



Federated Architecture Blueprint

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 

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. |
| 2.2 | 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 | 9 |
| Acknowledgements | 9 |
| How to read this document | 10 |
| 1. Executive summary | 11 |
| 1.1. Overview..... | 11 |
| 1.2. The strategic case for federation | 12 |
| 1.3. Users and use-cases..... | 12 |
| 1.4. Federated architecture: infrastructure layer..... | 12 |
| 1.5. Federated architecture: data layer..... | 14 |
| 1.6. Federated architecture: organisational layer | 15 |
| 1.7. Development and delivery approach | 16 |
| 1.8. Summary and further work | 16 |
| 2. The strategic case for a federated architecture..... | 17 |
| 2.1. DARE UK Phase 1 recommendations | 17 |
| 2.2. The federation challenge | 18 |
| 2.2.1. Conceptual data space | 19 |
| 2.2.2. Data pooling..... | 20 |
| 2.2.3. Federated analytics..... | 20 |
| 2.3. Federated infrastructure: the state of the art..... | 22 |
| 2.3.1. TRE federation proofs-of-concept: the DARE UK driver projects..... | 23 |
| 2.4. A federation blueprint | 25 |
| 2.4.1. Scope | 26 |
| 2.4.2. Design principles | 26 |
| 2.5. Summary..... | 27 |
| 3. Users and use-cases | 28 |
| 3.1. Rachel's journey: 2022 | 28 |
| 3.2. User personas..... | 32 |

FOR CONSULTATION & COMMENT

- 3.2.1. Federation actors and roles 33
- 3.2.2. Other stakeholders 34
- 3.3. User stories and requirements mapping 34
- 3.4. Future work 39
- 4. Federated architecture: infrastructure layer 40
 - 4.1. Notation 40
 - 4.1.1. Symbols 40
 - 4.1.2. Colours 40
 - 4.2. Actors and roles 42
 - 4.2.1. Researcher (actor) 42
 - 4.2.2. Information Governance (actor) 42
 - 4.2.3. Data Controller (actor) 43
 - 4.2.4. TRE Operator (actor) 43
 - 4.3. Participants 43
 - 4.3.1. Trusted research environment (TRE) 43
 - 4.3.2. Index Service 47
 - 4.3.3. Discovery Service 47
 - 4.3.4. Job Submission Service 48
 - 4.3.5. Software Service 49
 - 4.4. Interface types 50
 - 4.4.1. Query (Direct) 50
 - 4.4.2. Query (Indirect) 50
 - 4.4.3. Response 51
 - 4.4.4. Data Ingress and Data Egress 51
 - 4.4.5. Index 51
 - 4.4.6. Software 51
 - 4.4.7. Sync 52
 - 4.5. Structured data objects 52
 - 4.5.1. Data Extract Object 52
 - 4.5.2. Index Object 53
 - 4.5.3. Query Object 53

FOR CONSULTATION & COMMENT

| | | |
|--------|---|----|
| 4.5.4. | Job Request Object | 54 |
| 4.5.5. | Job Payload Artifact | 54 |
| 4.5.6. | Response Object | 55 |
| 4.5.7. | Environment Software Artifact | 55 |
| 4.5.8. | Project Sync Object | 55 |
| 4.6. | SDRI core services..... | 56 |
| 4.6.1. | Federation Services | 56 |
| 4.6.2. | Security Server..... | 57 |
| 4.7. | Related concepts | 57 |
| 4.7.1. | Projects..... | 57 |
| 4.7.2. | Federation identities..... | 58 |
| 4.7.3. | Authentication and authorisation | 59 |
| 5. | Federated architecture: data layer | 60 |
| 5.1. | Classifying sensitive data | 60 |
| 5.1.1. | A seven-point scale..... | 60 |
| 5.2. | Federation metadata..... | 61 |
| 5.2.1. | Infrastructure metadata | 62 |
| 5.2.2. | Content metadata | 63 |
| 5.2.3. | Governance metadata | 64 |
| 5.2.4. | Structured data packaging formats | 66 |
| 5.2.5. | Other considerations..... | 67 |
| 5.3. | Data findability..... | 67 |
| 5.3.1. | Discovery metadata..... | 67 |
| 5.4. | Data accessibility | 69 |
| 5.5. | Data interoperability..... | 70 |
| 5.5.1. | Syntactic interoperability..... | 70 |
| 5.5.2. | Terminological interoperability | 70 |
| 5.5.3. | Semantic interoperability..... | 71 |
| 5.5.4. | Data linkage..... | 71 |
| 5.6. | Data reusability..... | 71 |
| 6. | Federated architecture: organisational layer..... | 73 |

FOR CONSULTATION & COMMENT

- 6.1. Centralised vs distributed vs decentralised..... 74
- 7. Development and delivery approach..... 77
 - 7.1. Prototyping and technology selection 77
 - 7.1.1. Core services: technology evaluation 77
 - 7.1.2. Interfaces and other services: community driver projects 77
 - 7.2. Technology proof-of-concept..... 77
 - 7.2.1. Scenario 1: basic data exchange 78
 - 7.2.2. Scenario 2: linked data exchange..... 78
 - 7.2.3. Scenario 3a: remote direct query (single)..... 78
 - 7.2.4. Scenario 3b: remote direct query (federated) 79
 - 7.2.5. Scenario 4a: remote indirect query (single)..... 79
 - 7.2.6. Scenario 4b: remote indirect query (federated)..... 79
 - 7.3. Minimal viable product..... 80
 - 7.4. Test and validation 80
 - 7.5. Evolution 80
- 8. Summary and further work..... 81
- 9. References..... 82
- A A comparison of contemporary federated data architectures 85
- B Usage patterns 87
 - B.1 “Classic” TRE inter-operation..... 87
 - B.2 Francis Crick Institute federation model 89
 - B.3 OpenSAFELY 91
 - B.4 TELEPORT federation with pop-up TREs..... 93
 - B.5 TRE-FX federation with stand-alone job submission 95
 - B.6 TRE-FX federation with TRE-hosted job submission..... 97
- C Scenario analysis of the federated landscape 99
 - C.1 Four quadrants..... 99
 - C.1.1 Low numbers of TREs and low data mobility 99
 - C.1.2 Low numbers of TREs and high data mobility 100
 - C.1.3 High numbers of TREs and low data mobility 100

FOR CONSULTATION & COMMENT

| | | |
|-------|--|-----|
| C.1.4 | High numbers of TREs and high data mobility | 100 |
| C.2 | Observations | 101 |
| D | Master requirements table..... | 102 |
| E | Acknowledgements..... | 118 |
| E.1 | Federated architecture blueprint: direct feedback..... | 118 |
| E.2 | Phase 1b persona development | 118 |
| E.3 | DRI landscape review and community conversations | 119 |

FOR CONSULTATION & COMMENT

About document versions

This document is 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 E has a full list of acknowledgements.

FOR CONSULTATION & COMMENT

How to read this document

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

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

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

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

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

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

“All models are wrong, but some models are useful.”

George E.P. Box

FOR CONSULTATION & COMMENT

1. Executive summary

1.1. Overview

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

Analytics services for researchers working with sensitive data are typically—and increasingly—provided in trusted research environments (TREs), secure computer systems wrapped in information governance practices and processes modelled on the Five Safes approach developed by the Office for National Statistics (ONS¹). These cast the technical systems needed to support sensitive data research as one part (the “safe setting”) of a broader set of procedures designed to manage risk and create an overall trustworthy environment.

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

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

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

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

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

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

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

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

FOR CONSULTATION & COMMENT

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

1.2. The strategic case for federation

Chapter 2 sets out the strategic case for a standardised federation of multiple service providers: the Sensitive Data Research Infrastructure Federation (the SDRI Federation).

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

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

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

1.3. Users and use-cases

Chapter 3 motivates the federation blueprint with a collection of high-level user stories and usage patterns.

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

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

1.4. Federated architecture: infrastructure layer

Chapter 4 gets to the heart of the federated architecture blueprint, the infrastructure layer: how the Federation is realised through the exchange of Structured Data Objects between Participants over standardised connections, with Federation Services providing the cohesive “glue”.

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

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

The core of the SDRI Federation sits between the other services, with connections shown between the standardised **Security Servers** at each Participant, plus a single group of **Federation Services**. This core of

FOR CONSULTATION & COMMENT

Federation Services, Security Servers and connections together define the Federation. The Federation Services group comprises services for registry (of services, users, projects, etc.), trust (security certificate management and signing), management (of standard shared software), monitoring and accounting.

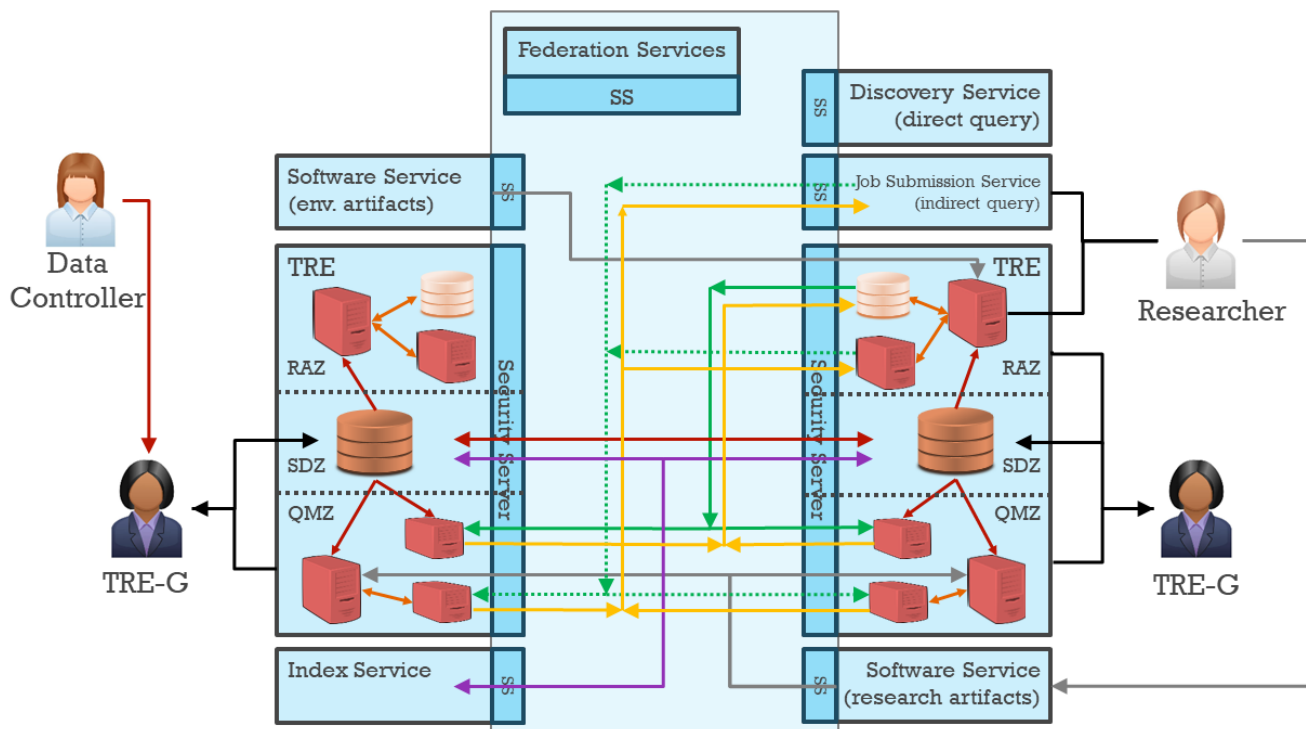


Figure 1. Simplified architectural sketch of the Sensitive Data Research Infrastructure Federation. Trusted Research Environments are denoted “TRE”. TREs are divided logically into three internal zones: a Research Analytics Zone (RAZ), a Secure Data Zone (SDZ), and a Query Management Zone (QMZ). Not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

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

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

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

Orange connections represent the responses returned by queries. While typically thought of as aggregate summaries, results do have the potential to contain disclosive information, depending on the query sent and the dataset queried. While results would only ever be sent through secure gateways (Security

FOR CONSULTATION & COMMENT

Servers) to other approved Participants within the closed Federation network, disclosure controls may be appropriate for certain kinds of results.

Red connections allow Data Controllers and TREs acting as data providers to send datasets and data extracts to governance authorities in TREs in standard, secure ways². Sensitive personal data are de-identified and approved for use in research but are nevertheless potentially disclosive and, despite the above remarks about secure gateways and closed Federation network still applying, disclosure controls are appropriate for red connections.

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

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

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

1.5. Federated architecture: data layer

*Chapter 5 looks at data from two angles: data **about** the Federation and data **within** the Federation.*

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

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

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

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

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

FOR CONSULTATION & COMMENT

- User metadata: information about each user of the Federation, the roles they have, the approvals they have, the Projects they are members of, and so on.
- Data Extract metadata: information about subsets or extracts of Datasets as used in Projects.

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

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

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

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

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

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

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

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

1.6. Federated architecture: organisational layer

Chapter 6 outlines some considerations and possible approaches to organisational arrangements without which this blueprint cannot proceed beyond a proof-of-concept stage.

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

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

FOR CONSULTATION & COMMENT

1.7. Development and delivery approach

Chapter 7 sketches a phased development and delivery approach to implementing this blueprint.

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

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

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

1.8. Summary and further work

Finally, Chapter 8 looks ahead.

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

FOR CONSULTATION & COMMENT

2. The strategic case for a federated architecture

“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 C *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.

2.1. 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.

2.2. 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 right³). 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.

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.

³ 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

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.

We have updated our original analysis of existing patterns of interaction between TREs with developments across the community during 2023. Appendix B uses published information about federation patterns from the current TRE landscape to illustrate how the SDRI Federation architecture has evolved from version 1.x to version 2.x.

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,
 - 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.

2.2.1. 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 2 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.

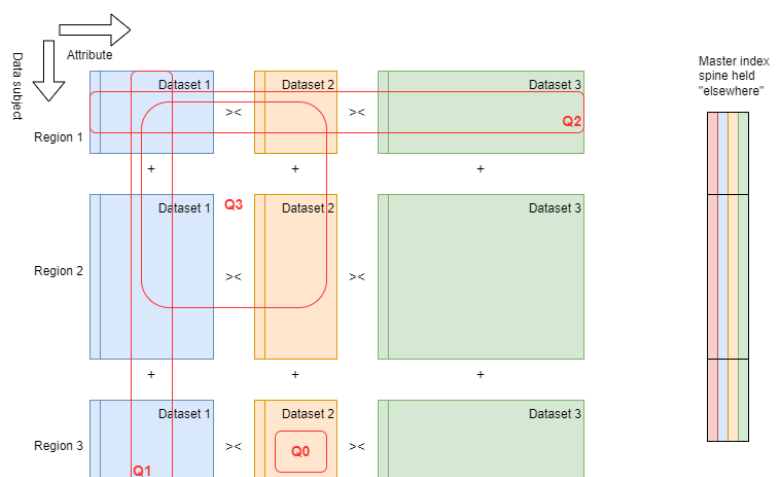


Figure 2. Conceptual dataspace for DARE UK

FOR CONSULTATION & COMMENT

Partitioning vertically splits the attribute space for the whole population. Doing both creates the picture in Figure 2.

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 2 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:

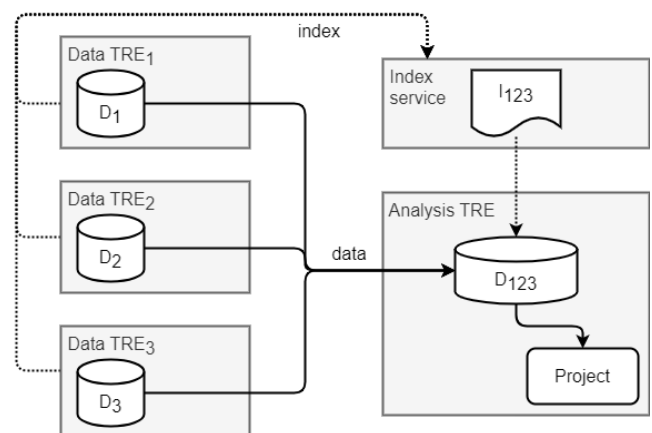
- Q0: a query within a single regional dataset.

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

2.2.2. 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.



2.2.3. 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.

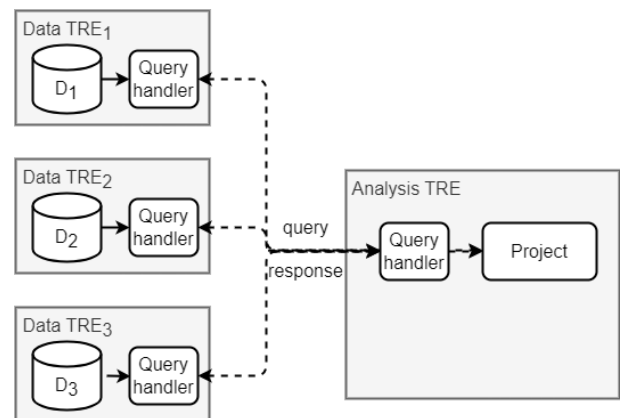
FOR CONSULTATION & COMMENT

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 (q.v. and cf. Section 4.3.3) which support distributed metadata discovery from *outside* the Federation – although because this use connects internal Federation queries to the outside world, Discovery Services must be designed with disclosure control in place and with careful governance oversight.

2.2.3.1. 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.

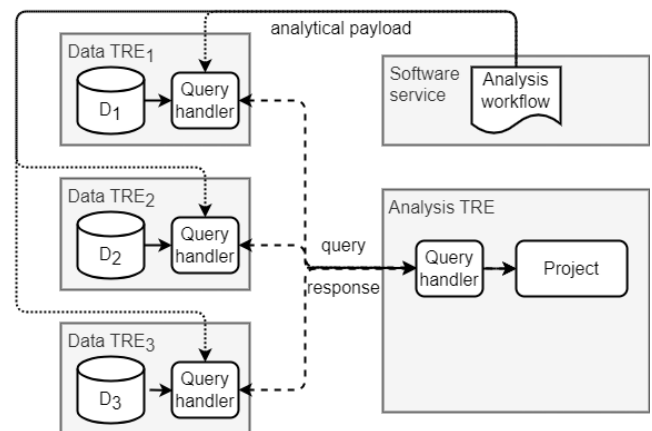
An example implementation of direct query can be found in the TELEPORT project [29]. TELEPORT uses the Trino SQL execution engine⁴ to connect remote data sources within one TRE to a “single pane of glass” user-view in another. To the research user, this has the appearance, and consequent utility, of a single database table, while behind the scenes queries and results are exchanged between participating TREs.

⁴ “Trino, a query engine that runs at ludicrous speed”. See <https://trino.io/>

FOR CONSULTATION & COMMENT

2.2.3.2. *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 (D₁, D₂ and D₃) 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.

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 software⁵, 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. Sections 4.5 and 5.2.4 and footnote 17).

2.3. 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

⁵ Hutch, a federated analytics execution agent. See <https://health-informatics-uon.github.io/hutch/>

FOR CONSULTATION & COMMENT

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].

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.

2.3.1. 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 website⁶.

⁶ DARE UK 2023 Driver Projects, <https://dareuk.org.uk/our-work/phase-1-driver-projects/>

FOR CONSULTATION & COMMENT

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 Bitfount⁷ and DataSHIELD⁸ 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.

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.

⁷ Bitfount federated AI and data science platform. See <https://www.bitfount.com/>

⁸ DataSHIELD secure bioscience collaboration. See <https://datashield.org/>

FOR CONSULTATION & COMMENT

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 Appendix D, *Master Requirements Table*.

This new version of the federated architecture blueprint models these developments much more accurately than did version 1. (cf. Appendix B).

2.4. 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

FOR CONSULTATION & COMMENT

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.

2.4.1. 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.

2.4.2. 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.)

FOR CONSULTATION & COMMENT

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.)

2.5. 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

3. Users and use-cases

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.

3.1. 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.

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.

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?

- 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#).

FOR CONSULTATION & COMMENT

- Health outcomes, collected by Peter and available for research but only for particular cohorts. Rachel will have to ask explicitly for what she needs.

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:

- 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.

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

Iain sends the ID1 and education index mapping to Preethi.

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

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.

These identifiers are not particularly sensitive of themselves but nevertheless sending

FOR CONSULTATION & COMMENT

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.

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.

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 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.

Rachel gets access to her approved linked data inside the TRE, and she’s off!

documents between different parties needs to be done securely.

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).

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.

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).

Finally!

Figure 3 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”. The colour scheme follows that of Section 4.1.2. 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

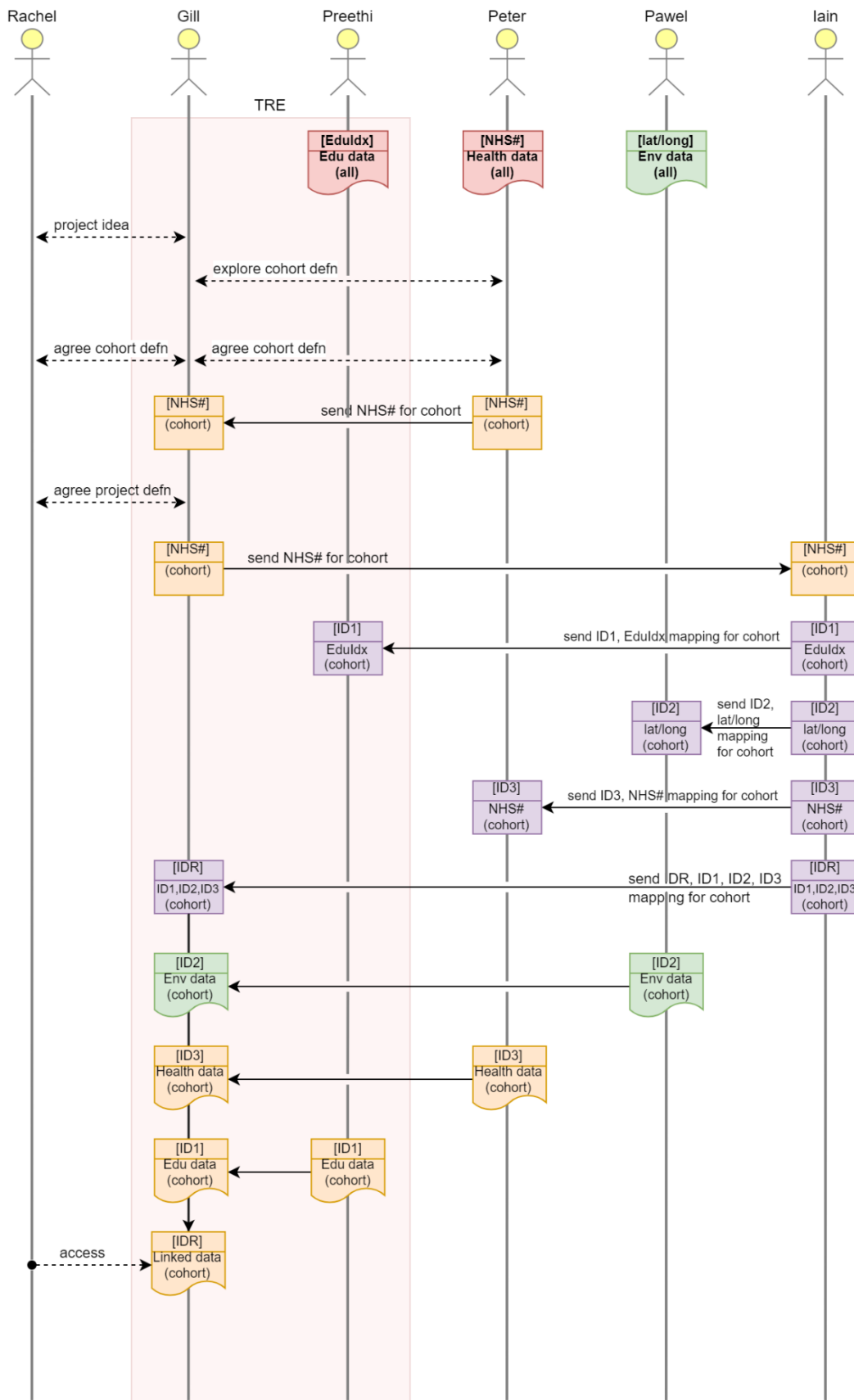


Figure 3. Rachel's Journey as a sequence diagram.

FOR CONSULTATION & COMMENT

3.2. 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 focused 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 |

FOR CONSULTATION & COMMENT

3.2.1. 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, by construction, is new.

3.2.1.1. 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.

3.2.1.2. 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.

3.2.1.3. 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.

FOR CONSULTATION & COMMENT

3.2.2. 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.

3.3. 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 D.

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 [32]. 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 | |

FOR CONSULTATION & COMMENT

| | | | |
|-----|--|------------|---|
| 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 |

FOR CONSULTATION & COMMENT

| | | | |
|-----|---|--------|--|
| 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 |

FOR CONSULTATION & COMMENT

| | | | |
|-----|--|--------|---|
| 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 |

FOR CONSULTATION & COMMENT

| | | | |
|-----|---|------------|--|
| 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 cardio-vascular disease and developing targeted interventions for high-risk groups | SG, PN | Data: Health data, Social data, Economic data, Lifestyle data, Consumer data |

FOR CONSULTATION & COMMENT

3.4. 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.

FOR CONSULTATION & COMMENT

4. Federated architecture: infrastructure layer

This blueprint draws on current best practice in secure data exchange environments but also reflects the design principle of “start from where you are”. This architecture proposes the minimum necessary new infrastructure to create the required trustworthy federation while causing the least disruption to TREs and data services already in use. It is also explicitly a “back end” architecture that connects TREs to other TREs. Adherence to the principle that all research with sensitive data take place within a TRE means that Researchers will interact only with TREs and never with the Federation infrastructure directly.

Figure 4 depicts the high-level architecture of the SDRI Federation. It shows a number of Federation Participants—TREs and supporting services—and indicates the principal information flows between them. For illustrative purposes we show two TREs and single versions of other services. In practice there will be more of each. A single set of Federation Services hold a record of all Federation Participants and provide a set of trust services that together create the required trustworthy environment.

4.1. Notation

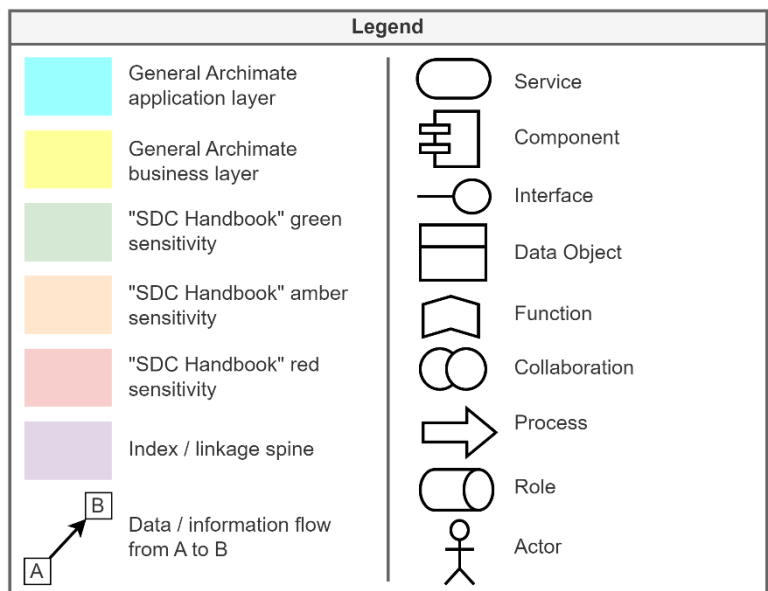
The diagrams follow the ArchiMate standard (version 3.1) [40] according to the following legend.

4.1.1. Symbols

The diagram elements have their usual ArchiMate meanings (right-hand column) with the exception of connecting lines.

Solid connecting lines indicate channels of data or information flow, with arrows indicating direction. Importantly, the absence of an arrow indicates that there is no data flow in that direction.

Dotted lines indicate an (unspecified) relationship between the connected elements.



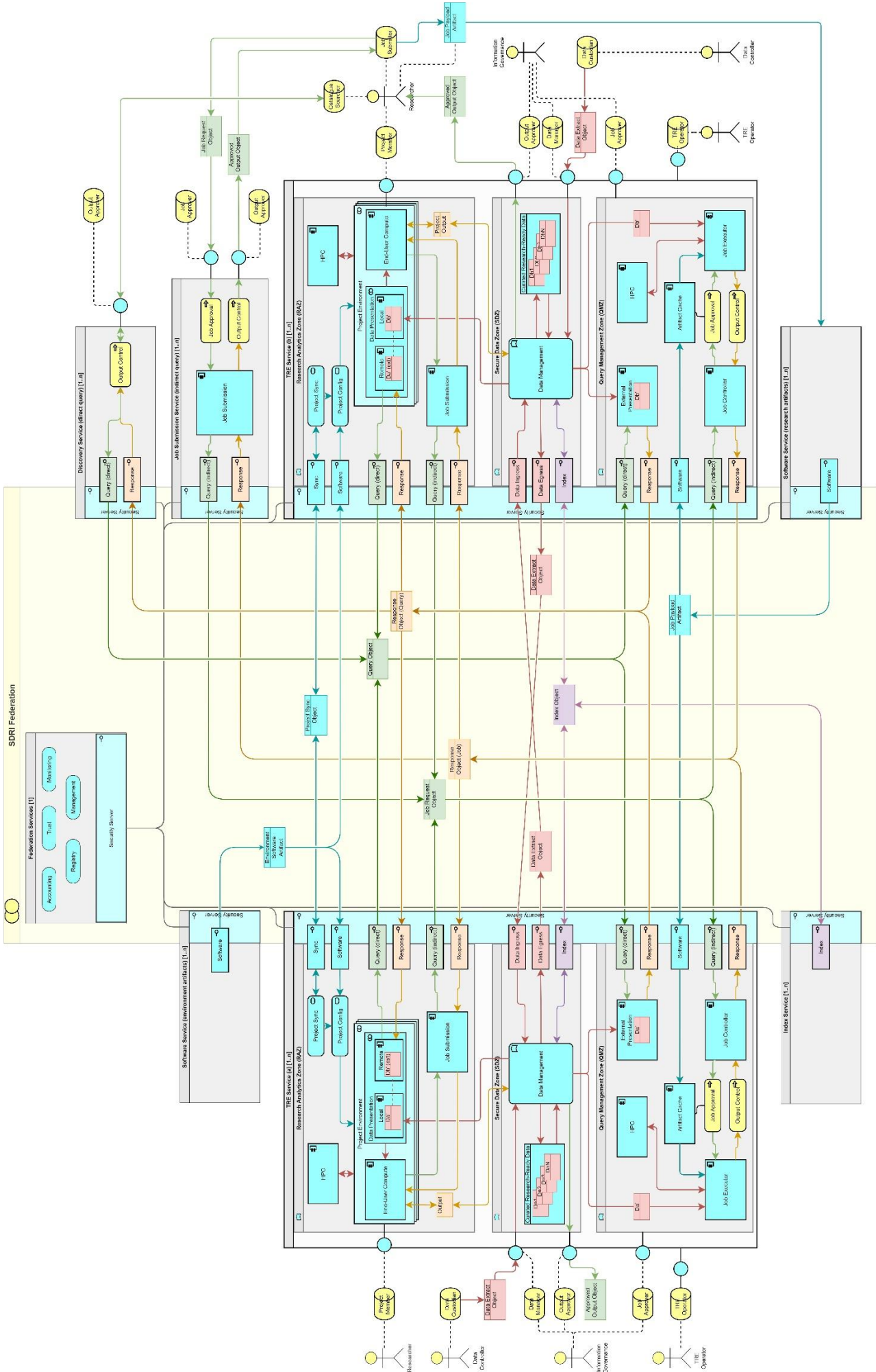
4.1.2. Colours

Yellow and cyan colours indicate elements in default ArchiMate architectural layers – the higher business layer and lower application layer respectively.

We use green, amber and red colours to indicate “data sensitivity” in the sense of potentially disclosive, aligning with the terminology used in the Statistical Disclosure Control Handbook [41].

Purple indicates indexing or linkage spine data, what might be termed “sensitive metadata”.

FOR CONSULTATION & COMMENT



FOR CONSULTATION & COMMENT

Figure 4 (previous page). Architectural diagram of the infrastructure layer of the SDRI Federation network. The notation broadly follows the ArchiMate v3.1 standard [40], although we use colour in a different way (see above). Note that the scope of the core federation is captured in the central "yellow collaboration" element and associated "blue box" security servers. Please refer to the key in Section 4.1 for definitions of the diagram elements.

4.2. Actors and roles

We resolve the federation users identified in Chapter 3 into *actors* and *roles* in the infrastructure picture. *Actors* are actual individuals or small teams. *Roles* capture specific activities or responsibilities taken on by actors.

4.2.1. Researcher (actor)

Researchers take on a number of roles within the overall federated system. We use these roles to model their interactions with TREs and other services.

4.2.1.1. Project Member (role)

A Researcher may become an approved and authorised member of one or more Projects (see below), and in that role (and in the context of these Projects) will interact with specially provided project environments within one or more TREs. Access to data within a TRE is granted to a Researcher in their role as a Project Member, on the basis of their individual project authorisations.

4.2.1.2. Job Submitter (role)

(Possibly a sub-role of Project Member.)

In contrast to the direct interaction of a Project Member, the Job Submitter role interacts indirectly with TREs through externally accessible Job Submission and Software services.

4.2.1.3. Catalogue Searcher (role)

The Catalogue Searcher role interacts with externally accessible data discovery and catalogue services. In this role the Researcher need not be a member of an existing project, and thus may not have approvals to access sensitive data of any sort.

4.2.2. Information Governance (actor)

Information Governance is a shorthand for the team of people charged with overseeing a TRE and the research that happens within it. The Information Governance actor takes a number of data-related roles. They may also take the role of TRE Operator, or may delegate it (see below).

4.2.2.1. Data Manager (role)

This role covers a wide range of data management tasks within a TRE, from curating and maintaining long-term copies of research-ready data, to receiving data extracts from other Data Managers in other TREs, linking them and providing them onwards to research Project Members within the TRE. The Data Manager role will typically work closely with the Data Custodian role (see below).

This role could be further broken down into finer-grained sub-roles.

FOR CONSULTATION & COMMENT

4.2.2.2. Output Approver (role)

Output Approvers are responsible for checking any and all research outputs to be released from the TRE to the “outside world” (rather than sent to another TRE).

4.2.2.3. Job Approver (role)

Job Approvers review computational jobs submitted by Researchers (in their roles as Job Submitters) for their safety and suitability to run inside the Job Approver’s TRE.

4.2.3. Data Controller (actor)

4.2.3.1. Data Custodian (role)

In the TRE domain, Data Controllers take the role of Data Custodians, releasing data approved for research to projects via TRE Data Managers.

4.2.4. TRE Operator (actor)

4.2.4.1. TRE Operator (role)

The TRE Operator runs the TRE technical service day-to-day. This role may be taken by the Information Governance actor, or it may be delegated by them to a different actor (the TRE Operator actor) under their direction.

4.3. Participants

Participants is the general name for the services connected together to form the SDRI Federation.

In this document we focus on securing the connections **between** Participants within a federated network. We must be aware that **any** Participant judged (by Federation governance processes) “good enough” to join the Federation must have an appropriate level of security around all participating service elements. This may mean that all Federation Participants must demonstrate a certain level of secure hosting and management, not merely deploy a Federation Security Server. This will form one aspect of the governing rules for the Federation.

4.3.1. Trusted research environment (TRE)

TREs are the main vehicles for delivering sensitive data to Researchers in secure, controlled and approved ways.

In developing this architecture we have tried to avoid specifying in too much detail what a TRE “is” and what it “isn’t”. Nevertheless, the linking of TREs into cooperating services capable of supporting federated analytics imposes certain requirements on the internal structures of TREs. We model this using a number of “zones” within a TRE (see Figure 5).

Different TREs can offer different capabilities, and so not every TRE needs to support the functions of every zone. Figure 5 illustrates the *maximal* TRE, which includes every zone.

The zones are illustrated with gaps between them. This is deliberate: the zones require different levels of governance and approval for the roles accessing them, and in particular, movement of data between them

FOR CONSULTATION & COMMENT

should be subject to appropriate controls and potential “air-gapping” to manage the related disclosure risks.

4.3.1.1. *Research Analytics Zone (RAZ)*

This zone provides the means for a Project Member to gain direct access to the data their project is approved to use, in an environment suitable for the analyses their research requires. This is often realised as a virtual desktop environment, a computational notebook or similar. There is often a strict requirement that project environments be completely isolated from one another.

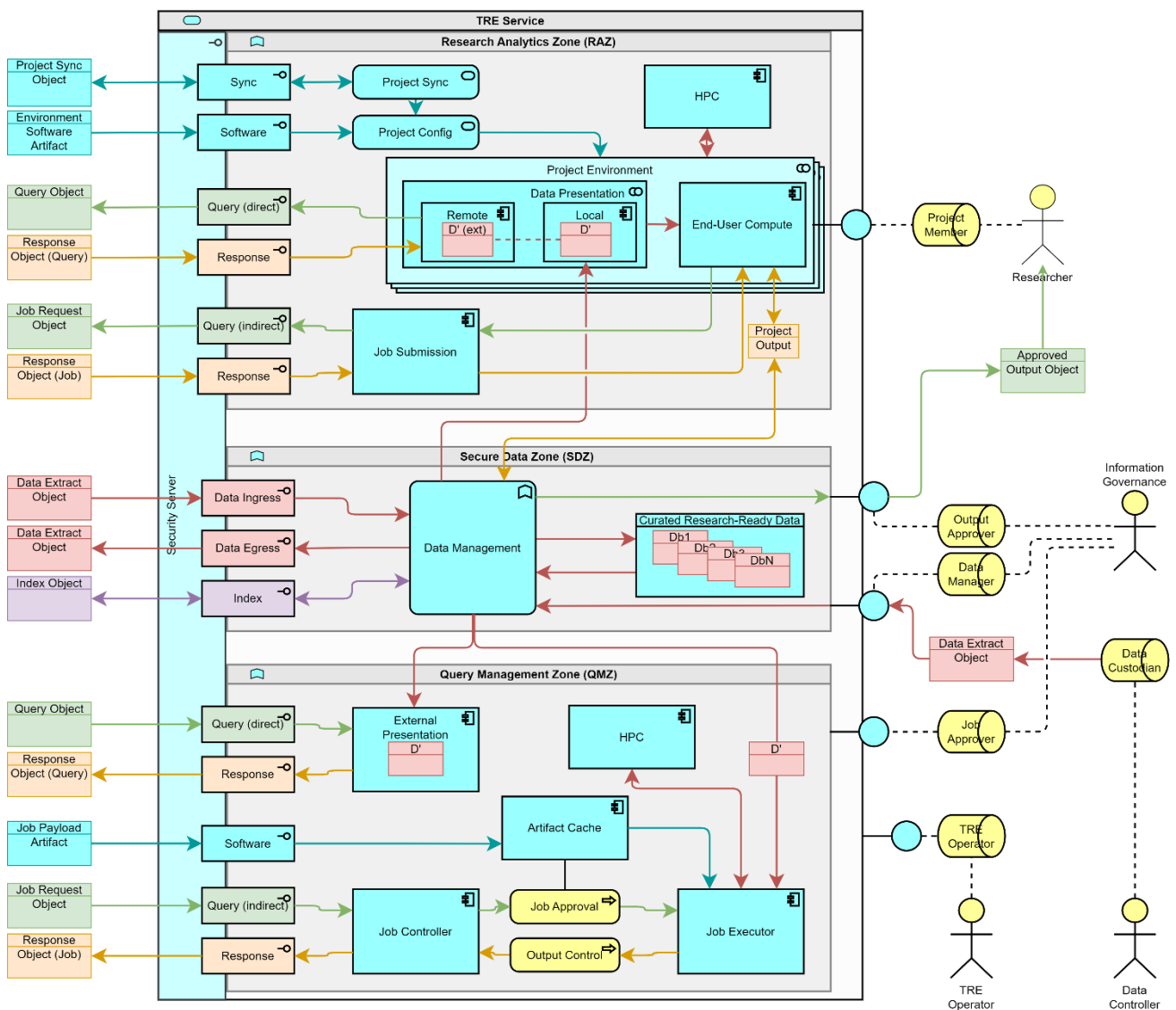


Figure 5. An expanded view of the TRE service from Figure 4.

FOR CONSULTATION & COMMENT

NB: A TRE need not have an RAZ. Instead it may operate as a pure data provider (with just a Secure Data Zone), or as a “headless” TRE able to run queries against data it hosts (with both a Secure Data Zone and a Query Management Zone).

An RAZ has a number of elements, not all of which need be present.

An RAZ **MUST** have one or more **Project Environments**. Project Environments **MUST** be suitable for the kinds of research the TRE supports and **SHOULD** be configured in standard and repeatable ways, modelled here by a relationship with a **Project Config** service. The Project Config service **MAY** connect to approved external software repositories, in which case the RAZ **SHOULD** support the **Software** interface type.

A Project Environment **MAY** be provisioned and managed dynamically and kept in sync with an agreed and approved project state (the “pop-up TRE” model). This project state may be shared between a number of participating TREs (strictly, between the participating TRE Governance actors) and synchronisation may require continually maintained connections between the participating TREs, modelled as a control relationship between a **Project Sync** service and the Project Config service. In this case the RAZ **MUST** also support the **Sync** interface type.

Each Project Environment is a combination (modelled as a collaboration) of an **End-User Compute** component and a **Data Presentation** component. The Data Presentation component **MAY** be composed of a **Local** data view (e.g., a file), **OR** a **Remote** data view (e.g., a representation of a remote resource in a web browser), **OR** a combination of the two (e.g., a polystore representation of two or more databases).

If the RAZ supports Remote data views then it **MUST** support the outgoing **Query (direct)** and incoming **Response** interface types (q.v.).

An RAZ **MAY** support indirect queries against remote TREs by providing a **Job Submission** component accessible directly from Project Environments. In this case the RAZ **MUST** support the outgoing **Query (indirect)** and incoming **Response** interface types (q.v.).

An RAZ **MAY** provide high-performance or advanced computing capabilities, modelled as an **HPC** component. This component **SHOULD** be accessible from the Project Environments and **MAY** be provisioned as a shared service, in which case special care must be taken in maintaining the strict isolation between projects.

The underpinning hardware for this component may overlap with – or indeed be the same as – the HPC component provided within a query management zone (cf. below). Its double inclusion in the diagram reflects the possibility of different modes of user access – interactive access directly from a Project Environment, or batch access via a job queue and potentially an internal Job Submission component.

4.3.1.2. Secure Data Zone (SDZ)

This zone supports the ingress, egress, management, linkage, curation and provision of research-ready sensitive datasets. TRE Governance actors with roles Data Manager and Output Approver **SHALL** be granted access to the SDZ; all other roles **SHALL NOT** be granted access.

FOR CONSULTATION & COMMENT

NB: A TRE need not have an SDZ. Instead it may operate as a pure analytics environment, with an RAZ supporting Project Environments with purely Remote data views, or with access solely to a Job Submission layer.

An SDZ has a number of elements, not all of which need be present.

An SDZ **MUST** have a **Data Management** function. The details of this function are largely out of scope, but its presence defines the core of an SDZ. All movement of data from the SDZ to other parts of the TRE, to other TREs or Index services, or to the outside world via an Output Approver **SHALL** pass through the Data Management function.

An SDZ **MAY** host and curate one or more datasets as **Curated Research-Ready Data**. Via its Data Management function it may provide these to local Project Members within Project Environments, to remote Project Members via external queries managed through the Query Management Zone, or to other Data Managers in remote TREs and Index services.

An SDZ **SHOULD** support the **Data Egress** and **Data Ingress** interface types for sending and receiving Data Extract Objects to and from remote TREs and Index services. (In practice, data ingress and egress may be managed through less formalised interfaces available to TRE Governance Data Managers.)

An SDZ that supports data linkage **SHOULD** support the **Index** interface type.

In this version of the blueprint a TRE with only an SDZ is equivalent to the Data Provider service in versions 1.x.

4.3.1.3. Query Management Zone (QMZ)

This zone handles queries sent to the TRE from other, remote TREs or external Job Submission services. Typically it sits alongside an SDZ and provides different methods of access to approved research-ready datasets stored within the SDZ.

NB: A TRE need not have a QMZ. Instead it may operate as a “classic” TRE, with an RAZ supporting Project Environments and an SDZ supporting data hosting, ingress and linkage, or as a pure analytics environment, with an RAZ supporting Project Environments with purely Remote data views.

A QMZ **MAY** support direct queries, where the received query object contains the actual runnable analytical artifact as a payload (e.g., an SQL query); or it **MAY** support indirect queries, where the received query object contains a reference to a runnable artifact held within an external repository of some kind; or it **MAY** support both.

A QMZ supporting direct queries **MUST** have an **External Presentation** component which can provide the approved dataset for the querying Project Member in a way that matches the query payload (e.g., as a project-specific database view for an SQL query). It **MUST** also support the incoming **Query (direct)** and outgoing **Response** interface types.

A QMZ supporting indirect queries **MUST** have a **Job Controller** component which can receive the incoming Job Request object. The Job Request **MUST** pass through a **Job Approval** process which **SHOULD** import the matching Job Payload Artifact from its remote repository, or take it from an internal **Artifact Cache**.

FOR CONSULTATION & COMMENT

Approved Job Payload Artifacts shall be passed to a **Job Executor** component which is able to execute them against the project dataset approved for access by the querying Project Member. Any results from the job’s execution shall be returned to the Job Controller via an **Output Control** process.

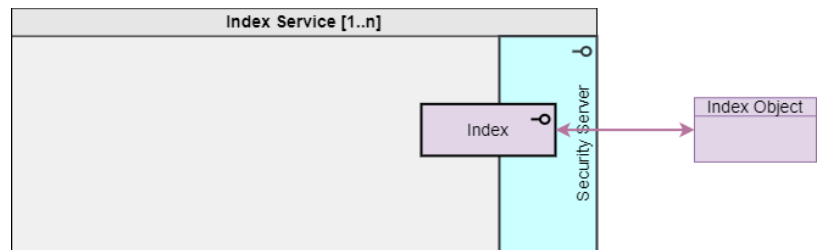
Note that either or both of the Job Approval and Output Control processes may involve manual inspection and assessment by TRE Governance Job Approver or Output Approver roles – hence their modelling as business processes rather than components or services.

A QMZ supporting indirect queries **MUST** also support the incoming **Query (indirect)** and outgoing **Response** interface types.

A QMZ **MAY** provide high-performance or advanced computing capabilities, modelled as an **HPC** component, in particular to support the execution of indirect query jobs. This component **SHOULD** be accessible from the Job Executor component and **MAY** be provisioned as a shared service, in which case special care must be taken in maintaining strict isolation between running jobs.

4.3.2. Index Service

An Index Service creates linkage spines for different Datasets. How a given service does this will depend first and foremost on the principal index key in question. For personal data, for example, the Index service will create depersonalised linkage spines by converting between “bare” personal identifiers and project-specific linkage keys.

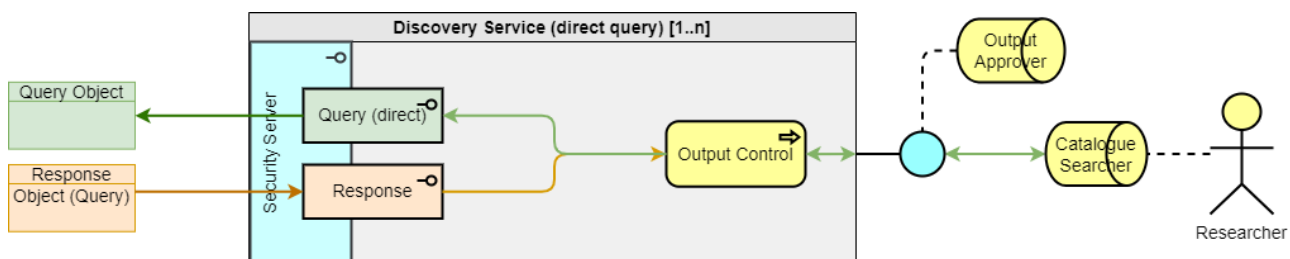


The Federation may include many Indexing Services, each perhaps specialising in a different kind of index.

Index Services **MUST** be trustworthy enough potentially to handle personal identifiers by which vertically partitioned datasets might be linked together. How indexes for such identifiers might be constructed is out of scope for this architecture. For a fuller treatment on how the *exchange* of indexes or linkage spines could be realised within the architecture see Chapter 5 *Federated Architecture: Data Layer*.

Indexing Services **SHALL** interact with other Federation participants solely through Indexing interface service calls.

4.3.3. Discovery Service



A Discovery Service provides information (metadata) about features of the Federation to users outside the Federation. It may achieve this by querying the Registry or other services within the Federation.

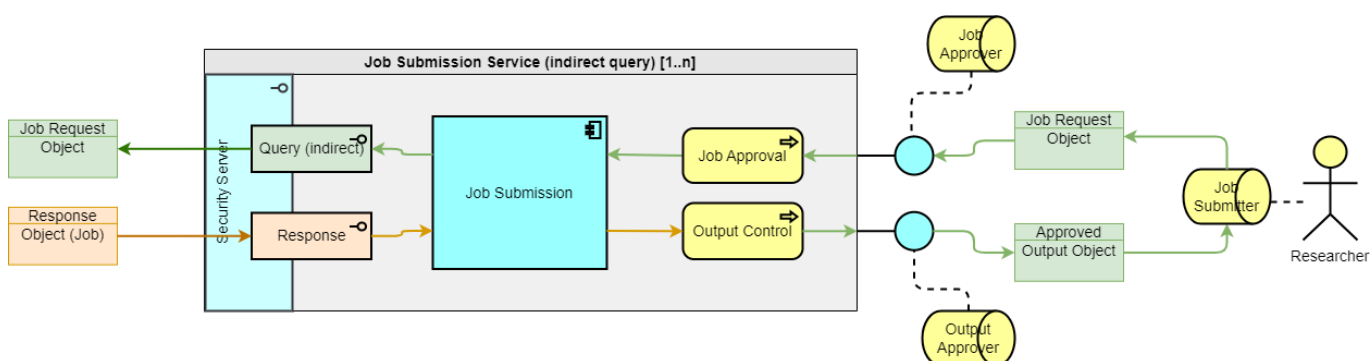
FOR CONSULTATION & COMMENT

The Federation may include many Discovery Services, perhaps specialising in different kinds of data.

A Discovery Service which enables dynamic discovery of metadata by querying other Federation services **MUST** support the outgoing **Query (direct)** and incoming **Response** interface types. Because Query interface services encompass a range of capabilities, Discovery Services are not restricted to static lists of metadata. They can range from simple high-level data or service discoverability to dynamic cohort discovery and “Beacon-like” services [51].

This dual “inward-outward” facing role will need careful security design; any outward-facing catalogue, for instance, **MUST** be air gapped or otherwise isolated from any other zone within the service. We model this with an **Output Control** process on the outward-facing interfaces.

4.3.4. Job Submission Service



A Job Submission Service combines the inward-outward facing nature of a Discovery Service with the indirect query capability of an RAZ. Job Submission Services are Federation Participants in their own rights, independent of any one TRE.

A Job Submission Service receives job requests from Job Submitters. These requests may need to be approved before being executed and so **MUST** pass through a **Job Approval** process overseen by a Job Approver role.

Approved job requests shall be passed to a **Job Submission** component which shall package them into standard Job Request Objects, forward them to the requested TREs and handle the Job Response Objects as they are returned. Handling the responses may involve composing or assembling them into a unified output object (e.g., aggregating the partial results from a federated query).

NB: how the requested TREs are made aware of job requests is undefined at this stage. They might choose to poll Job Submission Services that support a (currently undefined) polling interface, meaning that every TRE in the Federation might need to poll every Job Submission Service regularly. Or they might “listen” on their QMZ’s incoming Query (indirect) interface, requiring this interface to be open to incoming traffic from other Federation services.

Any unified output object **MUST** pass through an **Output Control** process overseen by an Output Approver role before it can be returned to the originating Job Submitter.

A Job Submission Service **MUST** support the outgoing **Query (indirect)** and incoming **Response** interface types.

FOR CONSULTATION & COMMENT

4.3.5. Software Service

A Software Service provides access for Federation participants to sources of software from outside the Federation.

A Software Service may:

- act as a direct network proxy for Internet-based third-party software services (e.g., CRAN⁹);
- act as an independently curated, high-assurance mirror service for popular software packages (e.g., Anaconda Python Enterprise¹⁰);
- act as a proxy for defined and approved user accounts on a public open-source software repository (e.g., GitHub¹¹);
- act as a proxy for Researcher workflows or analytical scripts stored in external repositories (e.g., WorkflowHub¹²) to be used as payloads for indirect queries;

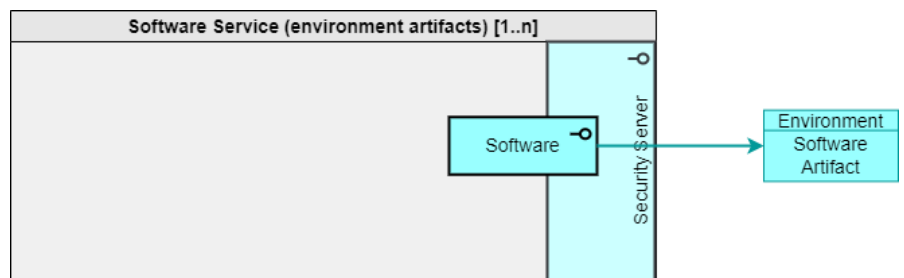
and so on.

Software Services MUST support the **Software** interface type.

As suggested, the Federation may have many Software Services, some specialising in particular kinds of software, language packages and so on. Two kinds are described here.

4.3.5.1. Environment Artifacts

To provision and configure Project Environments a Project Config service within a TRE's Research Analytics Zone should connect to a Software Service (environment artifacts). This service shall act as a proxy to approved sources of "environmental software" from which to build Project Environments – a Harbour repository of assured Docker containers; an approved source of Python packages, etc.



A Software Service (environment artifacts) supplies Environment Software Artifacts to requesting TREs.

⁹ The Comprehensive R Archive Network. See <https://cran.r-project.org/>

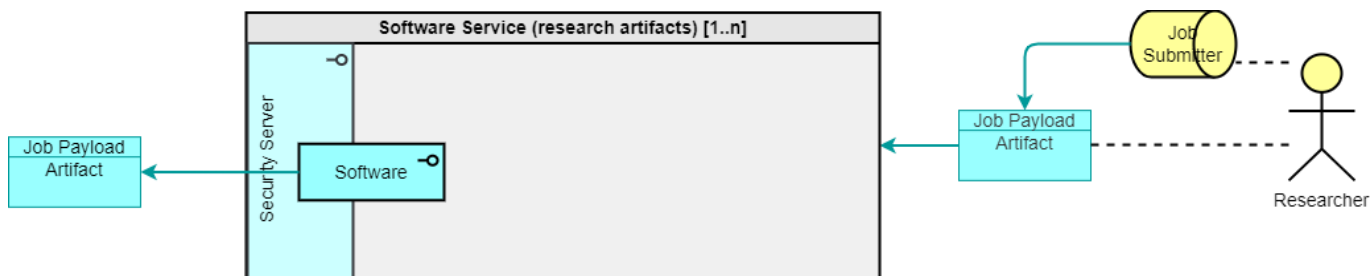
¹⁰ Anaconda Python Enterprise DS Platform. See <https://www.anaconda.com/products/enterprise>

¹¹ GitHub. See <https://github.com/>

¹² WorkflowHub. See <https://workflowhub.eu/>

FOR CONSULTATION & COMMENT

4.3.5.2. *Research Artifacts*



Indirect queries sent as Job Request Objects from Job Submission components within TREs or Job Submission Services include “pointers” to external analytics objects held in repositories, rather than actual query payloads. A Software Service (research artifacts) acts as a proxy for such external repositories, handling requests from components within TRE Query Management Zones and returning the requested research artifact (workflow, container or script, as examples) as a Job Payload Artifact.

4.4. Interface types

Interface services expose various capabilities for use by other members of the Federation. Note that traffic to and from all interface services route first through the Security Servers of the host Participant (q.v.).

At this level of the architecture we do not specify the details of individual interface calls but instead classify interface services into a small number of types, each of which will have a defined security context. We leave open the definitions of particular interfaces to promote innovation and expansion within the Federation, while providing an overall framework within which services can be placed. For example, an interface service that moves datasets between TREs MUST NOT be usable by Researcher system actors.

Note also that we use the terms “incoming” and “outgoing” to mean “incoming from another Federation Participant” and “outgoing to another Federation Participant”. Interface services do not connect Federation Participants to the wider Internet.

4.4.1. Query (Direct)

The Query (Direct) interface type supports queries between TREs and other Federation services. These interfaces produce and consume Query Objects, where the executable part of the query is fully contained in the object payload.

Query (Direct) interface services MUST connect solely to other Query (Direct) interface services.

4.4.2. Query (Indirect)

The Query (Indirect) interface type supports queries between TREs and other Federation services. These interfaces produce and consume Job Request Objects, where the executable part of the query is not contained in the object payload itself but is instead hosted in an external repository and only referred to by the Job Request Object.

Query (Indirect) interface services MUST connect solely to other Query (Indirect) interface services.

FOR CONSULTATION & COMMENT

4.4.3. Response

Responses are data generated by the execution of a query across several Federation services, whether direct or indirect. The model of federated queries assumed in the SDRI is entirely asynchronous and so we make a clear distinction between query and response interface types and do not assume a synchronous interaction between them. In practice, implementations of direct queries are very likely to require tight coupling between query and response interfaces (e.g., the coherent representation of remote datasets in a “single pane of glass” within a TRE Project Environment).

Response interface types produce and consume Response Objects.

The invocation of a Response interface service is triggered indirectly by the prior invocation of a Query service. Our working assumption is that, within the network of trust created by the SDRI Federation, responses can be returned safely to a querying entity without the need for IG intervention.

Response interface services **MUST** connect solely to other Response interface services.

4.4.4. Data Ingress and Data Egress

In contrast to returning query results, Data Ingress and Egress services move complete sensitive Datasets (or large extracts of Datasets) between Federation Participants. This places them in a different security context to query/response interfaces. Data Ingress and Egress services must only be accessible to TRE Governance actors in Data Manager roles.

Data Egress services produce Data Extract Objects, which Data Ingress services can consume.

Data Egress services **MUST** connect solely to Data Ingress services.

Conversely, Data Ingress services **MUST** connect solely to Data Egress services.

System actors with Data Manager roles **SHALL** be able to invoke Data Ingress/Egress services.

System actors without Data Manager roles **SHALL NOT** be able to invoke Data Ingress/Egress services.

4.4.5. Index

Index interface services provide a mechanism for TRE Governance roles and Index Services to exchange lists of personal identifiers, corresponding lists of depersonalised identifiers and master linkage spines for different Datasets. For more information see Section 4.3.2 *Index Service*.

Index interface services **MUST** connect solely to Index interface services.

As with Data Ingress/Egress services, system actors in Data Manager roles **SHALL** be able to invoke Index services.

System actors not in Data Manager roles **SHALL NOT** be able to invoke Index services.

4.4.6. Software

Software interface services provide a mechanism for TRE Operator roles to download and import approved software from a Federation Software Service. As described under Software Service, this may

FOR CONSULTATION & COMMENT

include environment software such as system components, standard analytics runtimes and packages, or research artifacts developed by Researchers and invoked via indirect queries.

Software interfaces produce and consume objects which encapsulate the approved software artifact.

Software interface services **MUST** connect solely to Software interface services.

System actors in TRE Operator roles **SHALL** be able to invoke Software interface services.

System actors not in TRE Operator roles **SHALL NOT** be able to invoke Software interface services.

4.4.7. Sync

Sync interface services provide a mechanism to maintain synchronisation of configuration state for Project Environments between multiple TREs. Details will be quite implementation specific but it is possible to model some general features.

Sync interfaces produce and consume Project Sync Objects which encapsulate the required configuration state.

Sync interface services **MUST** connect solely to Sync interface services.

System actors in TRE Operator roles **SHALL** be able to invoke Sync services.

System actors not in TRE Operator roles **SHALL NOT** be able to invoke Sync services.

4.5. Structured data objects

Participants in the Federation communicate by exchanging structured data objects over a common data exchange layer. The common data exchange layer provides the required technical security controls for exchange between Participants (see Section 4.6.2 *Security Server*) but additional security controls may be applied to certain types of objects.

Certain object types are closely associated with certain interface service types (see Section 4.4 *Interface Types*) and are produced and consumed by those interface services. Others are associated with Federation security control and are produced and consumed by underpinning security services.

The contents of structured data objects will depend on the particular interface services that produce or consume them.

The Federation requires that all objects to be exchanged between Participants **MUST** be packaged in a standard way. In this regard, we suggest the use of the “Five Safes” RO-Crate standard as the packaging format for all structured data objects in the Federation (cf. Section 5.2.4 and footnote 17).

4.5.1. Data Extract Object

Data Extract Objects are datasets or subsets of datasets that have been approved by a Data Controller for specific uses within the Federation. Data Extract Objects will typically contain sensitive data, often de-identified but individual-level personal data. Data Extract Objects are exchanged by TREs via Data Egress and Data Ingress interface services. Use of Data Ingress and Egress interface services must be restricted to TRE Governance actors in roles of Data Manager.

FOR CONSULTATION & COMMENT

Data Extract Objects are designated “SDC Red” in the architecture, meaning that, were they to be offered as candidates for release to the outside world, they would attract the most stringent statistical disclosure checks and would likely fail them. We reiterate, however, that the SDRI Federation is, by design, a closed environment and Data Extract Objects are only ever exchanged between TREs. Nevertheless, exchange of Data Extract Objects will require approvals from Data Controllers (in their roles as Data Custodians) to be in place, and **MUST** be overseen by TRE Governance Data Managers.

4.5.2. Index Object

Index Objects are exchanged by TRE Data Managers and Index Services via Index interface services. Index Objects do not contain sensitive data but could be said to contain “sensitive metadata”. Indexing individuals means that Index Objects will contain lists of personal identifiers and their exchange must be governed accordingly.

Index Objects are needed for certain kinds of data linkage. See Section 5.5.4 *Data Linkage* for a fuller treatment.

4.5.3. Query Object

Query Objects encapsulate direct queries and are produced and consumed by the Query (direct) interface type.

Direct queries can originate from Project Members working in Project Environments within a TRE, or from Discovery Services external to any TRE. In both cases they are targeted at one or more data resources remote from the calling service (i.e., hosted by another TRE).

Where a query originates from a Project Member the Query Object **MUST** contain enough information for the receiving TRE to be able to make the necessary authorisation decisions. This information includes, but is not limited to:

- Project Identity, in a form recognisable by the receiving TRE, indicating the project context this query is in;
- Project Member Identity, in a form recognisable by the receiving TRE, indicating who submitted the query;
- The target Dataset or Data Extract, in a form recognisable by the receiving TRE.

Where a query originates from a Discovery Service without an obvious Project context, how it is handled becomes a governance question to be codified in the Federation rulebook.

Query Objects are exchanged by Query (direct) interface services. Query Objects contain the full executable query for the remote data resource (e.g., as an SQL statement) and are not expected to contain sensitive data. In the architecture they are designated “SDC green”, meaning no form of output control is necessary before they can leave their originating environment.

A significant caveat to this last point arises where Query Objects might encapsulate partially trained deep neural networks in a federated machine learning setting, in which case they would be extremely likely to be sensitive at certain stages.

Again, though, we reiterate that Query Objects are exchanged between Federation Participants and not with the “outside world”. Thus, like all other structured data objects described here, their confidentiality,

FOR CONSULTATION & COMMENT

integrity and traceability are guaranteed by the secure data exchange layer common to all Federation Participants.

Note that we use “query” in a broad sense to encompass both the trivial (a microservice API call) and the complex (an encapsulated SQL script). **In all cases, though, everything the receiving TRE needs to execute the query and create an appropriate response is encapsulated in the Query Object.**

4.5.4. Job Request Object

Job Request Objects encapsulate indirect queries and are produced and consumed by the Query (indirect) interface type.

Indirect queries originate from Job Submission components, originated either by Project Members working in Project Environments within a TRE, or from Job Submission Services external to any TRE. In both cases they are targeted at one or more data resources remote from the calling service (i.e., hosted by another TRE).

As with direct queries, where the job request originates from a Project Member the Job Request Object **MUST** contain enough information for the receiving TRE to be able to make the necessary authorisation decisions. This information includes, but is not limited to:

- Project Identity, in a form recognisable by the receiving TRE, indicating the project context this query is in;
- Project Member Identity, in a form recognisable by the receiving TRE, indicating who submitted the query;
- The target Dataset or Data Extract, in a form recognisable by the receiving TRE.

Where a job request originates from a Discovery Service without an obvious Project context, how it is handled becomes a governance question to be codified in the Federation rulebook.

Job Request Objects are exchanged by Query (indirect) interface services. Job Request Objects do not contain executable payloads but instead contain “pointers” to executable artifacts held in external repositories (e.g., the URL of a CWL workflow)¹³.

As with Query Objects, Job Request Objects are not expected to contain sensitive data and are designated “SDC green”, meaning no form of output control is necessary before they can leave their originating environment.

4.5.5. Job Payload Artifact

Job Payload Artifacts encapsulate the executable artifacts referenced in Job Request Objects – the workflows, containerised applications or scripts prepared by Researchers in their role as Job Submitters and deposited in Internet-hosted repositories of some kind.

The artifacts themselves are retrieved from their repositories by Software Services which then package them into Job Payload Artifacts and return them to the requesting TREs via the Software interface.

¹³ It is not necessary that the TREs receiving a Job Request Object be able to resolve these payload URLs. Instead, TREs will request the payload artifact from a known, trusted Software Service (research artifacts) (or an internally cached version of same), and will receive in return a Job Payload Artifact object.

FOR CONSULTATION & COMMENT

Job Payload Artifacts MUST be subject to a receiving TRE's Job Approval process and MUST encapsulate sufficient information to enable the receiving TRE to assess their safety, in terms of the acceptability of their risk of execution. Because of these requirements it is possible, if not likely, that Job Payload Artifacts will be retrieved by TRE operations ahead of time, subjected to Job Approval and, if approved, cached locally within the TRE's Artifact Cache in readiness for matching indirect queries. It is thus not safe to assume there is a synchronous connection between receipt of a Job Request and retrieval of a Job Payload.

4.5.6. Response Object

Response Objects encapsulate the “answers” to queries submitted to TREs and are produced and consumed by the Response interface type.

Response Objects SHOULD have the same encapsulation structure for direct queries and indirect queries.

Response Objects may well contain data of high sensitivity: a direct query equivalent to “`SELECT * FROM Person_table`” will result in a Response Object equivalent to a Data Extract Object, for instance. In the architecture they are designated “SDC amber” but what level of oversight would be needed before a Response Object can leave its environment will depend on the context in which it was created. There are two scenarios we should consider.

1. Response Objects created in response to queries from an approved Project cannot, by definition, include data not already authorised for use by the Project Members. In this case Response Objects will either be returned to a Project Environment within a TRE, or to a secure Job Submission Service with an Output Control process in place. In neither sub-case is onward dissemination to the “outside world” possible without passing the Project's approved disclosure control.
2. Response Objects created in response to queries from a Discovery Service do not have an equivalent Project context, and are destined, by construction, to be disseminated to the “outside world” (this is a *Discovery Service*, after all). They must be handled differently, almost certainly handed directly to the Discovery Service's Output Control process.

4.5.7. Environment Software Artifact

Environment Software Artifacts encapsulate software artifacts used to construct Project Environments and are exchanged by Software interfaces.

In constructing and configuring Project Environments, TREs, rather than “downloading from source”, SHOULD request software artifacts from a Software Service. Not only does this provide an audit trail (the Software Service is a Federation Participant with a Security Server) but it also enables the Software Service to augment the software artifact with additional metadata and encapsulate everything in the Environment Software Artifact object.

4.5.8. Project Sync Object

Project Sync Objects encapsulate information about required Project-Environment configuration state and are produced and consumed by Sync interface types.

FOR CONSULTATION & COMMENT

4.6. SDRI core services

Core Services are a number of common services that together define the SDRI Federation. They include a set of Federation Services and a number of distributed Security Servers, one per Federation Participant.

All Core Service **MUST** be connected in a secure network which is independent of the Federation data exchange network.

4.6.1. Federation Services

The SDRI's Federation Services provide the coordinating functions and gatekeeping, registration and discovery services which, taken together, define the SDRI Federation. The lowest level of the Federation layer is agnostic towards both the nature of any exchanged objects and the purposes for which they are exchanged (see *Structured Data Objects* above).

There is only one set of Federation Services.

Federation Services **MUST** be highly available.

4.6.1.1. Accounting

Accounting services provide the means to track and record resource use across the Federation. In scenarios where remotely-executed queries may become complex, long-running workflows, a view of what costs are incurred where will become important.

4.6.1.2. Management

Management services provide the necessary tools for the operators of the Federation to maintain and run it to its agreed levels of service.

Management services **MUST** support mechanisms to ensure Security Servers across the Federation are up-to-date and synchronised with the currently agreed and approved global configuration.

4.6.1.3. Monitoring

Monitoring services include infrastructure monitoring for service availability and general system health and operational monitoring of the data exchange layer to ensure the necessary levels of confidentiality, integrity and auditability are being met.

4.6.1.4. Registry

Registry services record information about the different pieces of the Federation. There are a number of key records that **MUST** be recorded in the Registry:

- Federation Participants. Which Participants, defined by their security servers (qv), are part of the Federation. There are five kinds:
 - TREs;
 - Job Submission Services;
 - Software Services;
 - Discovery Services;
 - Index Services.

FOR CONSULTATION & COMMENT

- Datasets. Datasets are provided by Data Custodians and made available for use in TREs.
 - See Data topics later.
- Projects. In Federation terms Projects provide contexts which encapsulate Researcher users and Datasets into approved pieces of work.
- Users. Each and every user of the federation must be registered.

4.6.1.5. Trust

Trust services provide the necessary services for securing the foundational data exchange layer of the Federation. These services support the key security requirements of confidentiality, integrity, non-repudiation and availability. Trust services may include timestamping, encryption key management, security certificate management and so on.

In any implementation, trust services may be provided by trusted third-party suppliers¹⁴.

4.6.2. Security Server

Security servers act as the gateways of every Federation Participant and are the only components of the Federation that interact directly with each other and with the other Federation Services. The security features required of a Federation Participant are as far as possible abstracted into the Security Server. In particular the Security Servers provide the agency for the secure data exchange layer and hence are the guarantors of the confidentiality, integrity and auditability of inter-Participant exchanges within the Federation.

Every Federation Participant **MUST** run a Security Server.

Security Servers **MUST** operate to an agreed and approved global configuration.

Security Servers **MUST** support a mechanism to synchronise their configuration with the agreed global configuration.

If control-plane connectivity to Federation Management Services is interrupted, Security Servers **MUST** be able to continue operating independently.

4.7. Related concepts

4.7.1. Projects

The Project is a key concept in the use of the SDRI Federation. A Project defines a context for an approved research activity, including the Project Members involved, information about the data they are authorised to use, the TRE that hosts it, its duration and so on. A Project defines an authorisation context which provides a key piece of information for overall SDRI governance (cf. Chapter 6).

All Projects **MUST** be registered with the Federation's Registry services. An example of the kind of metadata required in a Project's Registry entry is offered in Section 5.2.3.1 *Project metadata*.

¹⁴ For a good discussion of trust services in the context of the UK eIDAS regulation, see the relevant pages at the UK Information Commissioner's Office, <https://ico.org.uk/for-organisations/guide-to-eidas/>.

FOR CONSULTATION & COMMENT

Simple Projects, typical of most current projects across the UK TRE landscape, will follow the data pooling pattern of access (Section 2.2.2). They involve one TRE with a Research Analytics Zone (the host), a number of TREs acting as data providers and, potentially, a trusted third-party Index Service.

More complex Projects will follow the federated analytics pattern (Section 2.2.3) and involve direct and indirect queries across multiple TREs with Query Management Zones capable of processing incoming query objects. For the purposes of governance and authorisation context, one TRE MUST be designated as the “host” or “instigator” of the Project.

The most complex Projects will potentially require a mix of data pooling—perhaps in an initial exploratory or development phase—and federated analytics—a “full production run” across remote data. For such Projects, one TRE should be designated as the host for the data pooling phases and, by construction, the “host” for the Project overall. This complex pattern anticipates large-scale federated machine learning across complex datasets (such as medical image stores).

4.7.2. Federation identities

Many elements of the SDRI Federation will have an *identity* and a number of *attributes* that can be used by system components and other system actors to reason about them. For example, a research user could have an identity and an associated list of active projects of which they were a member. Taken together, this information could be used by a remote data provider to decide whether or not to allow a query from that user to run in a particular project context.

These “Federation identities” must be unique within the Federation but do not necessarily need to have meaning outside the Federation. For the user example, the user’s Federation identity could be implemented as an SSO Token, for instance. This is further discussed in Section 4.7.3 below.

Implementation details are not dealt with here, but the table illustrates some of the required identities and some possible attributes for them. Attributes like this should be captured and recorded in metadata (cf. Section 5.2).

| Identity type | Example attributes |
|-----------------------------|---|
| Participant | Name; List of interfaces supported; List of capabilities accessible to the Federation; etc. |
| Researcher / Project Member | Name; Home institution (organisation vouching for their bona fides); Home TRE (TRE vouching for their access to the Federation); List of projects they are currently associated with (“currently” requires each membership be time-bound); etc. |
| Project | Name; List of current members (using their Federation identities; again, “current” requires these be time-bound); List of datasets associated with the project; Agreed disclosure control strategy; etc. |
| Dataset | Name; Data controller; Home service (Federation identity of the service regarded as the canonical source for this dataset); etc. |
| Data Extract | Name; Data controller; creation criteria (e.g., cohort definition); etc. |
| Linkage Spine | Identity of associated project; List of identities of associated datasets; etc. |

FOR CONSULTATION & COMMENT

4.7.3. Authentication and authorisation

The authentication of Researchers' identities and their subsequent authorisation to access Projects, Datasets and other Federation resources are split into two stages. This two-tier approach is not uncommon in large-scale federated environments (cf., for example, Appendix III of the *Architecture Vision* of the proposed EU Smart Middleware Platform [23]). To support a rich ecosystem of participants deploying different technology stacks, it is also necessary.

The sequence of events runs like this.

1. Two TREs establish a trust relationship, brokered by the Federation Services and using the Federation's foundational trust services. This "server to server" trust relationship is a standard approach to securing services across the Internet and is typically implemented using X.509 certificates and a public key encryption infrastructure. (We do not cover the details here.) At a foundational level, this is what joining the Federation as a Participant means.
2. A Researcher then authenticates themselves to "their" TRE using the TRE's locally preferred authentication mechanism. This may be Microsoft Active Directory, Linux LDAP/X509, OpenID Connect or a number of other technologies. The TRE may support more than one authentication mechanism for different kinds of user identity (federated identity management).
3. The authenticated Researcher's local identity is mapped onto an internal Federation identity using a common format which all participants in the Federation agree to support. Attributes associated with this identity can then be used by other Federation participants to reason about the Researcher, to make, for instance, authorisation decisions about granting the Researcher access to Projects, Datasets or other resources (single sign-on).

This division also helps enforce the principle of "no TRE, no data": Researchers access Datasets only through TREs, never directly. It also follows from "start from where we are" and "a standards-based ecosystem", allowing TREs to continue to serve their user communities in the best way while providing common back-office connections to federated resources.

FOR CONSULTATION & COMMENT

5. Federated architecture: data layer

In this chapter we discuss the data layer of the Federation from the angles of metadata and the FAIR principles of findability, accessibility, interoperability and reusability.

5.1. Classifying sensitive data

There is no generally agreed definition of “sensitive data”. Most working classifications are built around three considerations: the subject of a given dataset; the organisation responsible for custody of a given dataset; and the potential harm, to either subject or custodian organisation (or both), from unauthorised disclosure of the dataset.

The nature of a dataset’s subject often requires a particular legal or regulatory approach to classification. In the UK, for example, data about living natural persons is covered extensively in the UK GDPR [36]. A firm’s intellectual property may fall under the Copyright Designs and Patents Act [42]. Where the data subject is an endangered species, its treatment may be covered by international treaty such as CITES [43]. Still other subjects may require certain approaches because of cultural sensitivity¹⁵.

Organisations responsible for collecting or holding potentially sensitive data typically apply their own classification criteria. As responsible custodians, the impact of unauthorised disclosure will likely fall on them, making good data classification part of good corporate risk management practice.

In the interests of manageability, organisational risk management approaches tend to aim for a handful of sensitive data classes only. UK Government (and the US Government) apply three [44] (OFFICIAL, SECRET and TOP SECRET), or four if the UK’s OFFICIAL-SENSITIVE is counted separately. The International Information System Security Certification Consortium (ISC)² defines five in its standard commercial scale [45]. Work at the Alan Turing Institute has developed a five-tier classification model [46]. The NHS in England has an extensive example-driven list of over a dozen but these map onto just two on the UK Government scale [47].

The principal reason for an organisation to classify sensitive data is to help it decide how to manage them. This makes it possible to divorce the “why” from the “how”: why a particular dataset has been classified as “sensitive” doesn’t matter when it comes to storing and protecting it as a sensitive dataset. This is the approach taken in the Harvard Datatags system [48].

5.1.1. A seven-point scale

DARE UK aims to facilitate the combination and linkage of datasets from any and all possible sources. Linked data typically carry higher disclosure risk than their individual constituents, so some comparative scale will be useful. We recommend that datasets used within the SDRI Federation be recorded with two key pieces of information and a number from 1-6 on a “scale of harm”.

In assessing risk of harm, we assume that any unauthorised disclosure of data brings the chance of the data falling into the hands of someone in a position to cause harm to either the data subject or data

¹⁵ For example certain world cultures have, over the years, expanded traditional taboos on naming the recently deceased in speech to include electronic recordings, including digital photographs. See https://en.wikipedia.org/wiki/Taboo_on_the_dead and references within.

FOR CONSULTATION & COMMENT

custodian. Thus, we do not distinguish between data release to a small group and data release to everyone.

Datasets should be classified by:

- Data subject (what it’s about): individuals; firms; locations; intellectual property; ...
- Data custodian (who’s responsible for sharing it);
- “Harm”, which can mean physical, mental, emotional, economic or reputational, depending on the context.

| Category | Harm | UK Gov | GDPR | (ISC) ² | Turing |
|----------|---|--------------------|------------------|--------------------|--------|
| | None | Public | Public | Public | Tier 0 |
| 1 | Inconvenience | - | - | Proprietary | Tier 1 |
| 2 | Distress, embarrassment, some reputational damage | OFFICIAL | Personal | Private | Tier 2 |
| 3 | Actual harm | OFFICIAL-SENSITIVE | Personal | Confidential | - |
| 4 | Serious harm | OFFICIAL-SENSITIVE | Special Category | Sensitive | Tier 3 |
| 5 | Loss of life | SECRET | - | - | Tier 4 |
| 6 | Widespread loss of life | TOP SECRET | - | - | - |

NB: It must be emphasised that data classification in this manner is not a simple badge-it-and-forget affair. The sensitivity of a given dataset (whether Dataset or Data Extract) can and will change depending on the context it is in. The classifications themselves are also something of a blunt instrument: “John has asthma” and “John has HIV” are both personal health data (GDPR Special Category), but one could cause far more harm than the other if disclosed. It is far better to use this kind of classification only as a starting point and always consider the use of sensitive data within a “Five Safes” context, managing risk holistically across a number of dimensions.

5.2. Federation metadata

Our concept of Federation metadata covers high-level descriptions of all the elements of the SDRI Federation, from the services that comprise its infrastructure to the data, users and projects that make it useful. It is descriptive of the Federation and its activities, and provides a very high-level view of data assets within the Federation, but does not include rich, detailed descriptions of these data assets. How best the technical infrastructure could support rich discovery and exploration of datasets from many different disciplines is a challenging question; we offer some thoughts in Sections 5.3, 5.4 and 5.5 below.

For now, we divide Federation metadata into three groups: metadata that capture information about the Federation itself (infrastructure metadata); metadata that capture information about the datasets accessible within the Federation (content metadata); and metadata that capture information about what purposes the Federation is being used for (we can call this governance metadata). By construction, these map to the three layers of the SDRI Federation.

FOR CONSULTATION & COMMENT

In general, the visibility of metadata—private to a Participant, private to the Federation as a whole, or public—should be determined and agreed by Federation governance rules, perhaps following a “need to know” approach. Some examples:

- Public: names of Participants in the Federation; names of Datasets available within the Federation; counts and names of active Projects; counts of active Researchers; ...
- Federation-private: Federation identities of Participants and other entities and artifacts; service capabilities; Project risk-management information; ...
- Participant-private: Researchers’ and other users’ contact details; ...

5.2.1. Infrastructure metadata

Our definition of infrastructure metadata is best captured by the answer to the question: if the Federation had no users at all, what metadata would we still need to describe it? We divide this further into static descriptive metadata that describe the Federation “at rest” and dynamic operational metadata that describe it “in motion”.

5.2.1.1. Descriptive metadata (Federation at rest)

The Participants – the services described in Chapter 4 – require machine-readable descriptions which shall be recorded in the Registry Services, and which provide enough information to be reasoned about (e.g., for the purposes of automation).

Examples of descriptive metadata are:

- Basic metadata: name, Federation identity, ...
- Capabilities: available computation; available software; ...
- Datasets hosted (persistently available not project-specific): count; list of Federation identities; ...
- Indexes hosted (types of linkage available): list of Federation identities; ...

Most descriptive metadata should be visible within the Federation.

Some may be visible publicly (meaning able to be published rather than exposed directly from within the Federation to the public Internet!).

5.2.1.2. Operational metadata (Federation in motion)

Operational metadata are metadata captured and recorded through the operation of the Federation and its Participants. Operational metadata notably include information on data exchange logged by the Participant Security Servers and by the Federation Services.

Clear governance rules must be established around the use of operational metadata. It must be clear, for instance, which metadata logged within a Participant’s Security Server are private to the Participant, which may be shared with Federation Services, and which might be visible to other Federation Participants.

No operational metadata should ever be visible to the public.

FOR CONSULTATION & COMMENT

5.2.2. Content metadata

Content metadata describe, at a high level, the Datasets the Federation supports. When structuring metadata to describe such concepts steps should be taken to eliminate or reduce any duplication of information that would risk drift, divergence or fragmentation.

5.2.2.1. Dataset metadata

Datasets, while treated as dynamic, are potentially persistent and long-lived. Dataset metadata should record information about the data themselves, including the Data Controllers accountable for their use, but not things like where they can be accessed. The latter information should be left to the hosting Participant to advertise, and to the Federation Registry and Discovery Services to collate for search and discovery purposes. For example:

- Dataset record:
 - Name: Covid-19 self-reported symptoms in London, 2020
 - Federation identity: ee6574ac-8ad7-440c-8200-ca86dd556bbf
 - Data controller: ...
- TRE record:
 - Name: SAIL Databank
 - Federation identity: 5756f2c9-c6f3-4fcf-8d81-c4647e2a12bb
 - Datasets hosted: {ee6574ac-8ad7-440c-8200-ca86dd556bbf; ...}
 - ...

The dynamic nature of datasets arises not from their ephemerality or their movement around the Federation but from their changeability. Datasets are updated (new entries made, old entries pruned) and their schemas or formats change (more slowly). How different versions of a dataset should be managed and recorded is out of scope, but we would recommend that its Federation identity remain unchanged, just as its name would.

Summary metadata for a Dataset will be public, perhaps conforming to a common high-level catalogue schema. As a current starting point for defining the required fields in these high-level metadata records we would recommend Appendix A of the UK Statistics Authority DEA Data Capability Guidance [49]. We use this to derive the example metadata records below, our goal being to design defensively and align Federation metadata as closely as possible with anticipated governance or accreditation requirements¹⁶. Particular data domains may, of course, introduce their own standards, and commonality will need to be distilled and agreed accordingly. (In the health domain, for example, the HDR Alliance have defined a useful standard for data use registers [50].)

Some detailed Dataset metadata will be Federation-private.

¹⁶ The UK Statistics Authority Digital Economy Act scheme for UK-based processors of statistical data is a rigorous approach to accreditation but does not cover health-related data. However it has been announced (June 2023; see <https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/data-saves-lives-implementation-update/>) that the UK NHS and Statistics Authority will work together to co-design an updated version of the DEA scheme suitable for both statistical and health data.

FOR CONSULTATION & COMMENT

| Example metadata record: Dataset | |
|----------------------------------|---|
| Id | A unique Federation identity number for the Dataset. |
| Data name | A unique name provided to identify the Dataset. |
| Data description | A short description. |
| Data classification | The type of data (perhaps using a controlled terminology such as Dublin Core, eg, household survey data, administrative data, open data). |
| Data keywords | A set of related keywords. |
| Data supplier | The owner or supplier of the data. For personal data this should be the data controller. |
| Time coverage – start | The first point in time the data covers. |
| Time coverage – end | The latest point in time the data covers. |
| Data frequency | Where the data have a temporal frequency. |
| Update frequency | Where data are updated in their hosting provider environment. |
| Geography | The levels of geography included in the data. |

5.2.3. Governance metadata

What we term governance metadata covers the users of the Federation and the activities they carry out. Central to this idea is the concept of the Project: any and all research activities across the Federation are conducted within the contexts of Projects.

Governance metadata should be viewed as a machine-readable form of the record-keeping required of TREs, data providers and researchers under research approval and accreditation regimes. As with Dataset metadata above, we recommend making use of prevailing information governance requirements to drive the metadata standards within the Federation. Where multiple accreditation regimes exist a degree of harmonisation or duplication will be required in metadata records (cf. footnote 16 on previous page). As before, we use the current DEA standard to derive the example metadata records below.

5.2.3.1. Project metadata

As discussed in Chapter 4 the Project is a strong concept within the Federation. Projects are conceived outside the Federation and, once approvals are in place, are instantiated in a hosting TRE. At the point of Project instantiation, the hosting TRE should register the Project’s existence with the Federation Registry Service.

A Project’s metadata should encapsulate its scope including its hosting TRE, the Datasets or Data Extracts it has permissions to work with, the Researchers permitted to work on it, its start and end dates and so on. It should be detailed enough that authorisation or disclosure decisions can be taken by Federation Participants, for example upon receipt of a remote query.

Most metadata for a Project will be public.

Some detailed Project metadata may be Federation-private, and some may be Participant-private (e.g., held by the instantiating TRE).

FOR CONSULTATION & COMMENT

| Example metadata record: Project | |
|---|--|
| Id | A unique Federation identity number for the Project. |
| Project title | The official Project title as approved . |
| Project abstract | A short paragraph summarising the purpose of the Project. |
| Expected public benefits | A short paragraph summarising the expected public benefits. |
| Project keywords | A set of keywords describing the Project. |
| Project start date | The date this Project started in the hosting TRE. A Project is considered to start in a TRE when Researchers have access to the TRE and all data as approved in the Project application. |
| Project end date | The expected end date of the Project. |
| Host research environment | The name of the hosting TRE where research will take place. In the case of a Project involving federated analytics this should be the TRE which instantiates the Project. |
| Research environments | A list of any and all other TREs involved in the research—for instance in the case of a Project involving federated analytics. |
| Research sponsor | The name(s) of the organisations sponsoring this research. |
| Project approval on | The date this Project was approved by its governing authority. |
| Ethical approval on | The date ethical consideration/approval was given to this Project. |
| Ethical approval by | Who provided ethical approval for the Project. |
| Ethical restrictions | Any restrictions identified as part of the ethical approval. |
| Research Lead | The Federation identity for the lead Researcher (often termed "principal investigator" in academic projects). |
| Researchers | A list of Federation identities for all other Researchers on the Project. |
| People restrictions | Any restrictions on the people involved in this project identified as part of the Project accreditation. |
| Data used | A list of Federation identities for all Data Extracts used in this Project. |
| Data restrictions | Any restrictions on the data available to the project identified as part of the Project accreditation. |
| Dissemination restrictions | Any restrictions on the dissemination of research outputs identified as part of the Project accreditation. |

5.2.3.2. *User metadata*

Researchers within the Federation may be, and indeed should be able to be, involved in multiple Projects concurrently. Not all of these Projects need be hosted by the same participating TRE; thus Researchers will need a registered Federation identity which is common across all Participants. This echoes current best practice in DEA-accredited TREs where a researcher’s “identity number” is provided centrally by the UK Statistics Authority (the accreditation authority for DEA standards).

The primary reason we classify metadata for Researchers (and other users such as TRE Operators or Data Service Operators) under “governance” is that best practice captured in the Five Safes phrase “Safe People” requires all users or handlers of sensitive data to be trained or accredited to an acceptable level.

FOR CONSULTATION & COMMENT

“Acceptable” here typically means acceptable to the governance authority concerned with the data in question. The DEA record required for a researcher places a strong emphasis on this aspect, suggesting the example metadata record below.

| Example metadata record: Researcher | |
|-------------------------------------|---|
| Id | A unique Federation identity number for the Researcher. |
| Full name | This should include any middle names as recorded in official documents. |
| Research affiliation(s) | Where the Researcher has affiliation to an organisation or organisations all these affiliations should be recorded. |
| Type of accreditation | Provisional/Full accreditation. |
| Training course | The name of the training course attended as part of the accreditation. |
| Course provider | The organisation responsible for delivering this course. |
| Trained on | The date the researcher attended the training course. |
| Assessed on | The date the researcher completed the assessment. |
| Accredited on | The date the researcher was accredited. |

5.2.3.3. *Data Extract Metadata*

We define Data Extracts as snapshots created from Datasets according to some query—a cohort definition, for instance.

Data Extracts are one kind of structured data exchanged between Participants.

Metadata for Data Extracts will be logged by the secure data exchange layer and so must prove useful in that context (e.g., for audit purposes). Attributes could include: Data Controller; “parent” Dataset; version or timestamp of parent Dataset at extract creation; etc.

5.2.4. Structured data packaging formats

While the contents of metadata records will be driven by governance requirements, the format into which they are packaged for exchange between TREs or other services is an entirely technical decision.

The 2023 DARE UK Driver Projects pioneered the development and use of a “Five Safes” profile of the international RO-Crate standard for structured data packaging¹⁷. RO-Crate (“Research Objects + DataCrates”) extends the BagIt file packaging format¹⁸ to include standard representation for machine-actionable metadata. The “Five Safes” RO-Crate profile adds additional metadata structure useful in the TRE context.

The “Five Safes” RO-Crate profile has been demonstrated as fit for purpose in prototype implementations of this architectural blueprint and so we suggest that “Five Safes” RO-Crate be adopted as the packaging format for all structured data objects in the Federation (cf. Requirement R123 in Appendix D).

¹⁷ See <https://trefx.uk/5s-crate/0.4/> and <https://www.researchobject.org/ro-crate/1.2-DRAFT/>

¹⁸ See <https://datatracker.ietf.org/doc/html/rfc8493>

FOR CONSULTATION & COMMENT

5.2.5. Other considerations

Many exchanges of structured data within the Federation will occur in a Project context: an initial Data Extract sent at Project instantiation (see above); a Linkage Spine created to connect extracts to create a Project's working dataset; a query, sent from a TRE to one or more remote data providers.

We RECOMMEND that all such exchanges of structured data objects be tagged with a metadata record indicating this Project context.

5.3. Data findability

The Federation "content metadata" records introduced in the previous section are examples of the types of information that need to be captured and recorded in the Registry services of the Federation, but are largely useless in helping researchers find what data might actually be available to support their research within the Federation. As described in Section 3.1 *Rachel's Journey* this data discovery needs to happen outside the Federation, before a researcher has even defined the project they might ultimately propose.

A consequence of this is that data findability, or discovery, is not a core use-case for the SDRI Federation. The Federation does, however, have a role to play in supporting data discovery where it can—maintaining a record of what datasets from which providers are available in which TRE with what linkages available—and ensuring that such information can be accessed usefully in standard ways from outside the Federation without compromising its secure perimeter.

The Federation architecture as proposed does permit the exposure, via query interfaces, of metadata from the Federation to the public Internet. By this statement we mean there is nothing proposed in the architecture that renders this impossible. Whether and in what form it might be realised is currently left as a question of governance and of implementation. Possible approaches to exposing public metadata from controlled environments can be found in the GA4GH Beacon work [51] and in the HDR-UK CO-CONNECT work [52].

5.3.1. Discovery metadata

The ELIXIR Ontology Lookup Service hosts 280 life-science ontologies¹⁹. The NHS list of approved national information standards²⁰ counts 90 standards and twice as many collections, while the NHS Data Model and Dictionary describes over 2,750 data elements²¹. The INSPIRE Technical Guidelines on metadata implementation for geospatial data run to 99 pages²².

Harmonising data discovery in such a landscape is simply intractable. The best we can hope for across a federation of data resources and analysis environments is to adopt common basic discovery metadata which is aligned with metadata standards used by the likely largest sensitive data providers. In our terms this means looking at catalogue-level standards mandated within UK Government and health services.

¹⁹ See <https://www.ebi.ac.uk/ols/ontologies>

²⁰ See <https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections>

²¹ See https://www.datadictionary.nhs.uk/data_elements/overview.html

²² See <https://inspire.ec.europa.eu/documents/inspire-metadata-implementing-rules-technical-guidelines-based-en-iso-19115-and-en-iso-1>

FOR CONSULTATION & COMMENT

From an architectural perspective the SDRI Federation is an “overlay” on top of Web standards, notably HTTPS, XML and JSON. Hence we favour “Web facing” formats for metadata over internally-oriented standards.

5.3.1.1. Recommended standards

Our three key reference sites for metadata standards are:

- Central Digital and Data Office: *Open standards for government data and technology*.
 - <https://www.gov.uk/government/collections/open-standards-for-government-data-and-technology>
- Department of Health and Social Care: *A guide to good practice for digital and data-driven health technologies* (particularly section 10).
 - <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology#section-10>
- Office for National Statistics: *Data Standards*,
 - <https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datastandards>

There are a number of common themes across these sources. For interoperability and easier linkage we recommend that new, born-digital data aim for compatibility with the standards below—with the caveat that these may change or evolve over time.

- General discovery metadata:
 - W3C DCAT data catalogue standard²³ and extended application profiles;
 - schema.org²⁴;
 - schema.org includes definitions for data catalogue and dataset drawn directly from DCAT;
 - the serialisation of schema.org markup into JSON-LD format provides a compact, machine-readable version ideal for data exchange and query results.
 - CSVW²⁵ for CSV files published on the Web;
 - DCMI Dublin core metadata²⁶ where there is no current match in schema.org.
- Place metadata:
 - UPRN²⁷ unique property reference number for addressable locations in the UK;
 - ETRS89²⁸ European terrestrial reference system for locations in Europe;
 - WGS84²⁹ world geodetic system for global locations.
- Date/time metadata:

²³ See <https://www.w3.org/TR/vocab-dcat-3/> for the current version 3 working draft.

²⁴ See <https://schema.org> and <https://www.w3.org/wiki/WebSchemas/Datasets> for its extensions from DCAT.

²⁵ See <https://csvw.org/>.

²⁶ See <https://www.dublincore.org/>.

²⁷ See <https://www.geoplace.co.uk/addresses/uprn/>.

²⁸ See <http://etrs89.ensg.ign.fr/>.

²⁹ See <http://earth-info.nga.mil/GandG/update/index.php?action=home>.

FOR CONSULTATION & COMMENT

- ISO-8601³⁰ for dates and times.
- Health-related metadata:
 - OMOP observational medical observations partnership standard³¹ for electronic health records and similar;
 - **NB:** while the DHSC guidelines cited above recommend NHS use of the HL7 FHIR standard³² for data interchange, the HDRA White Paper of 2021 [53] considers both and leans towards OMOP as a more appropriate data model for research use. Further, at time of writing, the NHS Data for Research and Development programme is settling on use of OMOP as a common standard for research-ready versions of health records and hospital observation data across its planned network of secure data environments (SDE, an NHS term synonymous with TRE)³³.
 - DICOM³⁴ image storage format for medical images.

5.4. Data accessibility

Easier and more streamlined access to sensitive data is the *raison d'être* of the DARE UK programme and of the Federation described here. We adhere strongly to the principle of “no TRE, no data”—data access in a secure environment over data distribution to researchers’ local systems—which offers a far greater degree of data security but does place some new restrictions on data access.

One particular consideration is “data understanding”. Most models of analysis for any datasets bar the very smallest introduce an “understanding” or “exploratory” step between discovery and full-blown production analysis. A good illustration of this is the CRISP-DM process³⁵, a widely-used industry standard dating back to the 1990s. It introduces both “business [domain] understanding” and “data understanding” as steps before “data preparation” and “modelling” but crucially emphasises the iterative nature of the process. These steps are cyclic, not serial.

The data access model of TREs introduces a hard serialisation into the end-to-end data research process, especially where information governance requires a researcher to request in advance of their project being approved only the data elements they will need to answer their particular research question. Without an initial exploratory phase that request can be difficult to get right.

A proper understanding of the “linkability” of two or more datasets can also be difficult to achieve without some level of access to both datasets in advance (see also Section 5.5 *Data interoperability* below). Full data harmonisation of this nature (especially across our broadest possible definitions of sensitive data) is out of scope for this architecture. However, the restrictions introduced by the “no TRE, no data” principle are worthy of consideration: are there changes at architectural level that could facilitate a data harmonisation step?

³⁰ See https://en.wikipedia.org/wiki/ISO_8601 for a good discussion.

³¹ See <https://ohdsi.github.io/CommonDataModel/> and <https://www.ohdsi.org/>.

³² See <https://www.hl7.org/fhir/summary.html>.

³³ NHS Data for Research and Development Technology and Data Working Group, *working documents*.

³⁴ See <https://www.dicomstandard.org/>.

³⁵ See <https://www.datascience-pm.com/crisp-dm-2/>.

FOR CONSULTATION & COMMENT

OpenSAFELY [28] have shown that, for certain kinds of well-structured data, the majority of the algorithmic development and data exploration work can be done outside a TRE, on “fake data” that match the sensitive data schema and terminology sets but which contain random values. OpenSAFELY couples this development stage with an indirect query job submission model to deploy a researcher’s analysis code into the TRE without needing to grant them as an individual any kind of secure access. The “fake data” development model could be extended to other data sources even if the actual analysis step were to follow the “traditional” TRE model of secure access over remote desktop.

Enabling this degree of data exploration (or at least schema exploration) could be supported by additional Discovery Services sitting on the edge of the Federation.

5.5. Data interoperability

So far within the architecture we have recognised the fundamental importance of data interoperability in the form of data linkage but our treatment has been deliberately naïve. There are multiple levels on which to consider data interoperability and most of these are out of the scope of a federated architecture. Nevertheless we note them here and may expand on them in future iterations.

5.5.1. Syntactic interoperability

The most straightforward level of interoperability is syntactic or schema-level: are the datasets to be connected the same shape in at least one of their dimensions? In the horizontally and vertically partitioned dataspace we introduced in Section 0 there are two strong assumptions:

- EITHER the datasets have the same set of data subjects in the same order (e.g., different sets of attributes about the same group of people, ordered the same way);
- OR the datasets have the same set of attributes in the same order (e.g., the same set of attributes about two different groups of people).

Connecting datasets by these criteria is reasonably straightforward; relational databases are very good at exactly this kind of thing. Even differences in the ordering are easy to manage, by sorting, for example. We may need to define rules to handle gaps in the resulting dataset (either common rules or context-specific ones) but again, this is a well-understood area.

It is feasible to imagine an Index Service which could automate the linkage of two datasets under these conditions.

5.5.2. Terminological interoperability

Simple syntactic joining becomes harder when two datasets are probably interoperable but have been put together with slightly different terms. For example:

- Surname; Christian Name; Age;
- Given Name; Family Name; Age;
- Nom; Prénom; Age.

Human experience tells us that these three datasets most likely record the same information (even with the transposition of name parts and dual languages in play). An equivalent level of experience for an automated service could be created using terminology bases, in much the same way that computer-

FOR CONSULTATION & COMMENT

assisted translation tools work today. (The proposed EU Smart Middleware Platform architecture includes just such a vocabulary service [23].)

By introducing one or more terminology services, it is feasible to imagine an Index Service which could automate the linkage of two datasets under these conditions.

5.5.3. Semantic interoperability

By far the most complex level of interoperability is semantic: two data items may have the same name but the way they were recorded might be very different. Different people, in different contexts, under different time pressures, might record nominally identical data items in subtly different ways which make them non-interoperable in ways almost impossible for an automated system to identify.

Another semantic variant arises in linkage between two or more datasets which each contain a number of data elements that, either alone or combined, mean *nearly* the same thing. Here, human intervention can harmonise the datasets, perhaps by introducing a new, common element, constructed differently in different datasets but which is nevertheless equivalent between them. Whether this kind of harmonisation could be achieved outside the TRE, working purely with dataset schemas and terminology sets (cf. Section 5.4), is likely to be highly case dependent.

It is difficult to imagine a scenario in which an Index Service could automate the linkage of two datasets under these conditions.

5.5.4. Data linkage

With the caveats noted above we have introduced a model of data linkage within the federated architecture which can, in principle, be automated (at least to some extent). Our model makes three design assumptions:

- Linkage between Data Extracts for a Project is done using a common linkage spine, which may be created explicitly for the Project or may be persistent.
- Linkage spines are created and maintained by Indexing Services which are trusted third-parties (“TTPs”) independent of TREs or a Project’s information governance.
- Identifiers used in the linkage spines are transformed as part of the linkage process into Project-specific identifiers. Such identifiers have no meaning outside the Project and thus cannot be used, by themselves, to link to anything else.

Linkage spines are exchanged between Federation Participants as structured documents.

5.6. Data reusability

Reusability in a sensitive data environment has to be balanced against governance principles which restrict use of data to pre-approved purposes only. We can draw two broad categories of reusability:

1. Reuse under original approvals. Assembled datasets and analyses derived from them (including computer programs) may result in a model for which evidence must be preserved for many years (for example clinical trials or medical devices). The datasets and analyses must be preserved in a way that could be checked and re-validated in the future, but all within the same purpose for which approvals were originally granted (and all within the same, or an equivalent, TRE). This then

FOR CONSULTATION & COMMENT

becomes the challenge of preserving long-term a digital object that is quite possibly encrypted. Specialised archive services could be developed that would do this (many already exist).

2. Reuse for new research. Whether a new research project—perhaps under a new team, perhaps linking in additional data—could be authorised to build on the full results of another is clearly a governance question. (By “full results” we mean the full linked data and analysis environment that remains within the TRE, not the summary results approved for egress.)

In technical terms, a service which preserved the TRE environment for the purposes in (1) would serve equally to support those in (2). We do not expand on the details of such a service here.

FOR CONSULTATION & COMMENT

6. Federated architecture: organisational layer

Chapters 4 and 5 have attempted to distil and write down the technical specifics of a federated architecture for TRE services – the “what” of the SDRI Federation. This chapter is much more open. Reaching agreement on an organisational model to manage the required new elements of standards and core services – the “how” – must be done through wider community processes.

The functionality required of a federated architecture implies a certain logical organisational structure, as captured in the preceding chapters. However, there is flexibility in how that logical structure could be realised in practice, depending on how the community of potential Participants might agree on its setup and operation.

To meet the public need for a more standardised, more trustworthy environment, the Federation needs to be real, in the sense of some kind of membership organisation with rules and standards. For the Federation to be real it will need a Federation Authority (FA) to act at least as gatekeeper and maintainer of standards.

The rules of participation for the Federation need to be agreed by all relevant stakeholders, and captured in a “Federation Rulebook”. The role of the FA then becomes one of maintaining the Rulebook and overseeing its implementation. The Rulebook should cover the “how to” for at least the following:

- Agree baseline technical standards for the Federation (as described in Chapters 4 and 5). This may involve defining or approving invitations to tender for technology suppliers of Federation services.
- Agree baseline procedures for key events: onboarding a new Participant; offloading a departing Participant; etc.
- Agree baseline maturity or accreditation standards for Federation Participants. This could involve setting minimum capabilities for new Participants accompanied by continual improvement plans towards nationally-agreed standards.
- Agree the setup and operation of trust services, trust anchors and frameworks – essentially who is able to vouch for and sign identity assertions made by Participants, and how.
- Agree the setup and operation of registry services.
- Agree baseline training or accreditation standards for Federation users, including service operators, Researcher PIs and other researchers.
- Approve new Participants joining the Federation.
- Approve Participants leaving the Federation. (This may be trumped by contractual arrangements arising from the joining process.)
- Approve technical changes with implications for, or impact on, part or the whole of the Federation, including:
 - changes to Federation standard software, for instance changes to Federation Services software;
 - changes to data exchange protocols or formats;
 - changes to metadata standards.
- Oversee regular audit and accreditation for the Federation as a whole.

Note that the governance focus of the FA is emphatically on what here is *new*: interoperability standards, service onboarding, coordinated change management and incident response. The existing stakeholders already have governance arrangements in place to enable research with sensitive data within TREs. The

FOR CONSULTATION & COMMENT

FA *should not* disrupt existing data governance arrangements for participants wanting to join, but should instead complement them.

6.1. Centralised vs distributed vs decentralised

How could the Federation Authority be realised?

Chapter 3 of the “IDSA Rulebook” [54] on the creation and operation of data spaces, published by the International Data Spaces Association, offers a good discussion of the pros and cons of centralised vs decentralised models of federated governance for Data Spaces. We adapt that discussion here for the SDRI Federation.

One way to group the key services required of an FA is as follows:

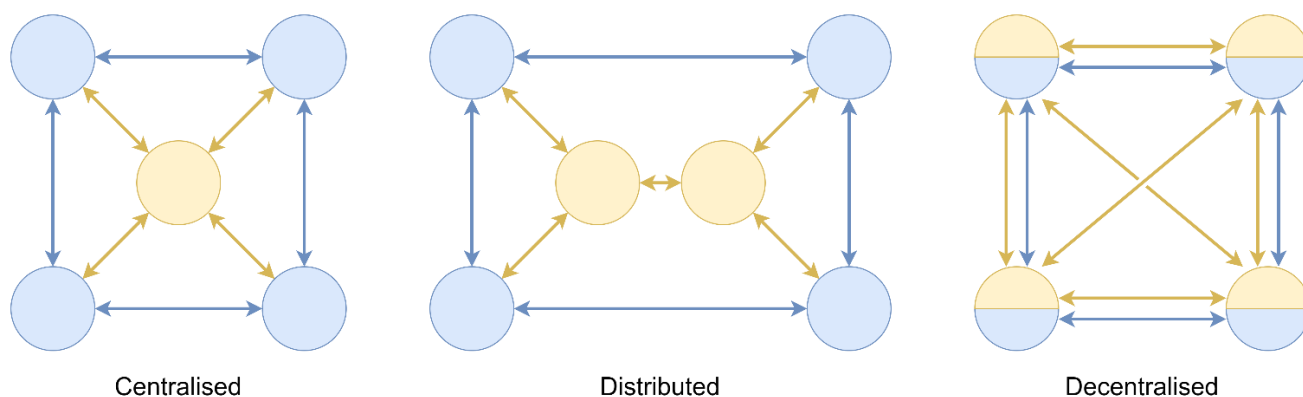
- Rules & Policies – underpinning agreements about what joining the Federation means, and technical implementations of them that enable digital handling.
- Trust, Identity & Certification – agents and methods for vouching for digital identities.
- Registry (Participants) – queryable records of who is a member of the Federation.

We could also add:

- Catalogue (data) – queryable records of what data assets could be accessible within the Federation, and how to go about applying for access.
- Observability – records of datasets exchanged by Participants, in which Project contexts, with what authorisation.

Each of these functions of the FA could be provided using different models of centralisation. Some functions fit certain models better than others.

The diagram below shows idealised models of centralisation, from fully centralised to fully decentralised.



Federation Participants and data exchanges between them are shown in blue (the Federation “data plane”), while services provided by the FA are indicated in yellow (the “control plane”). We illustrate three organisational models – centralised, distributed and decentralised – that could be used to provide FA services.

FOR CONSULTATION & COMMENT

Centralised. With a centralised FA, a central node runs all required Federation Services, including services to identify, verify, onboard and register Participants. In a maximally centralised model it could also run a single data discovery catalogue for the whole Federation.

Every Participant requires one control-plane connection to the central FA node.

Pros: Simplicity, familiarity and maturity of implementation and operation; advantages for observability and discovery; minimal attack surface for key FA security services.

Cons: Single point of failure and single point of attack; may be viewed as ceding too much sovereignty to a single entity; a single bad-faith operator could disrupt the activities of Participants arbitrarily.

Distributed. The distributed model retains some degree of centralised control but addresses the single point of failure challenges. Functional roles are distributed among a few synchronised nodes, enabling multiple entities to share responsibility for providing FA services.

Every Participant requires one control-plane connection to their “nearest” FA node. “Nearest” can be interpreted in flexible ways.

Pros: Greater flexibility in service deployment over centralised; more resilient to single-node failure; more resistant to bad-faith FA actors; small attack surface for key FA security services

Cons: Technically more complex to implement and run, requiring synchronisation protocols between FA nodes; observability and discovery become more complex; only partially addresses the sovereignty issue.

Decentralised. A decentralised design creates the highest levels of autonomy and sovereignty, notably around identity. A decentralised identity system requires that each Participant maintain identity information that can be verified by other Participants in ways that meet the agreed FA rules and policies. The operation of other required FA services – notably registry – also falls to the Participants.

Every Participant requires one control-plane connection to every other Participant.

Pros: Maximises individual Participant sovereignty; highly resilient to single-node failure; highly resistant to bad-faith FA actors.

Cons: Technically very complex to implement and run, requiring synchronisation protocols between all Participant nodes; observability and discovery become challenging; maximal attack surface for key FA security services.

These models are not exclusive. Different models can be used for the different service functions required of the Federation and Federation Authority. Trust and identity services, for instance, could be realised centrally, while data discovery through catalogues may be much easier to realise as a distributed service or set of services.

It’s worth highlighting that the choice of model here impacts only the *control plane* of the Federation. Data exchange connections between Participants are the same in each case – direct and point-to-point. The functions of the control plane determine only *how* the connection is made, not where it goes.

FOR CONSULTATION & COMMENT

Following [54], the figure below shows the three organisational models on a single radar diagram against axes representing six desirable properties.

Sovereignty. The first goal of the Federation is to improve data sharing for research while maintaining, or even enhancing, sovereignty for data providers. Sovereignty is partly a function of autonomy, trust and transparency: is the decision to share this dataset mine? Do I trust the recipient I'm sharing with? Do I retain sight of where and how my dataset is being used? In our use of the term, sovereignty sits with Federation Participants, particularly data providers, in contrast to "control" below.

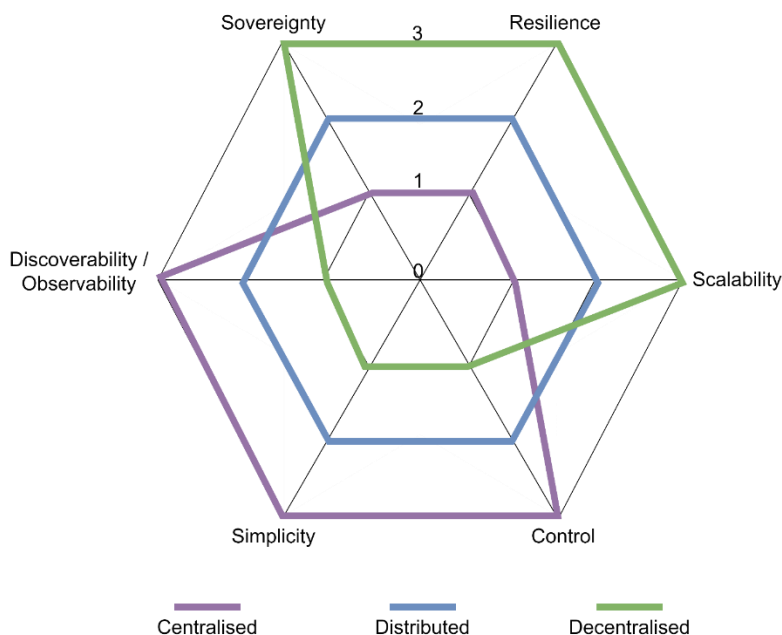
Resilience. Resilience is the ability of the overall Federation ecosystem to continue functioning in the event of unforeseen problems, such as the failure of a service node.

Scalability. In Federation terms scalability is not about the volume of data exchanged but about the number of Participants and number of concurrent Projects. These factors determine the potential load upon FA services.

Control. In Federation terms, control sits in contrast to sovereignty: what level of control does the FA have over the core federation services, not only in terms of their content but also in terms of who is allowed to access them? How much "say" does the FA have over day-to-day operations within the ecosystem?

Simplicity. In terms of both building the Federation and running it, mature, well-established technologies and architecture models are easier to deploy and operate.

Discoverability / Observability. This describes the overall transparency of the Federation, in terms of its content (essentially, the discoverability of data) and its operation (can everyone see what they need to see to retain overall trust?).



FOR CONSULTATION & COMMENT

7. Development and delivery approach

Our adopted design philosophy favours an incremental approach to delivering the federated network architecture: introducing (and enforcing) a common low-level foundation while aiming for minimal disruption to existing services and supporting maximum innovation at application level. This chapter sketches a phased delivery approach to building the first operational version of a federated TRE and data network. Note that we do not cover the long-term operational funding model of such a federated network of TREs here, although we do observe that such a model is absolutely critical for this network to succeed long-term in delivering better research outcomes for the UK.

7.1. Prototyping and technology selection

Suitable technologies to deliver the Federation core services should first be explored and selected. Two different approaches can be used, depending on the technology readiness level (TRL) required³⁶.

7.1.1. Core services: technology evaluation

Core services provide the secure, trustworthy backbone of the entire Federation. These should be selected from existing solutions, proven in operation (i.e., TRL 9).

We recommend convening a community-wide panel to draw up a shortlist of potential solutions and then commissioning a series of evaluation projects against a common “proof-of-concept” brief. Some candidate open-source technologies have been discussed throughout this report (cf. Appendix A).

7.1.2. Interfaces and other services: community driver projects

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

We recommend commissioning research and development projects to investigate different technological approaches to the required core services. DARE UK’s Phase 1b Driver Projects (2023) are a model approach³⁷.

7.2. Technology proof-of-concept

Using selected technologies, a proof-of-concept (PoC) system can be deployed against a number of test scenarios. Note that functionality and correct operation can be tested in all these scenarios without the need for any sensitive data.

Scenarios 1 and 2 below cover “traditional” TRE operation where data are moved into a secure environment for analysis. Scenarios 3 and 4 develop the newer remote-query model.

³⁶ Technology Readiness Levels. See https://en.wikipedia.org/wiki/Technology_readiness_level

³⁷ See <https://dareuk.org.uk/our-work/phase-1-driver-projects/> and Section 2.3.1.

FOR CONSULTATION & COMMENT

Note that all these scenarios are technical proofs-of-concept that demonstrate the required functionality of foundational and core components. They do not address data interoperability or information governance.

7.2.1. Scenario 1: basic data exchange

This is the base scenario involving the core Federation Services and secure data exchange between two TREs, one acting as a data provider and one as an analytical service.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); SDZ (interfaces: Data Ingress);
- 1 x TRE: Security Server (Core); SDZ (interfaces: Data Egress).

Required concepts:

- Identities: Participant; Project; Dataset; Data Extract;
- Structured Data Objects: Data Extract.

7.2.2. Scenario 2: linked data exchange

This scenario extends the first with an additional data provider TRE and introduces an Index Service.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); SDZ (interfaces: Data Ingress, Index);
- 2 x TRE: Security Server (Core); SDZ (interfaces: Data Egress, Index);
- 1 x Index Service: Security Server (Core); interfaces: Index.

Required concepts:

- Identities: Participant; Project; Dataset; Data Extract; Linkage Spine;
- Structured Data Objects: Data Extract; Linkage Spine.

7.2.3. Scenario 3a: remote direct query (single)

This scenario exercises the movement of direct queries rather than the movement of data and can be viewed as complementary to Scenario 1. Recall that “direct queries” are fully encapsulated within Query data objects.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); RAZ (interfaces: Query (direct), Response);
- 1 x TRE: Security Server (Core); SDZ; QMZ (interfaces: Query (direct), Response).

Required concepts:

- Identities: Participant; Project; Dataset;
- Structured Data Objects: Query; Response (query).

FOR CONSULTATION & COMMENT

7.2.4. Scenario 3b: remote direct query (federated)

This scenario extends Scenario 3a to include a second data provider and tests the splitting of a query to run against each independently. Note that this requires more sophisticated data presentation and query handling than Scenario 3a.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); RAZ (interfaces: Query (direct), Response);
- 2 x TRE: Security Server (Core); SDZ; QMZ (interfaces: Query (direct), Response).

Required concepts:

- Identities: Participant; Project; Dataset;
- Structured Data Objects: Query; Response (query).

7.2.5. Scenario 4a: remote indirect query (single)

This scenario exercises the movement of indirect queries rather than the movement of data. Recall that “indirect queries” are **not** fully encapsulated within Job Request data objects but instead refer to (or “point to”) an analysis workload hosted on a third-party Software Service which must be retrieved by the participating TREs prior to execution.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); RAZ (interfaces: Query (indirect), Response);
- 1 x TRE: Security Server (Core); SDZ; QMZ (interfaces: Query (indirect), Response; Software);
- 1 x Software Service (research artifacts): Security Server (Core); interfaces: Software.

Required concepts:

- Identities: Participant; Project; Dataset;
- Structured Data Objects: Job Request; Response (job); Job Payload Artifact.

7.2.6. Scenario 4b: remote indirect query (federated)

This scenario exercises the movement of indirect queries rather than the movement of data.

This scenario extends Scenario 4a to include a second data provider and tests the splitting of a query to run against each independently. Note that this requires more sophisticated job and query handling than Scenario 4a.

Required components:

- 1 x Federation Services (Core);
- 1 x TRE: Security Server (Core); RAZ (interfaces: Query (indirect), Response);
- 2 x TRE: Security Server (Core); SDZ; QMZ (interfaces: Query (indirect), Response; Software);
- 1 x Software Service (research artifacts): Security Server (Core); interfaces: Software.

Required concepts:

FOR CONSULTATION & COMMENT

- Identities: Participant; Project; Dataset;
- Structured Data Objects: Job Request; Response (job); Job Payload Artifact.

7.3. Minimal viable product

A successful technology proof-of-concept for (at least) scenarios 1 and 2 should be developed into a minimal viable product (MVP). Scenarios 3 and 4 (and other functionality) can be introduced later through evolution and improvement.

Note that MVP development here is not principally a technical activity. The journey from proof-of-concept to MVP should focus on developing the required governance framework around data exchange, linkage and project identities.

The end product of this phase is a limited deployment of a federated TRE network suitable for use with real data.

7.4. Test and validation

Alongside the development of an MVP a test and validation approach should be developed. This should involve the deployment of a mirror version of the PoC system and the instigation of a dedicated adversarial test team (a “red team” in security engineering jargon³⁸).

We recommend including a dedicated red team testing component in future operational plans for the SDRI Federation.

7.5. Evolution

Once in place the MVP can be expanded and extended incrementally in scope and functionality:

- Scope: new TREs and other services can be added to the network by deploying Security Servers and supporting appropriate interface Services;
- Functionality: new interface Services can be developed and incorporated into the Federation’s “working set” as technology evolves.

In both cases, how changes are made and approved are key decisions required of Federation governance.

³⁸ See https://csrc.nist.gov/glossary/term/red_team for a definition of “red team”. The NCSC also has a good discussion of red-teaming in machine-learning system design at <https://www.ncsc.gov.uk/collection/machine-learning/requirements-and-development/design-for-security>

FOR CONSULTATION & COMMENT

8. Summary and further work

This blueprint addresses the challenge of connecting researchers and resources within the UK's existing landscape of digital research infrastructure by proposing a secure, managed federation of data and service providers. By proposing a foundational layer of secure data exchange and broad classes of interface services we seek to create the necessary trustworthy environment while imposing as few operational restrictions on service providers as possible.

This technical architecture supports current models of data linkage through the indexing and assembly of disparate datasets into one secure setting, and also newer models of remote and federated analytics where complex “query objects” can be submitted securely to remote data services (directly or indirectly).

We describe the architecture in three layers: infrastructure, data and governance. This version 2.1 covers significant refinements to the infrastructure layer set out in detail in the “initial” version (April 2023), expands on discussion of the data layer but covers the governance layers in less detail. We invite comment from the broader UK research community on the ideas and approaches presented here.

FOR CONSULTATION & COMMENT

9. 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. Rijnveld, 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 (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).
- [39] S. Bradner, B. Leiba; *BCP14; The Internet Engineering Taskforce Best Current Practice*; <https://www.ietf.org/rfc/bcp/bcp14.html> (accessed 01/12/2023).
- [40] The Open Group; *ArchiMate 3.1 Specification*; <https://pubs.opengroup.org/architecture/archimate3-doc/toc.html> (accessed 20/03/2023).
- [41] Welpton, Richard (2019). *SDC Handbook*. Figshare. Book. <https://doi.org/10.6084/m9.figshare.9958520.v1> (accessed 29/11/2023).
- [42] UK Government; *Copyright, Designs and Patents Act 1988*; <https://www.gov.uk/government/publications/copyright-acts-and-related-laws> (accessed 20/03/2023).
- [43] CITES; *Convention on International Trade in Endangered Species of Wild Fauna and Flora*; <https://cites.org/eng> (accessed 20/03/2023).

FOR CONSULTATION & COMMENT

- [44] UK Government; *Government Security Classifications*; May 2018; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715778/May-2018_Government-Security-Classifications-2.pdf (accessed 20/03/2023).
- [45] (ISC)²; *Certified Information Systems Security Professional*; <https://www.isc2.org/Certifications/CISSP> (accessed 20/03/2023).
- [46] D. Arenas, J. Atkins et al; *Design choices for productive, secure, data-intensive research at scale in the cloud*; arXiv:1908.08737v2 [cs.CR] 15 Sep 2019; <https://arxiv.org/pdf/1908.08737.pdf> (accessed 25/04/2023).
- [47] NHS Digital; *Health and Social Care Cloud Risk Framework*, Chapter Dimensions that affect risk; 14 October 2021; <https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services/cloud-risk-framework> (accessed 20/03/2023).
- [48] L. Sweeney, M. Crosas, M. Bar-Sinai; *Sharing Sensitive Data with Confidence: The Datatags System*; Technology Science, 2015101601. October 15, 2015; <https://techscience.org/a/2015101601/> (accessed 20/03/2023).
- [49] UK Statistics Authority; *Data Capability Guidance*, v1.0; September 2022; available via <https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/> (accessed 12/07/2023).
- [50] N. Karrar, S.K. Khan, S. Manohar, P. Quattroni, D. Seymour, S. Varma, & The UK Health Data Research Alliance. (2022). *Improving transparency in the use of health data for research: Recommendations for a data use register standard*. Zenodo. <https://doi.org/10.5281/zenodo.5902743>
- [51] GA4GH Beacon Group; *Beacon v2 standard*; <https://docs.genomebeacons.org/> (accessed 23/03/2023).
- [52] HDR-UK; *The CO-CONNECT Project*; <https://www.hdruk.ac.uk/projects/co-connect/> (accessed 23/03/2023).
- [53] UK Health Data Research Alliance; *Recommendations for Data Standards in Health Data Research*; November 2021; <https://ukhealthdata.org/wp-content/uploads/2021/12/211124-White-Paper-Recommendations-of-Data-Standards-v2-1.pdf> (accessed 13/07/2023).
- [54] International Data Spaces Association, *IDSA Rulebook*, https://docs.internationaldataspaces.org/ids-knowledgebase/v/idsa-rulebook/idsa-rulebook/3_functional_requirements (accessed 30/01/2024)
- [55] P. Barnsley, J. Fleming; (2023). *Trusted Research Environments – federating data to complete research*. The Francis Crick Institute. Report. <https://doi.org/10.25418/crick.23626653.v1>

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 Usage patterns

How well does this architecture model existing patterns of inter-TRE communication and federation?

This second version has been guided by work ongoing through 2023 and by interactions with key stakeholders and service operators through the UK TRE community³⁹.

Below we map published information about other patterns of TRE federation against the architecture picture in Chapter 4.

B.1 “Classic” TRE inter-operation

This model is an amalgamation of many current TREs which feature virtual desktop access to project environments and access to approved datasets.

Features:

- Data pooling model.
- Isolated research projects.

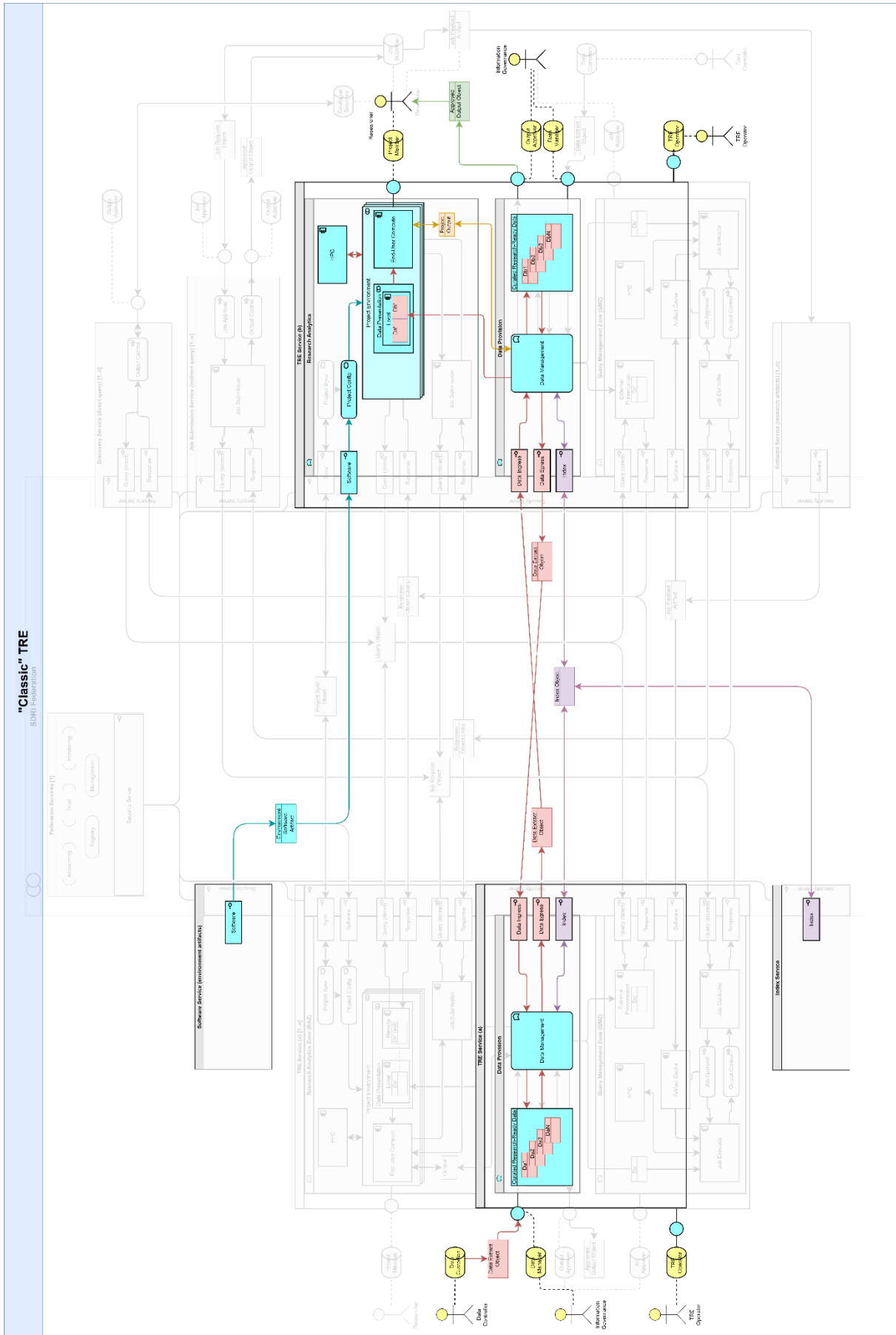
Elements:

- TRE “a” (left), acting purely as a data provider.
- TRE “b” (right), acting as both a data provider and analytics service provider.
- Index service, providing linkage spines.
- Software service, providing packages and other software components for the analytical project environments.

Diagram: (next page)

³⁹ The UK Trusted Research Environment Community. See <https://www.uktre.org/>

FOR CONSULTATION & COMMENT



FOR CONSULTATION & COMMENT

B.2 Francis Crick Institute federation model

Reference:

- The Francis Crick Institute, *Trusted Research Environments – federating data to complete research* [55].

Features:

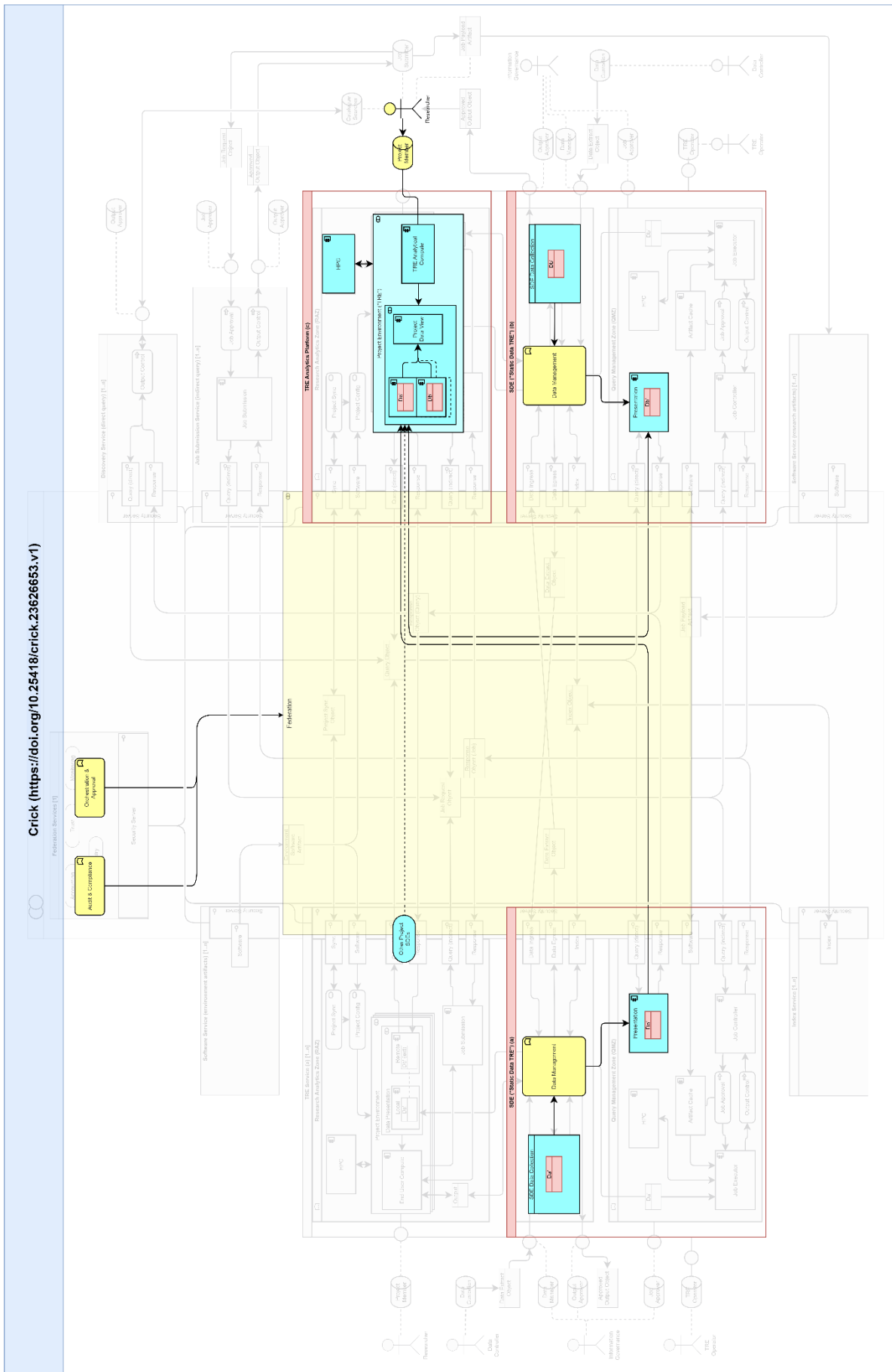
- TRE federation realised dynamically on a per-project basis.
- Separation of project analytics from data sources.
- Direct query model.

Elements:

- TRE “c” (upper right), acting purely as an analytical project environment, presenting a view of remote data to project members.
- SDE “b” (lower right) (“secure data environment”, a “static data TRE” in [1]), acting purely as a data provider with a remote presentation of data to TRE “c”.
- SDE “a” (left), acting purely as a data provider with a remote presentation of data to TRE “c”.
- Other project-specific SDEs (not illustrated in full).
- Federation between these elements on a per-project basis.

Diagram: (next page)

FOR CONSULTATION & COMMENT



Crick (<https://doi.org/10.25418/crick.23626653.v1>)

FOR CONSULTATION & COMMENT

B.3 OpenSAFELY

Reference:

- OpenSAFELY, *The OpenSAFELY Secure Analytics Platform* [28] and particularly <https://docs.opensafely.org/images/c4-container.svg>

Features:

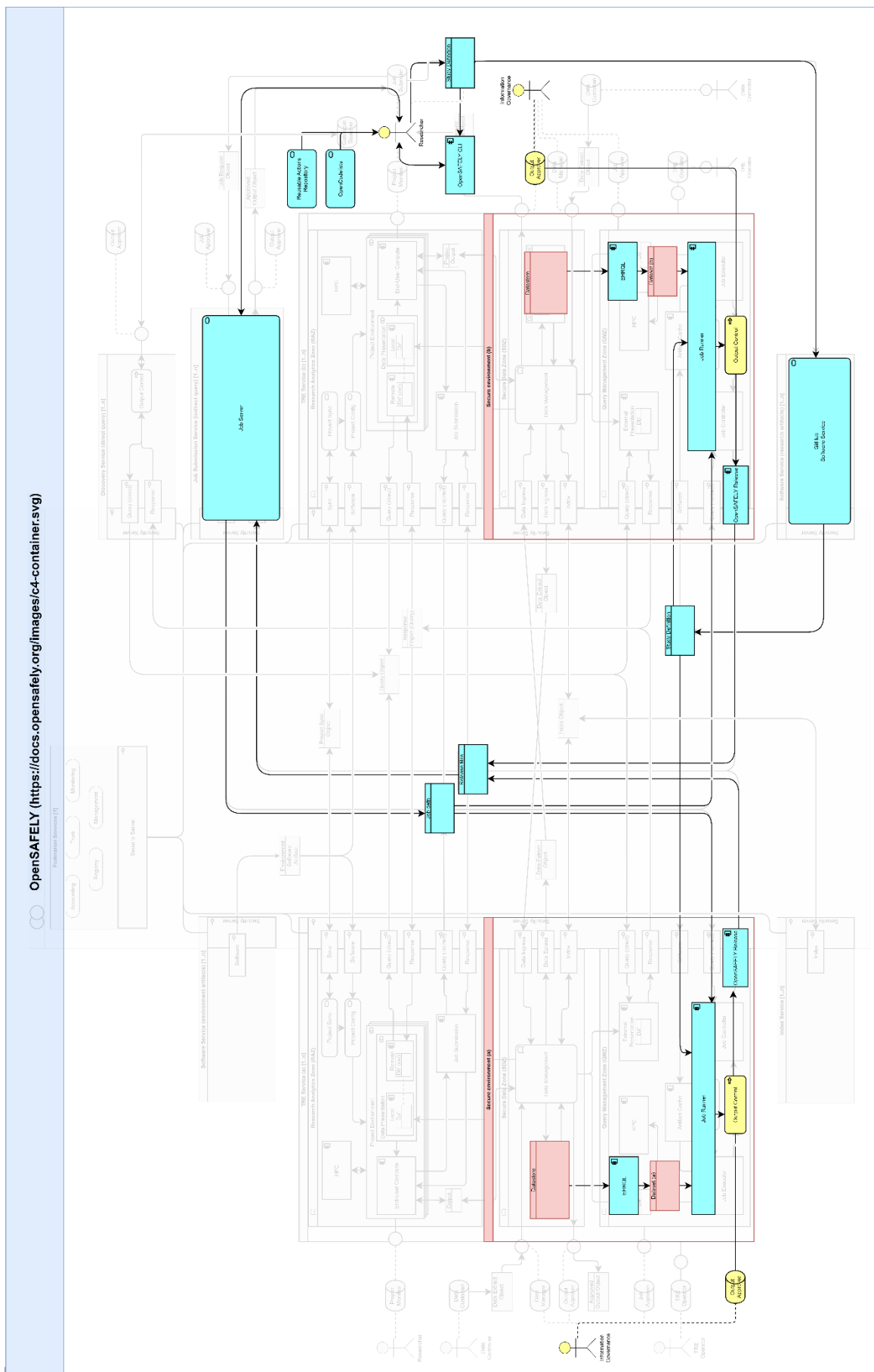
- Indirect query model.
- Researcher code and job development via local development tools (OpenSAFELY command line interface) and public repositories (notably GitHub, OpenCodelists).
- Job submission from outside TRE environments.

Elements:

- Job Server (upper right), acting as the point of interaction between researchers and TREs.
- Secure environment “a” (left), acting as a data provider and job handler.
- Secure environment “b” (right), acting as a data provider and job handler.
- GitHub (lower right), acting as a software service.

Diagram: (next page)

FOR CONSULTATION & COMMENT



FOR CONSULTATION & COMMENT

B.4 TELEPORT federation with pop-up TREs

Reference:

- C. Orton, et al. *TELEPORT: Connecting Researchers to Big Data at Light Speed* [29].

Features:

- Direct query model using polystore presentation.
- Dynamically-provisioned pop-up TREs with keep-alive sync to “mutually approved” state.
- GitOps synchronisation between participating TREs.

Elements:

- TRE “a” (left), acting as both data provider with remote data presentation, and potential provider of analytical project environments.
- TRE “b” (right), acting as both data provider with remote data presentation, and potential provider of analytical project environments.
- Package repo (upper left), providing software components for dynamic provisioning of project environments to mutually approved state.
- Continual policy sync between TREs “a” and “b”.

Diagram: (next page)

FOR CONSULTATION & COMMENT

B.5 TRE-FX federation with stand-alone job submission

References:

- T. Giles, et al. *TRE-FX: Delivering a Federated Network of Trusted Research Environments to Enable Safe Data Analytics* [27].
- T. Giles, et al, TRE-FX primary implementation report
<https://docs.google.com/document/d/1FxrwXoYjx5aUI3MQyrnHs7xigvATJME/>

Features:

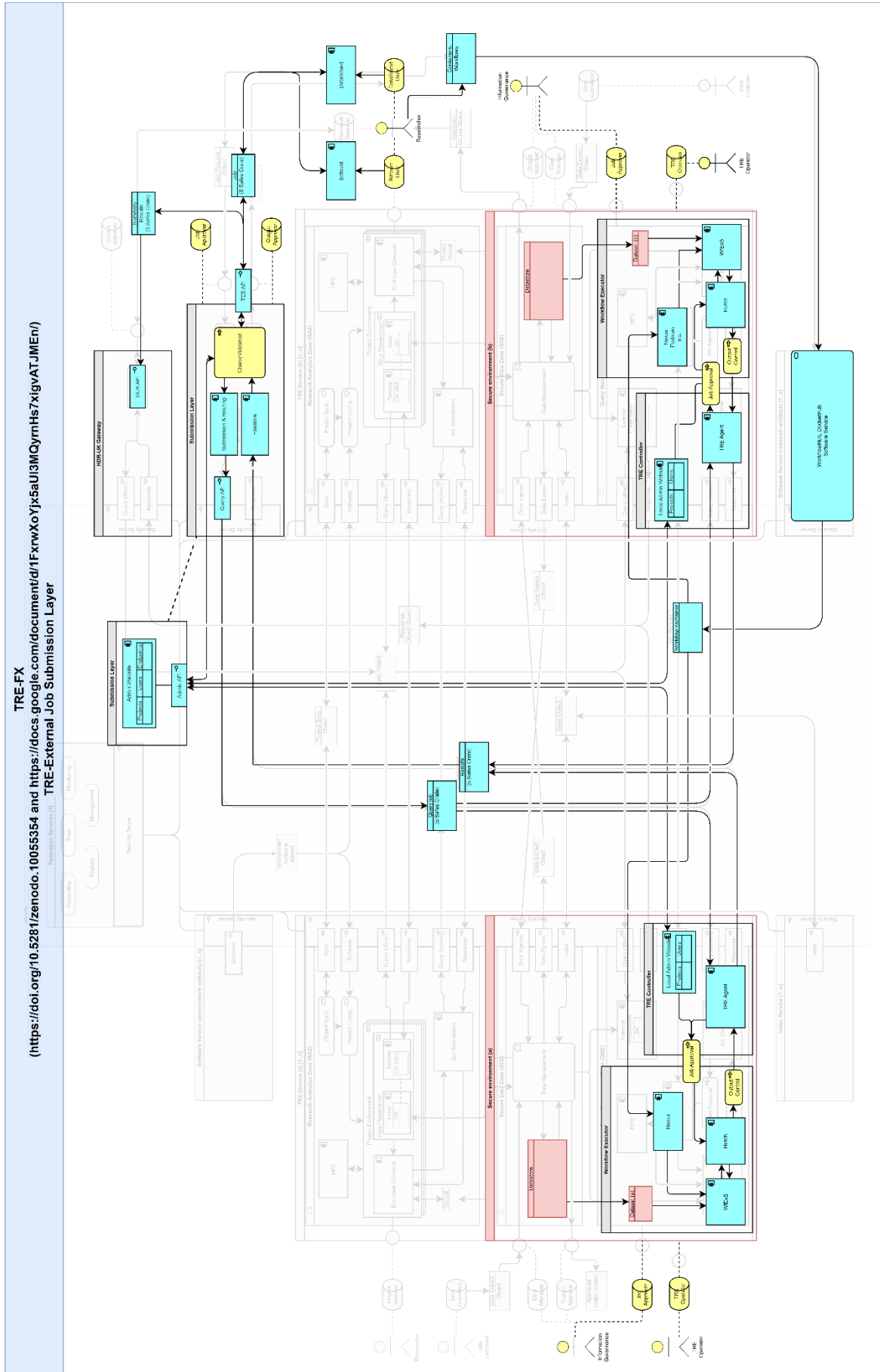
- Indirect query model.
- Researcher code and job development via local development tools (Bitfount, DataSHIELD) and public repositories (notably WorkflowHub and DockerHub).
- Standardised packaging methods for exchanged digital objects.
- Job submission from outside TRE environments.
- Single registry of projects and users.

Elements:

- Secure environment “a” (left), acting as both a data host and job handler.
- Secure environment “b” (right), acting as both a data host and job handler.
- Submission layer (upper right), acting as both a job submission service and a common lookup-registry for projects, users and data.
- WorkflowHub and DockerHub, acting as software services for researcher-developed artifacts.

Diagram: (next page)

FOR CONSULTATION & COMMENT



FOR CONSULTATION & COMMENT

B.6 TRE-FX federation with TRE-hosted job submission

References:

- T. Giles, et al. *TRE-FX: Delivering a Federated Network of Trusted Research Environments to Enable Safe Data Analytics* [27].
- T. Giles, et al, TRE-FX primary implementation report
<https://docs.google.com/document/d/1FxrwXoYjx5aUI3MQyrnHs7xigvATJME/>

Features:

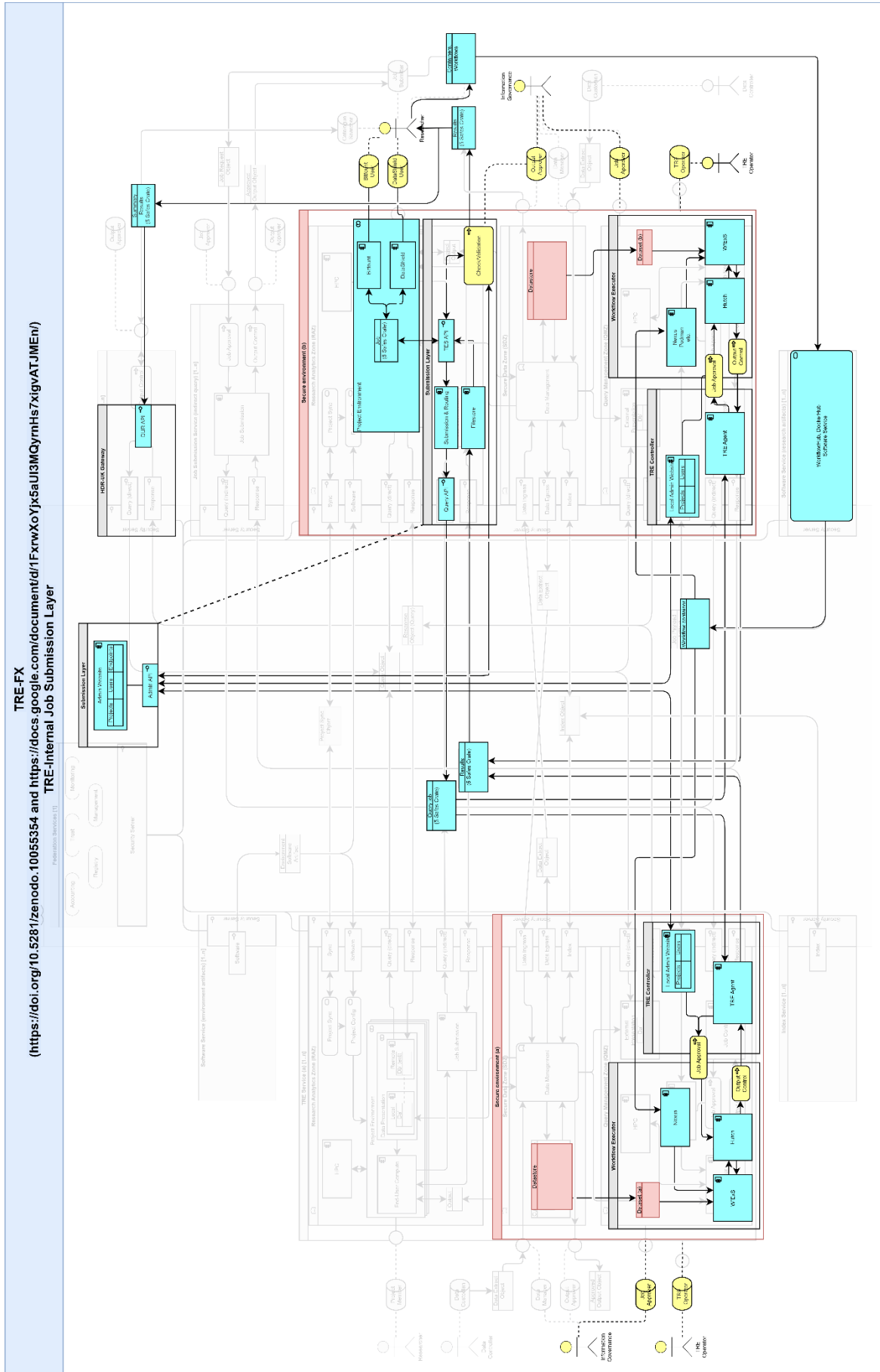
- Indirect query model.
- Researcher code and job development via local development tools (Bitfount, DataSHIELD) and public repositories (notably WorkflowHub and DockerHub).
- Standardised packaging methods for exchanged digital objects.
- Job submission from inside TRE project environments.
- Single registry of projects and users.

Elements:

- Secure environment “a” (left), acting as both a data host and job handler.
- Secure environment “b” (right), acting as an analytical project environment, a data host and job handler.
- Submission layer (upper right), acting as both a job submission service and a common lookup-registry for projects, users and data.
- WorkflowHub and DockerHub, acting as software services for researcher-developed artifacts.

Diagram: (next page)

FOR CONSULTATION & COMMENT



TRE-FX
(<https://doi.org/10.5281/zenodo.10055354> and <https://docs.google.com/document/d/1FxrWx0Yjx5aU13MqymHs7xigvATJMEn/>)
TRE-Internal Job Submission Layer

FOR CONSULTATION & COMMENT

C 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).

C.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.

C.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.

C.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.

C.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 friction⁴⁰.

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

C.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

⁴⁰ 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.

C.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.

FOR CONSULTATION & COMMENT

D Master requirements table

In the table below we follow the conventions of IETF BCP 14 (RFC2119 & RFC8174) [39], 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 from Chapter 3 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 architectures⁴¹. A key output of the SATRE project [25] 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.

⁴¹ See <https://satre-specification.readthedocs.io/en/v1.0.0/index.html> for an online version of v1.0.0 of the SATRE specification.

FOR CONSULTATION & COMMENT

| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|---|----------------------------|----------------------------|--|---------------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|---|-----------------------------|----------------------------|---|-----------------------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|---|---------------------|----------------|--|---------------|
| | | 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|---|---------------------------|-----------------------------|-------------------------|--------------|
| 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 | | |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|----|------------------------------------|------------------------------------|----------|----------------------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|----|------------------------------|------------------------------------|------------------------------------|----------------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|----|----------------------------------|----------------------------------|--|-----------------------------|
| 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) | | |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|----|----------------------------------|-----------------------------------|-------------------------|--------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|---|-------------------------|-----------------------------|--|--------------|
| 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|---|------------------------------|-----------------------------|-------------------------|----------------|
| 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 | | |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|---|---|---|---------------------------------------|-----------------------------|-------------------------|----------------|
| | | 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 |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|---|------------------------------|------------------------------|--|----------------------|
| 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 | | | |

FOR CONSULTATION & COMMENT

| | | | | | | | | |
|------|-----|--|---|---|------------------------------|------------------------------|--|--|
| 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 | | |

FOR CONSULTATION & COMMENT

FOR CONSULTATION & COMMENT

E Acknowledgements

E.1 Federated architecture blueprint: direct feedback

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

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

E.2 Phase 1b persona development

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

- The Alan Turing Institute

FOR CONSULTATION & COMMENT

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

E.3 DRI landscape review and community conversations

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

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

FOR CONSULTATION & COMMENT

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

FOR CONSULTATION & COMMENT