Federated Architecture Blueprint

DARE UK Delivery Team

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

Document Control

Version	Date	Authors	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.

Federated Architecture Blueprint

Contents

About Document Versions

The Federated Architecture Blueprint for DARE UK will develop and evolve over the course of 2023. We plan three iterations, approximately one per calendar quarter: “initial” at the end of Q1, “interim” at the end of Q2, and “final” in early Q4.

This version is “initial”. It proposes a model of a federated network infrastructure based on community needs assessed in DARE UK Phase 1a and subsequent consultation and evolution. It focuses on the “infrastructure layer” with an initial look at the “data layer” and “governance layer”. Later versions will further develop these themes and will assimilate feedback from community consultation on earlier versions.

1 Executive Summary

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 modelled on the Five Safes practices developed by ONS. 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.

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 gathering all data into a central location will not happen quickly, if ever. Thus we expect the sensitive data landscape to remain distributed and accordingly propose a federated approach to connecting TREs, data providers and other services together in a way that is standardised but as minimally intrusive to the good practice already in use.

We propose a managed federation formed from a set of coordinating central registry services and a network of secure interface services deployed at each federation participant. Together these services create a backbone for secure document 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 strong governance oversight.

Running on top of this backbone we propose a set of application services in a small number of different classes. We identify needs for services 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 APIs 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. We highlight instead 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. We highlight the key relationships and accountabilities within the proposed federation.

Finally, we note that this blueprint is an "initial" version. Two further iterations are planned ("interim", due in June, and "final", due towards the end of 2023) which will incorporate comment, feedback and changes identified through a planned community consultation process.

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. While the numbers and locations of data sources and services within this landscape will ebb and flow (see Scenario Thinking) 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 [1] that shape our approach to a federated architecture for trusted research environments (TREs) across the UK.

2.1.1 Data and discovery

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.

2.1.2 Core federation services

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.

2.1.3 Capability and capacity

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

2.2 A Managed Federation

While there are many ways to define "sensitive data" one particularly important definition is "individual-level public data". 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. Whatever the source, any use of public data for research must have public trust 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 [2]. To enable the exchange of sensitive data – in particular public data – the federation must be trustworthy.

The World-Wide Web is an excellent example of a federation of information resources connected by common protocols and standards. Unfortunately it cannot reasonably, as a whole, be described as trustworthy. Common standards and protocols are necessary for a trustworthy federation but insufficient in themselves. Trustworthy federations – online banking, e-government, corporate intranets – are routinely layered on top of the World-Wide Web through the introduction of additional technologies (typically security-related) and a managing organisation – the bank, the government, the firm.

Our proposal for the future of DARE UK is the creation of just such a managed federation, layered on top of the World-Wide Web. We envisage an ecosystem of sensitive data providers and analytical service providers of differing sizes and capabilities connected using common protocols and standards with a central set of registry services acting as the federation gatekeeper. The trustworthy federation is defined by a common, low-level set of security protocols and standards for secure document exchange, on top of which is built a rich set of application protocols and standards to support different analytical use-cases. The low-level protocols and standards define what it means to join the federation, and participants joining the federation are approved and registered centrally. In this blueprint we develop the ideas on federation touched on in the 2020 Health Data Research Alliance Green Paper on TREs [3].

Governance of the federation can be designed in a number of ways; what is key is that the federation has a single gatekeeper where participants are registered. The federation governance, registry services and low-level document exchange protocols must be designed to ensure that all members of the federation trust one another and that, once a participant has joined, they enjoy the same levels of trust as all other participants.

To achieve its goals, DARE UK must be designed from the beginning as a managed federation.

2.3 The State of the Art

Managed federations have been a staple of the UK research landscape since the early Noughties and the drivers of the UK e-Science Core Programme [4]. The World-wide LHC Compute Grid (WLCG [5]) and the International Virtual Observatory Alliance (IVOA [6]) 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 Noughties rise of "Grid computing" (a forerunner of cloud computing) are more than 15 European research infrastructures spanning health and social sciences [9]. Notable examples include ELIXIR [7], BBMRI [8], CESSDA [10] and ESS [11]. 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).

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.

X-Road [12], managed by the Nordic Institute for Interoperability Solutions [13] 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" [14].

GAIA-X [15], 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 [16]. It shares many similar concepts with X-Road and with SiMPl (qv). 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 most recent work in this space is perhaps the launch of an invitation to tender for the European Smart Middle Platform (variously SiMPl or SMP) [17]. SiMPl 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 [18] and the development of "data spaces". The published timetable for SiMPl suggests a minimal viable product should be released "at the beginning of 2024".

As noted, the proposed SiMPl architecture shares many common features with both X-Road and GAIA-X; these three initiatives do collaborate at various levels. Appendix A provides a comparison of these three initiatives, alongside similar concepts from the proposed DARE UK federated architecture.

2.4 Scenario Thinking

DARE UK could look different under different future scenarios, depending on a certain number of external policy drivers. Initial thinking pulls up two principal drivers: the number of TREs and their capabilities (call it the "Goldacre axis"); and mobility of data (the "DPA axis").

1. The number of TREs. The Goldacre review [19] 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 (eg, 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 [20] and the UK GDPR [21], perhaps creating governance counter-pressures towards more mobile data. Note that this doesn't address the "gravity" around very large datasets.

2.4.1 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.

2.4.1.1 Low-Low

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 or high-performance computing capabilities or both, 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. Such a TRE can accumulate expertise in working with the datasets in question, but in this scenario linkage between datasets becomes difficult. If legal agreements for data linkage are the bottleneck for sharing data, then the incentives on TREs towards technical interoperability are that much weaker.

For budgetary and technical reasons there are unlikely to be many TREs providing the specialised capabilities for working with large, complex data, so some kind of low-low scenario is quite likely in any of the futures we consider here.

2.4.1.2 Low-High

If the gravity of large, complex datasets means a low number of highly capable TREs exist, then these TREs are also available to process smaller, neater datasets. 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.

2.4.1.3 High-Low

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 (eg, a clinical trial dataset) or for individual research organisations (eg, a university or university department). In this scenario linkage remains difficult and the data landscape is even more fragmented than in the low-low scenario. If data sharing is difficult for governance reasons then there are few incentives for these TREs to maintain any level of technical interoperability or adhere strictly to any particular standard if doing so might constrain the TRE's core research purpose. The risk of technical drift between TRE environments is high with a consequent dissipation of expertise and increased friction.

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

2.4.1.4 High-High

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 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.

2.4.2 Analysis

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.

2.5 Summary

That the proposed DARE UK federated 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 participants must be managed. Central 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 document 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 core 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.

3 Scope

The DARE UK programme is about enabling broader use of public data in research, safely and at scale. The programme has not arisen in a vacuum: the UK has a number of existing trusted research environments (TREs) operating today, ranging in size and scope from national-scale to individual clinical trials. Many of the existing TREs – notably the larger ones – serve research in the health domain. Some provide secure access to government administrative data, survey micro-data and other non-health sensitive datasets. A few do both.

DARE UK's principal technical challenge is to create a federated architecture that can connect existing and future TREs, data providers, information governance authorities and researchers into something greater than the sum of its parts.

3.1 Design Principles

DARE UK's approach to the design and build of a federated network for research with sensitive data follows these principles.

1. Public trust first, last and always. The strongest design voice should come from the "public persona".
2. No TRE, no data. Reinforcing a recommendation from the Goldacre Review [19], require that any and all analysis of sensitive data take place within a TRE, and design accordingly.
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. Adopt the Five Safes framework as a guiding principle. Processes and governance are as important as infrastructure, and infrastructure choices should reflect this.
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.
7. Be as FAIR as possible. Findabilty, accessibility, interoperability and reusability are excellent qualities to maintain even in a sensitive data environment [22].
8. The "IETF principle" [23]: 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.

3.2 Objectives

1. Our focus is defining a federated network architecture in terms of required functionality and overall structure. We do not make any technological recommendations and we discuss particular technologies only in the context of related work in the area.
2. Our focus is first and foremost the "federated" part – connecting existing and future TREs together in consistent ways to enable cross-communication and interoperable working. Our aim is not to over-specify the internals of any one TRE.

4 Federation Drivers

By any measure the UK already has a research landscape for sensitive data and it is, in some ways, already 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 DARE UK federation is thus 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 DARE UK 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, managed federation 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.

4.1 Rachel's Journey

Rachel is a researcher. Here is an account of her journey from an idea to a start of a project built around that idea. 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;
- Dave, Della and Debra, three data providers.

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

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

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

Rachel has identified three datasets she needs:

- Education data, already collected by Debra 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 Dave and all openly available for research.
- Health outcomes, collected by Della and available for research but only for particular cohorts. Rachel will have to ask explicitly for what she needs.

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

Rachel understands she'll need to conduct her research in a TRE. Seeing that at least one of her datasets of interest is 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. Gill contacts the three data providers, Debra, Della and Dave. Della's health outcomes data is

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

the biggest constraint; Della can release a cohort set for research so defining the cohort is key. Gill, Rachel and Della work up a cohort definition for the project.

Rachel and Gill have agreed a definition for the project:

- Della has approved the cohort of health outcomes data, indexed at individual level by NHS#.
- Debra has approved access to the education data already within the TRE, already indexed at individual level with a unique "education data index".
- Dave is happy to provide access to the environmental datasets for the areas inhabited by Rachel's cohort. Dave's data can be indexed by lat/long or equivalent geospatial coordinates.

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.

Gill sends the set of NHS#s to Iain. Using the central 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 NSH#s 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".

Iain sends the ID1 and education index mapping to Debra.

Iain sends the ID2 and lat/long mapping to Dave.

Iain sends the ID3 and NSH# mapping to Della.

Iain sends the "master index" straight to Gill at the TRE.

Dave 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.

Dave sends this dataset to Gill, marked "for Rachel's project".

Della prepares the health outcomes data extract using the set of NHS#s, but she replaces NHS# with ID3 in Rachel's version of the dataset.

Della sends this dataset to Gill, marked "for Rachel's project".

Debra chooses to prepare the education data as an extract using the set of education data indexes and replaces education data index with ID1 in Rachel's version of the dataset.

Debra 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!

"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.

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

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

Some indirect mapping is required:

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

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

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

Debra 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 remaining in the linked dataset is the "IDR" which is unique to Rachel's project (and doesn't mean anything to anyone else).

Finally!

Rachel's research journey, while synthetic, is rooted very much in current practice of sensitive data research in the UK. It helps us tease out the key drivers for DARE UK, and in doing this we take two perspectives. The first we derive from potential users of the federation, from researchers like Rachel to system operators and data custodians. The other we derive from the existing landscape of services across the UK 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 DARE UK federation.

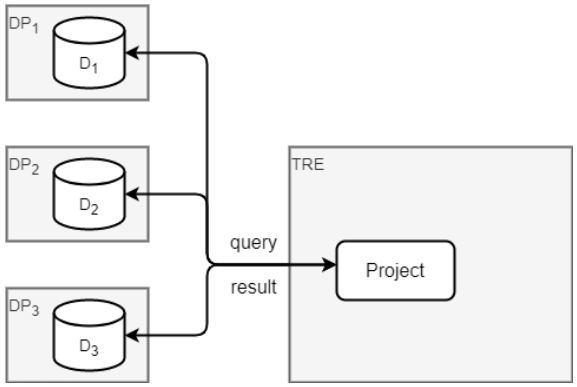
4.2 Landscape Review: Data Usage Patterns

We have analysed many of the existing patterns of interaction between TREs and data providers and have captured them as a series of high-level data usage diagrams (Appendix B). These diagrams are the "landscape equivalent" of user personas: sources of requirements based not on idealised users but on characteristic representations of services within the existing TRE ecosystem.

From this analysis we derive two essential use-cases, federated data and federated query. (Since our interest is in the federation of TREs and data providers at the organisational level we do not delve into the details of data provision to researchers within a TRE.)

4.2.1 Federated Query

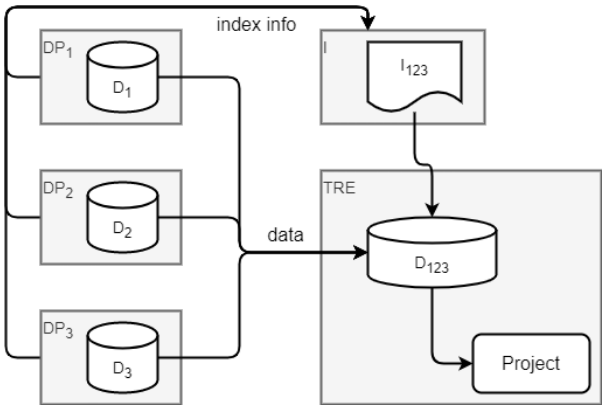
Federated query is the simpler pattern but covers the fewest concrete use-cases. Here, datasets (D_1, D_2 and D_3) remain within their data provider organisations (DP_1, DP_2, DP_3) and queries across them are sent from a project within a TRE. Results are returned to the project but not necessarily synchronously: query results may need to be disclosure checked before they are permitted to leave the data provider.



This pattern can work well when data are "vertically partitioned" but otherwise uniform (e.g., census data divided by region).

4.2.2 Federated Data

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

4.2.3 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 Providers. Each block in Figure 1 is conceptually held by a different organisation.

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

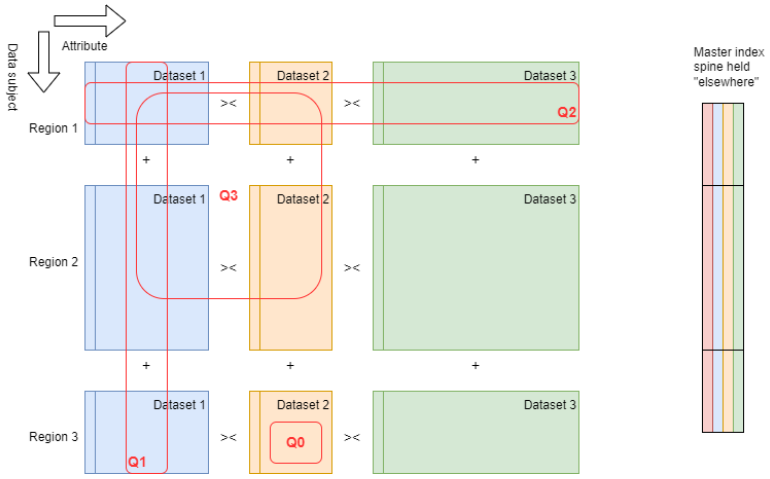


Figure 1. Conceptual dataspace for DARE UK

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

Q1: a query across a single Dataset but spanning multiple Regions to include a larger population than is available at any individual Data Provider. 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.

4.3 User Persona Development: Federation Roles and Actors

The scope of the DARE UK federated infrastructure gives rise, broadly speaking, to three groups of participants or "actors" – people (or potentially automated systems) which interact with the federation in different ways: Data Providers, Data Consumers and people who connect the two (call them "Connectors").

Most of these roles already exist in practice, making it feasible to develop supporting user personas (see below) around them. Almost by definition, though, the Connector role(s) required to run the DARE UK federation itself are new.

We give each role an abbreviation for later use.

4.3.1 Data Providers

Roles in the Data Providers group include:

- members of the public (P), as ultimate providers of their data for research in the public benefit;
- data controllers (DC), 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 (DX), responsible for curating and maintaining complete, accurate and useful sets of public data (a technical branch of data controllership, perhaps).

4.3.2 Data Consumers

Roles in the Data Consumers group include:

- academic researchers (R), 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 (R), 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 DARE UK federation we can treat them as Researchers.

4.3.3 Connectors

Roles in the "Connectors" group are more diverse than the other two and include the following:

- information governance (IG) professionals (abbreviated G) 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;
- indexers (I) and linkers (L) 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.
- data service operators (SO) 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;
- TRE operators (TO) 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 service operators (FO) 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.

4.3.4 Other stakeholders

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

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

4.3.5 User Personas

DARE UK works with relevant community groups across the UK to develop user personas to represent classes of users [1]. Personas give voice and motivation to the abstract "roles" discussed above 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 safe? how quickly?) than abstract system roles.

Table 1 summarises DARE UK's user personas and maps them to the system "actor" roles discussed so far. Phase 1a focussed on developing Data Provider and Data Consumer-class personas. Phase 1b will fill out the personas for the "Connectors".

Often it is easy to associate a particular persona with a single role; sometimes it is not. Some personas may take on more than one role, particularly roles within the "Connectors" group: a persona representing someone running a TRE service that also hosts important datasets will have both TRE operator and data service operator roles.

It is worth highlighting that **all** personas developed here **always** have the role of member of the public, even if they specialise elsewhere!

Table 1. DARE UK User Personas and their principal features.

Persona	Has Role	Key Motivation	Key Concern	Abbr
Grace Opedemi, member of the public	P	Understand how best use is being made of public sector research investments.	Keeping my data safe from unauthorised, unethical or other "bad" uses.	GO
Peter Shaw, data custodian	DC	Share and link my data with others.	Safety! (Don't break the law!) Poor data quality (terminology, linkage)	PS
Pritesh Navdra, techie data scientist	R	Keep on the leading edge of data science, while doing some good!	Poor data quality (terminology, linkage); poor tooling.	PN
Sharon Wakefield, researcher entering public health	R	Create more impactful research through greater access to linked data.	Ease of access to restricted data (skills, quality, linkage).	SW
Sarah Greenshaw, university public health research PI	R	Grow the research power and outward recognition of her group.	Competition from elsewhere, being left behind.	SG
Jeremy Foster, ed-tech business product manager	R	Generate ROI through accessing and sharing sensitive data.	Ease of access to restricted data (skills, quality).	JF
To do	G			
To do	TO			
To do	SO			
To do	FO			
To do	F			

4.4 High-level requirements

Analysis of data usage patterns and the first set of user personas has identified a number of key requirements for the overall federation and for individual services within it [1]. Some of these requirements are functional use-cases, others are non-functional constraints and still others are higher level "user stories" to be followed up in later stages of this work. A complete list of requirements can be found in Appendix C.

We weight requirements according to the number of personas mentioning them and the strength of that need (ie, whether a "must have", a "should have" etc.). When describing requirements, we follow the conventions of RFC2119 [24], vis:

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

From each requirement we identify the need for a potential system component ("Components" in the following tables) or an information context ("Contexts"), or both. Where requirements point towards needed cooperation between several possible system components and information contexts we use the catch-all "Federation Services".

4.4.1 Use-cases: functional requirements

Use-cases point towards required functionality and hence functional components within the overall architecture.

ID	Description	Wt	Components	Contexts
R005	The Federation MUST enable linkage between syntactically similar data	9	Federation Services; Index Service; Data Provider Service	Linkage Spine
R006	The Federation MUST enable linkage between syntactically dissimilar data	9	Federation Services; Index Service; Data Provider Service	Linkage Spine
R046	The Federation MUST support a "federated query" analysis pattern	9	TRE; Data Provider	
R047	The Federation MUST support a "linked-data assembly" analysis pattern	9	TRE; Data Provider	Linkage Spine
R014	The Federation MUST ensure research use is appropriately recorded in metadata records	6	Federation Services; TRE	Metadata
R020	The Federation MUST ensure data controllers are appropriately recorded in metadata records	3	Federation Services	Metadata
R024	The Federation MUST facilitate data discovery across the network	3	Discovery Services	Metadata
R018	Data Providers SHOULD provide tooling for pseudonymising data	2	Data Provider Service	Security
R023	The Federation SHOULD enable discovery of and access to modern data science computational capabilities	2	Discovery Services	Metadata
R025	TREs SHOULD provide metadata on access charges and running costs	2	TRE	Usage Costs

4.4.2 Constraints: non-functional requirements

Constraints point towards how well certain architectural features or concepts need to perform in their designed roles.

ID	Description	Wt	Components	Contexts
R003	The Federation MUST ensure the confidentiality of data storage	9	TRE; Data Provider Service	Security
R004	The Federation MUST ensure the confidentiality of data exchange	9	Federation Services	Security
R008	The Federation MUST reduce the barriers to data access	9	Federation Services	Usability
R009	The Federation MUST ensure the integrity of data exchange	6	Federation Services;	Security
R010	TREs MUST ensure the security of data access and use	6	TRE	Security
R021	The Data Provider MUST make data sharing as easy as possible	3	Data Provider Service	
R019	Data Providers SHOULD provide tooling for assessing data anonymity	2	Data Provider Service	Security

4.5 Future Work

Later versions of this document will complete the missing user personas from Table 1 and expand the use-case and constraints tables.

5 Federated Architecture: Concepts

5.1 Layers

In the following chapters we divide the DARE UK 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 DARE UK Federation, rather than network hardware or any particular technology. In ArchiMate terms [24] this is best thought of as the application layer.
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. In ArchiMate terms this is best thought of as another view of the application layer with elements of the business layer.
3. Governance. The highest level considered here, governance defines whether and how data may be used and thus drives the requirements of the lower two layers. In ArchiMate terms this is best thought of as the business layer with some elements of the strategic.

5.2 Concept Map

Figure 2 is a map of the key concepts used in this architecture discussion and the relationships between them.

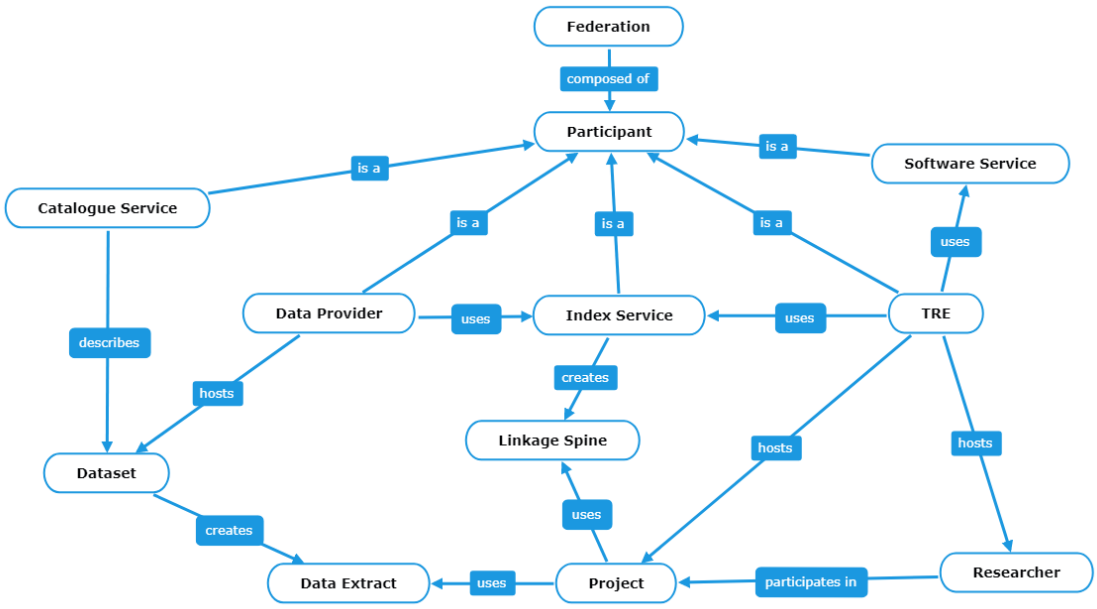


Figure 2. Concept map for the DARE UK Federation.

The map usefully illustrates the relationships between the static parts of the Federation – the Participants – and the dynamic or transitory parts – datasets, projects, researcher users and so on.

5.3 Participants

At a basic level the Federation connects Participants over a common, managed, secured document exchange network. Organisationally participants are approved to join the Federation by an organisation with overall

governance responsibility for it (cf. Chapter 8). Technically participants connect to the Federation through a single standard set of interfaces.

Participants join the Federation with one, or at most two, of five roles: as TREs, providing analytics services to Federation users; as Data Providers, providing datasets for use within the Federation; as Indexers, providing specialised services for creating linkage spines and common keys between datasets; as Catalogue Services, enabling users to discover datasets and other services available within the Federation; or as Software Services, providing an approved route to third-party software. A participant may combine the first two roles (a TRE can also host and provide data). Following the principle of separation of concerns the other roles should not be combined with any other.

Participants may offer a wide variety of services to users of the Federation. These capabilities should be advertised in a structured way (see below) and registered with the central registry service. They may also be advertised "externally" through an Internet-facing central catalogue.

5.4 Connections

Building on the principle of "separation of concerns" connections between participants are not unrestricted; only certain kinds of connection between certain kinds of participant are permitted. We use this principle to apply additional constraint requirements to each of the architectural components described in Chapter 6.

5.5 Structured Documents

Participants in the Federation communicate by exchanging structured documents over a common document exchange layer. The common document exchange layer provides the required technical security controls for exchange between participants (see Section 6.1.1.5 *Security Server*) but additional security controls may be applied to certain types of documents. In Figure 2, Data Extracts and Linkage Spines are two examples of structured documents.

Certain document types are closely associated with certain API service classes (see Section 6.1.2 *API Services*) and are produced and consumed by those API services. Others are associated with Federation security control and are produced and consumed by underpinning security services.

Documents associated with API service classes will encompass a very wide range. The Federation does not specify (nor should it) the capabilities or detailed service catalogues provided by individual participants, but it does require API services to be classified into a small number of classes. The reasoning here is that a class of API service can be mapped to a type of system actor and thus a particular security posture. As an example, an API service that moves datasets between data providers and TREs MUST NOT be usable by system actors with the role of Researcher.

The contents of structured documents will depend on the particular API services that produce or consume them. The Federation requires that all documents to be exchanged between participants be packaged in a standard way.

5.6 Federation Identities

Most of the concepts sketched in Figure 2 will, in practice, need to be identified uniquely. Each of these things 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 5.7 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 7.2).

Identity type	Example attributes
Participant	Name; List of APIs supported; List of capabilities accessible to the Federation; etc.
Researcher	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; 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.

5.7 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 [17]). To support a rich ecosystem of participants deploying different technology stacks, it is also necessary.

The sequence of events runs like this.

1. A TRE and a Data Provider establish a trust relationship, brokered by the central 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 the 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 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.

6 Federated Architecture: Infrastructure Layer

Figure 3 is a functional block diagram of the DARE UK federated architecture. It sketches a number of Federation participants – TREs, Data Providers and supporting services – and indicates the principal information flows between them. One TRE and two Data Providers are shown. In practice there will, of course, be many more.

The aim is to create a secured, trustworthy environment that connects all these participants (and any other future participants) in a common way. A single set of Federation Services hold a central record of all Federation participants and provide a set of trust services that together create the required trustworthy environment.

We have argued (Chapter 2) that DARE UK be built as a managed federation. This design draws on current best practice in secure data exchange environments but also reflects the design principle of "start 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 Data Providers and 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.

Federation Roles (System Actors)

Federation roles are represented as "stick figures" in Figure 3.

P	Member of the **Public** as ultimate provider of data.
DC	A **Data Controller** responsible for curating one or more public Datasets and accountable to the Public for their use.
R	An approved **Researcher** looking to access one or more Datasets – possibly linked together – for an approved purpose.
G	An **Information Governance** professional, acting with authorisation from DCs, responsible for oversight of data use, approved research projects and disclosure control within a given TRE.
DO	A **Data Provider Service Operator** responsible for providing necessary technical services within a Data Provider organisation to enable Data Controllers to share data onwards.
TO	A **TRE Operator** responsible for the operation of a given TRE under a given Governance authority.
CO	A **Catalogue** or **Discovery Service Operator** responsible for providing catalogue, search or other discovery services for datasets and other services available within the Federation.
SWO	A **Software Service Operator** responsible for providing access to software sources from outside the Federation.
FO	The **Federation Operator** responsible for maintaining and operating the federating infrastructure between TREs and Data Providers.
I	An **Indexer** responsible for creating and providing the cross-dataset keys or spines necessary for linking datasets together.
L	A **Linker** responsible for applying linkage spines or keys to approved project-specific – or more generally curated – Datasets to create a single linked Dataset.

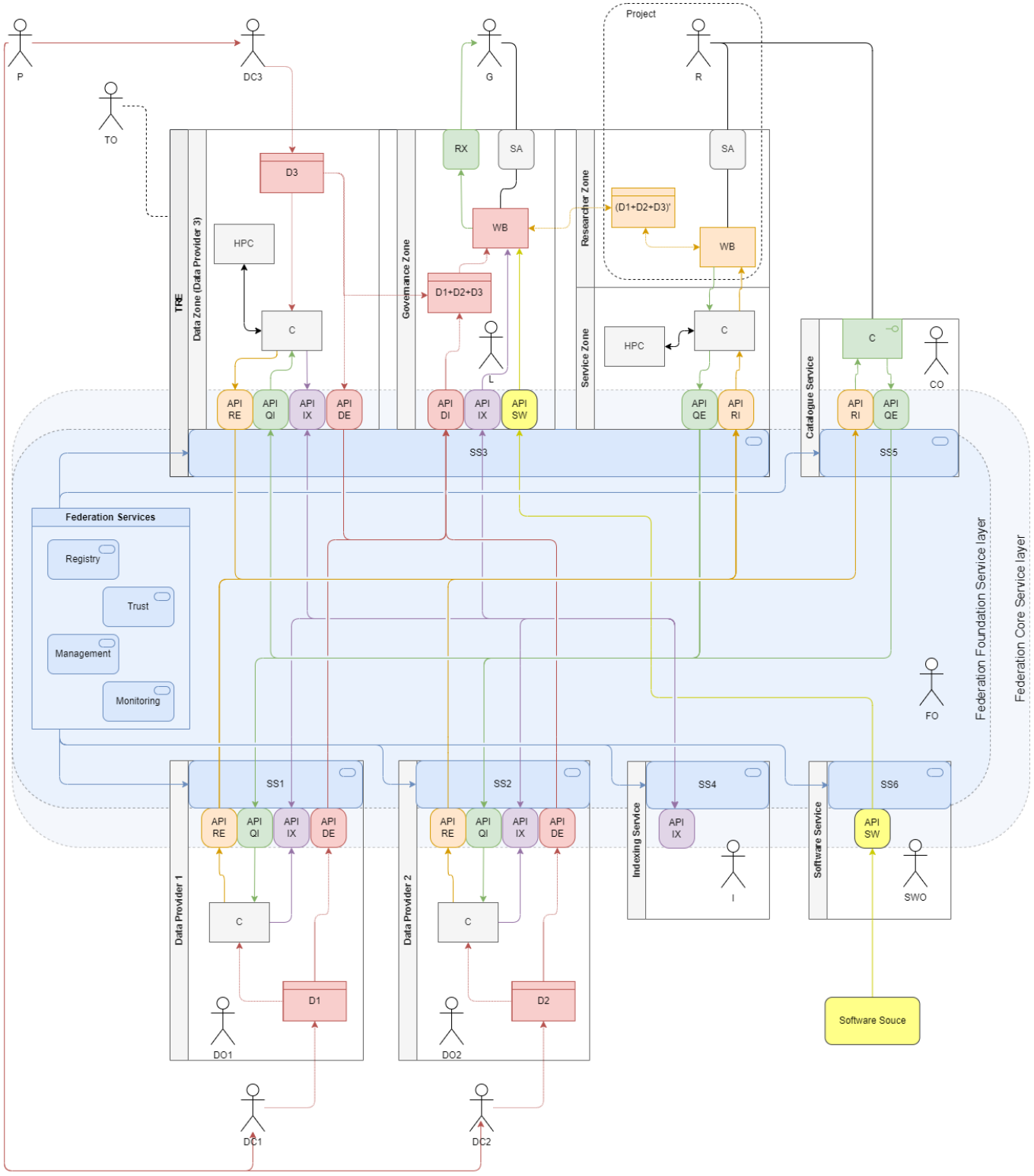


Figure 3, Functional block diagram of the infrastructure layer of the DARE UK federated network architecture. The notation broadly follows the ArchiMate v3.1 standard [22].

Federation Participants

Federation participants are represented as headed boxes with the heading typically on the left-hand side. The TRE participant is further subdivided into a number of internal zones.

TRE	A **Trusted Research Environment**. **TRE Researcher Zone**, a zone within the TRE where Researcher Workbenches can be provisioned with approved datasets for approved project work. **TRE Governance Zone**, a zone within the TRE where information Governance professionals may work with sensitive datasets. The Governance Zone and Researcher Zone SHOULD be air gapped or otherwise isolated from each other (e.g., by TRE standard operating procedures [SOPs]). **TRE Data Zone**, a zone within the TRE which enables it to act as a Data Provider for an approved Dataset. The Data Zone SHOULD be air gapped or otherwise isolated from any other zone within the TRE (e.g., by TRE SOPs). **TRE Service Zone**, a zone within the TRE providing additional computational services (e.g., query aggregators, virtual databases) that can be accessed directly by Researchers from within the TRE Researcher Zone. May include additional HPC capability.
Data Provider	An organisation overseen by a DC providing one or more sensitive Datasets. Note that a TRE MAY function as a Data Provider by providing a TRE Data Zone.
Catalogue Service	A service which enables discovery of datasets and computational services available within the Federation. This service should provide a view of available data and services to external users (i.e., to the Internet). It may achieve this by querying Registry or other services within the Federation. This dual "inward-outward" facing role will need careful security design; any outward-facing catalogue SHOULD be air gapped or otherwise isolated from any other zone within the service (e.g., by Catalogue Service SOPs).
Indexing Service	A service (potentially part of the Federation Services domain) which can create pseudonymous linkage spines from sets of personal identifiers.
Software Service	A service (potentially part of the Federation Services domain) which provides approved software packages to TREs. A Software Service MAY act in effect as a proxy to Internet-based software repositories.
Federation Services	A number of common services that together define the DARE UK federation. **Registry Services**, core user, TRE and dataset registration; hold the common identities of participants and other entities within the Federation. **Trust Services**, a set of services providing key security features such as encryption key management, certificate management and document exchange timestamping. **Monitoring Services**, providing monitoring for both "system health" and document exchange within the Federation. **Management Services**, providing the necessary services for Operators to manage the Federation.

API Services

All interactions between federation participants are conducted through API Services. API Services in turn route through common security services.

API	A collective term for a number of **Application Programming Interface** services through which any and all interactions between federated entities MUST route.
API QE	**Query Egress**, a category of API services that permit queries against a remote data service to leave a federated entity.

API QI	**Query Ingress**, a category of API services that receive and handle queries sent from QE services.
API RE	**Result Egress**, a category of API services that permit results from ingress queries to be returned to the calling entity.
API RI	**Result Ingress**, a category of API services that receive and handle results sent from RE services.
API DE	**Data Egress**, a category of API services that permit datasets D (or subsets thereof) to egress a federated entity for ingest by a DI service at another federated entity. Roles DC and G are the only permitted users of DE services; any and all other roles MUST NOT be permitted to use them.
API DI	**Data Ingress**, a category of API services complementary to DE that receive and handle ingressed datasets from other federated entities. Roles DC and G are the only permitted users of DI services; any and all other roles MUST NOT be permitted to use them.
RX	**Results Export,** a category of "API" services enabling the export of approved results from a TRE. Services in this class may not necessarily be API services at all but instead be managed export services (such as managed file transfer) requiring specific interactions from the Information Governance role.
API IX	**Indexing**, a category of API services that exchange both "bare" and pseudonymised personal identifiers to create linkage spines for data projects, but which handle no other data.
API SW	**Software**, a category of APIs which connect TREs to Software services. Software APIs are available to Information Governance users and allow software to be imported into a TRE.
SS	A **Security Server**, the only component of a federation participant permitted to communicate with other participants, including the federation Registry Services. SS receive and handle any and all messages (document exchange, queries etc.) sent by other federated entities. SS handle and dispatch any and all messages to be sent to other federated entities. Any and all communications to and from API services MUST route through both sending and receiving SSs.
WB	Virtual **Workbenches**, computational environments provided to **Researchers** or **Information Governance** professionals within a TRE, suitable for analytics or IG tasks accordingly.
SA	A **Secure-Access interface**, a remote-access service enabling controlled access to TRE **AWB** services for approved **Researchers**, and to **IGWB** services for approved information **Governance** professionals. SA disallows data egress and provides additional secure features as required by Information Governance.

Colour Schemes

Blue	The scope of DARE UK: federation services and their endpoints within TREs (especially the SS).
Red	Sensitive **Datasets** or services handling or transmitting sensitive **Datasets**. Red components MUST be treated to the highest levels of security control.
Orange	Services handling or transmitting query results which should be treated as sensitive.
Green	Services handling or transmitting queries which are unlikely of themselves to be sensitive.
Purple	Services handling what we might term "sensitive metadata", particularly lists of "bare" identifiers (NHS number, CHI, NI number etc.), but no associated data.

Other Abbreviations

D*n*	A "sensitive" **Dataset**. We use "sensitive" here to indicate in particular "individual-level". We expect Datasets to be de-identified (i.e., stripped of personally identifiable information or otherwise obfuscated) but nevertheless still individual-level and thus potentially re-identifiable.
D*n* + D*m*	A combination of **Datasets *n*** and ***m*** (e.g., linked using a common key or spine).
{D*n* + D*m*}'	A prepared subset of combined **Datasets *n*** and ***m*** suitable for provision to an approved **Researcher**.

C*n*	Some unit of **Computing** capability (e.g., a virtual machine).
HPC	A unit of **High-Performance Computing** capability or equivalent additional processing resource above and beyond that provided by Cs (e.g., a GPU system).

6.1 Federation Core Services

Core services encompass the **Foundation Services** that define the Federation itself and the **API Services** that enable application-level interactions between Federation participants.

6.1.1 Foundation Services

Foundation services provide a secure document exchange layer and set of gatekeeping, registration and discovery services which, taken together, define the DARE UK federation. The lowest level of the federation layer is agnostic towards both the nature of any exchanged documents and the purposes for which they are exchanged (see Structured Documents).

6.1.1.1 Registry Services

Registry services record information about the different pieces of the federation. Each of the *Contexts* described earlier defines an information class that should be recorded in a Registry. At a basic level there are four kinds:

- Federation participants. Which participants, defined by their security servers (qv), are part of the Federation. There are five kinds:
  - Data Providers;
  - TREs;
  - Software Services;
  - Catalogue Services;
  - Indexing Services.
- Datasets. Datasets are provided by Data Providers 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.

6.1.1.2 Trust Services

Trust services provide the necessary services for securing the foundational document 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.

6.1.1.3 Monitoring Services

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

6.1.1.4 Management Services

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

6.1.1.5 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 Foundation 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 document exchange layer and hence are the guarantors of the confidentiality, integrity and auditability of inter-participant exchanges within the Federation.

6.1.2 API Services

API services expose various capabilities for use by other members of the Federation. Note that traffic to and from all API Services route first through the Security Servers of the host Participant.

Note also that we use the terms “ingress” and “egress” to mean “ingress from another Federation participant” and “egress to another Federation participant”. Core Federation Services do not, by definition, connect Federation participants to the wider Internet.

At this level of the architecture we do not specify the details of individual API calls but rather seek to classify API services into a small number of types, each of which will have a defined security context. Our intention is to leave open the definitions of particular APIs to promote innovation and expansion within the Federation, while providing an overall framework within which services can be placed.

6.1.2.1 Query Egress/Ingress (API QE/QI)

Query Egress/Ingress pairs. Queries can range from the simple (e.g., a single invocation of a remote API) to the complex (e.g., a federated query submitted more like a computational job).

Query services can be invoked by Researchers from within a TRE, or by discovery services from within a suitably configured Catalogue Service (cf. Section 6.3).

Query Egress services MUST connect solely to Query Ingress services. (R032)

Conversely, Query Ingress services MUST connect solely to Query Egress services. (R033)

Query Egress/Ingress service pairs exchange structured documents of type “Query” (see Section 6.6.1 *Queries*).

6.1.2.2 Results Egress/Ingress (API RE/RI)

Results Egress/Ingress pairs. Results are, broadly, data generated by executing a query within the federation. The invocation of a results API service is triggered indirectly by the prior invocation of a query service. Our working assumption is that, within the secure Federation, results can be returned safely to a querying entity without the need for IG intervention.

Results Egress services MUST connect solely to Results Ingress services. (R034)

Conversely, Results Ingress services MUST connect solely to Results Egress services. (R035)

Results Egress/Ingress service pairs exchange structured documents of type “Results” (see Section 6.6.2 *Results*).

It is assumed that whatever information is received by a Results Ingress service is passed back to the originator of the triggering query (the Researcher or Catalogue Service). This context must be taken into account when designing what information should be transmitted by the pairing Results Egress service; whether this exchange can be fully automated is ultimately a case-by-case governance question.

6.1.2.3 Data Egress/Ingress (API DE/DI)

In contrast to returning 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/results APIs.

Data Egress services MUST connect solely to Data Ingress services. (R036)

Conversely, Data Ingress services MUST connect solely to Data Egress services. (R037)

System actors with roles of Information Governance or Data Provider Service Operator only SHALL be able to invoke Data Ingress/Egress services. (R038)

System actors with other roles SHALL NOT be able to invoke Data Ingress/Egress services. (R039)

6.1.2.4 Indexing (API IX)

Indexing API services provide a mechanism for Information Governance roles within a TRE, Data Providers and Indexing Services to exchange lists of personal identifiers, corresponding lists of depersonalised identifiers and master linkage spines for different Datasets. For more information see Section 6.2 *Indexing Service* below.

Indexing API services MUST connect solely to Indexing API services. (R040)

As with Data Ingress/Egress services, system actors with roles of Information Governance or Data Provider Service Operator only SHALL be able to invoke Indexing services. (R041)

System actors with other roles SHALL NOT be able to invoke Indexing services. (R042)

6.1.2.5 Software Ingress (API SW)

Software Ingress API services provide a mechanism for Information Governance roles within a TRE to download and import approved software from a Federation Software Service. Note that this differs from services, already realised in some extant TREs, that permit the import of software from approved or controlled sources available on the wider Internet.

Software Ingress API services MUST connect solely to Software Ingress API services. (R043)

System actors with roles of Information Governance only SHALL be able to invoke Software Ingress services. (R044)

System actors with other roles SHALL NOT be able to invoke Software Ingress services. (R045)

6.2 Indexing Service

An Indexing Service creates depersonalised linkage spines for different Datasets by converting between “bare” personal identifiers and project-specific linkage keys.

Indexing Services must be trustworthy enough potentially to handle personal identifiers by which horizontally 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 7 *Federated Architecture: Data Layer*.

Indexing Services interact with other Federation participants solely through Indexing API service calls.

6.3 Catalogue Service

A Catalogue Service provides information (metadata) about features of the Federation to users outside the Federation.

Catalogue Services are assumed to make use of the Query/Results API Services discussed above. This has two consequences:

- Because Query API Services encompass a range of capabilities Catalogue 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 [32].
- Because Catalogue Services potentially provide a window from the outside to the inside of the Federation the governance of any particular instance must be carefully considered.

6.4 Software Service

A Software Service provides centralised access for Federation participants to sources of software outside the Federation. Note that this differs from a TRE-based software import service which connects a TRE directly (by suitable network proxying, for example) to an Internet-based software repository.

Software Services may be implemented in a number of ways, but in all cases their expected use is to serve Software Ingress API invocations by Information Governance roles.

A Software Service may:

- act as a direct network proxy for Internet-based third-party software services (e.g., CRAN[1]);
- act as an independently curated, high-assurance mirror service for popular software packages (e.g., Anaconda Python Enterprise[2]);
- act as a proxy for defined and approved user accounts on a public open-source software repository (e.g., GitHub[3]);
- and so on.

Software Services interact with other Federation participants solely through Software Ingress API service calls.

6.5 TRE Components and Tools

We include this section for completeness, although we note that specification of the details of any particular TRE are out of scope of this architecture. TREs interact with each other through the Core Federation Service layer described above; this architecture does not prescribe how they might provide services to their users.

What is in scope are standard descriptions of TRE capabilities that can be registered with the Federation Registry Services as part of a TRE's induction and attachment to the Federation. These are discussed in Section 7.2.1 *Federation Metadata* below.

6.5.1 General Processing

As a target for remote query execution, we need a way to describe and register processing capabilities.

[1] The Comprehensive R Archive Network. See https://cran.r-project.org/
[2] Anaconda Python Enterprise DS Platform. See https://www.anaconda.com/products/enterprise
[3] GitHub. See https://github.com/

6.5.2 High-Performance Computing

As a target for remote query execution, we need a way to describe and register additional high-performance processing capabilities.

6.5.3 Information Governance Workbench

Any details are out of scope.

6.5.4 Analytical Workbench

Any details are out of scope.

6.5.5 Secure Access

Any details are out of scope.

6.6 Structured Document Types

6.6.1 Queries

Query documents can originate from Researcher users within a TRE. They are typically targeted at one or more data resources remote from the Researcher (Data Providers or another TRE). Query documents are sent via Query Egress API services and are consumed by Query Ingress API services at the remote participants. Query documents are not expected to contain sensitive data and are expected to be egressable into the Federation without disclosure control or governance oversight. Like all other structured documents described here their confidentiality, integrity and traceability is guaranteed by the secure document exchange layer common to all Federation participants.

6.6.2 Results

Result documents originate from Data Providers and are the “answers” to Query documents submitted to them. They are sent via Results Egress API services and are consumed by Results Ingress API services at the query originator’s participating TRE. The query originator is quite likely to be a Researcher user. Result documents may contain sensitive data (depending on the nature of the data resource queried) and their egress from the Data Provider may need disclosure control or intervention by information governance.

6.6.3 Datasets

Dataset documents originate from Data Providers and are datasets or extracts of datasets that have been approved by a Data Controller for specific uses within the Federation. Dataset documents will typically contain sensitive data, often de-identified but individual-level personal data. Dataset documents are sent via Data Egress API services and are consumed by Data Ingress API services. Use of Data Ingress and Egress API services must be restricted to Data Providers or Information Governance users only.

6.6.4 Indexes

Index documents are exchanged by Index API services between Information Governance, Data Provider and Index Service roles. Index documents do not contain sensitive data but could be said to contain “sensitive metadata”. Indexing individuals means that index documents will contain lists of personal identifiers and their exchange must be governed accordingly.

Index documents are needed for certain kinds of data linkage. See Section 7.5.4 *Data Linkage* for a fuller treatment.

7 Federated Architecture: Data Layer

This chapter is an outline. It will be further developed in later versions.

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.

7.1 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 GPDR [21]. A firm's intellectual property may fall under the Copyright Designs and Patents Act [26]. Where the data subject is an endangered species, its treatment may be covered by international treaty such as CITES [27]. Still other subjects may require certain approaches because of cultural sensitivity[4].

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 [28] (OFFICIAL, SECRET and TOP SECRET), or four if the UK's OFFICIAL-SENSITIVE is counted separately. (ISC)[2], the International Information System Security Certification Consortium, defines five in its standard commercial scale [29]. 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 [30]. Appendix D covers these in more detail.

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 [31].)

7.1.1 A Seven-Point Scale

DARE UK facilitates 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 DARE UK Federation be recorded with two key pieces of information and a number from 0-6 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 custodian. Thus, we do not distinguish between data release to a small group and data release to everyone.

Datasets should be classified by:

[4] 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.

- 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
0	None	Public	Public	Public
1	Inconvenience	-	-	Proprietary
2	Distress, embarrassment, some reputational damage	OFFICIAL	Personal	Private
3	Actual harm	OFFICIAL-SENSITIVE	Personal	Confidential
4	Serious harm	OFFICIAL-SENSITIVE	Special Category	Sensitive
5	Loss of life	SECRET	-	-
6	Widespread loss of life	TOP SECRET	-	-

7.2 Metadata

We can divide metadata into two groups: metadata that capture information about the Federation itself (Federation metadata); and metadata that capture information about the datasets accessible within the Federation (content metadata).

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; …
- Participant-private: Researchers' and other users' contact details; …

7.2.1 Federation Metadata

Our definition of Federation metadata is best captured in the answer to the question: if the Federation held no datasets 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".

7.2.1.1 Descriptive Metadata

The Participants, services, users and other entities described in Chapter 6 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).

Descriptive metadata for a Participant could be structured to reflect static and dynamic aspects:

- Static:
  - Basic metadata: name, Federation identity, …
  - Capabilities: available computation; available software; …
- Dynamic
  - Datasets hosted (persistently available not project-specific): count; list of Federation identities; …

- Projects hosted: count; list of names; list of Federation identities; …
- Indexes hosted (types of linkage available): list of Federation identities; …
- Users hosted (registered user accounts at this Participant): count; 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!).

7.2.1.2 Operational Metadata

Operational metadata are metadata captured and recorded through the operation of the Federation and its Participants. Operational metadata notably include information on document exchange logged by the Participant Security Servers and by the central 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 central Federation Services, and which might be visible to other Federation Participants.

No operational metadata should ever be visible to the public.

7.2.2 Content Metadata

Content metadata describe the data and projects the Federation supports. Note that, as illustrated in Figure 2, concepts like Dataset arise in multiple contexts. 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.

7.2.2.1 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 Catalogue Services to collate for search and discovery purposes.

As an 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.

Most detailed Dataset metadata will be Federation-private.

7.2.2.2 Project Metadata

As discussed in Chapter 5 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 decisions can be taken by Federation Participants, for example upon receipt of a remote query.

Summary metadata for a Project will be public.

Most detailed Project metadata will be Federation-private.

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

7.2.2.3 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 document exchanged between Participants.

Metadata for Data Extracts will be logged by the secure document 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.

7.2.2.4 Other Structured Documents

Many exchanges of structured documents 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 documents be tagged with a metadata record indicating this Project context.

7.3 Data Findability

As described in Section 4.1 *Rachel's Journey* data findability needs to happen outside the Federation, before a researcher has even defined the project they might ultimately propose.

Detailed metadata on Datasets is created and registered by Data Providers within the Federation.

The Federation architecture as proposed does permit the exposure, via query APIs, 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 it 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 [32] and in the HDR-UK CO-CONNECT work [33].

7.4 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.

7.5 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 this initial version of the architecture. We note them here and may expand on them in future iterations.

7.5.1 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 4.2 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 (are 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.

7.5.2 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 central terminology bases, in much the same way that computer-assisted translation tools work today. (The proposed EU Smart Middleware Platform architecture includes just such a central vocabulary service [17].)

By introducing a centralised terminology service, it is feasible to imagine an Index Service which could automate the linkage of two datasets under these conditions.

7.5.3 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.

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

7.5.4 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 Data Providers, 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.

Appendix D offers a sketch for how, under ideal but plausible conditions, an Indexing Service could be automated to provide project-specific linkage spines.

7.6 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 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). As with the details of particular services available with TREs (Section 6.5 *TRE Components and Tools*) we do not expand on the details of such a service here.

8 Federated Architecture: Governance Layer

This chapter is an outline. It will be further developed in later versions.

By its very nature the federated architecture presented in Chapters 5, 6 and 7 already implies a certain organisational structure and hence points towards certain governance needs. The Federation is defined by a common set of security standards and document exchange protocols orchestrated and managed to create a trustworthy network between existing TREs, data providers and other service providers.

The existing stakeholders already have governance arrangements in place to enable research with sensitive data within TREs. When considering governance arrangements for the Federation as a whole we adopt the principle that Federation governance should not disrupt existing governance arrangements for participants wanting to join. Any Federation governance should *extend* and not replace any existing governance arrangements.

8.1 The Project Model

Implicit throughout this document is the importance of the concept of "Project". This importance is highlighted in the governance discussions below.

A Project is a defined and approved piece of research work that enables analysis on agreed data resources (and possibly using agreed methods) for an agreed period of time. The data resources may be assembled from individual datasets, or they may be provided dynamically, as agreed. By definition Projects are time-limited, though they may be long-lived or easily extended. Section 4.1 *Rachel's Journey* describes how a Project can come to be defined.

Projects have a single designated owner, the Principal Investigator or PI (borrowing academic terminology). The PI may have a project team and may themselves do little or no actual work on the project, but they are responsible (indeed, accountable) for any and all activities of the Project and its staff, and for ensuring the Project meets any constraints or conditions specified in its approvals.

8.2 Stakeholder Map

In Figure 4 we map out the stakeholders involved in the DARE UK Federation and the key relationships between them. We use the standard RACI terminology to indicate the nature of each relationship, with the arrow direction indicating a "subject-to-object" relationship (e.g., a Researcher PI is accountable to Information Governance for the activities within her project).

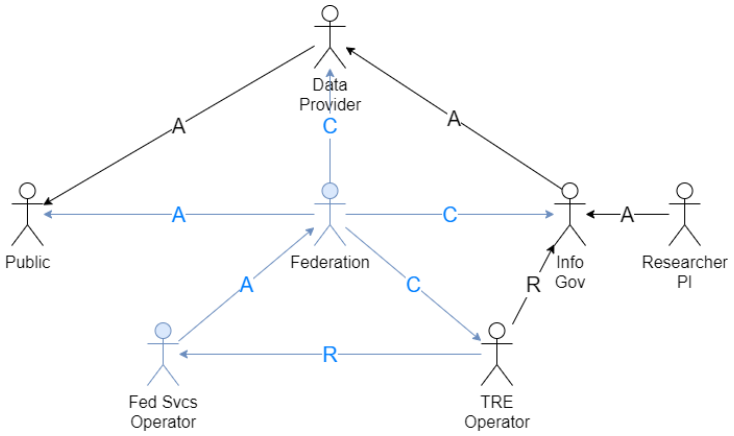
8.2.1 Existing Relationships

Data Providers, as custodians, controllers and curators of sensitive data, are ultimately accountable to the Public for the use of public data.

Information Governance (IG) bodies, as orchestrators of much of the sensitive data research landscape, are accountable to Data Providers for how they handle the data entrusted to them.

Researcher PIs, as owners of approved research Projects, are accountable to IG for any and all activities within those Projects.

TRE Operators run services on behalf of Information Governance and are responsible to them for the correct running of the TRE.



Responsible: the person, people or organisation responsible for correct execution – for getting the job done.

Accountable: the person (ideally an individual within an organisation) who has overall ownership of service quality and the end result.

Consulted: The people or organisations who are consulted and whose opinions are sought.

Informed: The people who are kept up-to-date on progress or status.

Figure 4. Map of accountabilities and responsibilities between stakeholders of the DARE UK Federation. New Federation entities and relationships are indicated in blue/light shading.

8.2.2 New Relationships

The Federation as a whole is represented by some body, called Federation in Figure 4, which is representative of all the other stakeholders (we indicate this by a "consults" relationship). The Federation is ultimately accountable to the Public.

The Federation Services Operator runs the central federation services discussed in early chapters and is accountable to the Federation for safe and successful operation.

TRE Operators, in joining the Federation and deploying a Federation Security Server (cf. Section 6.1.1.5), become responsible to the Federation Services Operator for the running of that Security Server.

8.2.3 Impact on Researchers

A possible exception to the "no impact on existing governance" principle may arise if higher levels of user accreditation are required for a TRE (say) to join the Federation. While research users will seldom if ever encounter the Federation first-hand, they may be required to undergo additional training if the Federation's currently agreed minimum is above the level required by their TRE. It is to be hoped that such impacts will be minimal, and rare.

8.3 Federation Governance Scope

Regardless of how the Federation governance is constructed it needs to be able to undertake a number of tasks, including but not limited to:

- Agree baseline technical standards for the Federation. This may involve defining or approving invitations to tender for technology suppliers of central 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 baseline training or accreditation standards for Federation users, including service operators, Researcher PIs and other researchers.
- Manage the appointment and oversight of the Federation Services Operator.
- 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 central Federation Services software;
    - changes to document exchange protocols or formats;
    - changes to metadata standards.
- Oversee the monitoring of Federation participants' progress towards achieving nationally-agreed standards of operation.
- Oversee regular audit and accreditation for the Federation as a whole. (The practicalities of this are the responsibility of the Federation Services Operator.)

9 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. This chapter sketches a phased delivery approach which is expanded further in the DARE UK Business Case and Federated Network Funding Model [34].

9.1 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) required[5].

9.1.1 Foundation Services: Technology Evaluation

Foundation 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 an expert 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).

9.1.2 API and other Services: Community Driver Projects

Securing the foundation layer allows for greater innovation at the API and application level without increasing risk. The core API services that run on top of the document exchange foundation can thus be drawn from a wider ecosystem.

We recommend commissioning research and development projects to investigate different technological approaches to the required core services. DARE UK's Phase 1b Driver Projects are a model approach[6].

9.2 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.

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.

9.2.1 Scenario 1: Basic Data Exchange

This is the base scenario involving the core Federation Services and secure document exchange between a Data Provider and a TRE.

Required components:

- 1 x Central Federation Services (Foundation);
- 1 x TRE: Security Server (Foundation); API DI (Core);

[5] Technology Readiness Levels. See https://en.wikipedia.org/wiki/Technology_readiness_level
[6] See https://dareuk.org.uk/our-work/phase-1-driver-projects/

- 1 x Data Provider: Security Server (Foundation); API DE (Core).

Required concepts:

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

9.2.2 Scenario 2: Linked Data Exchange

This scenario extends the first with an additional Data Provider and introduces an Indexing Service.

Required components:

- 1 x Central Federation Services (Foundation);
- 1 x TRE: Security Server (Foundation); API DI (Core); API IX (Core);
- 2 x Data Provider: Security Server (Foundation); API DE (Core); API IX (Core);
- 1 x Indexing Service: Security Server (Foundation); API IX (Core).

Required concepts:

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

9.2.3 Scenario 3: Remote Query

This scenario exercises the movement of queries rather than the movement of data and can be viewed as complementary to Scenario 1.

Required components:

- 1 x Central Federation Services (Foundation);
- 1 x TRE: Security Server (Foundation); API QE (Core); API RI (Core);
- 1 x Data Provider: Security Server (Foundation); API QI (Core); API RE (Core).

Required concepts:

- Identities: Participant; Project; Dataset;
- Structured Documents: Query; Results.

9.2.4 Scenario 4: Federated Query

This scenario extends the remote query scenario to include a second data provider and tests the splitting of a query to run against each independently. Note that the Query Egress API Service implementation required here is much more sophisticated than that in the simple remote query case.

Required components:

- 1 x Central Federation Services (Foundation);
- 1 x TRE: Security Server (Foundation); API QE (Core); API RI (Core);
- 2 x Data Provider: Security Server (Foundation); API QI (Core); API RE (Core).

Required concepts:

- Identities: Participant; Project; Dataset;
- Structured Documents: Query; Results.

9.3 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.

9.4 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 jargon[7]).

We recommend including a dedicated red team testing component in future operational plans for the DARE UK federation.

9.5 Evolution

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

- Scope: new TREs and Data Providers can be added to the network by deploying Security Servers and supporting appropriate API Services;
- Functionality: new API 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.

[7] 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

10 Summary and Further Work

This report 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 document exchange and broad classes of API 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 query where complex "query objects" can be submitted securely to remote data services.

We describe the architecture in three layers: infrastructure, data and governance. This "initial" version covers the infrastructure layer in some detail and the data and governance layers in less detail. We invite comment from the broader UK research community on the ideas and approaches presented here. Two further versions ("interim" and "final") will incorporate community feedback over the course of the year.

11 References

[1] DARE UK; *Initial Phase 1 Recommendations*; https://dareuk.org.uk/our-work/dare-uk-phase-1-recommendations/ (accessed 01/03/2023)

[2] 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).

[3] UK Health Data Research Alliance; *Trusted Research Environments (TRE), A strategy to build public trust and meet changing health data science needs*; draft Green Paper v1.0; 30 April 2020; https://ukhealthdata.org/wp-content/uploads/2020/04/200430-TRE-Green-Paper-v1.pdf

[4] 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

[5] The WLCG Collaboration; *The World-wide LHC Computing Grid*; https://wlcg.web.cern.ch/ (accessed 09/03/2023).

[6] The IVOA; *The International Virtual Observatory Alliance*; https://ivoa.net/ (accessed 09/03/2023).

[7] ELIXIR; *A distributed infrastructure for life science information*; https://elixir-europe.org/ (accessed 09/03/2023).

[8] BBMRI-ERIC; *A European research infrastructure for biobanking*; https://www.bbmri-eric.eu/ (accessed 09/03/2023).

[9] ESFRI; *The European Strategic Forum on Research Infrastructures*; https://www.esfri.eu/ (accessed 09/03/2023).

[10] CESSDA; *The Consortium of European Social Science Data Archives*; https://www.cessda.eu/ (accessed 09/03/2023).

[11] ESS-ERIC; *The European Social Survey*; https://www.europeansocialsurvey.org/ (accessed 09/03/2023).

[12] NIIS; *X-Road Architecture*; https://x-road.global/architecture (accessed 02/03/2023).

[13] NIIS; *The Nordic Institute for Interoperability Solutions*; https://www.niis.org/ (accessed 02/03/2023).

[14] Digital Nations; https://www.leadingdigitalgovs.org/ (accessed 09/03/2023).

[15] GAIA-X; *A Federated Secure Data Infrastructure*; https://gaia-x.eu/ (accessed 09/03/2023).

[16] 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).

[17] 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).

[18] European Commission; *A European Strategy for data*; policy paper; https://digital-strategy.ec.europa.eu/en/policies/strategy-data (accessed 09/03/2023).

[19] 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).

[20] UK Government; *Data Protection Act 2018*; https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (accessed 09/03/2023).

[21] UK Government; *The UK General Data Protection Regulation*; https://www.legislation.gov.uk/eur/2016/679/contents (accessed 09/03/2023).

[22] 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.

[23] IETF; *The Internet Engineering Taskforce*; https://www.ietf.org/ (accessed 20/03/2023).

[24] S. Bradner; *RFC 2119*; The Internet Engineering Taskforce Network Working Group; March 1997; https://www.rfc-editor.org/rfc/rfc2119 (accessed 13/03/2023).

[25] The Open Group; *ArchiMate 3.1 Specification*; https://pubs.opengroup.org/architecture/archimate3-doc/toc.html (accessed 20/03/2023).
[26] UK Government; *Copyright, Designs and Patents Act 1988*; https://www.gov.uk/government/publications/copyright-acts-and-related-laws (accessed 20/03/2023).
[27] CITES; *Convention on International Trade in Endangered Species of Wild Fauna and Flora*; https://cites.org/eng (accessed 20/03/2023).
[28] 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).
[29] (ISC)^2; *Certified Information Systems Security Professional*; https://www.isc2.org/Certifications/CISSP (accessed 20/03/2023).
[30] 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).
[31] Sweeney L, Crosas M, Bar-Sinai M; *Sharing Sensitive Data with Confidence: The Datatags System*; Technology Science, 2015101601. October 15, 2015; https://techscience.org/a/2015101601/ (accessed 20/03/2023).
[32] GA4GH Beacon Group*; Beacon v2 standard*; https://docs.genomebeacons.org/ (accessed 23/03/2023).
[33] HDR-UK; *The CO-CONNECT Project*; https://www.hdruk.ac.uk/projects/co-connect/ (accessed 23/03/2023).
[34] DARE UK; *Business Case and Federated Network Funding Model*; in preparation.
[35] OpenSAFELY; *The OpenSAFELY Secure Analytics Platform*; https://www.opensafely.org/ (accessed 23/03/2023)
[36] Intel; *Open Federated Learning (OpenFL)*; https://openfl.readthedocs.io/en/latest/index.html (accessed 28/03/2023)

A A Comparison of Contemporary Federated Data Architectures

Annex III of the *Recommendation Report* for the EU Smart Middleware Platform (SiMPl) [17] compares the concepts defined in the SiMPl architecture with those defined in the GAIA-X framework [16]. The table below extends this idea to include both concepts defined in this document and the equivalents from the X-Road architecture [12].

DARE UK	GAIA-X	SiMPl	X-Road	Notes
Participant	Participant	Organisation that deploys an SMP Agent	Organization	
Federation Services	Federator	Data Space governance	Central Services & Trust Services	
Security Server	Sovereign Data Exchange	SMP Agent	Security Server	The GAIA-X mapping is imprecise. It factors out a number of functions that are encapsulated in the other three models.
TRE	Consumer or Service instance	Composite of Application Provider and Infrastructure Provider	Service Consumer Information System	A DARE UK TRE has no direct equivalent but is a specialised example of a generic service.
Data Provider	Provider	Data provider	Service Provider Information System	
Researcher (User)	End User	End user	Data Requestor	
Catalogue Service	Catalogue	Data catalogue	Service Provider Information System	A catalogue 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 Information System	All DARE UK services can be modelled the same way in terms of their interaction with the federation structure.

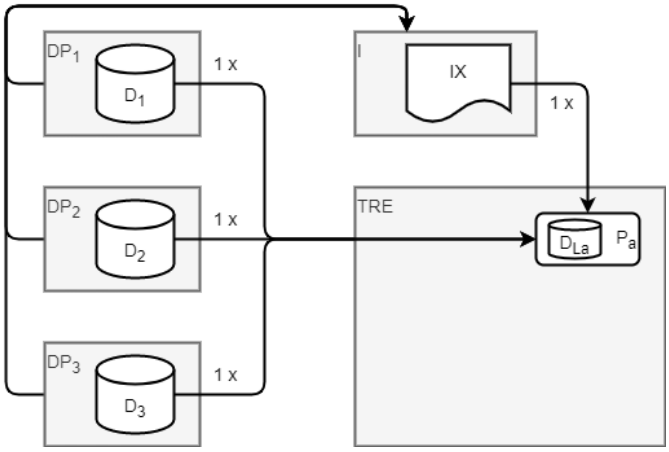
B Usage Patterns

These diagrams sketch patterns of interaction between trusted research environments (TREs), data provider (DP) services and trusted third-party Indexing services (I). We label datasets as D and use P to denote an abstract notion of a project, meaning an analysis environment created to answer a particular research question. Indexes or linkage spines, denoted IX, are created by Indexing services.

B.1 UP1. Transient data assembly, transient projects

For each approved project:

- Data are transferred into the TRE.
- A linkage index is transferred into the TRE.
- Data are linked in the project inside the TRE.
- The project environment, and its linked data, are discarded at project conclusion.



B.1.1 Current examples

- The Scottish National Safe Haven (pre-covid-19).
- The (original) Administrative Data Research Network.

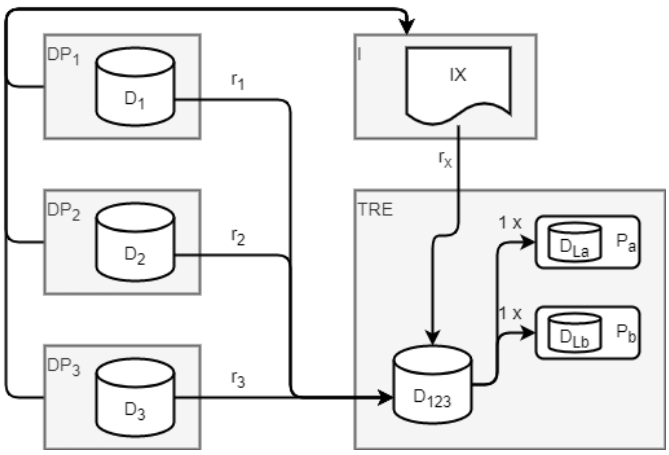
B.2 UP2. Persistent data assembly, transient projects

With some rates r_n of update:

- Data are transferred into the TRE & curated.
- A linkage index is transferred into the TRE.
- Data are linked inside the TRE.

For each approved project:

- Approved cohort data are provided to the project inside the TRE.
- The project environment, and its linked data, are discarded at project conclusion.



B.2.1 Current examples

- The Scottish National Safe Haven (through covid-19).
- The (newer) Administrative Data Research UK model.
- The ONS Secure Research Service.
- SAIL Databank.

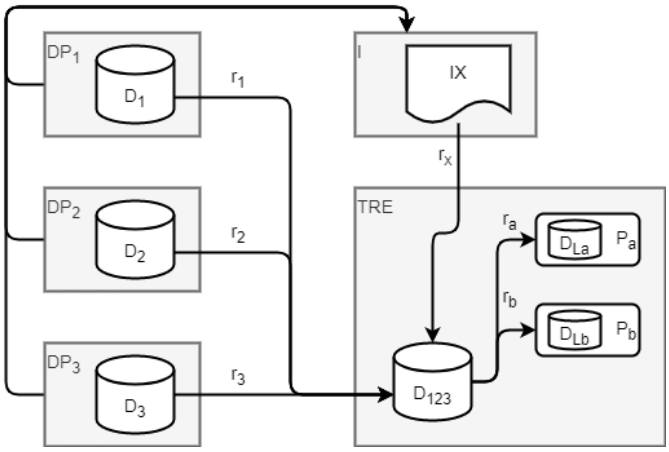
B.3 UP2 variant 1. Persistent data assembly, transient projects, refreshed data views

With some rates r_n of update:

- Data are transferred into the TRE & curated.
- A linkage index is transferred into the TRE.
- Data are linked inside the TRE.

For each approved project:

- Access to the linked dataset for approved cohort data is provided to the project inside the TRE.
- The project cohort data are refreshed at some rate r_m.
- The project environment, and its linked data, are discarded at project conclusion.



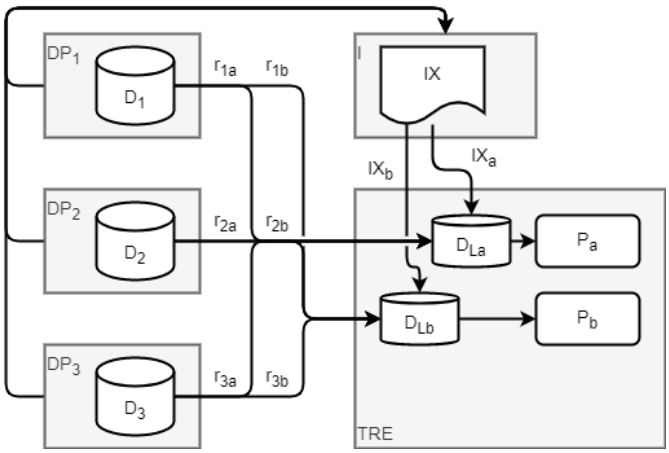
B.3.1 Current examples

- The Outbreak Data Analysis Platform.
- The Smart Data Foundry Research Environment Safe Haven (see https://smartdatafoundry.com/services/research)

B.4 UP3. Persistent data assembly, persistent projects

For each approved project:

- Once:
  - A linkage index is transferred into the TRE.
- At some rate r_n:
  - Data are added to the project inside the TRE.
  - The project and its linked data persist over a considerable period.



B.4.1 Current examples

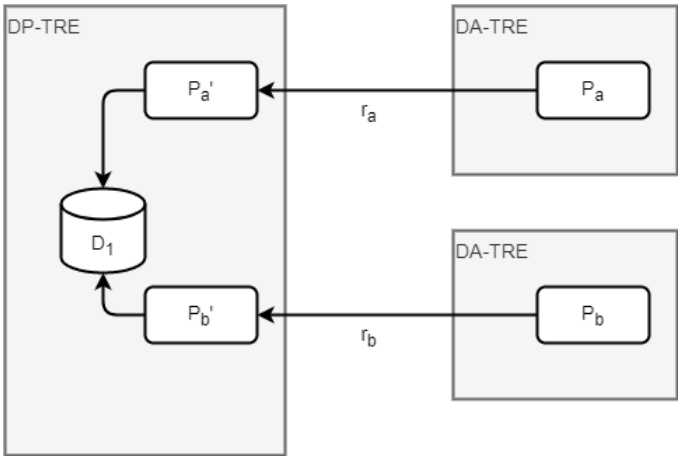
- Many clinical trials.
- The UK Longitudinal Linkage Collaboration.

B.5 UP5. Persistent data assembly, remote projects

The dataset in question may have been assembled from multiple data sources held elsewhere (cf. Pattern 2).

For each approved project at some rate r_m:

- Queries are sent from projects in DA-TRE (data analytics TREs) to a remote DP-TRE (data provider TRE).
- Each query is run against the dataset & results returned to the originating project in its DA-TRE.



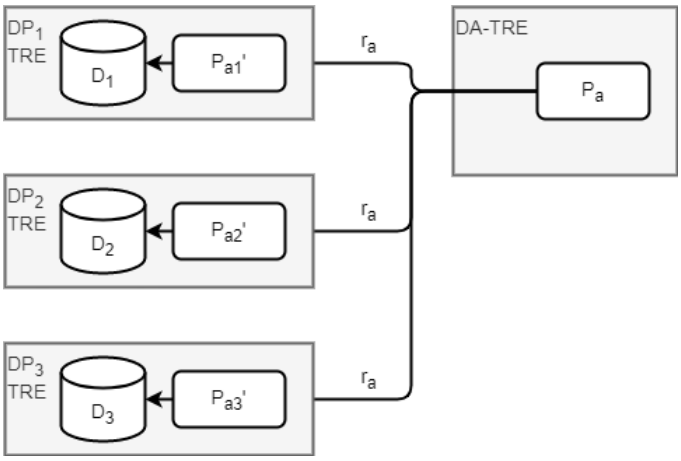
B.5.1 Current examples

- ?

B.6 UP6. Persistent data assembly, federated query

For each approved project at some rate r_m:

- Queries are sent from the project in a DA-TRE (data analytics TRE) to a number of remote DPTREs (data provider TREs).
- The query is run against each dataset & results returned.
- Results are combined within the project inside the originating DA-TRE.



B.6.1 Current examples

- OpenSAFELY.
- OpenFL [35].

C Master Requirements Table

Key:

- U/C/S/P (column 3): use-case / constraint / user story / programme-level.
- Wt (weight): calculated from number of personas or system-level requirement.
- MoSCoW: Must, Should, Could, Won't. Cf RFC 2119 [24].
- Sys/User: Whether the requirement arises from systems analysis of the current landscape, or from user persona analysis.
- # Pers.: for a User requirement, the number of personas raising this requirement.

ID	Description	U/ C/ S/ P	Wt	MoS CoW	Sys/ User	# Pers.	Tags/ headings
R001	The Programme MUST demonstrate the impact of the DARE UK federation	P	12	M		4	Transparency
R002	The Programme MUST communicate clearly and publicly on key concepts	P	9	M		3	Transparency
R003	The Federation MUST ensure the confidentiality of data storage	C	9	M	U	3	Security; TRE; DP
R004	The Federation MUST ensure the confidentiality of data exchange	C	9	M	U	3	Security; Fed Svcs
R005	The Federation MUST enable linkage between syntactically similar data	U	9	M	U	3	Fed Svcs; IX; DP
R006	The Federation MUST enable linkage between syntactically dissimilar data	U	9	M	U	3	Fed Svcs; IX; DP
R008	The Federation MUST reduce the barriers to data access	C	9	M	U	3	Fed Svcs
R009	The Federation MUST ensure the integrity of data exchange	C	6	M	U	2	Security; Fed Svcs
R010	TREs MUST ensure the security of data access and use	C	6	M	U	2	Security; TRE
R011	The Programme MUST demonstrate the security of data exchange practices	P	6	M		2	Security; Transparency
R012	The Programme MUST demonstrate the security of data storage practices	P	6	M		2	Security; Transparency
R013	The Programme MUST demonstrate the security of data access and use practices	P	6	M		2	Security; Transparency
R014	The Federation MUST ensure research use is appropriately recorded in metadata records	U	6	M	U	2	Fed Svcs; TRE
R007	I am fustrated by poor data quality	S	3		U	3	DP; Fed Svcs
R017	The Programme MUST provide clear public signposts to data used in the Federation	P	3	M		1	Transparency
R020	The Federation MUST ensure data controllers are appropriately recorded in metadata records	U	3	M	U	1	Fed Svcs; Metadata
R021	The Data Provider MUST make data sharing as easy as possible	C	3	M	U	1	DP
R024	The Federation MUST facilitate data discovery across the network	U	3	M	U	1	Discovery Svc; Metadata
R015	I am missing technical and data science skills	S	2		U	2	
R016	I find it challenging to access and build relevant collaborations	S	2		U	2	
R018	Data Providers SHOULD provide tooling for pseudonymising data	U	2	S	U	1	Security; DP
R019	Data Providers SHOULD provide tooling for assessing data anonymity	C	2	S	U	1	Security; DP
R023	The Federation SHOULD enable discovery of and access to modern data science computational capabilities	U	2	S	U	1	Discovery Svc; Metadata

R025	TREs SHOULD provide metadata on access charges and running costs	U	2	S	U	1	TRE; Usage costs
R022	I worry about understanding policies and regulations "correctly"	S	1		U	1	
R026	I find visualising large quantities of disparate data challenging	S	1		U	1	
R027	I stress about earning considerably lower income in the public sector	S	1		U	1	
R028	I want to speed up my workflow	S	1		U	1	
R029	I want to grow opportunities for my organisation	S	1		U	1	
R030	I want to be able to retain talent in my centre	S	1		U	1	
R031	I want to generate business value through data	S	1		U	1	
R032	Query egress services MUST connect solely to query ingress services	C	9	M	S		API Query
R033	Query ingress services MUST connect solely to query egress services	C	9	M	S		API Query
R034	Results egress services MUST connect solely to results ingress services	C	9	M	S		API Results
R035	Results ingress services MUST connect solely to results egress services	C	9	M	S		API Results
R036	Data Egress services MUST connect solely to Data Ingress services	C	9	M	S		API Data
R037	Data Ingress services MUST connect solely to Data Egress services	C	9	M	S		API Data
R038	System actors with roles of Information Governance or Data Provider Service Operator only SHALL be able to invoke Data Ingress/Egress services	C	9	M	S		API Data
R039	System actors with other roles SHALL NOT be able to invoke Data Ingress/Egress services	C	9	M	S		API Data
R040	Indexing API services MUST connect solely to Indexing API services	C	9	M	S		API Indexing
R041	System actors with roles of Information Governance or Data Provider Service Operator only SHALL be able to invoke Indexing services	C	9	M	S		API Indexing
R042	System actors with other roles SHALL NOT be able to invoke Indexing services	C	9	M	S		API Indexing
R043	Software Ingress API services MUST connect solely to Software Ingress API services	C	9	M	S		API Software
R044	System actors with roles of Information Governance only SHALL be able to invoke Software Ingress services	C	9	M	S		API Software
R045	System actors with other roles SHALL NOT be able to invoke Software Ingress services	C	9	M	S		API Software
R046	The Federation MUST support a "federated query" analysis pattern	U	9	M	S		TRE; DP
R047	The Federation MUST support a "linked-data assembly" analysis pattern	U	9	M	S		TRE; DP

D Comparison of "Sensitive Data" Definitions

D.1 UK Government classifications

UK Government standard security classification scheme.

Source:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715778/May-2018_Government-Security-Classifications-2.pdf

Classification	Description
OFFICIAL	ALL routine public sector business, operations and services should be treated as OFFICIAL - many departments and agencies will operate exclusively at this level. This includes a wide range of information, of differing value and sensitivity, which needs to be defended against [a particular] threat profile … and to comply with legal, regulatory and international obligations.
OFFICIAL-SENSITIVE	A limited subset of OFFICIAL information could have more damaging consequences (for individuals, an organisation or government generally) if it were lost, stolen or published in the media.
SECRET	Very sensitive HMG (or partner's) information that requires protection against [a] highly capable threat profile … AND where the effect of accidental or deliberate compromise would be likely to result in [serious damage or loss of life].
TOP SECRET	Exceptionally sensitive HMG (or partner's) information assets that directly support (or threaten) the national security of the UK or allies AND require extremely high assurance of protection from all threats … This includes where the effect of accidental or deliberate compromise would be likely to result in [extremely grave damage or widespread loss of life].

Organisations may apply a DESCRIPTOR to identify certain categories of sensitive information and indicate the need for common sense precautions to limit access. Where descriptors are permitted they must be supported by local policies and business processes. Descriptors should be used in conjunction with a security classification and applied in the format: 'OFFICIAL-SENSITIVE [DESCRIPTOR]'

D.2 Commercial Data Classifications from Highest to Lowest

$(ISC)^2$ standard commercial data classification scheme.

Source: https://www.isc2.org/

Classification	Description
Sensitive	Data that is to have the most limited access and requires a high degree of integrity. This is typically data that will do the most damage to the organization should it be disclosed.
Confidential	Data that might be less restrictive within the company but might cause damage if disclosed.
Private	Private data is usually compartmental data that might not do the company damage but must be keep private for other reasons. Human resources data is one example of data that can be classified as private.
Proprietary	Proprietary data is data that is disclosed outside the company on a limited basis or contains information that could reduce the company's competitive advantage, such as the technical specifications of a new product.

Public	Public data is the least sensitive data used by the company and would cause the least harm if disclosed. This could be anything from data used for marketing to the number of employees in the company.

D.3 NHS Digital Data Mappings

Example-driven mapping for a wide variety of data types developed by NHS Digital for use within the NHS. The embedded hyperlinks are the originals.

Source: 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/dimensions-that-affect-risk

Type	Sub-type	Description	Example
Publicly available information		Statistical material that is intended for public distribution. Identification from these materials, with or without any other materials, is not feasible.	The number of diabetics in Sheffield, or location information for health-care providers.
Synthetic (test) data		Synthetic (test) data is fictional data, engineered to be representative of real data, that is created in order to avoid the need to use real data when developing and testing IT systems. Synthetic data must pose zero risk of contributing to the revealing of any personal data.	Fabricated dummy Hospital Episode Statistics (HES) data set, used for testing purposes, risk assessed to ensure that there is no risk of the data contributing to the access to any personal data.
Aggregate data		Summarised and anonymised data, but which is not suitable for public distribution, for example due to the risk that it may be used with other material to contribute to the re-identification of individuals. The risk of such re-identification is not necessarily significant but does exist (especially in the presence of a sustained and skilled attack).	Summarised records of activity of a particular hospital.
Already encrypted materials		Materials that are already encrypted before they touch the cloud, using strong cryptography as defined by the current version of NIST SP800-57 and where the encryption keys are not stored with the cloud provider.	Scanned hospital patient notes which are encrypted by an application before being uploaded to the cloud for archive purposes.
Personal data (PID)		Information about an identified individual	
	Demographic data	Information about the individual rather than their clinical details.	A person's address details and NHS Number.

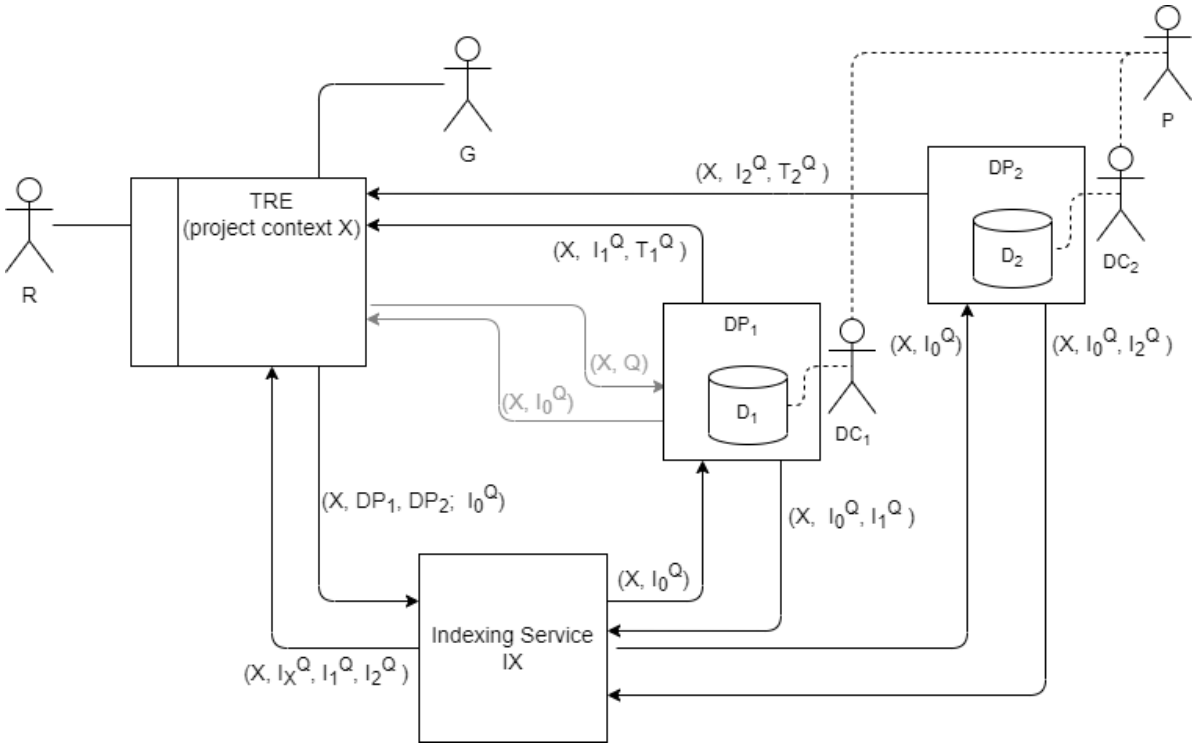
Type	Sub-type	Description	Example
	High risk demographic data	Demographic data where, in the event of a breach, there is a high risk of significant harm.	The address details of a person under the care of the UK Protected Persons Service, likely to be reflected in an S-flag applied to their PDS details.
	Personal confidential data (PCD)	PCD is based on the ICO definition of sensitive personal data, extended within health and social care to include deceased persons and information that is given in confidence and is owed a duty of care, such as: • social care records/child protection / housing assessments • DNA/finger prints • bank/financial/ credit card details • National Insurance number/tax, benefit or pension records • travel details (for example at immigration control, or Oyster records) • passport number/information on immigration status/travel records • work record or place of work/school attendance/records	
	Legally-restricted PCD	Sensitive personal data that are subject to additional regulations or statute, under either the: • Gender Recognition Act 2004 • Human Fertilisation & Embryology Act 2008	Details of a person's previous gender.
	Extra-delicate PCD	Sensitive personal data that are sometimes seen to be additionally delicate, but for which there are no legal restrictions. This determination is often not consistent, but is commonly held, and is often related to conditions that attract, or are considered to attract, stigma. For example, HIV status, mental health conditions, other conditions contained within the SCR 'sensitive code' list. Whilst many patients see information on these kinds of condition to be particularly private and not to be shared under any circumstances, others see them as important to share, and for any stigmas to be removed. Note that there is no legal distinction between PCD and extra-delicate PCD.	Details that a person has asked not to be shared.
Anonymised data		Sensitive personal data that has been subject to de-identification and/or other privacy-	Extract from a research database where all

Type	Sub-type	Description	Example
		enhancing techniques, in line with the ICO Anonymisation Code of Practice. Risk of re-identification is remote (and would be based on activities that are illegal and/or break contractual arrangements). No way of authorised linking with other data-sets.	pseudonyms have been removed.
Pseudonymised data		Sensitive personal data that has been subject to de-identification and/or other privacy-enhancing techniques, in line with the ICO Anonymisation Code of Practice, containing a pseudonym that allows for linking with other data-sets where that is permitted through business justification and legal basis. Otherwise, risk of unauthorised re-identification is remote (and would be based on activities that are illegal and/or break contractual arrangements).	HES data set.
	Reversibly pseudonymised data	Pseudonymised data where the pseudonym is also intended to be used to facilitate re-identification where that is supported by business purpose and legal basis.	Data dissemination to support risk stratification (where individuals may subsequently be usefully re-identified to support their direct care).
	Irreversibly pseudonymised data	Pseudonymised data where re-identification is not intended.	Data dissemination to support a research project that never requires re-identification.
Patient account data		Account credentials (including any recovery materials) for citizen accounts for patient-facing online health tools.	A person's account details for the NHS.UK website.
Patient choices		Statements/preferences made by patients regarding the use of their data.	A person's expressions of their wishes recorded in their GP's clinical system or on the Spine.
Patient meta-data (identifiable)		Information about how identified patients have used patient-facing online health tools.	History of an identified person's use of the NHS.UK website's symptom information.
Patient meta-data (linkable)		Information about how patients have used patient-facing online health tools (not identified, but linkable across sessions).	History of an unknown (but linkable) person's use of the NHS.UK website's symptom information.
Professional user account data		Account credentials (including any recovery materials) for professional user (such as a clinician) accounts that control access to any personal data (including PCD).	A clinical application logon.

Type	Sub-type	Description	Example
Professional account data (less-sensitive)		Account credentials (including any recovery materials) for professional user (such as a clinician) accounts that control access to anonymised information.	Authentication details to portal providing access to anonymised data.
Audit data		Data that records the use of a system and the provenance of the data that system manages	Clinical system audit trail
	Professional user meta-data	Information about how users have used clinical or administrative tools that process personal data.	History of a GP's use of their clinical system, or of summary care record (SCR)
	Audit data (personal)	Data describing the use of a clinical or administrative system that processes personal data, where that audit data itself includes or references PCD.	The audit trail of a GP system showing all users' interactions and use of the system.
	Audit data (non-personal)	Data describing the use of a clinical or administrative system, where that audit data itself does not include or reference PCD.	History of logins to a clinical system.
Key materials		Material that provides long-lived linkage between reversibly-pseudonymised data and personal data, or provides a similarly significant security function.	Look-up tables or decryption keys.
	Very short-lived	One-time decryption keys	A decryption key generated to support (and only usable within) a specific re-identification activity within an individual user session.
	Rotatable	Material that provides linkage between reversibly-pseudonymised data and personal data, that persists over time and over user sessions but is generally rotatable.	An encryption key used by a DSCRO to re-identify pseudonyms included in many data disseminations.
	Long-lived, persistent	Material that provides long-lived and persistent linkage between reversibly-pseudonymised data and personal data, or provides a significant security function.	A root certificate private key for a widespread PKI.

E Sketch Design for Data Linkage through Indexing Services

In this scenario information governance is able to assemble a linked dataset for a given project using an indexing service to construct a project-specific linkage spine and a consequentially minimalist linked dataset. The researcher can be granted access to the resulting linked dataset.



Researcher R has been approved to conduct project X, applying queries Q across Datasets D_1, D_2.

D_1 and D_2 are horizontally partitioned across different sets of individual attributes, governed respectively by Data Controllers DC_1 and DC_2.

D_1 is made available for research through a Data Provider service DP_1; D_2 likewise through DP_2.

A linkage spine needs to be constructed using identifiers for individuals.

We denote as I_0 a set of "bare" identifiers (NHS numbers, NI numbers, CHI numbers etc.) for all individuals in the scope of DC_1 and DC_2 (ie, all individuals in a given region).

We denote as I_0^Q a set of "bare" identifiers for all individuals in the project cohort, defined by query Q run against one of the datasets D_n or otherwise created elsewhere.

E.1 Workflow

1. R passes query(ies) Q to Information Governance G.
    a. (optionally) From within the TRE G passes the query and project context (X, Q) to one of the specified Data Providers DP_1 with a request to return a set of "bare" identifiers I_0^Q defining the project cohort.
    b. (otherwise) the set of identifiers for the project cohort is defined elsewhere and transmitted to G within the TRE environment.

2. G passes the project context, a list of required data services and the set of cohort identifiers (X, DP_1, DP_2; I_0^Q) to an Indexing Service IX.
3. For each DS_n IX sends the project context and set of cohort identifiers (X, I_0^Q) to the data service with a request to return a mapping table of "bare" identifiers to the data service's own identifiers (I_n^Q). (Note that the scope is the same as that of the project cohort defined by Q.)
4. Each DP_n returns a mapping in the project context (X, I_0^Q, I_n^Q) to IX.
5. IX creates its own project-specific mapping for I_0^Q, I_X^Q.
6. IX sends the combined mapping (the "master index file") (X, I_X^Q, I_1^Q, I_2^Q) to G in the TRE.
7. Each DP_n extracts the cohort-specific subset of its dataset (using the set I_0^Q) to create S_n^Q, and de-identifies it to create the TRE-friendly subset T_n^Q.
8. Each DP_n (independently) passes the project context, data-set specific identifiers and de-identified subset data (X, I_n^Q, T_n^Q) to G within the TRE.
9. Using the master index file G assembles the project dataset $T^Q = \{I_X^Q, T_1^Q \times T_2^Q\}$ and makes it available to researcher R.

In principle, for well-formed cohort identifiers, the tasks carried out by IX require no manual intervention. IX could thus operate as an automated service.