draft.
0.7	31/03/2023	Fergus McDonald, Hans-Erik Aronson	DARE UK internal review.
1.0 initial	13/04/2023	Rob Baxter	For publication and public comment.
1.1	03/08/2023	Rob Baxter	Updated. Feedback until end June 2023 incorporated.
1.2	11/08/2023	Rob Baxter	Version for internal review.
1.3	15/08/2023	Fergus McDonald, Emily Jefferson	DARE UK & HDR-UK internal review.
1.4	25/08/2023	Rob Baxter	Updated. Greatly expanded Executive Summary. Version for internal review.
1.5	04/10/2023	Fergus McDonald, Emily Jefferson	DARE UK & HDR-UK internal review.
1.6 interim	18/10/2023	Rob Baxter	For broader circulation and comment.
2.0 draft	11/12/2023	Rob Baxter	Incorporated revisions and lessons learned from Driver Projects and wider engagements.
2.0A draft	12/12/2023	Rob Baxter	Incorporated review feedback from SACRO project PI.
2.0B draft	08/01/2024	Rob Baxter	Incorporated review feedback from TRE-FX project PIs.
2.0C draft	29/02/2024	Fergus McDonald, Emily Jefferson	DARE UK & HDR-UK internal review.
2.0D draft	28/03/2024	Rob Baxter	Final tidy-up, incorporating research use-cases from February 2024 workshop.
2.0E draft	13/06/2024	Fergus McDonald, Emily Jefferson, Caole Goble, Phil Quinlan, Simon Thompson	Partner review.
2.0F draft	05/08/2024	Rob Baxter, Heikki Lehväslaiho	Fixed error in Chapter 8, prototype descriptions.
2.1 draft	30/08/2024	Rob Baxter	Restructuring across Chapters 2-4; realignment and rationalisation of user roles.
2.1 Part 1..5	19/09/2024	Rob Baxter	Separation into multiple parts for release.
2.2	31/10/2024	Emily Jefferson	DARE UK & HDR-UK internal review.
2.2 final	11/11 2024	DARE UK	For release.

FOR CONSULTATION & COMMENT

Contents

About document versions

This document is Part 1 (Executive Summary) of the *Federated Architecture Blueprint* for DARE UK. It defines a potential approach for an overall architecture for a network of sensitive data sources and secure analytical services in terms which are broadly—and deliberately—**technology neutral**. Choices of implementation technology are not dealt with here, nor are details of costs, benefits and delivery plan.

This document covers architecture version 2. It refines the model of a federated network infrastructure from the "initial" and "interim" versions, builds further on the "data layer" and most significantly draws in lessons and learnings from the 2023 DARE UK Driver Project programme.

Acknowledgements

Our thanks go to the many individuals and organisations that have engaged with us, both around this architecture and more generally, over the course of this work to date. Appendix A has a full list of acknowledgements.

1. How to read this document

This suite of documents is intended for a specialist audience of technologists and experts who have knowledge of the application, purpose, creation and architecture of UK wide federated services for research. We hope that this Executive Summary is broadly accessible, but the details of federating trusted research environments is unavoidably complex and the companions to this document are quite technical.

The goal of this document is to capture and distil the sensitive data research infrastructure ecosystem into a single, overarching architecture that defines, to a necessary level of detail and useful level of abstraction, the fundamental elements of the ecosystem and how they should or could interact in a federated context. It attempts to capture those elements and interactions of the ecosystem that exist today as well as those that will need to exist in future if the DARE UK vision for cross-domain sensitive data research in the public interest and at scale is to be realised.

On the premise of a sensitive data landscape that is and will remain distributed, this document proposes a federated approach that connects organisations together under a common set of rules and standards that are as minimally intrusive to the good practice already in use. The purpose of this document is to establish a holistic, system-wide description of a UK-wide federation of sensitive data research infrastructures that:

- enables shared understanding across the various communities in the ecosystem.
- is collectively owned, managed, and maintained by the various communities in the ecosystem, evolving over time alongside the ecosystem.
- is a model around which the various communities can surface, propose, discuss, and establish consensus around strategic issues, tensions, and questions.
- provides a framework for strategic investments in sensitive data research infrastructure, particularly around the concept of cross-domain sensitive data linkage and analysis in a distributed infrastructure landscape.

While this document draws on existing best practice (see section 2.3) and provides some early thoughts on what a delivery approach could look like (see section 8), there are still fundamental questions that the UK sensitive data research community need to tackle. The intent is that this document provides a catalyst and framing for taking those questions forward, describing the various pieces of the puzzle that need to fit together to realise a UK sensitive data research infrastructure federation.

To that end, this document is open to constructive challenge and critique that is in the spirit of advancing the UK's vision to be a global exemplar of harnessing data for the public good, by assembling a scalable, reliable and trustworthy cross-sectoral data ecosystem for research .

> "All models are wrong, but some models are useful."
>
> George E.P. Box

2. Overview

Research with sensitive data already happens in the UK, in pockets of good practice connected by ad hoc technical processes. Alongside "classic" sensitive data from health and government sources there is increasing research interest in bringing other kinds of data into a common framework. This fragmented landscape suffers from attendant frictions and bottlenecks in data sharing and is a significant drag on researcher productivity.

Analytics services for researchers working with sensitive data are typically—and increasingly—provided in trusted research environments (TREs), secure computer systems wrapped in information governance practices and processes modelled on the Five Safes approach developed by the Office for National Statistics (ONS[1]). These cast the technical systems needed to support sensitive data research as one part (the "safe setting") of a broader set of procedures designed to manage risk and create an overall trustworthy environment.

To introduce standardisation and additional trustworthiness to the existing – and future – network of TREs and data providers, we propose the idea of a Secure Data Research Infrastructure Federation, with three key capabilities:

- common, standardised security and privacy controls for individual TREs and other participating services;
- common, standardised collaborative data communication between participating services;
- a common TRE trust domain, including certifications and required levels of compliance.

Together these capabilities create a backbone for secure information exchange between all participants, with strong guarantees of confidentiality, integrity and availability. By this means we can connect TREs, data providers and other service providers together in a high-assurance network with common trust and strong governance oversight.

Running on top of this backbone we envisage a set of application services in a small number of different classes. We identify needs for service classes for:

- the exchange of data extracts;
- the exchange of linkage spines;
- the exchange of queries and results;
- and the download of approved software from controlled sources.

We deliberately discuss these services in the abstract, as classes of interfaces exchanging structured documents in separately secured contexts. In this way we seek not to over-specify what functionality an innovative network of TREs can and cannot offer but rather to highlight the need for descriptive metadata standards for a range of entities and concepts within the federation network.

Governance of the overall Federation follows the same principles as the technical approach: augment what is already in place without disrupting it. We highlight the key relationships and accountabilities

[1] The UK Office for National Statistics, https://www.ons.gov.uk/

within the proposed Federation, and introduce first ideas for the process-set necessary to govern a UK-wide federation for sensitive data research.

3. The strategic case for federation

The needs of independent information governance (for instance, between the four nations of the UK) and the practicalities of data movement in some cases (in large environmental datasets, for example) mean that data will and should remain distributed. On the premise of a sensitive data landscape that is distributed we accordingly propose a federated approach to connecting TREs and other services together in a way that is standardised but as minimally intrusive to the good practice already in use.

In our context, we use federation in its broadest sense of connecting organisations together under a common set of rules and standards. This provides the framework for research patterns which either involve moving analyses to distributed datasets ("**federated analytics**") or moving datasets into a single location for analysis ("**data pooling**"). We observe that the Federation must support both.

In parallel with the development of this architecture the DARE UK programme has supported **five "driver projects"**, each of which explored possible technologies and tools that could be used in later implementation work. We summarise these briefly and describe their impact on version 2.0.

4. Users and use-cases

We introduce **ten user personas** derived from hosted workshops in 2022 and 2023, representing archetypal users, from research through TRE service provision to data custodianship and including a "member of the public" persona. From these personas we enumerate **61 high-level user stories** as sources of requirements for TREs, data providers and the Federation itself.

We observe that both current practice and future use will require an architecture that supports both the data pooling and federated analytics patterns.

5. Federated architecture: infrastructure layer

The picture below is a simplification of the detailed infrastructure diagram from Chapter 5 and illustrates the essence of the Federation.

Federation Participants are shown in blue: TREs and supporting services. We show two **TREs**, two **Software Services** and one each of **Index**, **Job Submission** and **Discovery** for illustration. In the actual Federation there will be many of each kind, specialising in different kinds of data or analytical capability.

The core of the SDRI Federation sits between the other services, with connections shown between the standardised **Security Servers** at each Participant, plus a single group of **Federation Services**. This core of Federation Services, Security Servers and connections together define the Federation. The Federation Services group comprises services for registry (of services, users, projects, etc.), trust (security certificate management and signing), management (of standard shared software), monitoring and accounting.

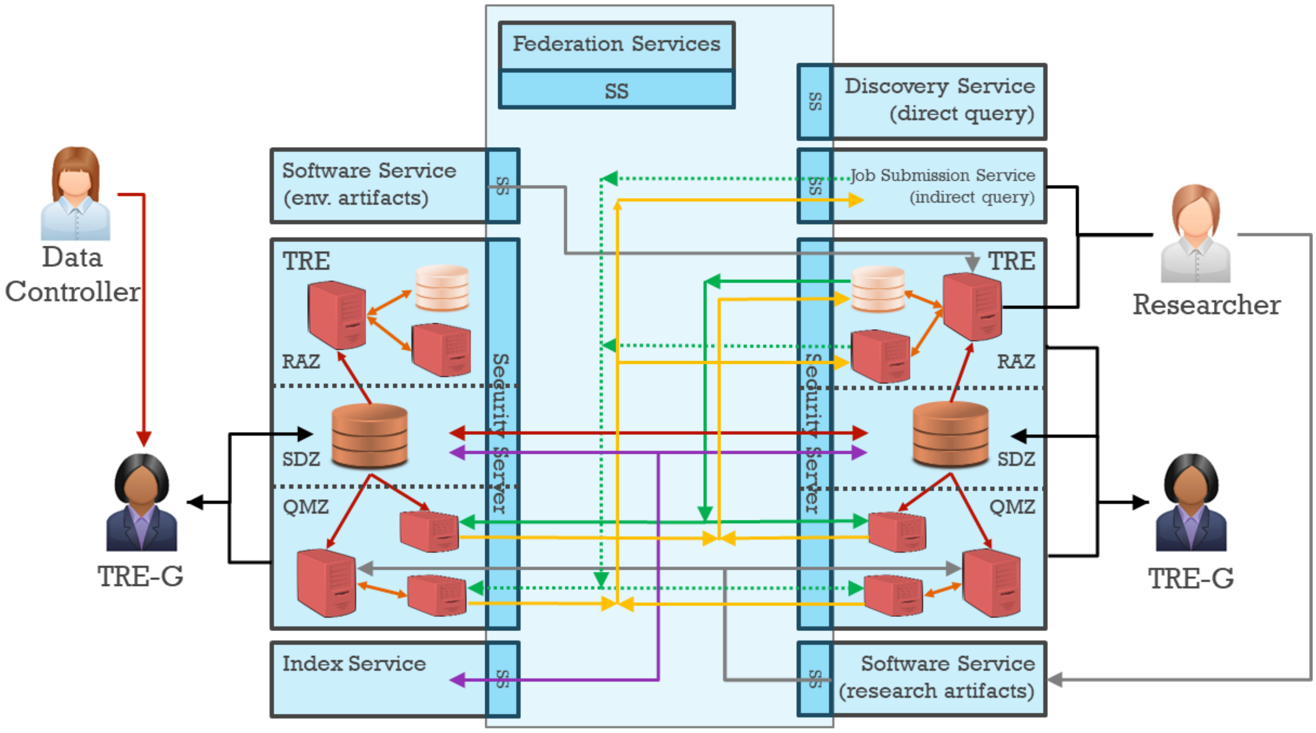
Federation Services
SS
Discovery Service (direct query)
SS
Software Service (env. artifacts)
SS
Job Submission Service (indirect query)
SS
Data Controller
TRE
RAZ
SDZ
QMZ
Security Server
TRE
RAZ
SDZ
QMZ
Security Server
Researcher
TRE-G
TRE-G
Index Service
SS
Software Service (research artifacts)
SS

Figure 1. Simplified architectural sketch of the Sensitive Data Research Infrastructure Federation. Trusted Research Environments are denoted "TRE". TREs are divided logically into three internal zones: a Research Analytics Zone (RAZ), a Secure Data Zone (SDZ), and a Query Management Zone (QMZ). Not all zones need be present in any given TRE."SS" = Security Server, a secure common gateway for all inter-TRE traffic. "TRE-G", TRE Governance, is shorthand for all those responsible for the security and integrity of running a TRE.

Different-coloured connections between Participants are shown, with the colours representing the different types of connections allowed within the Federation. Note that these connections run directly between Participants, not through any Federation Services hub. The Federation control plane and data plane are independent, touching only at individual Participants' Security Servers.

The arrows in the diagram are significant and indicate the direction of flow of information.

Green connections allow Participants to send "queries" to other Participants. These "queries" come in two forms: "direct queries" (solid green) which include all the necessary information for the query receiver to run it (an SQL statement, for example); and "indirect queries" (dotted green) which indicate that a TRE needs to download additional software (workflows, scripts or containers, for example) in order to execute it. Queries are, of themselves, unlikely to be disclosive and so may be treated with low levels of disclosure control.

Orange connections represent the responses returned by queries. While typically thought of as aggregate summaries, results do have the potential to contain disclosive information, depending on the query sent and the dataset queried. While results would only ever be sent through secure gateways (Security Servers) to other approved Participants within the closed Federation network, disclosure controls may be appropriate for certain kinds of results.

Red connections allow Data Controllers and TREs acting as data providers to send datasets and data extracts to governance authorities in TREs in standard, secure ways[2]. Sensitive personal data are de-identified and approved for use in research but are nevertheless potentially disclosive and, despite the above remarks about secure gateways and closed Federation network still applying, disclosure controls are appropriate for red connections.

The other connections shown are purple for index services, which create linkage spines for data linkage, and grey for software artifacts delivered by software services (the workflows used in indirect queries are an example).

The architecture only specifies what is strictly necessary to meet the needs of the different methods of federation described in Chapter 3: data pooling, and federated analytics with both direct and indirect queries. To this end our model of a TRE has three distinct zones: a **research analytics zone** (RAZ), a **secure data zone** (SDZ) and a **query management zone** (QMZ). We observe that not every TRE need support every zone.

We conclude this chapter with definitions of some additional key concepts, including **projects**, **identities** and **authorisation**.

6. Federated architecture: data layer

We provide a simple cross-comparison of current data classification schemes (e.g. GDPR, UK Government) mapped to a single seven-point scale which could be used as a standard designation across the Federation.

The introduction of registry services raises the need for a **common metadata model** of the Federation itself. In discussing this we use the same layering as the architecture itself and produce the following model:

Federation metadata: what the Federation actually *is*, comprising:

- Infrastructure metadata: what the service layer looks like, comprising:
  - Descriptive metadata: static information about Participants, their service types, capabilities and so on.
  - Operational metadata: dynamic information, especially logging data from Security Servers.
- Content metadata: what "content" is in the Federation, comprising:
  - Dataset metadata: high-level (catalogue-level) information about each dataset available for potential research use within the Federation.
- Governance metadata: who has access to Federation assets for what purposes, comprising:
  - Project metadata: information defining each current or completed research project.
  - User metadata: information about each user of the Federation, the roles they have, the approvals they have, the Projects they are members of, and so on.

[2] Throughout, we use "TRE Governance" as a shorthand for the team of people charged with running a TRE, including technical administrators, data analysts, statistical disclosure control experts and other information governance professionals.

- Data Extract metadata: information about subsets or extracts of Datasets as used in Projects.

Where possible we illustrate these concepts with examples drawn from existing sources, notably the metadata records required of services seeking to acquire accreditation as data processors under the Digital Economy Act 2017.

We observe that the creation of a single registry with this kind of metadata model also enables some form of **publicly accessible presentation** of what research projects are active right now, using which datasets – with obvious exciting opportunities for greater public transparency.

Strictly speaking, the Federation metadata model introduced here should define the limit of our scope with respect to any broader discussion of data standards. Nevertheless, we go on to discuss a number of concepts that will be the focus of Discovery Services and Index Services (q.v.) yet to be developed.

We use the FAIR principles of findability, accessibility, interoperability and reusability to frame this discussion.

For **findability** we recommend agreeing and adopting within the Federation existing standards for high-level metadata, highlighting current recommendations from UK Government and National Health Service sources: DCAT, schema.org, Dublin Core, UPRN, ISO 8601, OMOP, and so on.

For **accessibility** we highlight the need to find the right mix of data pooling vs federated query for complex projects. Projects involving initial, iterative "exploratory data analytics" on small-scale data samples are difficult to realise in a purely federated analytics environment, for instance.

For **interoperability** we focus on data linkage and discuss three areas of increasing challenge to automating linkage and Index Services across the Federation. This kind of categorisation should support incremental development of discovery and indexing services of increasing sophistication.

For **reusability** we observe simply that reuse of sensitive data from one project in another is much more a governance question than a technical challenge.

7. Federated architecture: organisational layer

We note that the design of the operational model of the Federation must be **community-led**, and the organisational structures of the Federation must be comprised of the set, or an agreed core sub-set, of the Federation Participants (TREs and their governance bodies, other services).

We introduce the idea of a **Federation Authority** (FA) as an oversight body, and discuss the pros and cons of delivering different aspects of the FA's functions through **centralised**, **distributed** or **decentralised models**. We draw no conclusions but offer this up as a starting point for broader community dialogue.

8. Development and delivery approach

We observe that our separation of concerns into Federation foundation services on the one hand, and application-level services on the other leads to a two-speed approach to technology selection and development. Software for the foundation services should be selected from existing solutions already proven in operation (technology readiness level 9 in the standard industry jargon); it should NOT be commissioned from new research work.

This encapsulation of essential security features in the foundation layer means that application services which run “on top” can be more innovative and even experimental without compromising overall Federation security.

We sketch a number of **small pilot scenarios** which can build on each other to realise a running system which can be **scaled out incrementally** without the need for a single “big bang”.

9. Summary and further work

This blueprint is version 2.2. How future versions may evolve is currently in planning and may change based on feedback from the community, stakeholders and/or DARE UK programme governance structures.

A Acknowledgements

A.1 Federated architecture blueprint: direct feedback

Thanks to the organisations, groups and individuals who have provided direct feedback on earlier versions of this document. In many, many cases feedback was comprehensive, detailed and thoughtful, and we thank those groups in particular for the time they invested. We have done our best to incorporate the multiple angles and viewpoints; we may not have succeeded completely but without doubt this document is the better for it.

- Alison Kennedy, Director, The Hartree Centre
- Professor Ben Goldacre OBE, Director, Bennett Institute for Applied Data Science, University of Oxford
- Canon Medical Research Europe
- Professor Carole Goble CBE, University of Manchester; Head of ELIXIR UK
- Dr Claire Bloomfield, Deputy Director, Data for Research and Development, NHS England
- Professor David DeRoure, Director, Oxford e-Research Centre, University of Oxford
- Professor David Ford, Swansea University; Director, SAIL Databank and SeRP
- DEA Research Assessment Panel
- Professor Elena Simperl, Kings College London
- Professor Emily Jefferson, University of Dundee; CTO, HDR-UK
- Heikki Lehväslaiho, CSC IT Centre for Science, Finland
- Professor Jim Smith, University of the West of England
- Lifebit Biotech
- medConfidential
- Dr Olly Butters, Institute of Population Health, University of Liverpool
- OurFutureHealth
- PA Consulting
- Dr Pete Barnsley, Head of Special Projects, The Francis Crick Institute
- Dr Phil Quinlan, Director of Health Informatics, University of Nottingham
- Research Data Scotland
- Professor Søren Holm, University of Manchester
- Dr Steven Newhouse, Deputy CIO Precision Medicine, Barts Heath
- Professor Tim Hubbard, Kings College London; ELIXIR
- Professor Tony Brooks, University of Leicester; Global Alliance 4 Genomics & Health and EPND
- Will Crocombe, RISG Consulting
- Dr William Viney, Patient Experience Research Centre, Imperial College
- … and…
- Members of the public

A.2 Phase 1b persona development

Thanks to attendees of the July workshop at the Wellcome Trust in London, including representatives of:

UK Research and Innovation

HDRUK Health Data Research UK

ADR UK

- The Alan Turing Institute
- Amazon Web Services
- The Bennett Institute for Applied Data Science
- Connected By Data
- The Francis Crick Institute
- HDR-UK
- InnovateUK KTN
- MRC
- The Office for National Statistics
- Research Data Scotland
- RISG Consulting
- SAIL Databank/UK SeRP
- Secure Data Access Professionals Group
- STFC
- UK Data Service
- UK Health Security Agency
- UK Longitudinal Linkage Collaboration
- UK Statistics Authority

A.3 DRI landscape review and community conversations

Our thanks also to the organisations, groups and individuals who engaged with our surveys, follow-ups and ad-hoc conversations over the course of 2023. All these engagements have helped shape and steer our thinking.

- AIMES TRE
- Akrivia Health Clinical Research Interactive Search (CRIS)
- Alan Turing Institute Data Safe Haven
- Aridhia DRE
- AWS Service Workbench
- Barts Health Precision Medicine Platform
- BHF Data Science Centre instance of NHS England TRE/SDE
- Big Data and Analytical Unit Secure Environment (BDAU SE), Imperial College
- British Ocean Sediment Core Research Facility
- Centre for Macaques, Medical Research Council
- Centre for Rapid Online Analysis of Reactions (ROAR)
- CLARIN
- Clinoverse
- Connected By Data
- Consumer Data Research Centre (Leeds)
- DAFNI - Data and Analytics Facility for National Infrastructure
- DataLoch
- Edinburgh International Data Facility
- Electron beam lithography facilities, University of Cambridge

- EPND (European Platform for Neurodegenerative Diseases)
- FAIRDOM
- FAIRDOM-SEEK
- Gates Ventures
- Genomics England RE
- GG&C Safe Haven
- Grampian Data Safe Haven, University of Aberdeen & NHS Grampian
- Health Informatics Centre, University of Dundee
- InterConnect and MRC Epidemiology Unit in-reach system
- JASMIN
- Leeds Analytic Secure Environment for Research (LASER)
- Lifebit Federated Trusted Research Environment
- Microsoft AzureTRE
- National Survey of Sexual Attitudes and Lifestyles (Natsal)
- Natural History Museum
- NDORMS
- NERC Digital Solutions
- NHS England SN SDE Network Technology & Infrastructure Working Group
- NI Honest Broker Service
- NIHR BioResource
- NURTuRE
- ONS Integrated Data Service
- ONS Secure Research Service
- OpenSAFELY in OpenSAFELY-TPP and OpenSAFELY-EMIS
- OurFutureHealth TRE
- Personalised Medicine Centre, Ulster University
- Royal Botanic Gardens Kew
- SAIL Databank
- Scottish National Safe Haven
- Secure eResearch Platform (Serp)
- Sir Peter Mansfield Imaging Centre
- Software Sustainability Institute
- STFC Scientific Computing Department
- The Francis Crick Institute
- The GW4 Isambard Tier-2 HPC service
- UK Data Service
- UK Health Security Agency (UKHSA)
- UK Longitudinal Linkage Collaboration
- UKAEA
- UKAEA Materials Research Facility
- UKRI - Medical Research Council - Mary Lyon Centre at MRC Harwell
- United Kingdom Multiple Sclerosis Register
- University of Liverpool

- University of Portsmouth
- University of Sheffield Sensitive Data Service