TRE-FX

Giles, Thomas ; Soiland-Reyes, Stian; Couldridge, Jonathan; Wheater, Stuart; Thomson, Blaise; Beggs, Jillian

DOI:
10.5281/zenodo.10055353

Publication date:
2023

DARE UK

TRE-FX

Final Project Report

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

DARE UK

TRE-FX

Final Project Report

1. Lay summary (max 350 words)

Trusted Research Environments (TREs) are secure locations in which data are placed for researchers to analyse. TREs host administrative data, hospital data or any other data that needs to remain securely isolated, **but it is hard for a researcher to perform an analysis across multiple TREs**, requesting and gathering the outputs from each one. This is a common problem in the UK's devolved healthcare system of geographical and governance boundaries.

There are different ways of implementing TREs and the analysis tools that use them. A solution must be straightforward for existing, independent systems to adopt, must cope with the variety of system implementations, and must work within the "Five Safes" framework that enables data services to provide safe research access to data.

TRE-FX assembled leading infrastructure researchers, analysis tool makers, TRE providers and public engagement specialists to streamline the exchange of data requests and results. The "Five Safes RO-Crate" standard packages up (Crates) the Objects needed for Research requests and results with the information needed for the tools and TRE providers to ensure that the crates are reviewed and processed according to Five Safes principles. TRE-FX showed how this works using software components and an end to end demonstrator implemented by a TRE in Wales. Two other TREs, in Scotland and England, are preparing to follow suit. Two analysis tool providers (Bitfount and DataSHIELD) modified their systems to use the RO-Crates. The next step is practical implementation as part of the HDR UK programme. Two large European projects will develop the approach further.

TRE-FX shows that it is possible to streamline how analysis tools access multiple TREs while enabling the TREs to ensure that the access is safe. The approach scales as more TREs are added and can be adopted by established systems. Researchers will then be able perform an analysis across multiple TREs much more easily, widening the scope of their research and making more effective use of the UK's data. If we had had this for COVID-19 data analysis it would have super-charged researchers to be able to quickly answer pressing questions across the UK.

2. Project outputs

2.1 Overview

The TRE-FX project (February-September 2023) aimed to address challenges associated with executing data analysis across multiple Trusted Research Environments (TREs) with differing geographical or governance boundaries. A collaboration was formed with leading technology providers from HDR UK and ELIXIR-UK (the national node of the European Research Infrastructure for Life Science Data), three TRE providers (UKSerp/SAIL - Wales/on premise stack, TREEHOOSE - Scotland/AWS stack, PIONEER - England/Azure stack), and two TRE Client analysis platforms (DataSHIELD - open-source project, BitFount - commercial).

The TRE-FX concept is that federated analysis and learning tools, now and future, need to access multiple TREs to analyse data from each. However, currently each toolkit implements its own unique exchange standards and each TRE develops a bespoke solution for the toolkits to access, all that must be independently tested against the Five Safes framework that governs and safeguards GDPR-compliant data. This federated yet highly autonomous and heterogeneous arrangement is likely to persist, even in the face of proposed blueprints. It is complicated, a blocker to interoperability, and scales poorly as more and more TREs are added to the requests.

Acknowledging the heterogeneity of established analysis platforms and TREs, TRE-FX set out to streamline the flow of metadata needed for federated analysis using existing tooling and open standards to avoid unnecessary reinvention. Scalable interoperability depends upon standardised objects flowing between services, APIs and modularised component services that can be interchanged and mapped to multiple (existing) implementations:

- secure RO-Crate Research Objects support data request/results transit between Clients and TREs, collecting and carrying the metadata and elements required to operate a multi-phase Five Safes review process;
- a technology agnostic service runs pre-approved workflows for answering data queries within the TREs;
- a component-driven architecture with reference implementations demonstrates appropriate modularised services and the viability of the approach.

Our website (https://trefx.uk/) showcases our results and continued developments.

TRE-FX outputs against objectives:

Objective 1: Showcase technical interoperability between cross-sector TREs and existing Federated Analytic solutions. A component driven architecture, using approved workflows for query execution within a TRE, has been developed using existing standards such as GA4GH Task Execution Service and RO-Crates. An open-source end to end reference Minimum Viable Product has been implemented within a test area of SAIL (see video) and has been assessed by the other partner TREs, with further test implementation pending. BitFount, a federated learning client, has modified its services to demonstrate interoperability with our components.

Objective 2: **Use ELIXIR RO-Crate Research Objects as a Five Safes interoperability layer.** A "Five Safes RO-Crate" Digital Object profile captures the required information to request a workflow to be run with associated Five Safes information embedded, and holds the results and provenance of the workflow execution. Following an eight phase Five Safes RO-Crate review process, TREs use the crate elements to validate all client requests before they action them within their TREs and check results before disclosure, publishing and receipt by those clients. At the time of writing we have demonstrated a Five Safes RO-Crate being used in three federated platform implementations (TRE-FX primary implementation, BitFount and DataShield) modifying their services.

Objective 3: Demonstrate dynamic analytics without moving data. An existing ELIXIR Workflow Execution System (WfExS) has been used, extended, and integrated, to execute pre-approved analytical processes within the TRE. ELIXIR's WorkflowHub Registry has been used to register the workflows for transparency.

Objective 4: Evidence public's trust in the technical methodologies. Our PPIE team has been involved in design decisions throughout development. The public were supportive and encouraged by the efforts being made. The comments and questions raised demonstrated an insightful critique of the approach, and the findings will help shape the future direction. The clear message was that whilst the Five Safes framework and the federated approach can give reassurance, the details of implementation will really matter in order to gain public trust.

Earlier investments by DARE UK (TREEHOOSE, FED-NET), UKRI/NIHR (CO-CONNECT) and ELIXIR (RO-Crate, WfExS, WorkflowHub) have been reused and extended where necessary. Adoption of GA4GH standards and ELIXIR technologies aligns TRE-FX with the EOSC-Life European Science Cluster, the European Open Science, the FAIR Digital Object Forum and Horizon Europe projects BY-COVID and EOSC-ENTRUST.

Although TRE-FX has built upon standards and technologies originally developed for health data, the framework and reference implementations are applicable to ALL sensitive data types, i.e. not health data specific. UKSeRP for instance is a TRE platform which hosts both administrative and health data, and was one of the reference implementations of the framework.

As a viable reference architecture for federated analytics TRE-FX provides a valuable step towards fostering the development of a federated network of TREs within HDR-UK, future DARE UK programmes and internationally. The Five Safes RO-Crate is a promising, flexible streamlining mechanism that can be adopted by established platforms and TREs. Immediate next steps are to complete the deployment of TRE-FX in the remaining TRE partners and outstanding PIE tasks, which were delayed due to time and resource stresses when running a sprint across multiple partners participating in multiple sprints. TREEHOOSE and PIONEER have purchased the reserved infrastructure instances required to support TRE-FX deployment and the project partners are committed to deploying and testing this technology in their environments.

2.3 Technical Achievements

The TRE-FX project has made significant technical achievements. It has established:

- The Five Safes RO-Crate, a new Digital Object format for standardising and streamlining the flow of requests, results and metadata between federated analysis platforms and TREs for workflow-based computational analysis, with three reference implementations;
- An eight step crate review process for analysis clients and TREs;
- A component-driven architecture, using approved workflows for query execution within a TRE, with an open source reference implementation in one TRE.

Three reference implementations demonstrate the utility and adaptability of the Five Safes RO-Crate:

- Our primary reference implementation of the component architecture prototyped within SeRP;
- Bitfount existing technology modified to implement the Five Safes RO-Crate and to demonstrate interoperability with our components;
- DataSHIELD existing technology modified to implement the Five Safes RO-Crate.

2.3.1 Standard Digital Objects for Federated Analytics

RO-Crate is a community-based specification for packaging and describing research outputs, based on FAIR linked data standards. The approach has been adopted by a variety of research domains [Soiland-Reyes 2022] with specialisation in different *profiles* to combine generic and domain-specific metadata. RO-Crate uses open Web standards including schema.org. Recently, the Workflow Run RO-Crate profile has been developed and implemented by more than 6 workflow engines including CWL and Galaxy [De Geest 2022; Leo 2023]. The Five Safes RO-Crate, developed through collaboration with a wide range of stakeholders including national and international TRE providers, builds upon the Workflow Run RO-Crate profile.

The Five Safes RO-Crate, defines a standard Digital Object that represents a unit of computational workflow-based access to sensitive information managed in accordance with a set of principles conforming to the Five Safe framework. A compliant crate encapsulates information about data, workflows, and provenance in a single package that is both machine and (via translation) human readable. Metadata provides the necessary context for

evaluating the safety and appropriateness of both data access and analysis. A standard wrapper for the analytics request would enable interoperability across federated platforms whilst allowing TREs full discretion over whether they accept to process the Crate.

The aim is to enable trusted workflow execution in a TRE, from an authenticated workflow run request, through approval and review processes to a completed workflow execution. A protocol of multiple phases internal to the TRE, includes: check, validation, workflow retrieval, sign-off, workflow execution, disclosure control, publishing and receiving. At the later stages the crate also conforms to the Workflow Run RO-Crate Profile for return to the user, and a derived public version (possibly redacted) can be published in the Data Use Register to document the analysis. The pre-approved workflows may be registered in WorkflowHub (TRE-FX Team).

A compliant crate is not itself inherently safe: its role is to streamline the flow of information by standardising the metadata it collects and carries to support the Five Safes processes of the TREs and their issuing/receiving clients. The structured format of the Five Safes RO-Crate allows for a standardised audit trail, facilitating regulatory compliance checks and adherence to industry standards like BagIt and the RO-Crate specifications further enhances interoperability.

A crate operates in a pre-determined and controlled setting: the workflow to be executed within the TRE to answer the request must be pre-approved and executed in a secure deployment, and the TRE services to manage the crate review phases must adhere to the Five Safes. The initial crate with a workflow run request references a pre-approved workflow and project details for manual and automated assessment according to the TRE's policy. Further details can be found in Technical Documentation - Five Safes RO-Crate.

2.3.2 Defining a component driven architecture

The components, their APIs and interactions set out a framework for conducting analytics across multiple, isolated data environments while maintaining strict compliance with security and governance protocols. Our architecture separates out the key stages of processing, the security around those stages and the use of the Five Safes RO-Crate as the payload to facilitate communication between the layers. Following best practice, each component is designed to be modular, (easing integration, scalability, and maintenance) and the architecture is layered to separate concerns.

- **Submission Layer:** Existing software solutions, or researchers, securely submit their GA4GH TES request, which contains a link to the Five Safes RO-Crate. It must be accessible to both those submitting requests and to the TREs. This could either be within a secured network or a more public area, depending on infrastructure requirements. A secure API receives an GA4GH TES message, provides a secure API endpoint for the TRE Controller to access and update requests that have been queued for each TRE. Other functions include status updates (including error reporting) for each request and user initiated request cancellations. It is likely that one submission layer is utilised by a number of TREs.
- **TRE Controller Layer:** An interface between the Submission layer and the Workflow Executor, located inside the TRE but an area separated from any project data. This component acts as an air-gap between the outside world where the jobs are submitted and an internal queue where approved requests are staged for processing. The TRE Controller is responsible for ensuring that the request made is in line with local policy, such as, to confirm if this is for an approved project and has the request come from a user named on the project. The main purpose is to pull down relevant TES requests from the Submission Layer, unpack the RO-Crate and store it within an internal queue. It must provide APIs to the Workflow Executor such that the executor can access and update requests.
- **Workflow Executor Layer:** Situated within the TRE and in an area that can access the data, the primary role is to process the crate, package up the results and submit them back to the TRE Controller Layer. It updates the crate to allow disclosure checks by staff or by semi-automated methods. After executing the jobs and

successfully passing the disclosure controls, the output is relayed to the TRE Controller Layer. This, in turn, facilitates the egress of the results back to the Submission Layer.

- **Transparency Layer:** An additional layer functions as a data use register, enhancing research oversight and auditability by facilitating the publishing of simplified Five Safes RO-Crate in the HDR UK Data Use Register, and the pre-approved workflows in WorkflowHub. As an entity external to the TREs, it offers a public interface where executed analytical tasks (but not the data or results) and the associated metadata can be viewed. The result is that a full record of the requests made and the details of the workflows can be accessed.

Primary TRE-FX Implementation: A fully open-source end to end reference implementation of all components, realised as a Minimum Viable Product within a test area of UKSeRP TRE. Users interact with the submission layer via a portal and/or a GA4GH TES API, requiring registration for authentication. We developed a TRE-Controller (can be run on a VM) that incorporates various dockerized software modules including a TRE Admin UI, offering an integrated view of ongoing tasks, logs, and system metrics and a basic disclosure control egress service acting as a safety measure to prevent unsafe data disclosure. In this implementation HUTCH serves as the workflow executor, retrieving approved workflows and containers from a local HTTP source (Sonatype Nexus). The actual workflow executions are orchestrated by the ELIXIR WfExS Workflow Execution Service which recruits containers to run analyses. HUTCH communicates with the TRE Controller using REST APIs, it posts notifications, such as status updates. Post-analysis, results are held by Hutch until disclosure control approval is received. The data is then moved to a location where it can be picked up by the TRE Controller for data egress and returned to the submission layer for collection by the query submitter. An HDR gateway API has also been developed to act as a transparency layer in this implementation. This is capable of handling the output Five Safes RO-Crates and posting these to the HDR Data Use Register. Further details in Technical Documentation - Primary Implementation.

2.3.3 Existing implementations converted to support Five Safes RO-Crate

"One size does not fit all" when ensuring sustainability, scalability and compliance with both existing and new TRE technologies. To demonstrate the power of the Five Safes RO-Crate to support interoperability with established federated analysis platforms, two further implementations were developed:

- Bitfount, adapted to implement the Five Safes RO-Crate and to use our Workflow Executor Layer
- DataSHIELD, adapted to implement the Five Safes RO-Crate.

Bitfount implementation: The commercial UI allows users to run federated learning which relies on rapid responses which would have been hampered by the manual disclosure control steps presented in the primary implementation above. As part of the TRE-FX project, Bitfount made plugins to their submission layer to dispatch Five Safes RO-Crates, and the Bitfount Pod (an open-source component of their stack) to serve as the TRE-Controller and made capable of interacting with HUTCH. Despite being an alternative implementation, Bitfount maintains potential interoperability with the primary submission layer. Further details in Technical Documentation - Bitfount Implementation.

DataSHIELD implementation: DataSHIELD explored and integrated new analysis techniques and interaction protocols without using the common components of the other two implementations, and instead adapted its own components. This demonstrates that RO-Crates can be used with different implementation choices and that the framework implementation can be modified and reconfigured. A lightweight testing framework replaced the submission layer and TRE agent. The software infrastructure used Quarkus and Kubernetes. Further details in Technical Documentation - DataSHIELD Implementation.

2.3.4 Integration with existing Workflow Execution Systems

The Workflow Execution Service WfExS has been developed by ELIXIR and Barcelona Supercomputing Centre to execute publicly available workflows in multiple languages including Common Workflow Language (CWL). In

TRE-FX it was adapted to operate in closed environments. Its use aligns TRE-FX with European programmes, promoting the adoption of standards, reusing existing technologies, and contributing to the growth of a collaborative ecosystem within the European research and data handling landscape. As an additional transparency layer, making analytic methods Findable, Accessible, Interoperable, and Reusable, TRE-FX uses ELIXIR's WorkflowHub.

DataSHIELD created a Docker container, triggered by WfExS to engage a separate analysis VM, while Bitfount containerized their analysis tool for direct execution via WfExS. The integration of both technology providers with WfExS demonstrates TRE-FX's flexibility, allowing diverse software vendors to seamlessly adapt their technologies to the architecture. Interoperability by way of CWL and RO-Crate make this technology platform attractive to a broader range of software vendors, ultimately fostering a more versatile and inclusive ecosystem.

2.3.5 Enriched Provenance Model for Five Safes validation

By offering transparency, the TRE-FX system enhances accountability and fosters confidence in the research ecosystem. The structured format of the RO-Crates allows for a standardised audit trail and can thus streamline TRE operations, ensuring compliance with TRE SOPs and data security policies. The Five Safes RO-Crate profile is built on the Workflow Run RO-Crate profile, and specifies the expected processing model for execution. At the final phase it holds the workflow execution provenance, tracking the provenance of the TRE validation, review and disclosure control, e.g. relating to policies and semi-automated processing.

The incrementally gathered metadata will be crucial parts for publication of the (possibly retracted) Five Safe RO-Crates into Data Usage registries (e.g. the HDR Innovation Gateway), which is useful for tracking and oversight and as an example of TRE usage, as the published Crate references the analytical workflow and sufficient parameters for recomputability. We are aligning the Five Safes RO-Crate requirements with the concurrently evolving HDR Data Usage Registry standard and have identified that the TRE side can inject additional metadata to complete the Five Safes picture. With the Horizon Europe project BY-COVID we are also exploring Common Provenance Model, which uses RO-Crate and PROV (the W3C standard for provenance) for a federated approach where parts of provenance can be sensitive/restricted.

2.4 PIE workstream

The PIE workstream of TRE-FX was fully integrated into the project, with a lay co-applicant chairing the weekly team meetings. There were several key outputs from the PIE workstream:

- Two co-developed brochures, structured to present the project's details in a manner easily comprehensible to the lay audience.
- Two explainer videos (final stages of production), co-developed with PIE representatives to explain the project to lay audiences. We aligned with the SATRE driver to use the same video production company to generate a more cohesive overall set of PIE material for the DARE UK programme.
- Eight interactive focus group meetings run by Alterline that captured the thoughts, feelings and experiences of the members of the public. Participants were citizens of the UK and were a representative mix of ages, genders, backgrounds, education levels and country they live in. The TRE-FX team developed the content and the questions to be asked, but had no involvement in the running of the events to ensure there is no indirect influence on participants. The content was reviewed by an external lobbying group prior to sharing with the focus groups.
- Two focus group reports, the outputs of which then influenced the project.
- A series of ad-hoc meetings with PIE representatives to discuss specific topics throughout the project.
- Conceived the driver programme shared glossary of terms.

PIE informed the direction of TRE-FX development. An example was our approach to automation. While a significant portion of the technical solution for our project could have been automated from the outset, the team recognised the importance of public trust and input. In response to public feedback and concerns about unchecked automation, specific stages of the technical pipeline were integrated with manual checks to ensure accuracy and reliability. This approach balanced efficiency with the public's desire for transparency and oversight.

Work with PIE representatives highlighted the need for clear and transparent communication, trust, accountability, and robust data governance in Federated Analytics initiatives to win over public trust. It underscored the importance of addressing concerns related to data security, privacy, and control, while emphasising the potential benefits of collaboration, improved research outcomes, and enhanced healthcare services.

3. Impact of TRE-FX and how this contributes to the design of phase two of the DARE UK programme

The TRE-FX project has made a significant step towards an adoptable streamlined, collaborative and secure data analytic environment acting across different domains. This can be applied in the UK and beyond in Europe.

3.1 National Impact

As a viable reference architecture for federated analytics, TRE-FX has pump-primed the HDR UK Federated Analytics (FA) work programme (started April 2023). The architecture is data domain agnostic, and although HDR UK FA programme will be focusing on health data exemplars, many of these will also require non health datasets held in non health focused TREs, providing an opportunity to test the architecture across a wider range of sensitive data platforms.

The framework's pragmatic approach to reuse of platforms and tools and standards is deliberate to enable the adoption by established TREs and analysis platforms as demonstrated by our reference implementations.

TRE-FX has also included other National outreach activities during the project. Examples include:

- **North West SDE**: collaboration meeting (23/05/23, Manchester) regarding an EPSRC proposal lead by Glenn Martin "Methodology for Developing and Validating Prediction Models Under Federated Analysis"
- **UK TRE RSE Community** RSECon 2023 TRE Satellite workshop (04/09/23, Swansea): presented TRE-FX and participated in TRE RSE meetings and discussions via a slack channel
- **EUCAN-Connect General Assembly/2023 DataSHIELD Conference** (09/10/23, Groningen): presented TRE-FX
- **HPC-AI Advisory Council Conference** (19/11/23, Leicester): presented TRE-FX, invited talk on TREs at Scale
- **UKRI DRI Community Congress** (6/3/23, Birmingham): presented TRE-FX, invited talk on The Power of DRI: A health data perspective.

3.2 European Impact

The UK is acknowledged as having advanced thinking in the area of multi-TRE federated analysis. As TRE-FX has utilised and enhanced a range of standards and solutions already widely used across Europe (GA4GH standards, RO-Crate, workflows and WfExS) this has positioned the TRE-FX framework to be easily adopted by European initiatives, notably **ELIXIR, European Research Infrastructure for Life Science Data**, (established 12/2013, 23 National Nodes, Hub based in EMBL-EBI, 250+ organisations). This intergovernmental organisation brings together life science resources from across Europe and aims to coordinate them so that they form a single infrastructure. Through ELIXIR we are able to promote the deployment of the Five Safes RO-Crate and workflows in the European Health Data Space, Genomics Data Infrastructure and national TREs.

The work is already being incorporated into the following Horizon Europe projects, both led by ELIXIR and both contributing to the European Open Science Cloud:

- **BY-COVID, Beyond COVID** (10/21 - 09/24, 53 partners, 19 countries), which aims to develop a framework for making data from infectious diseases open and accessible. WP4 connects the COVID-19 Data Portal to analysis tools and national resources, including Federated Analysis using workflows. Our work on Five Safes RO-Crate forms a significant component.

- **EOSC-ENTRUST, A European Network of TRUSTed research environments**, (03/24 - 02/27, 34 partners, 17 countries) which aims to develop an interoperability blueprint and establish the framework for large-scale collaborations on sensitive and restricted data via a network of composable TREs across Europe, linked to the European Open Science Cloud, EuroHPC and the European Health Data Space. WP7 is dedicated to RO-Crate and workflow processing, building entirely on the outputs of TRE-FX, with partners Manchester, Nottingham and Barcelona Supercomputing Centre who develop the WfExS.

European outreach activities

- ELIXIR All Hands 2023 Mini-Symposium: Human Genomics and Translational Data 2024-2028, 05/06/23, Dublin, invited talk
- ELIXIR Bioinformatics Industry Forum 2023, 21/11/23, London, invited speaker, round table representation
- Joint Federated Analytics Workshop, 11/10/23, Hybrid Barcelona, invited talk, demonstration, and session lead

3.3 International Impact

The TRE-FX approach will be promoted mainly through three approaches:

- **The TRE platforms**, notably: SeRP which has penetration in Canada and Australia.
- **International partnerships**, notably: a Wellcome Trust Discovery Award proposal that brings in partners from Canada, Australia, Singapore and Japan.
- **Standards organisations**, notably: the FAIR Digital Object Forum (FDOF) where RO-Crate is the premium means of implementing FAIR Digital Objects; The Research Data Alliance Trusted Research Environments for Sensitive Data: FAIRness for "Closed" Data and Processes Working Group; and the Global Alliance for Genomics and Health.

3.5 DARE Phase 2

The Five Safes RO-Crate, defines a standard Digital Object that significantly aids interoperability, streamlining information exchange and access to sensitive information managed in accordance with a set of principles conforming to the Five Safe framework. The Five Safes RO-Crate provides hooks to support potential integration with federated identity management, semi-automated disclosure control, and automated risk assessment for the output control components of the DARE UK programme. The TRE-FX reference implementations align to the DARE UK blueprint (section 4) and components developed by other DARE Sprints can be incorporated.

There is an extraordinary and timely opportunity for DARE UK to make a major step towards developing a full reference implementation of the blueprint by capitalising on their investments, bringing together the Sprint outputs and investing in consolidation and integration, beyond isolated components and proof of concept demonstrators.

- TRE-FX would like to have a working demonstrator system available for others to try and use. Federated Analytics can be a hard concept for researchers and TREs to grasp from powerpoint slides and reports.

Having a live system, even with artificial data, could act as a really important tool for the wider programme.

- TRE-FX would like to incorporate and test the ***SACRO semi-automated disclosure control software*** within our TRE Controller Layer and investigate how we would return information to the researcher if the result would be disclosive, so that they can modify the research object accordingly.
- TRE-FX would like to incorporate ***SARA's data provenance for semi-automated risk assessment*** within the Five Safes RO-Crate provenance that documents workflows and associated assess actions, ensuring each analysis step is traceable and complies with pre-approved standards.
- Collaboration with ***the SATRE driver*** could enable TRE-FX to promote the framework to the TRE community via their collaboration cafes and co-develop the next version of the SATRE specification which incorporates the TRE-FX components and Five Safes RO-Crates. TRE-FX focuses on "middleware" that is invisible to the users and is targeted at the analysis and TRE infrastructure developers. Capacity building within TREs and deployment support is needed to ramp up adoption across the spectrum of TRE capabilities.
- ***TELEPORT to use the TRE-FX external submission layer*.** Although architecturally different, with TELEPORT focusing on federation using "data pooling" where researchers can view the underpinning row level data and TRE-FX focusing on "sending algorithms to the data" without researcher visibility of the row level data, there were many overlapping components and patterns observed across the two drivers. Investigating the similarities across the TELEPORT and TRE-FX projects and developing a wider framework which incorporates both solutions would enable a clear strategy for supporting different types of federated projects using a standardised combined framework. This would be influenced by the DARE blueprint and also inform future versions of the blueprint, and TELEPORT is a potential gateway to easing adoption of TRE-FX.

Consequently, post the end of the driver programmes there are some relatively easy "quick wins" which could significantly increase the impact of the driver programmes with a modicum of additional investment (e.g. DARE phase 1c). There is an exciting opportunity for combining TRE-FX and TELEPORT into a powerful hybrid capable of addressing a range of Federated Analysis patterns addressing the spectrum of where the data is pooled and where the combined analysis occurs.

TRE-FX was a driver programme which built MPV reference implementations using dummy datasets. Collaborating with research projects seeking to utilise federation to robustly test the framework is another clear next step. With a modicum of resources (e.g. DARE phase 1c) the framework could be consolidated and tested by running "dummy" projects through the data governance application process. A much larger DARE phase 2 initiative could support cross data domain exemplar projects utilising the framework on real sensitive datasets. Such an exemplar would require major work packages on information governance, PPIE, TRE adoption and technical development to productionize the MVP software.

4. Alignment with DARE UK Federated Architecture Blueprint

TRE-FX aligns with the broader aspirations of the Federated Architecture Blueprint v1.2-interim under the DARE UK program. The blueprint envisages a managed Federation formed through a network of central registry services and secure interface services deployed at each Federation Participant, aiming to foster a high-assurance network for secure information exchange amongst TREs, data providers, and other service providers with rigorous governance oversight [1 Executive Summary, page 11]. The core essence of the TRE-FX project aligns with this vision, demonstrating a tangible framework to augment federated analytics across a network of TREs and data providers. The blueprint's emphasis on an open standards based ecosystem and avoidance of proprietary lock-in [3.1 Design Principles, page 18] aligns with the TRE-FX project's approach towards leveraging existing

open-standard technologies from ELIXIR and HDR-UK. TRE-FX promotes an ecosystem of varied services interoperating through agreed standards, thereby reducing barriers for researchers and data providers, and avoiding proprietary lock-in.

The blueprint underscores the need for services that are capable of managing “the exchange of queries and results” [1 Executive Summary, page 11]. By employing existing technologies, the TRE-FX project demonstrated a viable mechanism for federated analytics, evidenced through the three reference implementations, enabling the transmission of analytical scripts and workflows to datasets located across diverse secure physical locales [3.3 Related Work, page 19]. This amplifies the service exchange proficiency within the federation and is in sync with the blueprint’s overarching goal of connecting disparate data realms [2 The Strategic Case for A Federated Architecture, page 12].

The blueprint posits stringent governance principles, notably adherence to the Five Safes framework as a guiding principle and the mandate that any analysis of sensitive data should be confined within a TRE [3.1 Design Principles, page 18]. The TRE-FX project resonates with these principles by introducing a Five Safes RO-Crate Digital Object that can encapsulate the necessary provenance information for a query or analysis (implemented as a workflow in the TRE-FX architecture) to be assessed against the Five Safes framework. This would be valuable in the low data mobility quadrant analysis scenarios [2.4.1 Four quadrants, page 15]. Regardless of the number of TREs, the issue of data silos and restricted accessibility remains prominent. The implementations described by this project could help to overcome these barriers by providing references upon which productionisable tools can be built for decentralising data analysis, allowing researchers to execute analytic queries across different silos without the need to centralise the data.

On the technological front, the TRE-FX project provides a substantial proof-of-concept for the Remote Query and Federated analytics outlined scenario [9.2 Technology Proof-of-Concept, page 60]. By showcasing a viable mechanism for federated analytics, the TRE-FX project significantly contributes towards realising the blueprints proposed technological aspirations. It outlines an external submission layer [5 Federated Architecture: Concepts, 5.1 Layers, page 29] and a localised infrastructure with distinct zones (Researcher Zone, Governance Zone, and Data Zone) [6 Federated Architecture: Infrastructure Layer - Federation Participants, page 33] connected via APIs. The TRE-FX project provides a reference for some of these components. These include the external submission layer developed by SeRP that already has applications that span more than the TRE-FX project (it is being used by TELEPORT and other funded research initiatives) and a proposed component-based architecture that sits with the TREs and leverage APIs to facilitate not only seamless query and result exchanges but also interaction with existing TRE services and SOPs.

Acknowledgements

This work was funded by UK Research & Innovation [Grant Number MC_PC_23007] 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).

TRE-FX DOI: 10.5281/zenodo.10055354

Authors

- Thomas Giles, University of Nottingham, 0000-0003-1356-4289
- Stian Soiland-Reyes, The University of Manchester, 0000-0001-9842-9718
- Jonathan Couldridge, University of Nottingham, 0000-0003-4054-1146
- Stuart Wheater, DataSHIELD - Arjuna Technologies, 0009-0003-2419-1964
- Blaise Thomson, Bitfount
- Jillian Beggs, University of Dundee, 0000-0002-7591-3256
- Suzy Gallier, University Hospitals Birmingham NHS Foundation Trust, 0000-0003-1026-4125
- Sam Cox, University of Nottingham, 0000-0002-9841-9816
- Daniel Lea, University of Nottingham, 0000-0001-8152-0398
- Justin Biddle, University of Swansea
- Rima Doal, University Hospitals Birmingham NHS Foundation Trust
- Naaman Tammuz, Bitfount
- Becca Wilson, University of Liverpool, 0000-0003-2294-593X
- Christian Cole, University of Dundee, 0000-0002-2560-2484
- Elizabeth Sapey, The University of Birmingham, 0000-0003-3454-5482
- Simon Thompson, University of Swansea, 0000-0001-6799-0705
- Emily Jefferson, HDR UK, 0000-0003-2992-7582
- Philip Quinlan, University of Nottingham, 0000-0002-3012-6646
- Carole Goble, The University of Manchester, 0000-0003-1219-2137

Contributors

- Vasiliki Panagi , University of Nottingham, 0000-0002-4206-4024
- Andrew Percy, University Hospitals Birmingham NHS Foundation Trust
- Sharon Steele, Alterline
- Olly Butters, University of Liverpool, 0000-0003-0354-8461

Annex Milestones and Deliverables

Milestone **M1** was directly linked to the first Driver project showcase.

- The development of the initial infrastructure to handle RO-Crate processing (**WP3-D1**).
- Initial testing of RO-Crate with the existing federated analytic software vendor partners (**WP5-D1**).
- Establish collaborative partnerships with research organisations both within and outside of DARE UK.
- Bitfount and DataSHIELD insights on federated analysis of sensitive information across organisations.
- Understanding by the project partners of ELIXIR-UK technologies for workflows (WfExS) and RO-Crate.

Milestone **M2** (end of July 2023).

- The development of a primary submission layer, based on HUTCH, that sits outside of the TREs capable of interacting with a variety of alternative services and software solutions (**WP3-D2**).
- Initial integrations of the Five Safes RO-Crate within the technology stacks presented by both federated analytic software partners (**WP5-D1**).
- The establishment of further collaborative partnerships with research organisations both within and outside of DARE UK, including plans to integrate the TRE-FX glossary into the the SATRE project and aspirations to adopt SACRO checking within the DataSHIELD implementation (**out of original scope**).
- Interrogation of the Five Safes RO-Crate by the wider TRE community through a stakeholder engagement workshop (**WP2-D2**).
- Hosting of first PIE engagement event (**WP1-D1a**)
- Initial development of a TRE workflow executor (extending ELIXIR's WfExS Workflow Execution Service) for processing Five Safe RO-Crate objects (**WP3-D1**)

Milestone **M3** (end of October 2023) coincided with the project end date.

- Delivery of a TRE controller by SeRP that allows TREs to control access for users, endpoints and projects (**WP3-D1** - Crosses over with the teleport project as a common deliverable).
- Delivery of a platform tool for modifying Five Safe RO-Crate (HUTCH - HDR UK) (**WP4-D1**).
- Delivery and deployment of the modular federated technology platform including a MVP submission layer ((**WP3-D1+2).**
- Delivery of Bitfount and DataSHIELD workflows compatible with HUTCH (**WP5-D1**)
- Testing of a Bitfount and DataSHIELD queries with the SeRP submission layer (**WP5-D2**)
- Development of additional framework for integrating HUTCH directly with Bitfounts software UI (**out of scope of the original deliverables**).
- TRE-Controller and workflow executor at two UK sites (SAIL and Nottingham) (**WP3-D1+D2**).
- Implementation and testing of Five Safes RO-Crate driven analysis by both Bitfount and DataSHIELD (**WP5-D1 / D2)**, including the development of two (tool specific) alternative reference frameworks for processing queries that still align the Digital Object standards set out by the Five Safe RO-Crate.
- Delivery of integrations between the primary submission layer and the HDR UK data use register for transparency (**WP4 D2**).