DARE

DARE Teleport

Connecting big data to users at light speed

UK Research and Innovation

HDRUK
Health Data Research UK

ADR UK
Data-driven change

DARE UK

DARE Teleport Final Report

Federated data to support team science

Authors: Chris Orton, Simon Thompson, Alexandra Lee, Joss Whittle, Louise Clark, James Healy, Michael Jackson, Kostas Kavoussanakis, Carole Morris, Mark Parsons, Bianca Prodan, Donald Scobbie, Dionysis Vragkos

10.5281/zenodo.10055358

1. Executive Lay Summary

Health and administrative data about people resident in the UK are generated by the NHS, government and other organisations and subsequently held in different locations across the UK's four nations according to devolved legislation and governance. Trusted Research Environments (TREs) were developed by organisations in academia, NHS, government agencies, and in some cases commercial companies to safely store this data and control access to it for research purposes. De-identified data is typically available for research via these TREs, however, each houses specific data based at different institutions, which adds difficulty to a researcher's task of trying to gather a full picture of scientific outputs for the entire nation due to the replication needed to analyse all available data across the country.

A solution to improving data access and efficiency for researchers is **federated data access**, enabling parallel access to data housed in multiple physically separated environments (without data moving from their host environments, instead being accessible from its host location to a researcher in a single safe, secure environment) where researchers can see the data required for their research projects.

Teleport is starting the transformation of traditional TRE access to data, partnering with the custodians of national-level data in Wales and Scotland (SAIL Databank and Scottish National Safe Haven) and the technology providers of the TRE platforms (the Secure eResearch Platform (SeRP) at Swansea University and EPCC at the University of Edinburgh). By making data accessible by connecting TREs, researchers could, for example, have better-facilitated access to understand rare diseases at scale and generally increase the quality of research in common conditions by increasing the sample size accessible in a single environment, rather than having to run multiple disparate analyses. It is the starting place for efficient study across the UK, promoting the move away from duplicated research in siloed environments, instead enabling better scientific outputs and health outcomes due to the increase in scale, granularity, and connectivity of data available. Both sites in scope for Teleport have national-level data holdings of complementary scale and research infrastructure of equivalent maturity to deploy, test, and develop the proposed access solution.

UK Research and Innovation HDRUK Health Data Research UK ADR UK

2. Introduction to Teleport

As more secure and trusted environments (whether known as TREs, Safe Havens or Secure Data Environments (SDEs)) come online to serve the need for remote access to data for research, there will be a greater expectation on the technical providers of these environments to provide solutions to interconnect these environments across the UK enabling researchers to perform UK-wide analyses, without the need to replicate work in more than one location due to physical separation of data holdings. The COVID-19 pandemic demonstrated many examples of analyses conducted across data from the four nations of the UK, some for the first time. However, these analyses frequently entailed running an analysis distinctly four or more times in four or more distinct locations (one in each national TRE and other data provider environments), and then pooling these nation-specific results in one location to further analyse these and draw UK-wide cross nation conclusions. Whilst this option does provide the ability to curate data in each location, it could be more efficient as the same analysis has to be run multiple times before combining the results and then analysing these combined results.

Currently, the main two options being developed to facilitate more efficient use of data from multiple trusted research environments are as follows:
1) To enable access to data housed in multiple environments from a single access point without compromising existing security and governance measures and processes implemented by each environment by creating an extension environment where researchers can see and analyse all the data required for their research.
2) in enabling packaged analyses, using the same analysis methodology, to be sent to a network of environments in parallel through containerised workloads to draw back results from each environment, with no direct access to the underlying individual-level data other than in viewing its metadata and characteristics through a central workload delivery function.
Both options are examples of ‘federation’ – which typically would be analyses performed on decentralised data hosted in multiple locations.

Swansea University, The University of Edinburgh, and Public Health Scotland (PHS) were funded through the DARE Teleport project to pilot a solution aligned with option 1.

One of the issues with moving from traditional siloed TREs to federation across TREs, is the perceived compromise that TREs and data providers would have to make in terms of governance, means of access for researchers, and operations to deliver equivalent security, analytical experiences, and equivalent results across TREs.
Teleport does not introduce any substantial changes to the way TREs operate (in terms of access, governance, and data use regulations) and, therefore, reduces the impact on existing TRE architectures by building an enclosed extension environment which satisfies both of the partnering TRE requirements.

Currently, in both Wales and Scotland, the path for a researcher to access data is to apply to an independent governance panel that is responsible for the review of project applications on a public benefit, scientific rationale, and impact basis. A panel approves for the researcher to access the data for their scientific project based upon the nature of the project being undertaken, specifically whether it is for the public good and the appropriateness of the data for this research. If approved, researchers must prove they are competent in data science and are also ‘safe’ researchers by providing evidence of certification in both data protection and best practice in data analysis.

Once approved, access to data is provided through a secure connection to a remote environment, where individual-level de-identified health or administrative data can be used to derive results for publication. Results are subject to standard data release checks to ensure outputs are safe, i.e. appropriately aggregated, rules around the categorisation of certain variables and small-number suppression are followed, and there is very little risk of re-identification of individuals.

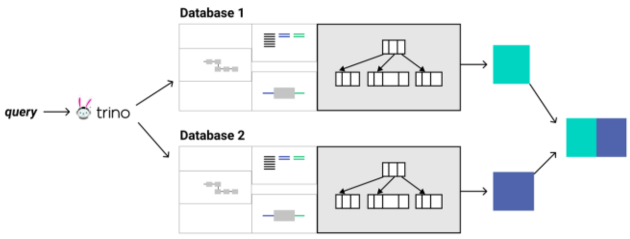
Whilst, as above, federation can lead to major changes being necessary in terms of how a researcher approaches their analytical tasks, Teleport is attempting to maintain the status quo for the majority of researchers in that individual-level data that is usually provisioned within TREs is still the focused product being accessed, permitting granular curation, modelling, and analysis to be undertaken much in the same way that current TREs enable through pre-configured virtual desktop infrastructures. In Teleport, whilst the method of logging into the proposed 'Pop-up TRE' will differ slightly from a researcher's usual experience of either SAIL or the Scottish National Safe Haven, the tooling within the TRE will still provide means to deploy queries to databases or files respectively, and then use curated, safe data within statistical tooling in languages and software that most statisticians and data scientists would be conversant with (e.g. R, Python, delivery through JupyterHub etc.).

The technical solutions being implemented, therefore, place most of the change and deployment prerequisites on the technical providers of the platforms being connected, so in the case of Teleport, the teams who technically develop and manage the SeRP and EPCC infrastructures in Swansea and Edinburgh, respectively.

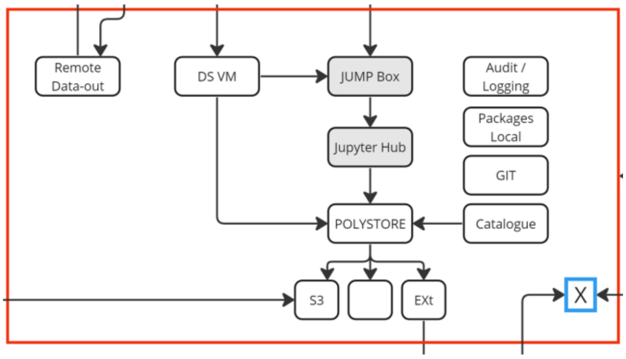
3. Technical Background and Design

Teleport pop-up TREs have several major components that work together to form the deployment. Despite the host environment being configured differently, the pop-up element will remain consistent across deployments and provide a basis upon which to ensure that the operation, security, and capability of the environment is fully defined. The user experience will initially vary as the method of gaining access to the host TRE will follow the usual method implemented by that TRE; once inside the TRE the user will access the pop-up TRE through a web interface, which will offer a research environment that is jointly controlled by all participating TREs, therefore that which is offered inside this pop-up TRE is naturally limited to the components developed by TELEPORT. It is believed that JupyterHub (a multi-tenant version of Jupyter Notebook) can offer a variety of developer IDEs (integrated development environment), which can be used to develop code that can be stored in Git. This code can interact with Trino as the SQL engine, which in turn can access both local and remote data stores.

The use of Trino is a key design aspect. Trino is a SQL execution engine that can connect to multiple data stores. It supports SQL pass-through, meaning that when joining data from multiple datastores, the engine will attempt to do as much filtering and processing of data in that datastore before being pooled for combining with the results from other datastores accessible from the pop-up. This, therefore, allows the minimum amount of data to be returned to satisfy the analysis whilst still enabling these results to be joined for further processing and refinement. The ability to join across data stores is a crucial capability.



The components:



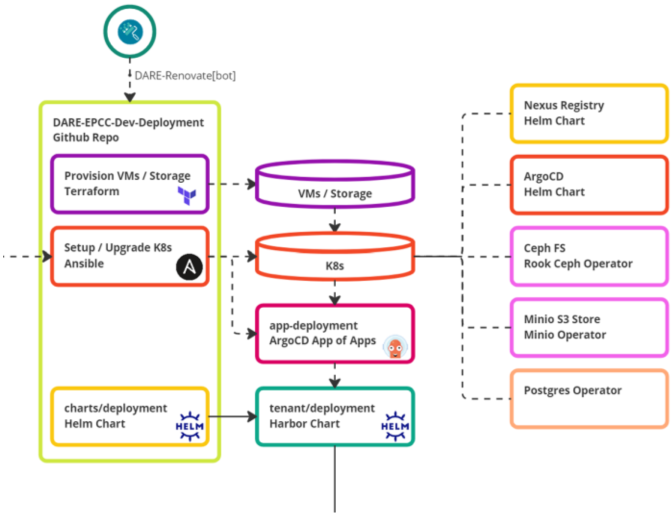
JUMP Box VMs are accessed with RDP (remote desktop protocol) from a Guacamole server, enabling pixel-only access through the air gap to Jupyterhub. JupyterHub will provide Python IDE, R Studio IDE, VS Code IDE, and access to data sources like Trino and GPUs (graphical processing units).

The POLYSTORE (implemented through Trino) can access back-end S3 buckets, PostgreSQL and remote data stores at other sites. Due to the polystore nature and interconnectors available in Trino there are a large number of data sources that could be integrated depending on what is available locally (DB2, MS SQL, etc ..).

An Open Policy Agent(OPA) based command and control system manages user and data access and can quickly and automatically revoke access if required. Audit logging across the TRE uses Grafana, Prometheus, and Loki. Gitlab and a Nexus Package Registry allow migrating software securely through the air gap and serving them within the TRE.

The entire solution depends on Kubernetes to achieve every aspect being software-defined and ensure that the solution can horizontally scale as required depending on the size and needs of the research project. Having the entire deployment controlled and monitored by a CI/CD (continuous integration/continuous deployment) engine is also a key design choice, allowing the pop-up TRE to be maintained in a consistently known state at all times and any deviation or alteration will be reported and corrected by the system.

Given that not all TREs have Kubernetes, the Teleport solution was developed to support a full end-to-end deployment, only requiring a number of Linux servers / virtualised servers. This is achieved by layer up technology:


DARE-Renovate[bot]
DARE-EPCC-Dev-Deployment Github Repo
Provision VMs / Storage Terraform
Setup / Upgrade K8s Ansible
charts/deployment Helm Chart
VMs / Storage
K8s
app-deployment ArgoCD App of Apps
tenant/deployment Harbor Chart
Nexus Registry Helm Chart
ArgoCD Helm Chart
Ceph FS Rook Ceph Operator
Minio S3 Store Minio Operator
Postgres Operator
HELM

1. Terraform for deployments of standardised infrastructure and storage. Terraform supports a large number of environments, and in this project, we tested with KVM (Kernel-based Virtual Machine), OpenStack, VMware ESXi, and minikube.
2. Ansible is then used against this infrastructure to stand up K3s (a Kubernetes distribution).
3. Ansible is used to deploy Kubernetes operators (helper infrastructure components), CephFS operators (data storage), Minio operators (S3 buckets) and Postgres Operator (PostgreSQL databases)
4. Ansible is used to deploy ArgoCD, a Kubernetes CI/CD engine that will deploy from Github and supports a concept called App of Apps, which enables a chaining of applications to form solutions.
5. Ansible deploys the host App of App project-specific deployment, which will trigger a chain of deployments and configurations.

4. Development

Each major component of Teleport is in a GitHub repository with its source code, container image definitions, and helm charts. A CI/CD system builds and tests components when they change and versions them accordingly. Containers and charts are signed, pushed to a secure Harbor registry, and scanned for vulnerabilities. TREs can then migrate them through the air gap into the Nexus registries if required.

The design is to build the components separately in different repositories. Deployments are themselves a GitHub repository that describes how these components should be deployed to create the required solution and architecture separation required by the host organisation.

Repository	Purpose	URL
DARE-Teleport	Main repository	https://github.com/SwanseaUniversityMedical/DARE-Teleport
DARE-Guacamole	Remote desktop server	https://github.com/SwanseaUniversityMedical/DARE-Guacamole
DARE-OPA	Open Policy Agent, policies	https://github.com/SwanseaUniversityMedical/DARE-OPA
DARE-Airflow	Airflow job scheduler	https://github.com/SwanseaUniversityMedical/DARE-Airflow
DARE-Trino	Trino warehouse	https://github.com/SwanseaUniversityMedical/DARE-Trino
DARE-Hive	Apache Hive	https://github.com/SwanseaUniversityMedical/DARE-Hive
DARE-Jupyter	Jupyter Hub Server	https://github.com/SwanseaUniversityMedical/DARE-Jupyter
DARE-Minikube-Deployment	Laptop deployment	https://github.com/SwanseaUniversityMedical/DARE-Minikube-Deployment
DARE-SeRP-Dev-Deployment	Main dev team deployment	https://github.com/SwanseaUniversityMedical/DARE-SeRP-Dev-Deployment
DARE-EPCC-Dev-Deployment	EPCC dev team deployment	

5. Implementation, Deployment, and Operation

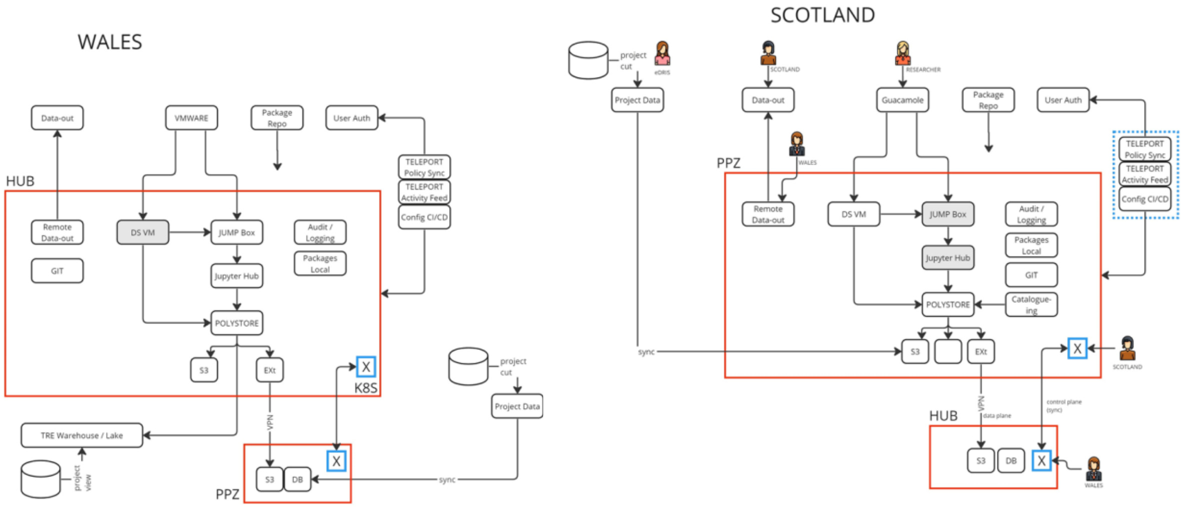
Deployment of a TRE in the context of Teleport follows a DevOps strategy, using a Git repository to manage a Helm Chart customising Teleport for the given environment. ArgoCD watches the deployment chart and provisions the Teleport storage and services into the cluster.

Each site manages its deployment in a multisite TRE between SeRP and EPCC, configuring only the components needed for the given environment. Generally, one site adopts the host role and deploys a full user-facing TRE, while the remote site deploys only data sources and command and control components.

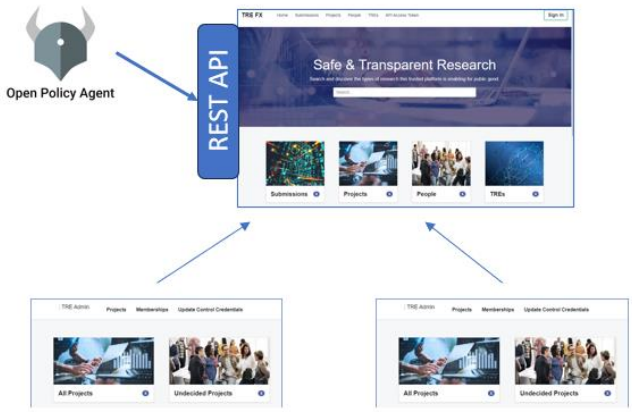
Multiple TREs deployed across SeRP clusters have demonstrated that the Pop-Up TRE concept allows processing data held in each TRE without wholly copying it from one TRE to another. EPCC has been testing deploying Teleport on VMs in the Safe Haven environments and studying site-to-site VPN usage.

Despite the TRE implementations being functionally similar, the deployments and chosen technologies differ. The main elements allowing the user authentication and initial remote access remain as is currently approved by each

TRE, ensuring that the accredited architecture that both TREs have is unaffected by the federated project. The pop-up TRE is inserted into the infrastructure (RED box below). The sites are interconnected by site-to-site VPN. Command and control are implemented next to the research infrastructure (BLUE Box) and utilise the same connectivity. Nothing is hosted in an internet-facing manner.



In order that the local infrastructure can be integrated, the deployment configuration is locally defined by the host TRE so that local services such as authentication can be specified upon deployment.

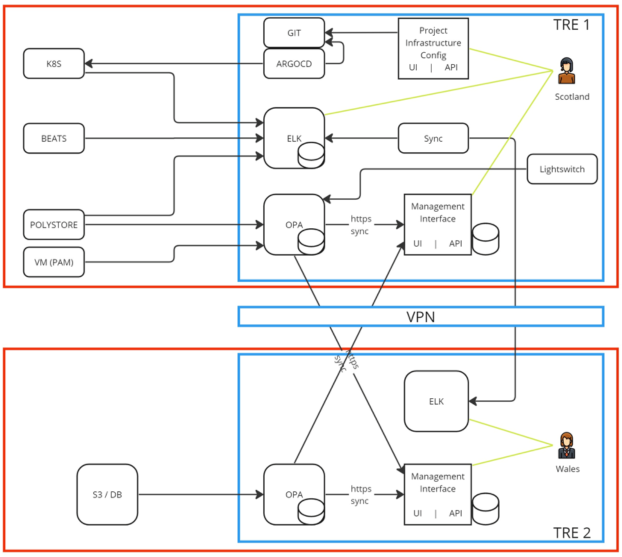


The development of the command and control (BLUE Box) is based on OPA and syncing OPA policy bundles across all sites. During the project, it was realised that a number of management interfaces and capabilities would need development. Swansea University was also co-developing the DARE TRE-FX project, which had a similar need for defining a federated project and the ability to assign researchers, which each TRE would then authorise in the consortium, it was therefore decided that these elements would be used across both projects and that the TRE-FX "submission layer" would include the ability to produce OPA Policy Bundles for integration with OPA and subsequently JupyterHub and Trino. Given the reliability of the other project's development timescale, this has been developed but is still awaiting end-to-end integration. However, the proof of concept (as demonstrated at the DARE final showcase event) has shown the OPA policy bundle synchronisation to be working as required.

All logs and activity are recorded to the SEIM, and implemented using Elasticsearch/Kibana/Logstash (ELK). All logs are synchronised to all endpoints so that all parties can audit or inspect the activity logs to ensure that the activities within the host are in line with what has been agreed upon or what is expected.

6. Teleport within the DARE UK Federated Architecture Blueprint

The concept of Teleport is that access to multiple TREs should be possible through a single entry point, subject to all connected TREs governance being followed, and their data being provisioned to be visible to the connectivity of the ‘Pop-up TRE’ engines.



Whilst the resultant networking and software deployment to achieve this may be deemed ‘data pooling’, because multiple datasets from multiple providers are accessible and usable from a single location, the data itself is nonetheless originally still resident in its TRE database, file store, or whichever chosen method of data provision is stood up by the providing TRE. This in itself creates a federated network of datasets, available for analysis through a centralised connected interface, which then permits curation, analysis, and workloads to be undertaken in a methodology much more akin to current research practices than perhaps compared to other federated analytics solutions (containerised workloads etc.). In order to provide researchers with the tooling and access they need to achieve study across multiple TREs, federation has been a necessity to establish the initial connections. Regardless of which TRE hosts the ‘pop-up’ infrastructure as the lead, access to data held in multiple disparate locations, but visible at the individual-level layer rather than just the metadata layer, provides the analytical ecosystem to run end-to-end research pipelines from curation through to research output. This is particularly important in managing a solution which permits concordance across the infrastructure, data, and governance layers as outlined by the federated architecture blueprint.

There is flexibility to the connectivity and software definition of Teleport, which also means it should be interoperable with designs from other DARE projects and their solutions for data federation and enhanced analytical capability. For instance, in providing networked connectivity between national TRE/SDEs where underlying individual-level data is available, this could promote the idea that ‘Teleport access’ could be the precursor to ‘TRE-FX’ analysis - whereby initial access fosters the curation, common data modelling, and preparation needed to then pivot to a workload deployed through the RO-Crate approach, having the analytical workflow execution as an option within the Teleport interfacing for query production and deployment.
Such a combinatory exploration could also have implications to improve the federated network where certain data providers/TREs prefer the more traditional federated analytics model (no direct visible access to underlying individual-level data but queries are developed primarily on metadata and deployed through to the payload beyond), in combination with those more used to direct data provisioning via databases or file stores within virtual desktop infrastructure constructs.

At a technical policy level, there has also been overlap, synergy, and collaboration between Teleport and TRE-FX. Both programs have similar requirements of the federated project being defined and then each participating TRE opting into the project and independently approving users. Due to the development of this aspect by the same team, it makes sense to extend this aspect to cover both programs. For Teleport, the production of Open Policy Agent(OPA) "Bundles" enables the shared elements to deliver the Teleport-specific requirements. The design of the system is to integrate OPA as the authorisation layer for the popup TRE. The ability to control the project and project membership locally, with this intent being synced to the research platform, enables the participating TRE to remain in control. This concept is fundamental for the multi-TRE model of Teleport in terms of each TRE having permission sets to grant access to relevant data through the pop-up TRE.

Similarly, as with any project that is initiated through the 'five-safes' framework, which is embedded at the heart of accessing data through TREs, SACRO technology could be embedded within the 'Pop-up TRE' construct of Teleport to provide a standardised method of output checking for any connected TRE through the Teleport system - given that research outputs have to be reviewed by any providing entity to the pop-up prior to its release out of this environment. Given that many TREs administer similar levels of scrutiny on research outputs in terms of their disclosure risk, content, and security, the initial semi-automated element of SACRO in adjudicating output suitability for egress could provide enhanced efficiency of review across such a platform, rather than siloed and bespoke review guidelines standing alone in making such decisions. Indeed, such a framework could also be used to assess the competency of workflow queries that are planned to be deployed through federated frameworks, such as the TRE-FX model, if there were any concerns that the containerised queries could pose any disclosure or analytical concerns before they are deployed.

7. Impact and Legacy of Teleport

The technical components and capabilities in Teleport will have a legacy impact.

1. A demonstration of a TRE that is 100% software-defined, self-healing, fully monitored, fully audit logging, MVP feature set.
2. Trino extensions to support OPA.
3. Trino extensions to support Trino as a data source.
4. Demonstration of CI/CD orchestrating a deployment of complex architecture with optional components.

The working relationship between EPCC/PHS and SAIL/SeRP. They bring together similar but different implementations, and finding the common ground and the synergies between the organisations has led to an acceptance that the implemented solutions still fulfils all the required governance and operational needs of mature and highly-regarded TREs.

Teleport has been deployed between Dementias Platform UK (DPUK) and SAIL in order to enable access to routine NHS data and the corresponding longitudinal cohort data while operating within the two application and governance procedures. At the time of writing, we are waiting for the researchers to be approved.

Teleport has also been discussed as a way of joining up the UK Dementia Platform (SeRP UK-hosted) and the Australian Dementia Platform (SeRP Monash University-hosted), which is likely to be established in early 2024.

8. Future Direction

Despite the fantastic achievement of establishing a proof of concept solution, it would be timely to move the platform to a production-ready state and consider a wider set of use cases beyond that of the initial project.

There are a few areas that could be expanded:

1. Formalisation and publication of a combined approach to information governance based on the project and its current partners

2. Medium article on the use of ArgoCD App of Apps deployment pattern in a complex architecture such as Teleport, including all the undocumented aspects that need to be addressed to achieve this
3. Integration of the POC OPA command and control elements into the main systems and include in CI/CD

The aspect that presented a minor challenge to the project was the secure networking or communications between the TREs. Given this project was a point-to-point implementation, this could easily be achieved through a site-to-site VPN. However, a more complex deployment could adopt this approach with the hosting TRE establishing site-to-site VPN with all endpoints, enabling these deployments at scale and then having a networking fabric that could support this use case and be dynamically configured on a per-project basis. Early POC projects such as JISC Safe Share (2016) were demonstrations of a management secure network that could enable endpoints to be dynamically interconnected securely. Recent developments such as Nebula (and others) network overlays allow this type of network to be established in a software-defined project-specific manner. The DARE technical blueprint also references a secure set of interconnected APIs in a trusted consortium so establishing a trusted interconnection or network overlay is the next technological improvement that should be established. This could form a sub-project of Teleport.

Conceptual integration between DARE Teleport and DARE TREFX would also be an area worthy of consideration. How do they compare, what scenarios does one provide a more tangible solution for, pre or post-data egress, virtual or horizontal data combination, federated analysis into pooled ephemeral analysis space for a two-phased egress approach, etc.

Development of an understanding of what national processes and approaches might be needed to support federated projects in the future. All major TREs have robust operating procedures for dealing with the scope of their TRE, but if there are to be projects involving many TREs concurrently, then how this is handled by the data-holding TREs is a valid question for development. Also, considering how the research user experience also becomes “federated” and where the notion of an agreed, approved and funded federated project across the UK or wider gets defined should be embodied in future work.

9. Work Package Overview

- **WP1 Project Governance and Engagement**
  The project teams met for full-day working meetings three times in person (Edinburgh March 2023, Swansea May 2023, Edinburgh October 2023) and once virtually (September 2023), as well as regularly communicating through Slack and through versioning and iteration in Github. Recruitment of posts was on time and led to the project beginning quickly and delivering outputs as showcased at the October DARE final showcase event.
- **WP2 Information Governance**
  Approval from governance panels to undertake the technical project was not needed due to the position of the principal and co-investigators in terms of their leadership of teams who administrate and develop the TREs in scope for Teleport. A briefing document on the Teleport project and its governance permutations was circulated to governance boards and lay panels in April and May 2023. Scoping on processes and governance procedures was undertaken and completed in May and June 2023, with synthetic data used in the development of the solution through to October 2023.
- **WP3 Infrastructure Establishment**
  Design and build phases of the infrastructure and deployment were carried out between February and June 2023, with initial implementation of the Teleport solution across both sites in July 2023. Fine-tuning and test runs of this enabled the solution demonstration at the DARE final showcase event in October 2023. The pilot release of Teleport has been established and is published through this report and the Github repositories listed above.

- **WP4 Shared Policy Framework/Platform POC**
  As part of the implementation of the Teleport solution, both SeRP and EPCC have coalesced around the same architecture for the pop-up TRE to instantiate. This is embedded into the Github repositories as linked above and is the framework for an all open-source software-defined TRE setup in any organisation that wishes to network systems accordingly.
- **WP5 Shared User Passport POC**
  This is a fundamental layer of Teleport and now TRE-FX, which is described above as the OPA policy bundle, which was successfully demonstrated (in beta) at the DARE final showcase and will need to move to tangible deployments to be released as a fully production product. However, the Teleport goal was to create a proof of concept product which is successfully operational.
- **WP6 Exemplar Science Project**
  Due to the timescales of the build and deployment of the main Teleport architecture, the system has been developed using representative synthetic data to date, and simulations of accessing data held in both SAIL and the Scottish National Safe Haven run on this (as shown at the DARE final showcase event). The systems and governance frameworks needed to use Teleport for a genuine scientific analysis are all in place. The pivot to real data and scientific use cases needs further integration between DARE projects and researchers with the capacity to undertake a project using Teleport and, indeed, other federated analytics solutions such as TRE-FX. Such plans will be possible through collaboration with the HDR UK Driver Programmes, and Teleport already has a scientific exemplar project about to go-live between SAIL Databank and Dementias Platform UK, as described in the impact and legacy section.

10. Acknowledgements and attributions

This work was funded by UK Research & Innovation [Grant Number MC_PC_23009] as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, delivered in partnership with Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK).

We thank DARE UK for facilitating the funding for the project, supporting the project throughout, and providing opportunities to collaborate with other projects who are developing solutions in this space.

We would like to thank the leadership team of Professor Simon Thompson as principal investigator from Swansea University, and Professor Mark Parsons and Carole Morris as co-investigators from the University of Edinburgh and Public Health Scotland respectively.

The following teams were instrumental in the design, development, and deployment of the Teleport architecture.
Swansea University: Alexandra Lee, Joss Whittle
University of Edinburgh (EPCC): Michael Jackson, Bianca Prodan, Donald Scobbie, Kostas Kavoussanakis
Public Health Scotland: Dionysis Vragkos, Louise Clark

Management of the project was provided by Chris Orton (Swansea University) and Bianca Prodan (University of Edinburgh/EPCC)

Public involvement and engagement were managed by Chris Orton and Claire Newman (Swansea University), with invaluable input from our lay lead James Healy, and the members of the SAIL Consumer Panel, the Scottish Public Benefit and Privacy Panel for Health and Social Care (HSC-PBPP), and Research Data Scotland.