



# Federated Architecture Blueprint - Part 2 (Strategic Case)

DARE UK Delivery Team


Version 2.2 final

November 2024



**FOR CONSULTATION & COMMENT**

**Licence**

This work © 2024 by HDR UK and other members of the DARE UK consortium is licensed under CC BY-NC-SA 4.0 The Creative Commons license icons, including the CC logo, a person icon (BY), a crossed-out dollar sign (NC), and a circular arrow (SA).

**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

Document control .....	3
Contents.....	4
About document versions .....	5
Acknowledgements .....	5
1. Introduction .....	6
1.1. DARE UK Phase 1 recommendations .....	6
2. The federation challenge.....	8
2.1. Conceptual data space .....	9
2.2. Data pooling .....	10
2.3. Federated analytics.....	10
2.3.1. Direct query .....	11
2.3.2. Indirect query.....	11
3. Federated infrastructure: the state of the art.....	13
3.1. TRE federation proofs-of-concept: the DARE UK driver projects .....	14
4. A federation blueprint.....	16
4.1. Scope.....	16
4.2. Design principles .....	17
5. Summary.....	18
6. References.....	19
A A comparison of contemporary federated data architectures .....	21
B Scenario analysis of the federated landscape .....	23
B.1 Four quadrants.....	23
B.1.1 Low numbers of TREs and low data mobility .....	23
B.1.2 Low numbers of TREs and high data mobility .....	24
B.1.3 High numbers of TREs and low data mobility .....	24
B.1.4 High numbers of TREs and high data mobility .....	24
B.2 Observations .....	25

## FOR CONSULTATION & COMMENT

### About document versions

---

This document is Part 2 (Strategic Case) 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 of Part 1 (Executive Summary) has a full list of acknowledgements.

## FOR CONSULTATION & COMMENT

# 1. Introduction

---

“The UK Research and Innovation DARE UK (Data and Analytics Research Environments UK) programme has been established to design and deliver a coordinated and trustworthy national data research infrastructure to support research at scale for public good. DARE UK is a cross-domain programme—its scope covers all types of sensitive data, including data about education, health, the environment and much more.”

DARE UK Phase 1 report: *Paving the way for a coordinated national infrastructure for sensitive data research*

The DARE UK programme is built on the concept of a UK sensitive data research landscape which is fundamentally distributed, both in its sources of available data and in the analytical services able to process them [1]. While the numbers and locations of data sources and services within this landscape will ebb and flow (see Appendix B *Scenario Analysis*) there is no likely future scenario which brings all data and all compute services together in one location. To enable researchers to work with data linked from multiple sources, a federated digital research infrastructure is needed.

## 1.1. DARE UK Phase 1 recommendations

There are ten key recommendations from the DARE UK Phase 1 report [2] that shape our approach to a federated architecture for trusted research environments (TREs) across the UK, and two from the DARE UK 2022 public dialogue [3].

### Data and discovery

From [2]:

1. Enhance the data lifecycle to support effective cross-domain sensitive data research.
2. Explore the implications of new data types on approaches to making these data available for research.
3. Develop guidelines on privacy enhancing technologies (PETs) for use by TREs.
4. Establish a UKRI-wide metadata standard working group.
5. Leverage existing Digital Object Identifier (DOI) minting services to provide persistent identifiers for all UKRI discoverable assets at UKRI-wide and council levels.

### Core federation services

From [2]:

1. Develop reference architecture(s) for TREs.
2. Assemble an API (application programming interface) library to support core federation services.
3. Run a competitive call for driver projects to utilise the new infrastructure services and validate that they are fit for purpose.
4. Establish an approach to business continuity and disaster recovery.

### Capability and capacity

From [2]:

## FOR CONSULTATION & COMMENT

4. Use automation to ensure data research infrastructure services are reliably secure, auditable and reproducible.

### Public engagement and dialogue

From [3]:

4. The processes and systems supporting data research across the UK should be unified in their approaches where possible.
5. Where feasible, processes enabling access to sensitive data for research should be standardised and centralised.

Of these 12 the strongest influence on this blueprint comes from the public dialogue Recommendations 4 and 5, the public view that trustworthiness will derive in no small part from standardised, centralised processes and systems, where feasible. These concepts sit at the heart of our proposed approach of a common federation for sensitive data research infrastructure.

**FOR CONSULTATION & COMMENT**

## 2. The federation challenge

While there are many ways to define “sensitive data” one important definition is “individual-level public data”, and particularly individual-level data defined as “special category” under the UK GDPR [36] (electronic health records, for example). The UK has rich sets of data about its citizens, both collected routinely through citizens’ interactions with government, health bodies and other administrative centres, and collected voluntarily through clinical trials, survey responses and so on. Making these data available for research at population scale, in joined-up ways, has tremendous potential for public good (see box right<sup>1</sup>). But whatever the source, any use of public data for research must have public trust, and benefit, at its heart.

The need to connect distributed data and distributed analytics services requires a federated approach: a common set of protocols and standards agreed by all participants enabling the “intelligent” exchange of data for research [5] and increasing the prospects of safe automation across the landscape. To enable the exchange of sensitive data—in particular public data—the federation must be trustworthy.

One aspect of the challenge we cannot ignore is that we do not start from scratch. The UK has a significant number of TREs, already delivering real scientific advances, as COALESCE illustrates. Any federation architecture must recognise the existing service infrastructure, whilst enhancing its trustworthiness and creating an environment where common standards create a platform for continued innovation.

Using this approach we derive three essential use-cases:

1. Data pooling, where approved datasets or data extracts are moved between TREs, pooled in a single location and optionally linked, before being provided to a research team as a project. Analysis tools and resources are provided at the pooling location to support the project.
2. Federated analytics, where approved datasets are held in situ and analytical “queries” are split into parts that can run independently on each of the remote datasets. This is further divided into:
  - a. Direct query, where an analytical query sent to the remote datasets is fully encapsulated in the request object and contains everything needed to execute the query on the data; and,

### *The first, but not the last*

In January 2024 the COALESCE consortium published the UK’s first whole-population analysis [4]. The study, of covid-19 under-vaccination and severe outcomes, was a meta-analysis across the separate, independent TREs of the UK’s four nations: the NHS England Secure Data Environment, the Scottish National Safe Haven, the SAIL Databank in Wales and the Northern Ireland Honest Broker Service. The meta-analysis method meant that comparable statistical analyses were performed separately inside each TRE, and the resulting statistics were knitted together afterwards. The study had to overcome challenges of data harmonisation and scale in four different ways, across four different secure environments.

One key goal of a technical and organisational federation of the UK’s TREs is to make future studies like COALESCE much easier to conduct.

<sup>1</sup> For more information on the ground-breaking COALESCE study, see <https://www.ed.ac.uk/usher/eave-ii/connected-projects/coalesce/uk-first-whole-population-analysis>



## FOR CONSULTATION & COMMENT

- b. Indirect query, where an analytical query sent to the remote datasets contains references to additional computational workflows, scripts or other software that must be downloaded from another service before the query can be executed.

Since our interest is in the federation of TREs and data providers at the organisational level we do not consider the details of data provision to researchers within a TRE.

### 2.1. Conceptual data space

We can bring these ideas together into a conceptual data space where different kinds of dataset are divided across different regional data custodians. Each block in Figure 1 is conceptually held by a different organisation.

This division works particularly well when considering individual-level health or administrative data which are held locally or regionally (by local authority or by health board, for instance). Generally, we assume there is a population of interest, defined by some primary key, which is divided into discrete regions. Within each region are a number of disjoint datasets about each population subset.

With the primary key running row-wise, partitioning the overall dataspace horizontally results in a number of sub-populations with common attributes.

Partitioning vertically splits the attribute space for the whole population. Doing both creates the picture in Figure 1.

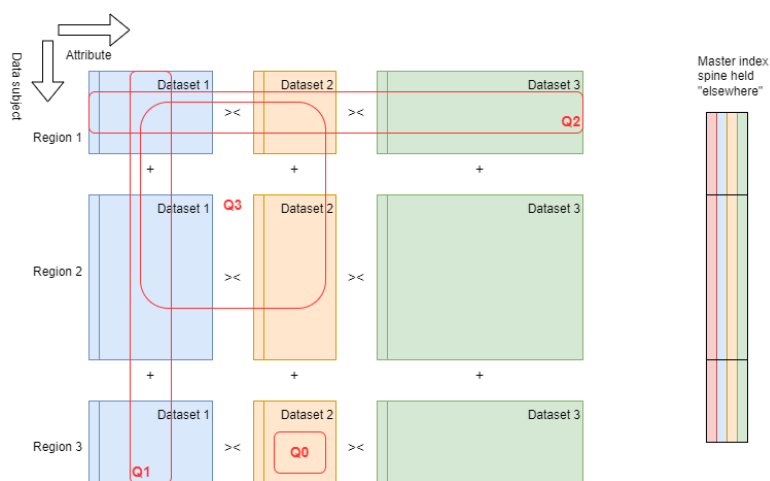


Figure 1. Conceptual dataspace for DARE UK

The reality of data combination is much messier than this picture suggests, of course; nevertheless a conceptual abstraction at this level is useful in categorising use-cases and identifying common requirements and functionality within a broad architecture. In particular it helps us characterise query patterns across the different dimensions, and hence understand what federation mechanisms will be needed to enable them. Figure 1 highlights four basic query patterns:

- Q1: a query across a single dataset but spanning multiple regions to include a larger population than is available at any individual data custodian. Queries of this kind can be run independently in each region and the results combined trivially.
- Q2: a query across the population of a single region but spanning multiple datasets. Queries of this kind (probably) cannot be run independently on each dataset but (probably) require the joining of schema-wise-different datasets by some kind of key representing individuals.
- Q3: a query combining the complexity of both Q1 and Q2, requiring joins across multiple datasets and combination across multiple regions.

For completeness there is also:

**FOR CONSULTATION & COMMENT**

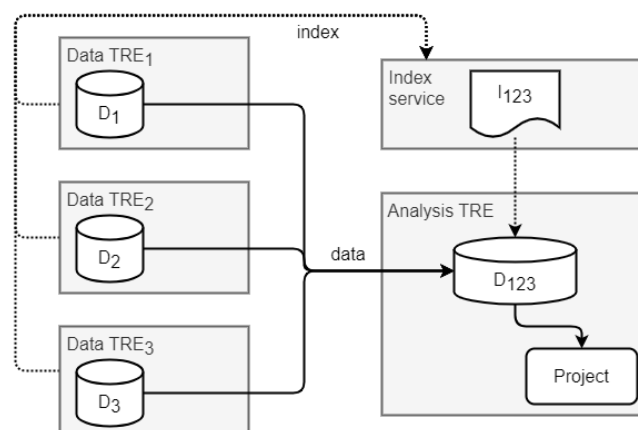
Q0: a query within a single regional dataset.

These high-level data patterns give rise to number of requirements that we note below.

## 2.2. Data pooling

The data pooling pattern occurs more often in current use. Here datasets are often vertically partitioned and need to be linked together using a common “master index” ( $I_{123}$ ). The index is created by a trusted third-party “index service” in a way that ensures that the resulting linked dataset ( $D_{123}$ ) is only ever created within the analysis TRE.

This pattern is needed to combine different kinds of data using a common spine such as individual-level identifiers, universal property reference numbers etc. and requires careful governance of both datasets and indexes.



## 2.3. Federated analytics

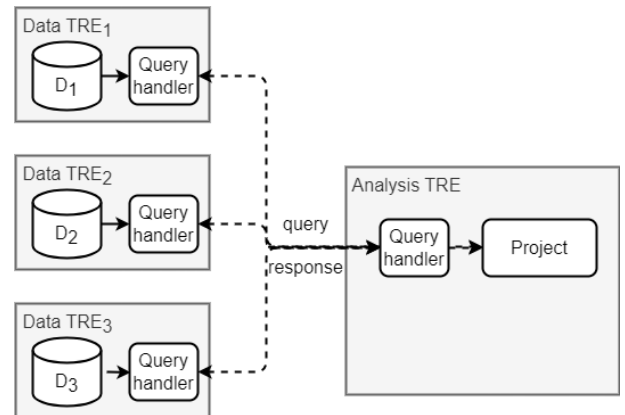
The federated analytics pattern works very well when data are horizontally partitioned but otherwise uniform (e.g., census data divided by region). It can be made to work when data are vertically partitioned, although it is technically more challenging to include the additional index service needed to make the join between the remotely calculated query results. In either use, the underpinning premise of the Federation – a trustworthy network between Participants – enables the exchange of queries and results in the context of an approved Project to happen without the need for “Federation-internal” disclosure control. All analytical queries and all results are maintained within the secured Federation network, and only move between TREs or other equivalently secured services.

Federated analytics can also be used as a mechanism to create Discovery Services (cf. Part 4 (Components)) which support distributed metadata discovery from *outside* the Federation – although because this usage connects internal Federation queries to the outside world, Discovery Services must be designed with disclosure control in place and with careful governance oversight.

## FOR CONSULTATION & COMMENT

### 2.3.1. Direct query

Of the two federated analytics patterns the direct query pattern is the simpler but covers the fewest concrete use-cases. Here, datasets ( $D_1$ ,  $D_2$  and  $D_3$ ) remain within their data provider organisations (“data TREs” 1, 2 and 3) and queries across them are sent from a project within an “analysis TRE”. The data TREs need to have the capability to handle the queries. Responses are returned to the project but not necessarily synchronously: query responses may need to be disclosure checked before they are permitted to leave the data TRE.

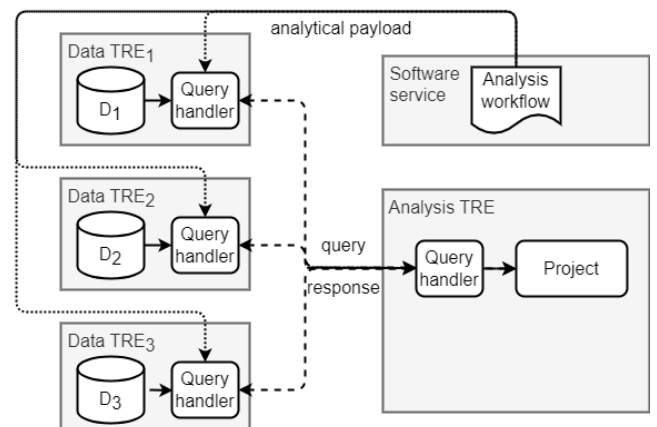


The “query” here is fully encapsulated in the request from the analysis TRE; no additional information or external software is needed by the data TREs to execute the query. The actual query may be simple (e.g., an SQL `COUNT`) or it may be a complex object containing partial training results from a machine learning model needing additional disclosure checks, but in all cases it must be fully encapsulated in the Query Object as received by the data TREs.

An example implementation of direct query can be found in the TELEPORT project [29]. TELEPORT uses the Trino SQL execution engine<sup>2</sup> to connect remote data sources within one TRE to a “single pane of glass” user-view in another. To the research user, this has the appearance, and consequent utility, of a single database table, while behind the scenes queries and results are exchanged between participating TREs.

### 2.3.2. Indirect query

The indirect query pattern captures the use-cases seen in federated analytics using job submission: a job request is created by researchers on a project and sent to participating “data TREs”. Again, the datasets ( $D_1$ ,  $D_2$  and  $D_3$ ) remain within their provider organisations. To execute the job query, the TREs must download the actual “analytical payload” (a workflow, for example) from another source, run it, and return the response to the originating service. (This download may need to be done in advance, and the contents of the payload risk-assessed before it can be executed within the TRE.) Each TRE must, of course, have the capability to handle the queries.



As with direct query, responses are returned to the project but not necessarily synchronously: job responses may need to be disclosure checked before they are permitted to leave the data TRE.

<sup>2</sup> “Trino, a query engine that runs at ludicrous speed”. See <https://trino.io/>

## FOR CONSULTATION & COMMENT

An example implementation that support both indirect and direct query can be found in the TRE-FX project [27]. TRE-FX uses the Hutch federated job execution software<sup>3</sup>, enabling researchers to request the execution of complex workflows within participating TREs. The workflows can either be fully encapsulated in the request object, mapping onto the direct query model, or be developed “out of band” by a researcher, uploaded to a trustworthy repository and then downloaded and screened for safety by operators at participating TREs, each acting independently and in accordance with their own risk profiles and policies. In both cases TRE-FX uses the same standard approach for object exchange between TREs, the RO-Crate packaging format (cf. Part 4 (Components)).

---

<sup>3</sup> Hutch, a federated analytics execution agent. See <https://health-informatics-uon.github.io/hutch/>

**FOR CONSULTATION & COMMENT**

### **3. Federated infrastructure: the state of the art**

---

Infrastructure federations have been a staple of the UK research landscape since the early 2000s and the drivers of the UK e-Science Core Programme [8]. The World-wide LHC Compute Grid (WLCG [9]) and the International Virtual Observatory Alliance (IVOA [10]) adopted techniques for managing “virtual organisations” developed in those early years and are now global science federations managing petabytes of natural science data.

Closer to the concept of sensitive data but also seeing roots in the e-science development of “Grid computing” (a forerunner of cloud computing) are more than 15 European research infrastructures spanning health and social sciences [13]. Notable examples include ELIXIR [11], BBMRI [12], CESSDA [14] and ESS [15]. Of these, ELIXIR operates as an international treaty organisation through its founding partner EMBL and the other three are incorporated as European Research Infrastructure Consortia (ERICs).

International ambition on the sharing and pooling of routine national “register” data for research is well illustrated in the Nordic Commons model proposed in Scandinavia [16]. With their strong traditions of good national record-keeping, and bound by the GDPR, the Nordic countries offer a blueprint for federated data sharing that is well worth studying.

UK research is thus not alone in seeking a federated solution to distributed resources in an environment that requires very high levels of trust. There are a number of current and emerging technology solutions which seek to build (or have built) federated environments between independent organisations with high levels of assurance and trustworthiness. All follow the same pattern of inter-service standards and many make use of a managing agency.

X-Road [17], managed by the Nordic Institute for Interoperability Solutions [18], is the open-source platform developed by the government of Estonia from the 1990s onwards to underpin the delivery of government services in the new nation that emerged from the Soviet Union. X-Road provides a secure infrastructure for document exchange between government agencies, police, health services and citizens. While X-Road is open source it remains the backbone of digital government in Estonia, Finland, Iceland and other nations and so its core development is managed by NIIS. Estonia, along with the UK, was one of the founders of the “Digital Five” advanced digital governments, now the “Digital Nations” [19].

GAIA-X [20], initiated in 2019 by the French and German economics ministries, is seeking to define a reference architecture and model implementations of a secure, federated infrastructure [21]. It shares many similar concepts with X-Road and with both IDSA and SiMPI (q.q.v.). GAIA-X’s designs and software implementations are open source but managed by the GAIA-X aisbl (a Belgium non-profit incorporation) which is open to join but requires a subscription fee. GAIA-X describe a number of “lighthouse projects”, federated infrastructures in operation using their architecture in sectors spanning agriculture, automotive and tourism.

The International Data Spaces Association (IDSA) is “a cross-industry, transnational coalition of more than 140 leading companies and research organizations” that has been developing concepts and standards for “data spaces” since 2016. Data spaces are federations of organisations created to enable the secure sharing of data between them, with a strong focus on contractual arrangements for commercial use. Version 4 of the IDSA Reference Architecture Model (“IDS-RAM”) is publicly available [22].

## FOR CONSULTATION & COMMENT

The most recent work in this space is perhaps the launch of an invitation to tender for the European Smart Middle Platform (variously SiMPI or SMP) [23]. SiMPI is designed to create an open standards-based approach to cloud interoperability and provisioning (“cloud-to-edge federation”) and to underpin the European Data Strategy [24] and the further development of data spaces. The published timetable for SiMPI suggests a minimal viable product should be released “at the end of 2024”.

As noted, the proposed SiMPI architecture shares many common features with X-Road, IDS-RAM and GAIA-X; these four initiatives do collaborate at various levels. Appendix A provides a comparison of these initiatives, alongside similar concepts from the proposed SDRI Federation architecture.

### 3.1. TRE federation proofs-of-concept: the DARE UK driver projects

During 2023 the DARE UK programme funded a portfolio of driver projects to explore potential technologies in this space, three of which in particular have a strong bearing on topics covered later in this blueprint. For an overview of these projects, see the DARE UK website<sup>4</sup>.

**SATRE** [25] compared openly available UK TREs hosting health, manufacturing, commercial, science and humanities data and aligned them into a standardised TRE reference architecture. SATRE’s scope was strongly intra-TRE, looking to answer the question: how do we specify what a TRE should be at a technical level? Answers are recorded in the project’s principal output, the “SATRE Specification” [26].

**TRE-FX** [27] demonstrated the use of existing technologies from ELIXIR and HDR-UK to support federated analytics across a network of TREs and data providers. Federated analytics—sending the analysis scripts or programs to the dataset, where the dataset is split across several physical locations—is one of a small number of key application types that would run on top of the core federation. TRE-FX applied the “job submission” approach to federated analytics also seen in OpenSAFELY [28] and numerous other solutions: request that a TRE download and run an analysis script developed “outside” the environment. TRE-FX developed a standard way to submit jobs that is “5 safes” compliant, and worked with partners from Bitfount<sup>5</sup> and DataSHIELD<sup>6</sup> to integrate these standards into their product suites.

**TELEPORT** [29] demonstrated how to offer a single query interface to users of a TRE that spans multiple remote datasets – a “single pane of glass” approach whereby a researcher can log into one TRE and see their approved project data from the other TREs as though it were all held within the same environment. Potentially data can be linked across the different TREs if an indexing service has provided the different TREs with the same pseudo-identifiers corresponding to the same individual. TELEPORT combined this data federation approach with the use of “pop-up TREs” or “TREs-within-TREs”, project-specific instances of TREs created virtually within a larger TRE infrastructure. By synchronising these “pop-up TREs” with overlapping governance “wrappers” defined by the TREs contributing data to the project in question, TELEPORT showed how federated querying can be made just as safe and secure as accessing data in a single location.

---

<sup>4</sup> DARE UK 2023 Driver Projects, <https://dareuk.org.uk/our-work/phase-1-driver-projects/>

<sup>5</sup> Bitfount federated AI and data science platform. See <https://www.bitfount.com/>

<sup>6</sup> DataSHIELD secure bioscience collaboration. See <https://datashield.org/>

## FOR CONSULTATION & COMMENT

Two additional projects developed enhanced tooling for assessing disclosure risk in datasets at the beginning and the end of the research process.

**SACRO** [30] sought to reduce the operating costs of TREs and the time taken to check and release research results by, among other things: producing a consolidated framework with a rigorous statistical basis that provides guidance for TREs to agree consistent, standard processes to assist in quality assurance; and, designing and implementing a semi-automated system for checks on common research outputs, with increasing levels of support for other types of output, such as AI (artificial intelligence).

**SARA** [31] focused on semi-automated tools to improve two areas of data risk assessment and monitoring: data provenance, describing the origins, actions performed and agents involved in data creation and transformation; and privacy assessment, minimising the risk of identifiable information in clinical free-text records (for example, GP letters and discharge summaries).

The five driver projects mapped well onto version 1.x of this blueprint but highlighted a missing distinction between “direct query” and “indirect query” in approaches to federated analytics, and a missing synchronisation interface for the pop-up TRE model.

Direct query—the TELEPORT approach—encapsulates everything a remote TRE might need to run the query across its hosted data and return a result. This single pane of glass is seen in a number of current products and is generalised in the polystore database concept.

Indirect query—the TRE-FX approach—uses a job submission model of query where the actual query payload must be retrieved from a software repository outside any of the participating TREs. As noted above, this approach is also used in other models.

TELEPORT’s approach to pop-up TREs relied on a “keep-alive” synchronisation channel between the two participating TREs. This channel provides continual monitoring of the running state of a multi-TRE (and hence multi-governance) project against a “known good”, mutually approved state. Deviations from the approved state, or failure of the keep-alive, can result in researcher access to the pop-up project environment being revoked—or in the entire virtual pop-up TRE being “rapidly deprovisioned”.

While this blueprint is concerned principally with connections *between* TREs, and the SATRE specification [26] is concerned with what it is to be a TRE, the two naturally touch. This blueprint meets the SATRE specification where it should. A detailed mapping between the Federation requirements and SATRE specification statements can be found in the *Master Requirements Table*, Appendix A of Part 3 (Use Cases).

This new version of the federated architecture blueprint models these developments much more accurately than did version 1.



**FOR CONSULTATION & COMMENT**

## 4. A federation blueprint

---

In the rest of this blueprint we describe a UK-wide federation of sensitive data research infrastructure—the SDRI Federation, or simply “the Federation”—built on common standards, with a small number of registry and coordination services, designed to support a wide, rich ecosystem of TREs and other services. The Federation is designed to be trustworthy, with a common set of low-level security protocols and standards for secure data exchange, on top of which is built a rich set of application protocols and standards to support different analytical use-cases—federated analytics, data pooling, federated machine learning or something else. It starts from where we are—an existing ecosystem of largely independent TREs—and builds on the ideas of federation touched on in the 2020 Health Data Research Alliance Green Paper on TREs [6] and expanded in a companion paper from 2021 [7].

The low-level protocols and standards would define, at a purely technical level, what it means to join the Federation—chapter one of its “rulebook”, if you will. Other rules of engagement should, in time, come to supplement the technical—should participants require certain levels of formal accreditation before they can join the Federation, for instance? Development of the Federation rulebook beyond the purely technical is fundamentally a question of governance and we only touch on it here where it has a direct bearing on the technical blueprint. How the Federation should be managed and run are decisions to be taken by the broadest stakeholder community.

The organisation of the Federation could be designed in a number of ways. A key requirement is that the Federation organisation and overseeing authority, any registry services and the low-level data exchange protocols must be designed to ensure that all members of the Federation can trust one another and that, once a Participant has joined, they enjoy the same levels of trust as all other Participants. This is our definition of *trustworthy*. Note that this statement applies to *service Participants* in the Federation, not to researchers or projects or access to sensitive datasets. Governance for approving projects, encapsulating data and researchers in authorised contexts, requires the same rigour in approval and access management as it does today. The organisation of the Federation is a new concept, not a replacement for existing data governance approaches.

### 4.1. Scope

In the following chapters we divide the SDRI Federation into three layers and consider each in turn. Each layer underpins each subsequent one.

1. **Infrastructure.** The lowest level we discuss, infrastructure considers the services and functionality necessary to realise the Federation, rather than network hardware or any particular technology.
2. **Data.** The infrastructure layer can exist perfectly well without data but would be uninteresting. The mechanisms by which data are discovered, linked and made accessible are considered within the data layer.
3. **Organisational.** The highest level considered here, we use “organisation” to refer to oversight of the Federation infrastructure, its operational model and the definition of the “rulebook” for service onboarding, technical standards and change management.

Most of the focus of this blueprint is on the infrastructure layer. Some discussion of data standards and technical governance is essential to set the infrastructure in context, but detailed treatments of these two topics are out of scope of this document.



## FOR CONSULTATION & COMMENT

### 4.2. Design principles

DARE UK's approach to the design and build of a federated network for research with sensitive data follows a number of principles, closely aligned with the SATRE principles.

1. Public trust first, last and always. The strongest design voice should come from the “public persona”. (SATRE: Maintaining public trust.)
2. No TRE, no data. Reinforcing a recommendation from the Goldacre Review [33], require that any and all analysis of sensitive data take place within a TRE, and design accordingly. (SATRE: Maintaining public trust.)
3. Start from where we are. Much of the service ecosystem already exists. Our blueprint must arise through co-design with existing and emerging practitioners.
4. Five Safes are better than one. Secure infrastructure is only one aspect of a TRE. Adopt the Five Safes framework [34] as a guiding principle. Processes and governance are as important as infrastructure, and infrastructure choices should reflect this. (SATRE: Maintaining public trust.)
5. Separation of concerns. Different system actors have very different “security clearances”. Their interactions should be segregated from one another as far as possible.
6. An open-standards-based ecosystem. We seek a rich ecosystem of varied services interoperating through agreed standards. (SATRE: Standardisation.)
7. Be as FAIR as possible. Findability, accessibility, interoperability and reusability are excellent qualities to maintain even in a sensitive data environment [37]. (SATRE: Usability.)
8. The “IETF principle” [38]: rough consensus and running code over rigid specifications and monolithic stacks. Nucleate advances in small groups and grow outwards.
9. Open source first. Seek as often as possible to avoid proprietary lock-in. Strictly, the scope of this principle is that of the networked components defining the federation core. Beyond this core scope, “open standards” (principle 6) is the better arbiter. (SATRE: Standardisation.)
10. Low barriers. Strive to reduce barriers for researchers and for data providers. (SATRE: Usability.)
11. Observability. Human initiated and automated processes resulting in change within the TRE network should be observable. (SATRE: Observability.)

**FOR CONSULTATION & COMMENT**

## 5. Summary

---

That the proposed SDRI Federation architecture shares similarities with past, present and future approaches to connecting data safely and securely with analytical resources is no coincidence. Where trust is paramount the exchange of sensitive information between parties must be done in a controlled environment with a common rulebook agreed by all participants. Registry services are necessary to keep track of which services are currently participating, what their capabilities are, what datasets might be available and so on. Secure data exchange that provides the necessary levels of confidentiality, integrity and traceability is an essential foundation but should not unduly restrict the kinds of application that run on top. The common federation provides a well-managed and safe set of tracks; beyond ensuring that trains don't crash into the wrong stations at the wrong times it has little to say about the rail services on top.

**FOR CONSULTATION & COMMENT**

## 6. References

---

- [1] DARE UK (2023); *UK Sensitive Data Research Infrastructure: A Landscape Review*; Zenodo; <https://doi.org/10.5281/zenodo.10082545>.
- [2] DARE UK; *Initial Phase 1 Recommendations*; <https://dareuk.org.uk/our-work/dare-uk-phase-1-recommendations/> (accessed 01/03/2023).
- [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> (accessed 27/06/2023).
- [4] The HDR UK COALESCE Consortium; *Undervaccination and severe COVID-19 outcomes: meta-analysis of national cohort studies in England, Northern Ireland, Scotland, and Wales*; January 15, 2024; DOI: [https://doi.org/10.1016/S0140-6736\(23\)02467-4](https://doi.org/10.1016/S0140-6736(23)02467-4); The Lancet, volume 403, issue 10426, P554-566, February 10, 2024
- [5] The Royal Society; *Science as an open enterprise*; The Royal Society Science Policy Centre report 02/12; <https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf> (accessed 09/03/2023).
- [6] T. Hubbard, G. Reilly, S. Varma, & D. Seymour (2020); *Trusted Research Environments (TRE) Green Paper* (2.0.); Zenodo; <https://doi.org/10.5281/zenodo.4594704> (accessed 10/05/2023).
- [7] 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> (accessed 10/05/2023).
- [8] T. Hey and A. E. Trefethen; *The UK e-Science Core Programme and the Grid*; Future Generation Computer Systems, Volume 18, Issue 8, 2002; [https://doi.org/10.1016/S0167-739X\(02\)00082-1](https://doi.org/10.1016/S0167-739X(02)00082-1)
- [9] The WLCG Collaboration; *The World-wide LHC Computing Grid*; <https://wlcg.web.cern.ch/> (accessed 09/03/2023).
- [10] The IVOA; *The International Virtual Observatory Alliance*; <https://ivoa.net/> (accessed 09/03/2023).
- [11] ELIXIR; *A distributed infrastructure for life science information*; <https://elixir-europe.org/> (accessed 09/03/2023).
- [12] BBMRI-ERIC; *A European research infrastructure for biobanking*; <https://www.bbmri-eric.eu/> (accessed 09/03/2023).
- [13] ESFRI; *The European Strategic Forum on Research Infrastructures*; <https://www.esfri.eu/> (accessed 09/03/2023).
- [14] CESSDA; *The Consortium of European Social Science Data Archives*; <https://www.CESSDA.eu/> (accessed 09/03/2023).
- [15] ESS-ERIC; *The European Social Survey*; <https://www.europeansocialsurvey.org/> (accessed 09/03/2023).
- [16] NordForsk; *A vision of a Nordic secure digital infrastructure for health data: The Nordic Commons*; ISSN 1504-8640 (2019); <http://norden.diva-portal.org/smash/get/diva2:1376735/FULLTEXT01.pdf> (accessed 10/05/2023).
- [17] NIIS; *X-Road Architecture*; <https://x-road.global/architecture> (accessed 02/03/2023).
- [18] NIIS; *The Nordic Institute for Interoperability Solutions*; <https://www.niis.org/> (accessed 02/03/2023).
- [19] Digital Nations; [https://en.wikipedia.org/wiki/Digital\\_Nations](https://en.wikipedia.org/wiki/Digital_Nations) (accessed 07/08/2024).
- [20] GAIA-X; *A Federated Secure Data Infrastructure*; <https://gaia-x.eu/> (accessed 09/03/2023).
- [21] GAIA-X Technical Committee; *Gaia-X Architecture Document*, v 22.10; 2022; <https://docs.gaia-x.eu/technical-committee/architecture-document/22.10/> (accessed 02/03/2023).

## FOR CONSULTATION & COMMENT

- [22] International Data Spaces Association; *IDS Reference Architecture Model*, v4., April 2022; <https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/> .
- [23] European Commission; *Simpl: cloud-to-edge federations and data spaces made simple*; news article, 24/02/2023; <https://digital-strategy.ec.europa.eu/en/news/simpl-cloud-edge-federations-and-data-spaces-made-simple> (accessed 02/03/2023).
- [24] European Commission; *A European Strategy for data*; policy paper; <https://digital-strategy.ec.europa.eu/en/policies/strategy-data> (accessed 09/03/2023).
- [25] C. Cole, et al; *SATRE: Standardised Architecture for Trusted Research Environments*. Zenodo, Oct. 30, 2023. Doi: 10.5281/zenodo.10055345.
- [26] SATRE: Standard Architecture for Trusted Research Environments, *specification v 1.0.0*, <https://satre-specification.readthedocs.io/en/v1.0.0/index.html>
- [27] T. Giles, et al. *TRE-FX: Delivering a Federated Network of Trusted Research Environments to Enable Safe Data Analytics*. Zenodo, 30 Oct. 2023, doi:10.5281/zenodo.10055354.
- [28] OpenSAFELY; *The OpenSAFELY Secure Analytics Platform*; <https://www.opensafely.org/> (accessed 23/03/2023)
- [29] C. Orton, et al. *TELEPORT: Connecting Researchers to Big Data at Light Speed*. Zenodo, 30 Oct. 2023, doi:10.5281/zenodo.10055358.
- [30] J. Smith, et al. *SACRO: Semi-automated Checking of Research Outputs*. Zenodo, 6 Nov. 2023, doi:10.5281/zenodo.10055365.
- [31] A. Casey, et al. *SARA: Semi-automated Risk Assessment of Data Provenance and Clinical Free-text in Trusted Research Environments*. Zenodo, 30 Oct. 2023, doi:10.5281/zenodo.10055362.
- [32] DARE UK and The PSC, *Scientific use-cases for cross-domain sensitive data research*, March 2024. *In preparation*.
- [33] B. Goldacre et al; *Better, broader, safer: using health data for research and analysis*; 7 April 2022; <https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis> (accessed 02/03/2023).
- [34] F. Ritchie (2016); *Five Safes: designing data access for research*; 10.13140/RG.2.1.3661.1604.
- [35] UK Government; *Data Protection Act 2018*; <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted> (accessed 09/03/2023).
- [36] UK Government; *The UK General Data Protection Regulation*; <https://www.legislation.gov.uk/eur/2016/679/contents> (accessed 09/03/2023).
- [37] M. Wilkinson, M. Dumontier, I. Aalbersberg et al; *The FAIR Guiding Principles for scientific data management and stewardship*; *Sci Data* 3, 160018 (2016); <https://doi.org/10.1038/sdata.2016.18>.
- [38] IETF; *The Internet Engineering Taskforce*; <https://www.ietf.org/> (accessed 20/03/2023).

**FOR CONSULTATION & COMMENT**

## A A comparison of contemporary federated data architectures

Annex III of the *Recommendation Report* for the EU Smart Middleware Platform (SiMPI) [23] compares the concepts defined in the SiMPI architecture with those defined in the GAIA-X framework [21]. The table below extends this idea to include the concepts defined in this document and the equivalents from both the IDSA reference architecture model (version 3.) [22] and the X-Road architecture [17].

DARE UK	GAIA-X	SiMPI	IDSA	X-Road	Notes
Participant	Participant	Organisation that deploys an SMP Agent	Core Participant (also Intermediary)	Organization	
Federation Services	Federator	Data Space governance	Intermediaries, especially Clearing House, Identity Provider and Vocabularly Provider	Central Services & Trust Services	
Security Server	Sovereign Data Exchange	SMP Agent	IDS Connector	Security Server	The GAIA-X mapping is imprecise. It factors out a number of functions that are encapsulated in the other four models.
TRE	Consumer or Service instance	Composite of Application Provider and Infrastructure Provider	Service Provider; Composite of Data Consumer and Data Provider	Service Consumer Information System	A DARE UK TRE has no direct equivalent but is a specialised example of a generic data consuming service.
Data Provider / Data Custodian	Provider	Data Provider	Data Provider	Service Provider Information System	
Researcher (User)	End User	End user	Data User	Data Requestor	
Discovery Service	Catalogue	Data catalogue	Broker Service Provider	Service Provider Information System	A catalogue or discovery service in X-Road would be a specialised kind of Information System hosted by a Service Provider.
Index Service; Software Service	Consumer or Service instance	Composite of Application Provider and Infrastructure Provider	Service Provider	Service Provider Information System	All DARE UK services can be modelled the same way in terms of their interaction with the federation structure.

**FOR CONSULTATION & COMMENT**

## FOR CONSULTATION & COMMENT

## B Scenario analysis of the federated landscape

The 2023 DARE UK survey and review of sensitive data research infrastructure [1] reveals a fragmented and rapidly changing landscape of data and service providers. The changeability is driven in part by a desire to build on the research and data sharing successes of the UK's response to covid-19, but what form the landscape will finally take is hard to predict. A federated network of trusted research environments could look quite different under different future scenarios, depending on a certain number of external policy drivers. In this section we try to explore some possible futures using a “scenario thinking” approach.

Initial thinking pulls up two principal external “landscape drivers”: the number of TREs and their capabilities; and the mobility of data.

1. The number and capabilities of TREs. The Goldacre review [33] argues for a small number of highly capable TREs; the current landscape has a fairly large number of TREs. Some of these are large and capable, supporting national and regional research projects; many more are smaller and support smaller university groups, individual clinical trials and so on. Assuming that there is one overall budget for TRE provision across the UK, larger numbers could mean each has limited capability, and vice versa.
2. Mobility of data. Governance concerns and consequential risk management approaches currently keep data close to home, tightly controlled with a data controller or data custodian. The increasing volumes of certain kinds of data (e.g., medical images, genomic data) also make it increasingly difficult to move them around. To mitigate the first of these concerns UK Government has consulted on possible changes to the Data Protection Act 2018 [35] and the UK GDPR [36], perhaps creating governance counter-pressures towards more mobile data. Note that this doesn't address the “gravity” around very large datasets (see below).

### B.1 Four quadrants

Using these two drivers we can sketch four possible future scenarios in which the DARE UK federation might look slightly different:

- Low numbers of TREs and low data mobility;
- Low numbers of TREs and high data mobility;
- High numbers of TREs and low data mobility;
- High numbers of TREs and high data mobility.

#### B.1.1 Low numbers of TREs and low data mobility

Low data mobility for governance reasons may be relaxed in the future but it's unlikely the same will be true for very large datasets (high-resolution Earth observation, medical imaging, genomic data etc.). Partly because of their size, but also often their complexity, working with datasets of this nature will typically require specialised tooling, high-performance computing capabilities, dedicated GPU or AI hardware, or all of these, and these capabilities typically grow “around” the datasets.

Low mobility for governance reasons leads to a similar scenario where TREs grow “around” the sensitive datasets (this is typically what is meant by “data gravity”). Such a TRE can accumulate expertise in working with the datasets in question, but in this scenario linkage between datasets becomes difficult. If



## FOR CONSULTATION & COMMENT

legal agreements for data linkage are the bottleneck for sharing data, then the incentives on TREs towards technical interoperability are that much weaker: if data move infrequently then current ad hoc methods of data movement may suffice.

### **B.1.2 Low numbers of TREs and high data mobility**

If the gravity of large, complex datasets means a low number of highly capable TREs grow up around them, then these TREs are also available to process smaller, neater, more mobile datasets from elsewhere. If an easing of governance pressures makes these smaller datasets more mobile this could in turn lead to an increase in demand on the small number of TREs. Provided these TREs can build the capacity to manage this increased demand this should not cause any problems.

High mobility of datasets should, in principle, make linkage between them easier. Agreements between data controllers on linkage spines, indexing etc. will be (legally) easier to come to (this almost defines what we mean by “easing of governance pressures” on data mobility) and the necessary data and tools can be sent to linkage teams within the TREs. This would require TREs to acquire additional capabilities in data linkage, and perhaps knowledge of different kinds of data, on top of the expertise they will have built around the datasets they curate themselves.

### **B.1.3 High numbers of TREs and low data mobility**

The volume and complexity argument suggests that a small number of highly capable TREs are likely to exist in all scenarios. But, if moving smaller, neater datasets remains difficult for governance or risk management reasons, this scenario pictures a large number of additional small-scale (even “pop-up”) TREs being created around individual datasets (e.g., a clinical trial dataset) or for individual research organisations (e.g., a university or university department). In this scenario linkage remains difficult and the data landscape is even more fragmented than in the low-low scenario. If data sharing is difficult for governance reasons then there are few incentives for these TREs to maintain any level of technical interoperability or adhere strictly to any particular standard if doing so might constrain the TRE’s core research purpose. The risk of technical drift between TRE environments is high with a consequent dissipation of expertise and increased friction<sup>7</sup>.

High numbers of TREs in a landscape of low data mobility is probably a scenario to be avoided if possible.

### **B.1.4 High numbers of TREs and high data mobility**

High numbers of TREs in a scenario of high data mobility is a very different prospect to the high-low picture. In this scenario, the relative ease of data sharing provides a real incentive for small-scale TREs to stick to interoperability standards—play the game and data linkage becomes much easier for your researchers. While the big, highly capable TREs are ever-present this scenario envisages a true ecosystem of TREs of many scales being able to exchange data relatively freely. Open standards are a key enabler for

---

<sup>7</sup> Imagine an extreme version of this scenario where hundreds of research groups end up with their own TREs, each of which has been built around the groups’ “traditional” bespoke analytics environments and domain-specific languages. The blockers to research are never technical interoperability with the neighbouring lab’s TRE and are always the slow and painful negotiation over data sharing – so why spend time on technical interoperability when you need to invest more in data negotiation?



## FOR CONSULTATION & COMMENT

this scenario, along with open software recipes to enable many groups to create their own readily interoperable TREs.

The biggest challenge in this scenario is governance, closely followed by a set of technical controls that span the whole ecosystem and maintain the necessary security posture across multiple organisations, data controllers and researchers.

### B.2 Observations

None of these scenarios expects to see a complete de-fragmentation of the distributed landscape. While some consolidation is desirable (e.g., to avoid the high-low scenario) it seems optimistic to expect a reduction in the numbers of centres of data gravity to one over the next 5-10 years. Thus we should expect that the federation of distributed data sets and computational services to remain a key challenge within the UK research landscape. This observation underlines the architectural approaches described in the blueprint.