DARE UK

Federated Architecture Blueprint - Part 3 (Use Cases)

DARE UK Delivery Team

Version 2.2 final

November 2024

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

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.

Contents

About document versions

This document is Part 3 (Use Cases) 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.

1. Introduction

By some measures the UK already has a research landscape for sensitive data that is federated. Data are distributed and distant from researchers, services are available to link datasets together and trusted research environments exist to bring all these things together. Federation is ad hoc, though, friction is high and end-to-end researcher productivity can be painfully low.

The SDRI Federation is not so much a new thing as the improvement of an existing thing. Our goal is to remove the ad hoc, reduce the friction and increase the baseline trustworthiness of connections between data providers, TREs and researchers. From a researcher's perspective the ideal SDRI Federation is something that they will never actually see; rather they will see its positive impact on their productivity.

With this view in mind, many of the important drivers of the Federation are non-functional rather than functional. They are about increasing trust and improving performance rather than adding new features per se. We advance the argument that a secure federation with an agreed rulebook and matching organisational model creates an environment which supports innovation, providing a common, trustworthy foundation which enables the development of new services and enhanced capabilities while maintaining the integrity and confidentiality of the whole.

2. Rachel's journey: 2022

Where do we start from?

Rachel is a researcher. As an illustration of the different roles and processes that are currently undertaken in setting up a research project with sensitive data, here is an account of her journey from an idea to the start of a project built around that idea. The time is late 2022, the setting our current sensitive data research landscape. We have a small cast of characters:

- Rachel, a researcher;
- Gill, an information governance professional in charge of a TRE;
- Iain, who provides an indexing service;
- Pawel, Peter and Preethi, three data providers.

We follow Rachel's journey below and make observations as we go.

Rachel has a research question she'd like to explore: "understanding environmental health impacts on educational achievement". She realises she'll need to bring together different kinds of data to answer this.

How does Rachel figure out what data she needs? Where does she look? How does she know whether the data she needs are stored as one, two or many datasets?

Rachel has identified three datasets she needs:

- Education data, already collected by Preethi for the whole population and available for research in a TRE run by Gill.
- Environmental data on air quality, groundwater quality—in fact loads of interesting variables—covering the whole country, collected by Pawel and all openly available for research.
- Health outcomes, collected by Peter and available for research but only for particular cohorts. Rachel will have to ask explicitly for what she needs.

- Education data use a special index based on name, address and data of birth.
- Environmental data are indexed by location, typically latitude/longitude, and a shape that defines the area they cover.
- Health outcomes data are indexed by NHS number (NHS#).

Rachel understands she'll need to conduct her research in a TRE. Seeing that at least one of her datasets of interest is already available in a TRE, she contacts Gill.

Rachel knows who to ask but would another researcher know where to go next?

Gill works with Rachel to define the project, including identifying how disclosure control of project results will need to be managed, given the different risk appetites of the data providers involved.

Managing disclosure risk is a really important topic to get right, right at the start of a project.

Gill liaises with the three data providers, Preethi, Peter and Pawel. Peter's health outcomes data is the biggest constraint; Peter can only release a specified cohort set for research so defining the cohort is key. Gill, Rachel and Peter work up a cohort definition for the project.

Cohort definition is manual and iterative here; is there any technical way to speed it up or smooth it out?

Rachel and Gill have agreed a definition for the project:

"Project" is a key concept. It ties together the researchers, the datasets they need and the

- Peter has approved the cohort of health outcomes data, indexed at individual level by NHS number (NHS#).
- Preethi has approved access to the education data already within the TRE, already indexed at individual level with a unique "education data index".
- Pawel is happy to provide access to the environmental datasets for the areas inhabited by Rachel's cohort. Pawel's data can be indexed by latitude/longitude or equivalent geospatial coordinates.

approvals they have, for a certain period of time and for a specific purpose.

Gill now orchestrates data assembly for Rachel's project within the TRE. Indexing the three datasets so they can be linked is key and she works with Iain, her trusted third-party indexer.

Here we assume that one indexer has "lookup tables" for all the key private data.

Gill sends the set of NHS#s to Iain. Using the registers that he looks after Iain creates four lookup tables for the project:

- A set of "education data index" numbers mapped to a set of unique but meaningless numbers called "ID1".
- A set of latitude/longitude pairs mapped to a set of unique but meaningless numbers called "ID2".
- The original set of NHS numbers mapped to a set of unique but meaningless numbers called "ID3".
- A "master index" mapping ID1, ID2 and ID3 to a set of numbers unique to Rachel's project called "IDR".

This approach is creating project-specific identifiers, which is good practice.

Some indirect mapping is required:

- NHS# maps to name, address and date of birth which map to education index (Iain knows how because he created the education index in the first place!).
- NHS# maps to an address which maps to a unique property reference number (UPRN) which maps to a lat/long pair.

Iain sends the ID1 and education index mapping to Preethi.

Iain sends the ID2 and latitude/longitude mapping to Pawel.

Iain sends the ID3 and NHS# mapping to Peter.

Iain sends the "master index" straight to Gill at the TRE. He uses an existing secure file transfer channel between his organisation and Gill's TRE.

These identifiers are not particularly sensitive of themselves but nevertheless sending documents between different parties needs to be done securely.

Pawel prepares the environmental data using the set of lat/long pairs, but he replaces lat/long with ID2 in Rachel's version of the dataset.

Pawel sends this dataset to Gill, marked "for Rachel's project". The dataset isn't particularly sensitive so he emails it to Gill as an encrypted zipfile.

Peter prepares the health outcomes data extract using the set of NHS#s, but he replaces NHS# (and any other personally identifying attributes) with ID3 in Rachel's version of the dataset.

Sending datasets between different parties definitely needs to be done securely.

Currently there are many different methods employed: managed file transfer of various kinds, secure email, occasional physical device transfer (a disk-drive passed literally form hand to hand).

Peter sends this dataset to Gill, marked "for Rachel's project". He does this using a managed file transfer service which is very secure but requires a bit of manual finessing at both ends.	
Preethi chooses to prepare the education data as an extract using the set of education data indexes. She removes all the personally identifiable attributes and replaces each education data index number with ID1 in Rachel's version of the dataset.	Preethi and Gill could choose to allow Rachel access to the full education dataset and give her a lookup table matching education data indexes to the set of "IDR" indexes.
Preethi passes this dataset to Gill (all within the TRE).	
Gill uses the three datasets and the "master index" from Iain to zip everything together into Rachel's final, approved linked dataset.	The only index number left in the linked dataset is the "IDR" which is unique to Rachel's project (and doesn't mean anything to anyone else).
Rachel gets access to her approved linked data inside the TRE, and she's off!	Finally!

Figure 1 on the next page illustrates the above narrative as a sequence diagram, showing interactions and data and metadata movement between the actors. Metadata objects are rectangular and datasets are "document shaped". Time runs from top to bottom.

Rachel's research journey, while synthetic, is rooted very much in current "data pooling" practice of sensitive data research in the UK. It helps us tease out the key drivers for the SDRI Federation, and in doing this we take two perspectives. The first perspective comes from potential users of the federation, from researchers like Rachel to system operators and data custodians. The second comes from the existing landscape of services across the UK [1] and how they currently interact with each other—Gill's TRE and Iain's indexing service, for example. In both cases we have distilled community interactions, desk research and expert knowledge into a series of user personas on the one hand and data usage patterns on the other. We use these two perspectives to identify the key requirements for the SDRI Federation.

FOR CONSULTATION & COMMENT

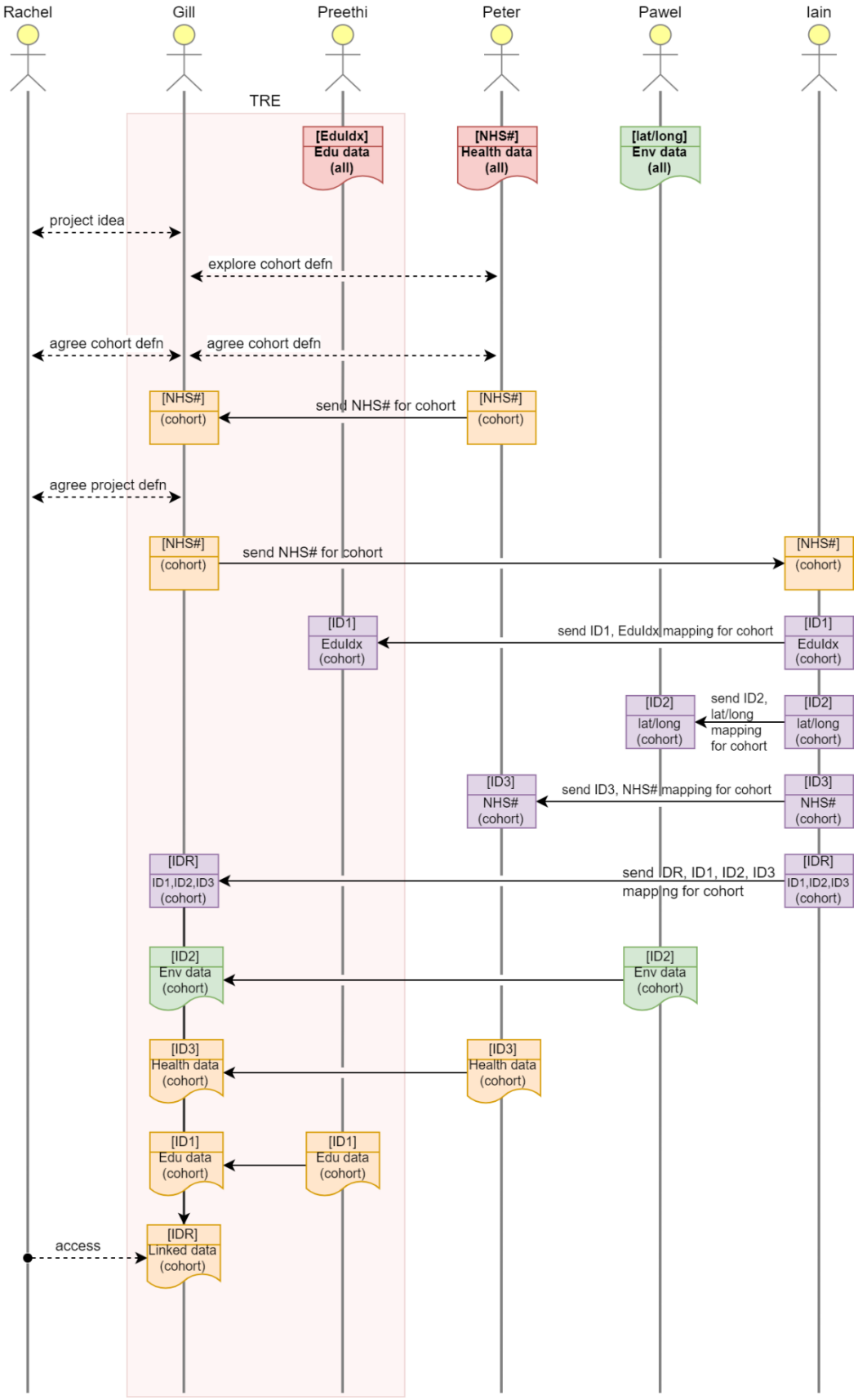

Figure 1. Rachel's Journey as a sequence diagram.

3. User personas

DARE UK has worked with relevant community groups across the UK to develop user personas to represent classes of users. Personas give voice and motivation to the abstract "actors" used later in our system architecture and consequently are a better source of genuine use-cases. In particular, a persona's needs and motivations can be a better tool to identify non-functional requirements (how safely? how quickly?) than abstract system roles.

Table 1 summarises DARE UK's user personas. Phase 1a focussed on developing data supplier and data consumer personas. Phase 1b filled out the personas for the service provider roles.

Often it is easy to associate a particular persona with a single type of actor; sometimes it is not. Some personas may act in multiple ways, particularly within the service provider group: a persona representing someone running a TRE service that also hosts important datasets will, at different times, act as both TRE operator and data provider.

Table 1. DARE UK User Personas and their principal features. For the definitions in the "Actor" column, see the following section.

Persona	Key Motivation	Key Concern	Actor
Grace Opedemi, member of the public	I want to understand how best use is being made of public sector research investments.	Keeping my data safe from unauthorised, unethical or other "bad" uses.	Public
Peter Shaw, data custodian	I want to share and link my data with others.	Safety! (Don't break the law!) Poor data quality (terminology, linkage)	Data Custodian
Pritesh Navdra, techie data scientist	I want to keep on the leading edge of data science, while doing some good!	Poor data quality (terminology, linkage); poor, "old" tooling.	Researcher
Rachel Wakefield, researcher entering socio-economic research	I want to create more impactful research through greater access to linked data.	Ease of access to restricted data (skills, quality, linkage).	Researcher
Sarah Greenshaw, university public health research PI	I want to grow the research power and outward recognition of my group.	Competition from elsewhere, being left behind.	Researcher
Jeremy Foster, ed-tech business product manager	I want to generate ROI through accessing and sharing sensitive data.	Ease of access to restricted data (skills, quality).	Researcher
Gill King, information governance professional	I want to be seen as empowering research instead of as a barrier to it.	Lack of standardisation, lack of automation, inefficient processes.	Information Governance
Helen Chow, TRE service owner	I want to be able to demonstrate the value of my TRE.	Sustainability! Maintaining multiple accreditations; implementing change is hard.	TRE Operator
Colin Iwobi, TRE admin and operator	I want to improve the user experience for our TRE users.	Supporting new software tools; slow safety approval process; interacting with frustrated users.	TRE Operator
Roy Bose, Federation operator	I want to support new and emerging analytical use-cases across the network.	Building & maintaining trust; keeping it simple & sustainable; making research more transparent.	Federation Operator

3.1. Federation actors and roles

We can group the different "actors" in the last column of the table into three groups: Data Providers, Data Consumers and Service Providers, the latter providing services that connect the former two.

Most of these roles already exist in practice, except for Federation Operator, which, *de facto*, is new.

3.1.1. Data Providers

Actors and roles in this group include:

- members of the Public, as ultimate providers of their data for research in the public benefit;
- Data Controllers, responsible for guarding access to public data, complying with data protection law and ethical guidance, and accountable to the public for the uses of their data;
- Data Custodians act as intermediaries between Data Controllers and Researchers. Data Custodians are the ones who provide sensitive data for research projects.

3.1.2. Data Consumers

Actors and roles in this group include:

- academic Researchers, looking for access to sensitive data to address particular research questions. Their requirements may be for linked datasets, or large datasets, or they may need significant computational analysis power or sophisticated software to carry out their research;
- commercial Researchers, looking for access to sensitive data to develop or test new products or services. Commercial researchers have different motivations to academic researchers but in terms of their interaction with the SDRI Federation we can treat them as Researchers.

3.1.3. Service Providers

Actors and roles in this group are more diverse than the other two and include the following:

- Information Governance (IG) professionals act as intermediaries between Data Providers and Data Consumers, ensuring all necessary ethical, data protection and legal approvals are in place for a research project to proceed. They also act as brokers between these two groups and the TRE and other technical service operators;
- Data Managers are responsible for providing the technical means to disseminate datasets approved by data controllers for release to IG for onwards sharing to data consumers. They are accountable to their data controllers (or data custodians) for the security and integrity of these technical dissemination mechanisms. In practical terms, data managers usually operate within TREs to provide research-ready data;
- Indexers and Linkers provide services to join different datasets together, particularly individual-level datasets that need to be joined using individual-level keys. These roles may be a subset of IG; certainly they are accountable to IG and to data controllers;
- TRE Operators are responsible for the running of a given TRE under its particular IG regime. This responsibility extends to all security controls required by IG;
- Federation Operators are responsible for running the technical services that connect TREs and data services together to form the federation. This responsibility extends to all the security controls required by the overall federation IG.

3.2. Other stakeholders

There are a small number of roles who don't interact directly with the federation but have a stake in its outcomes, including:

- Funders (F), responsible for seeing overall return on investment in the federation infrastructure.

4. User stories and requirements mapping

Analysis of both data usage patterns and user personas has identified a number of key requirements for the overall federation and for individual services within it [2]. Some of these requirements are functional use-cases, more are non-functional constraints, and many are higher level "user stories" to be followed up in later development stages.

Note that these requirements are by no means exhaustive. Nor are they detailed enough to begin software implementation. They are, however, sufficient to provide a framing set for this architecture. The "master list" of requirements can be found in Appendix A.

Below we list the user stories derived from our personas, ordered by the "most popular". We have incorporated a number of genuine, current cross-domain research projects as motivations for our "Researcher" personas. These were generated in a workshop held in February 2024 which brought together more than 50 UK based researchers and public participants to surface use cases for linking sensitive data [4]. These examples provide useful first insight into the spread of data types in use in research projects today, and are recorded from use-case U42 onwards. We have assigned them as closely as possible to one of our four researcher personas, and we note the types of data required by each.

Further requirements, derived from system-level considerations and an ongoing review of the existing landscape, appear throughout this document and are flagged as they arise.

UId	Requirement	Personas	Labels
U01	I want to help achieve the greater good and make an impact.	PN, GO, RW, JF, RB	Usability
U02	I want credit for the research I help create.	SG, GK, RW, DS	TRE
U03	I want to share data with others easily and securely.	JF, DS, GO, RB	Data quality, Security, TRE, Transparency
U04	I am frustrated by poor data quality.	RW, PN, JF	
U05	I want to understand how datasets vary semantically between providers.	RW, DS, PN	
U06	I find mapping legal regulations to TRE policy challenging.	GK, DS, CI	Usability
U07	I find data interoperability a big challenge.	DS, PN, RW	AI/ML, TRE, HPC
U08	I find it difficult to access the data I need.	PN, RW, JF	Transparency
U09	I want to keep my data safe!	GO, DS, JF	Usage costs, Transparency
U10	I want my engagements with stakeholders to be smoother than they are!	CI, GK	Federation Services
U11	I want more automated processes and tools.	GK, CI	
U12	I want more standardisation.	GK, HC	
U13	I find it challenging to access and build relevant collaborations.	JF, SG	TRE
U14	I am missing technical and data science skills.	RW, JF	
U15	I find implementing change is difficult.	HC, CI	Usability

U16	I want to be able to use/support users with the latest tools & software. I want to support new and emerging analytical use-cases across the network.	CI, PN, RB	
U17	I don't understand a lot of the jargon, or the policy and regulations.	GO, DS	
U18	I want to ensure the public purse is yielding good value for money.	HC, GO	Usability
U19	I find tracking projects from start to finish is opaque and difficult.	GK	TRE
U20	I want to be seen as empowering research instead of as a barrier to it.	GK	Data quality, Security, TRE, Transparency
U21	I want to generate business value through data.	JF	
U22	I want to grow opportunities for my organisation.	SG	
U23	I want to be able to retain talent in my centre.	SG	Usability
U24	I want to speed up my workflow.	RW	AI/ML, TRE, HPC
U25	I find visualising large quantities of disparate data challenging.	PN	Transparency
U26	I stress about earning considerably lower income in the public sector.	PN	Usage costs, Transparency
U27	I want to be able to demonstrate the value of the TRE.	HC	Usage costs, TRE, Transparency
U28	I find sustainability is a real challenge!	HC	Usage costs, TRE
U29	I find it challenging to gain and maintain accreditations across multiple schemes.	HC	TRE
U30	I find the lack of consistency in documentation and data frustrating.	HC	Data quality, Metadata, Usability
U31	I want to improve the user experience for our TRE users.	CI	Usability
U32	I'm frustrated that many of our users seem to have a poor UX.	CI	Usability, TRE
U33	I would like to see a practical-based "TRE driving licence" for users!	GK	TRE
U34	I want to discover data easily.	PN	Metadata, Data Provider, Data Discovery Service
U35	I worry about the costs of accessing lots of data.	PN	Metadata, Usage costs, TRE
U36	I worry about anonymising sensitive data "well enough".	DS	Security, Data Provider
U37	I don't know about, or how to find, relevant data about me.	GO	Metadata, Data Discovery Service, Transparency
U38	I worry about introducing single points of failure into the TRE network.	RB	Federation Services, Reliability
U39	I worry about adding more complexity to TRE operations!	RB	Usability, TRE
U40	I worry about software or platform vendor lock-in.	RB	Federation Services, Reliability, TRE, Data Provider

U41	I want to find ways to make research with sensitive data more transparent.	RB	Metadata, Federation Services, Transparency
U42	I want to understand how best to reduce bottlenecks within NHS service provision	SG	Data: Health data, Social data, Economic data, Longitudinal
U43	I want to understand how to improve children's health, education, and economic prospects through family level analysis and intervention	SG	Data: Health data, Social data, Economic data, Lifestyle data, Consumer data, Environmental data
U44	I want to identify areas of unmet need within the NHS and optimising resource allocation to meet needs.	SG	Data: Health data, Social data
U45	I want to understand the impact of transport systems and low emission zones on population health to design locally optimal interventions.	PN	Data: Health data, Social data, Lifestyle data, Environmental data, Geographic data
U46	I want to understand what is grown, transported, eaten, and wasted to deliver routes to action.	PN	Data: Health data, Economic data, Consumer data, Commercial data, Environmental data
U47	I want to train AI models to detect colon cancer more effectively using colonoscopic data.	SG, PN	Data: Health data, including imaging data
U48	I want to understand the interaction of work status on mental health and vice versa.	RW	Data: Health data, Economic data
U49	I want to understand what diets are both healthy and environmentally sustainable, not improving one at the expense of the other, and developing policies that encourage the population to close the gap between the current diet and the desired diet	RW	Data: Social data, Lifestyle data, Consumer data, Commercial data, Environmental data
U50	I want to understand the root causes of long-term unemployment to enable preventative and pro-active policy interventions.	RW	Data: Social care data, Social data, Economic data
U51	I want to understand the root causes of child and adolescent mental health challenges to identify high-risk groups and develop system-wide interventions.	RW	Data: Social care data, Health data, Social data, Justice data
U52	I want to understand the factors that influence vaccine uptake to improve public messaging and future pandemic policy.	SG	Data: Health data, Social data, Geographic data
U53	I want to quantify the impact of domestic violence, interventions, and policies.	RW	Data: Social care data, Health data, Social data, Justice data, Economic data
U54	I want to understand how to encourage economic growth (outcomes incomplete)	RW	Data: Health data, Social data, Economic data, Lifestyle data, Consumer data
U55	I want to understand the effect of housing on health outcomes	SG	Data: Health data, Social data
U56	I want to link education, employment, and mental health data to improve mental health and educational attainment	RW	Data: Health data, Social data, Economic data
U57	I want to change behaviours and challenging assumptions around energy use	PN	Data: Social data, Economic data, Environmental data

U58	I want to understand the factors affecting productivity	RW	Data: Social data, Economic data
U59	I want to understand the triggers for poor mental health and identifying at risk groups	RW	Data: Health data, Economic data, Lifestyle data, Environmental data, Geographic data
U60	I want to improve contraception to meet women's needs	SG	Data: Health data
U61	I want to understand the root causes of long-term ill health and developing more holistic solutions	RW	Data: Health data, Economic data, Lifestyle data, Consumer data
U62	I want to ensure the correct care gets to where it is needed and enable better self-management of conditions	SG	Data: Health data, Social data
U63	I want to understand causes of demand, improving access, and uptake for mental health services	SG	Data: Health data, Social data, Economic data, Geographic data
U64	I want to understand the effect of inequality in respite care on mental health	SG	Data: Social data, Health data
U65	I want to improve the speed and access to health services, and enhancing integrated care	RW	Data: Social care data, Health data, Social data, Economic data
U66	I want to understand the impact of different modes of transport on health	PN, SG	Data: Health data, Lifestyle data, Geographic data
U67	I want to understand online relationships and their influences on individuals and society	RW	Data: Health data, Lifestyle data
U68	I want to identify areas of discrimination and assessing how discrimination impacts quality of life	RW	Data: Social data, Economic data
U69	I want to understand reality vs what is reported to optimise health interventions	SG	Data: Health data
U70	I want to help graduates explore new jobs and improve search elements based upon skills and experience	RW	Data: Social data, Economic data
U71	I want to achieve economies of scope and scale for local councils	RW	Data: Social care data, Lifestyle data
U72	I want to improve monitoring and intervention of health conditions at home for the elderly	SG	Data: Health data, Economic data
U73	I want to provide comprehensive, tailored support for social and MH challenges via a single point of support	RW	Data: Health data, Justice data, Social data
U74	I want to understand relationship between food, education and health outcomes	RW	Data: Health data, Social data, Lifestyle data, Consumer data
U75	I want to understand the impact that energy and climate has on personal finances and health outcomes	PN, SG	Data: Health data, Economic data, Environmental data
U76	I want to understand the causes of obesity to help make more targeted interventions	SG, PN	Data: Health data, Economic data, Lifestyle data, Consumer data, Geographic data
U77	I want to understand of the short- and long-term impacts of vaping	SG, PN	Data: Health data, Consumer data, Geographic data

U78	I want to understand how children's environments affect health and education outcomes	SG, RW	Data: Social care data, Health data, Social data, Economic data
U79	I want to understand impact of screen time on children's short- and long-term outcomes, particularly mental health and social effects	RW, SG	Data: Health data, Social data, Economic data, Lifestyle data, Consumer data, Geographic data
U80	I want to understand affordability of Net Zero 2050 requirements and developing the optimal incentives to achieve this	RW, PN	Data: Social data, Economic data, Commercial data, Environmental data, Geographic data
U81	I want to understand businesses resilience to climate change, boosting productivity, and predicting the success/failures of businesses	PN, RW	Data: Commercial data, Environmental data
U82	I want to understand impact of income volatility on mental health	SG	Data: Health data, Economic data
U83	I want to understand the impact of low and ultra-low emission zones on different groups	SG, PN, RW	Data: Health data, Social data, Economic data, Lifestyle data, Environmental data, Geographic data
U84	I want to develop early interventions for mental health challenges	SG	Data: Health data, Social data, Justice data
U85	I want to understand the impacts of ASD and ADHD diagnoses on education outcomes and enhancing screening and support	SG	Data: Health data, Social data
U86	I want to understand the factors that contribute to inequalities in lung cancer care and impact access to care	SG	Data: Health data, Social data, Economic data, Lifestyle data
U87	I want to predict disease before any symptoms present and develop new treatments based on new biomarkers	SG	Data: Health data, Social data, Economic data, Lifestyle data, Consumer data
U88	I want to understand the consequences of court decisions	RW	Data: Health data, Justice data, Economic data
U89	I want to identify high risk groups for respiratory disease	SG, PN	Data: Health data, Social data, Economic data, Environmental data, Geographic data
U90	I want to understand how shopping habits affects health	RW	Data: Health data, Economic data, Lifestyle data, Consumer data
U91	I want to understand how the ability to travel to green space impacts upon mental health	RW, SG	Data: Health data, Social data, Lifestyle data
U92	I want to understand access to higher education for different groups and the factors that can enable or limit access	RW	Data: Health data, Social data, Justice data, Economic data
U93	I want to understand the factors that most influence cardo-vascular disease and developing targeted interventions for high-risk groups	SG, PN	Data: Health data, Social data, Economic data, Lifestyle data, Consumer data

5. Future work

Requirements analysis is an ongoing business in modern system design and build. These high-level personas and their stories and motivations ground the DARE UK programme nicely and, even at this remove, highlight some key needs of – and constraints on – the development of the Federation. Nevertheless, further detailed analysis will be required to break down the user stories into more digestible – and testable – requirements.

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] C. Cole, et al; *SATRE: Standardised Architecture for Trusted Research Environments*. Zenodo, Oct. 30, 2023. Doi: 10.5281/zenodo.10055345.

[4] DARE UK and The PSC, *Scientific use-cases for cross-domain sensitive data research*, March 2024. *In preparation*.

[5] S. Bradner, B. Leiba; *BCP14*; The Internet Engineering Taskforce Best Current Practice; https://www.ietf.org/rfc/bcp/bcp14.html (accessed 01/12/2023).

A Master requirements table

In the table below we follow the conventions of IETF BCP 14 (RFC2119 & RFC8174) [5], vis:

> The capitalised key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in BCP 14.

From each requirement we identify the need for one or more service capabilities or information contexts, or both, noted in the primary, secondary or tertiary “scope” columns. “F/N” codes for functional or non-functional requirements; “St” indicates the strength of the requirement using the above MoSCoW abbreviations. The second column indicates which user story raises this requirement. Some requirements arise in multiple stories and have multiple entries in this table.

In the final column we cross-reference each requirement with relevant statements from version 1.0.0 of the SATRE specification for standard TRE architectures[1]. A key output of the SATRE project [3] this specification “aims to standardise the capabilities of TREs, making it easier for users, operators, and developers to work with sensitive data, and making the operation of TREs more transparent to data owners and the general public”.

Many user stories as expressed at the moment are high level or require further analysis and have not yet been tagged with system level requirements. They are not presented in this table.

[1] See https://satre-specification.readthedocs.io/en/v1.0.0/index.html for an online version of v1.0.0 of the SATRE specification.

RId	UId	Requirement	F/N	St	1ry Scope	2ry Scope	3ry Scope	SATRE v1.0.0
R001	U01	The Federation MUST demonstrate impact on research	N	M	Core: Federation			2.2.14
R002	U01	The Federation MUST communicate clearly and publicly on key concepts	N	M	Core: Federation: Registry	Service: Discovery		4.8.1; 4.8.2
R003	U09	The Federation MUST ensure the confidentiality of data storage	N	M	TRE: SDZ			2.1.1; 2.1.3; 2.5.12; 2.5.16; 3.1.1
R004	U09	The Federation MUST ensure the confidentiality of data exchange	N	M	Interface: Data Egress	Interface: Data Ingress		2.5.13; 2.5.16; 3.1.1
R005	U07	The Federation MUST enable linkage between syntactically similar data	F	M	Service: Index	Interface: Index		
R006	U07	The Federation MUST enable linkage between syntactically dissimilar data	F	M	Service: Index	Interface: Index		
R007		- retired -						
R008	U08	The Federation MUST reduce the barriers to data access	N	M	Core: Federation			2.1.4; 2.1.5
R009	U09	The Federation MUST ensure the integrity of data exchange	N	M	Interface: Data Egress	Interface: Data Ingress		2.5.16; 3.1.1
R010	U09	TREs MUST ensure the security of data access and use	N	M	TRE: RAZ	TRE: QMZ		2.1.1; 2.1.8; 2.5.12; 2.5.16; 3.1.1
R011	U09	The Federation MUST demonstrate the security of data exchange practices	N	M	Core: Federation			2.5.13; 2.5.15
R012	U09	The Federation MUST demonstrate the security of data storage practices	N	M	Core: Federation			2.5.13; 2.5.15

R013	U09	The Federation MUST demonstrate the security of data access and use practices	N	M	Core: Federation			2.5.13; 2.5.15
R014	U02	The Federation MUST ensure research use is appropriately recorded in metadata records	F	M	Core: Federation: Registry			2.2.15; 4.8.2
R014	U19	As above						
R014	U41	As above						
R015	U14	- retired -						
R016	U13	-retired -						
R017	U41	The Federation MUST provide clear public signposts to data used	F	M	Core: Federation: Registry	Service: Discovery		3.4.1; 3.7.1
R018	U36	Data Custodians SHOULD provide tooling for pseudonymising data	F	S	Role: Data Custodian	TRE: SDZ		
R019	U36	Data Custodians SHOULD provide tooling for assessing data anonymity	F	S	Role: Data Custodian	TRE: SDZ		
R020	U36	The Federation MUST ensure data controllers are appropriately recorded in metadata records	F	M	Core: Federation: Registry			2.2.14
R021		- retired -						
R022		- retired -						
R023	U16	The Federation SHOULD enable discovery of and access to modern data science computational capabilities	F	S	Core: Federation: Registry			
R024	U34	The Federation MUST facilitate data discovery across the network	F	M	Service: Discovery	Core: Federation: Registry		3.4.1; 3.6.1; 3.7.1
R024	U37	As above						
R025	U27	As above						4.4.4
R025	U28	As above						4.4.3

R025	U35	TREs SHOULD provide metadata on access charges and running costs	F	S	TRE	Core: Federation: Registry	0	2.2.16; 2.3.1; 2.3.4
R026		- retired -						
R027		- retired -						
R028		- retired -						
R029		- retired -						
R030		- retired -						
R031		- retired -						
R032	U03	Query (direct) interface services MUST connect externally to Query (direct) interface services	F	M	Interface: Query (direct)			2.2.9
R033	U03	Query (indirect) interface services MUST connect externally to Query (indirect) interface services	F	M	Interface: Query (indirect)			2.2.9
R034	U03	Response interface services MUST connect externally to Response interface services	F	M	Interface: Response			2.2.9
R035	U03	- retired -						
R036	U03	Data Egress interface services MUST connect externally to Data Ingress interface services	F	M	Interface: Data Egress	Interface: Data Ingress		2.2.9; 3.1.4; 3.1.5; 3.1.12
R037	U03	Data Ingress interface services MUST connect externally to Data Egress interface services	F	M	Interface: Data Ingress	Interface: Data Egress		2.2.9; 3.1.4; 3.1.5; 3.1.12
R038	U03	System actors in the role of Data Manager SHALL be authorised to invoke Data Ingress/Egress interface services	N	M	Interface: Data Egress	Interface: Data Ingress		2.2.11; 3.1.6; 3.1.12
R039	U03	System actors not in the role of Data Manager SHALL NOT be authorised to	N	M	Interface: Data Egress	Interface: Data Ingress		2.2.11; 3.1.6; 3.1.12

		invoke Data Ingress/Egress interface services						
R040	U07	Index interface services MUST connect only to Index interface services	F	M	Interface: Index			2.2.9
R041	U11	System actors in the role of Data Manager SHALL be authorised to invoke Index interface services	N	M	Interface: Index	Service: Index		2.2.11
R042	U11	System actors not in the role of Data Manager SHALL NOT be authorised to invoke Index interface services	N	M	Interface: Index	Service: Index		2.2.11
R043	U11	Software interface types MUST connect only to Software interface types	F	M	Interface: Software			2.1.9; 2.2.9
R044	U11	System actors in the role of TRE Operator SHALL be authorised to invoke Software interface services	N	M	Interface: Software			2.1.9; 2.2.11
R045	U11	System actors not in the role of TRE Operator SHALL NOT be authorised to invoke Software interface services	N	M	Interface: Software			2.1.9; 2.2.11
R046	U16	The Federation MUST support a "federated analytics" analysis pattern	F	M	TRE: QMZ			
R047	U16	The Federation MUST support a "linked-data pooling" analysis pattern	F	M	TRE: SDZ	Service: Index		
R048	U16	- retired -						
R049	U16	System actors in the role of Output Approver SHALL be authorised to egress data objects from the TRE SDZ to the outside world	N	M	TRE: SDZ			2.1.1; 3.3.4

R050	U16	System actors not in the role of Output Approver SHALL NOT be authorised to egress data objects from the TRE SDZ to the outside world	N	M	TRE: SDZ			2.1.1; 3.3.4
R051	U40	Federation services MUST be interoperable with existing deployed service endpoints	F	M	Core: Federation			
R052	U40	Federation Service endpoints in TREs and other federated services (Security Servers) MUST be deployable on all existing TRE infrastructure platforms.	F	M	Core: Security Server			
R053	U39	Federation Service endpoints in TREs and other federated services (Security Servers) SHOULD be as encapsulated as possible.	F	S	Core: Security Server			
R054	U03	All data exchange between Federation participants MUST be encrypted.	N	M	Interface: Data Egress	Interface: Data Ingress		2.5.13
R055	U03	All query exchange between Federation participants MUST be encrypted.	N	M	Interface: Query (direct)	Interface: Query (indirect)		2.5.13
R056	U03	All query results exchange between Federation participants MUST be encrypted.	N	M	Interface: Response			2.5.13
R057	U03	All index data exchange between Federation participants MUST be encrypted.	N	M	Interface: Index			2.5.13
R058	U16	The Federation SHOULD support the "indirect query federated analytics" analysis pattern	F	S	TRE: QMZ	TRE: RAZ	Service: Job Submission	
R059	U16	The Federation SHOULD support the "direct query federated analytics" analysis pattern	F	S	TRE: RAZ	TRE: QMZ		

R060	U11	Sync interface types MUST connect externally to Sync interface types	F	M	Interface: Sync			2.1.9; 2.2.9
R061	U11	System actors in the role of TRE Operator SHALL be authorised to invoke Sync interface services	N	M	Interface: Sync			2.1.9; 2.2.11
R062	U11	System actors not in the role of TRE Operator SHALL NOT be authorised to invoke Sync interface services	N	M	Interface: Sync			2.1.9; 2.2.11
R063	U08	An RAZ MUST have one or more Project Environments	F	M	TRE: RAZ	Collaboration: Project Environment		
R064	U08	Project Environments MUST be suitable for the kinds of research the TRE supports	N	M	Collaboration: Project Environment	TRE: RAZ		2.1.2; 2.1.10
R065	U11	Project Environments SHOULD be configured (and configurable) in standard and repeatable ways	F	S	Collaboration: Project Environment	TRE: RAZ		2.1.2; 2.4.1; 2.4.2
R066	U11	Project Config services SHALL be used to configure Project Environments	F	M	Service: Project Config	Collaboration: Project Environment	TRE: RAZ	2.2.1; 2.2.2; 2.4.1; 2.4.2
R067	U11	Project Config services MAY connect to approved external repositories	F	O	Service: Project Config	TRE: RAZ		2.1.9
R068	U11	RAZ's with Project Config services which connect to external repositories SHOULD support the Software interface type.	F	S*	TRE: RAZ	Interface: Software		2.1.12
R069	U11	Project Environments MAY be provisioned and managed dynamically as "pop-up" environments	F	O	Collaboration: Project Environment	TRE: RAZ		2.2.2; 2.4.2
R070	U11	Where "pop-up" Project Environments are to be kept in runtime alignment with an approved state, the hosting RAZ MUST support the Sync interface type.	F	M*	TRE: RAZ	Interface: Sync		2.4.4

R071	U16	Data Presentation components MAY provide a view on remote data resources.	F	O	Component: Data Presentation	Collaboration: Project Environment		
R072	U16	If an RAZ supports Project Environments with remote Data Presentations then it MUST support the Query (direct) interface type.	F	M*	TRE: RAZ	Interface: Query (direct)	Collaboration: Project Environment	
R073	U16	If an RAZ supports Project Environments with remote Data Presentations then it MUST support the Response interface type.	F	M*	TRE: RAZ	Interface: Response	Collaboration: Project Environment	
R074	U16	An RAZ MAY provide a Job Submission component to support indirect queries against remote data resources	F	O	TRE: RAZ	Component: Job Submission		
R075	U16	If an RAZ provides a Job Submission component then it MUST support the Query (indirect) interface type	F	M*	TRE: RAZ	Interface: Query (indirect)		
R076	U16	If an RAZ provides a Job Submission component then it MUST support the Response interface type	F	M*	TRE: RAZ	Interface: Response		
R077	U16	An RAZ MAY provide an HPC component which offers additional, significant computing and analytical capability.	F	O	TRE: RAZ	Component: HPC		
R078	U09	A TRE MAY have An SDZ (secure data zone)	F	O	TRE: SDZ			
R079	U09	System actors in the role of Data Manager SHALL be granted access to the SDZ	N	M	TRE: SDZ			3.1.6; 3.1.12
R080	U09	System actors in the role of Output Approver SHALL be granted access to the SDZ	N	M	TRE: SDZ			3.1.6; 3.1.12; 3.3.4

R081	U09	System actors in neither Data Manager nor Output Approver roles SHALL NOT be granted access to the SDZ	N	M	TRE: SDZ			3.1.6; 3.1.12; 3.3.4
R082	U09	An SDZ MUST have a Data Management function	F	M	TRE: SDZ	Function: Data Management		3.1.4; 3.1.5; 3.1.12
R083	U03	All movements of data to or from the SDZ MUST pass through the Data Management function	N	M	TRE: SDZ	Function: Data Management		3.1.1; 3.1.4; 3.1.5; 3.1.12
R084	U03	An SDZ SHOULD support the Data Egress interface type	F	S	TRE: SDZ	Interface: Data Egress		3.1.5; 3.1.12
R085	U07	A Data Management function SHOULD support linkage of datasets from both local and remote data sources	F	S	Function: Data Management			
R086	U07	An SDZ that supports linkage of datasets (via its Data Management function) SHOULD support the Index interface type	F	S*	TRE: SDZ	Interface: Index		
R087	U03	An SDZ SHOULD support the Data Ingress interface type	F	S	TRE: SDZ	Interface: Data Ingress		3.1.4; 3.1.12
R088	U03	A TRE MAY have a QMZ	F	O	TRE: QMZ			2.1.3
R089	U03	A QMZ MUST support the Response interface type.	F	M	TRE: QMZ	Interface: Response		
R090	U03	A QMZ that supports direct queries MUST support the Query (direct) interface type.	F	M	TRE: QMZ	Interface: Query (direct)		
R091	U03	A QMZ MAY support direct queries via an External Presentation component.	F	O	TRE: QMZ	Component: External Presentation		
R092	U03	External Presentation components MUST connect internally to Query (direct) interface types.	F	M	Component: External Presentation	Interface: Query (direct)		

R093	U09	An SDZ MAY host (and curate) one or more datasets as a Curated Research-Ready Data collection.	F	O	TRE: SDZ	Data: Curated Research-Ready Data		3.1.8
R094	U03	External Presentation components MUST connect internally to Response interface types.	F	M	Component: External Presentation	Interface: Response		
R095	U03	A QMZ that does not support direct queries MUST support indirect queries via Job Controller and Job Executor components.	F	M*	TRE: QMZ	Component: Job Controller	Component: Job Executor	
R096	U03	A QMZ that supports indirect queries MUST support the Query (indirect) interface type.	F	M*	TRE: QMZ	Interface: Query (indirect)		
R097	U03	Job Controller components MUST connect internally to Query (indirect) interface types.	F	M	Component: Job Controller	Interface: Query (indirect)		
R098	U03	Job Controller components MUST connect internally to Response interface types.	F	M	Component: Job Controller	Interface: Response		
R099	U16	A QMZ MAY provide an HPC component which offers additional, significant computing and analytical capability.	F	O	TRE: QMZ	Component: HPC		
R100	U34	A Discovery Service MAY enable data discovery by querying other services (including Federation services) within the Federation	F	O	Service: Discovery			3.4.1; 3.7.1
R101	U34	A Discovery Service which queries other services (including Federation services) within the Federation MUST support the Query (direct) interface type	F	M	Interface: Query (direct)	Service: Discovery		2.2.9; 3.7.1

R102	U34	A Discovery Service which queries other services (including Federation services) within the Federation MUST support the Response interface type	F	M	Interface: Response	Service: Discovery		2.2.9; 3.7.1
R103	U34	A Discovery Service which queries other services within the Federation MUST implement an Output Control process to manage potential disclosure of confidential information from within the Federation	F	M	Service: Discovery	Process: Output Control		3.3.4
R104	U16	A Job Submission service MUST implement a Job Approval process for all received job requests.	F	M	Service: Job Submission	Process: Job Approval		
R105	U16	System actors in the role of Job Approver SHALL be authorised to access the Job Approval process.	N	M	Service: Job Submission	Process: Job Approval		
R106	U16	System actors not in the role of Job Approver SHALL NOT be authorised to access the Job Approval process.	N	M	Service: Job Submission	Process: Job Approval		
R107	U16	A Job Submission service MUST support the Query (indirect) interface type	F	M	Service: Job Submission	Interface: Query (indirect)		
R108	U16	A Job Submission service MUST support the Response interface type	F	M	Service: Job Submission	Interface: Response		
R109	U16	A Job Submission service MUST implement an Output Control process to approve the external release of any job response artifacts.	F	M	Service: Job Submission	Process: Output Control		3.3.4
R110	U16	System actors in the role of Output Approver SHALL be authorised to access the Output Control process.	N	M	Service: Job Submission	Process: Output Control		2.1.1; 3.3.4

R111	U16	System actors not in the role of Output Approver SHALL NOT be authorised to access the Output Control process.	N	M	Service: Job Submission	Process: Output Control		2.1.1; 3.3.4
R112	U16	A Software Service MUST support the Software interface type	F	M	Service: Software	Interface: Software		2.1.12; 2.1.13
R113	U12	Participant services within the Federation MUST run a standard Security Server	F	M	Core: Federation	Core: Security Server		
R114	U12	Federation services exchanging data extracts MUST use the Data Extract Object format (cf. R123)	N	M	Object: Data Extract	Interface: Data Egress	Interface: Data Ingress	
R115	U12	Federation services exchanging indexes or linkage spines MUST use the Index Object format (cf. R123)	N	M	Object: Index	Interface: Index		
R116	U12	Federation services sending direct queries to other services MUST use the Query Object format (cf. R123)	N	M	Object: Query	Interface: Query (direct)		
R117	U12	Federation services sending indirect queries to other services MUST use the Job Request Object format (cf. R123)	N	M	Object: Job Request	Interface: Query (indirect)		
R118	U12	Federation services returning direct query results to other services MUST use the Response (Query) Object format (cf. R123)	N	M	Object: Response (Query)	Interface: Response		
R119	U12	Federation services returning indirect query (job) results to other services MUST use the Response (Job) Object format (cf. R123)	N	M	Object: Response (Job)	Interface: Response		
R120	U12	Software Services returning research artifacts to other services MUST use	N	M	Object: Job Payload Artifact	Interface: Software		

		the Job Payload Artifact format (cf. R123)						
R121	U12	Software Services returning Environmental software artifacts to other services MUST use the Environment Software Artifact format (cf. R123)	N	M	Object: Environment Software Artifact	Interface: Software		
R122	U19	All Projects MUST be registered with the Federation Registry	N	M	Core: Federation: Registry			2.2.14; 4.8.2
R123	U12	All structured data objects exchanged by Federation Participants MUST be packaged in a standard way.	N	M	Object: all	Interface: all		
R124	U03	Release of Data Extract objects MUST occur via TREs' Data Management functions, overseen by Data Manager roles.	N	M	TRE: SDZ: Data Management	Object: Data Extract		
R125	U11	Query Objects MUST encapsulate all information necessary for a receiving TRE to execute a direct query	F	M	Object: Query	Interface: Query (direct)		
R126	U11	Job Payload Artifacts MUST encapsulate all information necessary for a receiving TRE to execute an indirect query	F	M	Object: Job Payload Artifact	Interface: Query (indirect)	Component: Job Executor	
R127	U12	Response Objects SHOULD have the same encapsulation structure for responses to direct or indirect queries	N	S	Object: Response			
R128	U11	TREs SHOULD download Environment Software Artifacts from Federation Software Services (rather than "download from source")	F	S	TRE: RAZ	Interface: Software	Service: Software	2.1.12; 2.1.13

R129	U12	Federation Core Services MUST be connected in a secure network control plane, independent of the data exchange network	F	M	Core			
R130	U12	Federation Services MUST be highly available	N	M	Core: Federation			
R131	U12	Federation Management Services MUST provide a mechanism to Security Servers are up-to-date and synchronised with the currently agreed and approved global configuration (cf. R133)	F	M	Core: Federation: Management	Core: Security Server		2.4.3; 2.4.4.; 2.4.5
R132	U12	Security Servers MUST operate to an agreed and approved global configuration	F	M	Core: Security Server	Core: Federation: Management		2.4.3
R133	U12	Security Servers MUST support a mechanism to synchronise their configuration with the agreed global configuration (cf. R131)	F	M	Core: Security Server	Core: Federation: Management		2.4.4; 2.4.5
R134	U12	Security Servers MUST be able to operate independently if their connection to Federation Services is interrupted	N	M	Core: Security Server	Core: Federation		
R135	U12	Content metadata (ie, about Datasets) within the Federation SHOULD align with UK Government standards and recommendations	N	S	Metadata			3.1.3
R136	U12	Governance metadata (ie, about Projects and Users) within the Federation SHOULD align with UK Government accreditation requirements	N	S	Metadata			

R137	U19	Where a Project spans multiple TREs (eg, one based on federated query patterns) one TRE MUST be designated as the Project host	N	M	TRE	Core: Federation: Registry		
R138	U12	Project Identities MUST be globally recognisable within the Federation	F	M	Metadata	Core: Federation: Registry		
R139	U12	Project Identities MUST be globally unique within the Federation	F	M	Metadata	Core: Federation: Registry		
R140	U12	Project Member Identities MUST be globally recognisable within the Federation	F	M	Metadata	Core: Federation: Registry		
R141	U12	Project Member Identities MUST be globally unique within the Federation	F	M	Metadata	Core: Federation: Registry		
R142	U12	Dataset Identities MUST be globally recognisable within the Federation	F	M	Metadata	Core: Federation: Registry		
R143	U12	Dataset Identities MUST be globally unique within the Federation	F	M	Metadata	Core: Federation: Registry		
R144	U12	All structured data objects exchanged by Federation Participants SHOULD include the appropriate Project Identity as context.	F	M	Metadata			
R145	U03	Job Payload Artifacts MUST be subject to the Job Approval Process of the receiving TRE	F	M	TRE: QMZ: Job Approval	Object: Job Payload Artifact		
R146	U03	Job Payload Artifacts MUST encapsulate all information necessary for a receiving TRE to evaluate the safety of an indirect query	N	M	Object: Job Payload Artifact	TRE: QMZ: Job Approval		