

DARE UK

The 2023 DARE UK Driver Projects Summaries and lessons learned

May 2024



Contents

Introduction	3
Safe data: the SARA project	6
Safe outputs: the SACRO project	8
Safe projects: the TRE-FX and TELEPORT projects	10
- Delivering a federated network of trusted research environments to enable safe data analytics (TRE-FX)	10
- Connecting researchers to big data at light speed (TELEPORT)	12
Safe settings: the SATRE project	13
Reflections	15

Introduction

Some of the most important research questions we can ask are the ones that affect us as people - about our health and its connections to our lifestyles and environments; about our children and how we can give them the best opportunities to grow and thrive; and about our societies and what makes them better, fairer, or safer. Answering these questions often requires sensitive data, which is often sensitive because it concerns us as people or has the potential to impact us as people. If we choose to use these data in research, we must handle them with care.

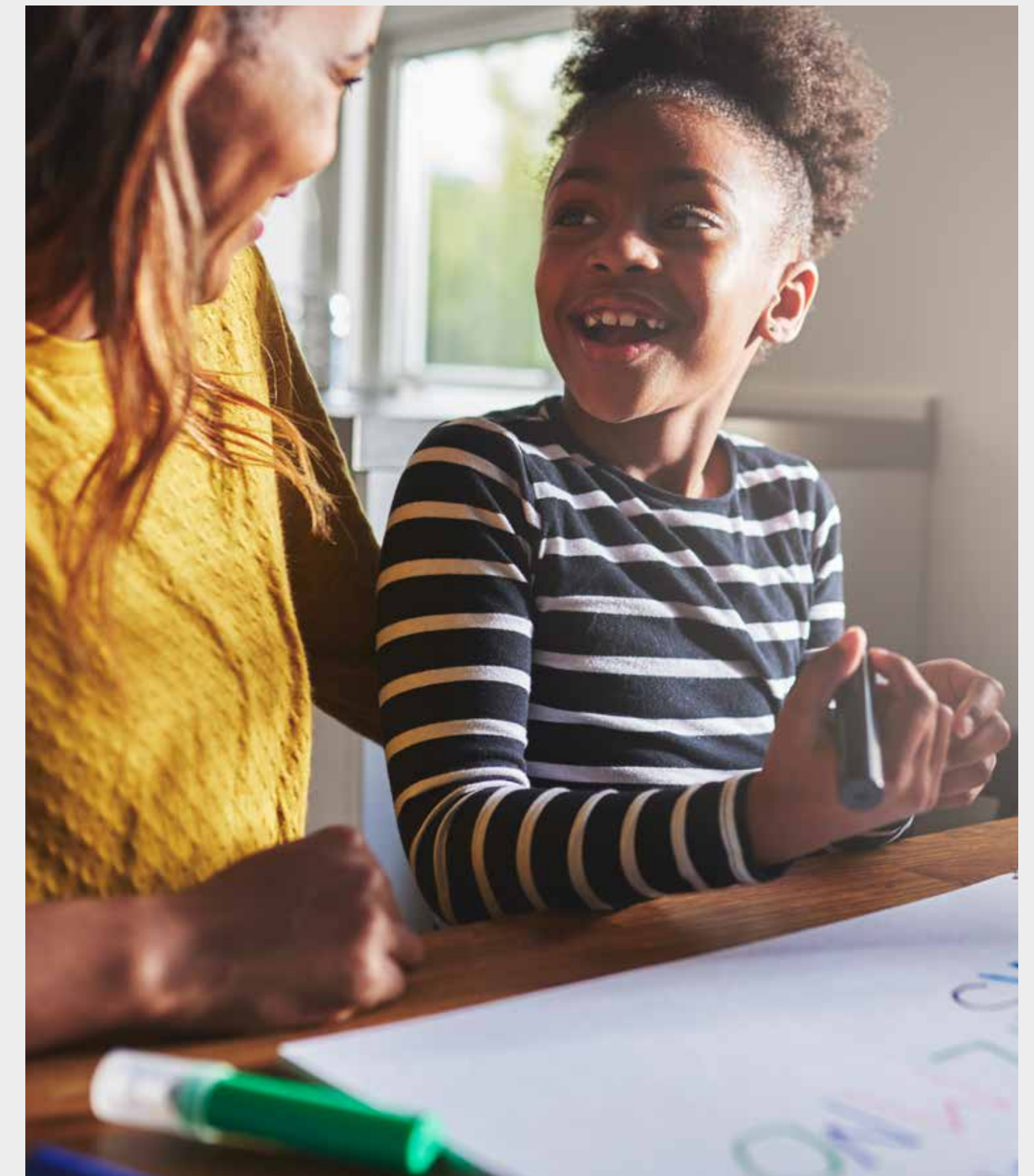
The notion of sensitive data also extends beyond the personal to include things like biodiversity, market pricing, critical infrastructure, and national security. In all cases, the data are sensitive because disclosing them to those not authorised to see them could result in harm to the data subjects, whether people, firms, or endangered species. The UK can and does use sensitive data in research, and it does so to some of the highest standards in the world¹. Over the last two decades, the UK research community has developed and adopted a common framework for handling sensitive data with care: the Five Safes².

The Five Safes approach is a deceptively simple way to think holistically about working with sensitive data. Its “safes” – usually posed as questions – are safe data (how can we minimise the amount of potentially disclosive data

we work with while still keeping it useful?), safe projects (is the use of the data to answer the proposed research question sensible, ethical, lawful and in the public interest?); safe people (are the researchers involved trustworthy and sufficiently skilled?); safe settings (is the research environment sufficiently secure?); and, safe outputs (do we have mechanisms in place to ensure that confidentiality is maintained in any final research outputs?).

¹ See the Global Data Governance Mapping Project Year Three report from the Digital Trade and Governance Hub at George Washington University: <https://globaldatagovernancemapping.org/images/DataGovHub-Year-3/Mapping%20Year%20Three.pdf>

² F. Ritchie (2016); Five Safes: designing data access for research; 10.13140/RG.2.1.3661.1604



The Five Safes should be thought of not as individual things but as aspects of a common approach to managing disclosure risk; tightening one “lever” may allow us to relax another while maintaining the same level of risk control. A modern embodiment of the Five Safes as a means to enable research with sensitive data is the trusted research environment or TRE. TREs are secure computing environments – safe settings - wrapped in information governance and risk management procedures. As interest has risen in research with sensitive data at increasing scale, including increased linkage between formerly disjoint datasets, interest in TREs has risen too.

The DARE UK³ (Data and Analytics Research Environments UK) programme is funded by UK Research and Innovation (UKRI) - the UK’s largest public funder of research and innovation - as part of its Digital Research Infrastructure⁴ portfolio of investments, which support the development of a coordinated vision for digital research infrastructure in the UK. DARE UK is a pan-UKRI, cross-domain programme whose scope covers all types of sensitive data, including data about education, health, the environment and much more. There is growing consensus that all sensitive data should only be accessed and analysed by researchers within a TRE. Central to the DARE UK programme’s ambition is to enable and support the development of a national interoperable secure network of TREs, laying the foundation for an ecosystem of next-generation TREs for advanced data linkage and research for the public good.

Phase 1 of the programme ran from July 2021 until the end of March 2024 focusing on ‘design and dialogue’, aimed

at understanding the challenges across the sensitive data research ecosystem and seeding early exploratory work addressing a range of challenges across the landscape. The work captured in the DARE UK Phase 1 recommendations⁵, initial landscape review⁶ and subsequent infrastructure landscape review⁷ evidence that there continue to be gaps across the sensitive data research ecosystem that make studies that require multiple data sharing agreements across a disparate number of data owners, or the ability to carry out ‘federated’ analyses across multiple TREs, infeasible today.

There are key challenges for researchers working with sensitive data today:

- Limited ability to link and analyse data held within different TREs – limiting the scale and questions which researchers can ask and answer
- Inability to install their own software and utilise data and tools available on the internet due to TREs rightly restricting open access to the internet
- Lack of high-performance computing (HPC) and graphics processing units (GPUs) availability to meet demands from researchers wanting to utilise compute intensive approaches (e.g. image processing, geospatial, sensor data from wearables)
- Almost no capability to support AI research on sensitive data within TREs. Researchers are not provided with modern AI development tools and TREs do not have mature processes for ensuring models that leave TRE environments are ‘safe’ or non-disclosive

- Requirement to learn new environments, technology stacks and processes for each TRE, placing a large overhead on researchers
- Requirement to clean and standardise data on a per project basis rather than this being done once and then shared across projects
- Export of results from TREs is a time-consuming manual process that hinders timely research impact and efficiency
- Cost and efficiency of standing up and running TREs
- Complexity and length of time for data governance applications from submission to approval
- Demonstrating both individual (i.e. per org, initiative) and collective (i.e. as an ecosystem) trustworthiness towards the public

³ Dare UK. See: <https://dareuk.org.uk>

⁴ UKRI Digital research infrastructure. See: <https://www.ukri.org/what-we-do/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/>

⁵ DARE UK. “Paving the way for a coordinated national infrastructure for sensitive data research”. (2022). Zenodo. See: <https://zenodo.org/records/7022440>

⁶ DARE UK. “A review of the UK data research infrastructure”. (2021). See: https://dareuk.org.uk/wp-content/uploads/2021/11/DARE_UK_Data_Research_Infrastructure_Landscape_Review_Oct_2021.pdf

⁷ DARE UK. “UK Sensitive Data Research Infrastructure: A Landscape Review”. Zenodo, Nov. 08, 2023. doi: 10.5281/zenodo.10082545



While the adoption of TREs is clearly positive, an over-proliferation of TREs might be argued as too much of a good thing. Some of these points are made in Better, broader, safer: using health data for research and analysis, perhaps better known colloquially as the Goldacre Review⁸. The current picture of TRE provision is one of bounty with a steady increase in TREs over the years that are poised, off the back of cloud-first technology approaches and the response to intersectional societal challenges such as the COVID-19 pandemic, for a period of evolution and growth. Many TREs are growing up around particular sensitive datasets, on the one hand providing secure gateways for research access to those data, but on the other, risking the creation of a large number of highly secure data silos. If we are to have a manifold landscape of TREs, we need to ensure they can interoperate and federate with each other so that research can be done between and across them without undermining the over-arching Five Safes principles.

For researchers with ambitions to deliver research outputs for public benefit, getting this right in the UK will mean:

- seamless data access processes for research
- population, national and regional scale research opportunities
- a more seamless research user experience across distributed, disparate TREs
- effective and efficient leading-edge infrastructure capabilities (e.g. HPC and GPU availability)
- and opportunities for more routine industrial collaborations with significant public benefit

To this end, alongside various other activities, DARE UK funded two portfolios of projects to begin exploring the challenges in sensitive data research in the UK, all addressed the challenges of advancing data research to enable better, broader and safer research with sensitive data, and all embraced the Five Safes.

In 2022 DARE UK funded a portfolio of nine Sprint Exemplar Projects⁹ in three broad themes: driver use cases, technology demonstrators, and establishing best practice. Following on from this in 2023, DARE UK funded a portfolio of five Driver Projects to work on different, complementary technology-centric aspects of TREs in three broad themes: standardising TREs, connecting TREs, and increasing automation around TREs. Some of these built on earlier work from the DARE UK Sprint Exemplar Projects, and others were new.

This report serves to briefly summarise each project, structured around the Five Safes principle the project most closely, though not exclusively, aligns with. Note that these are summary outlines of the work each project undertook and delivered, for more detailed descriptions refer to each project's individual reports and outputs (these are referenced throughout this document).

⁸ See: <https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis/better-broader-safer-using-health-data-for-research-and-analysis>

⁹ DARE UK Sprint Exemplar Projects. See: <https://dareuk.org.uk/our-work/sprint-exemplar-projects/>

Safe data: the SARA project

Semi-Automated Risk Assessment of Data Provenance and Clinical Free-Text in TREs (SARA)¹⁰ explored two aspects of safe data, asking: Is it possible to use machine learning techniques to better understand privacy risks in free-text data? And, for any data brought into a TRE, can we capture and record more information about how it has been processed to give researchers a better picture of the provenance of the data they ultimately use?

The first question is a particular challenge in assessing free text in health data – radiological reports, hospital discharge letters, and so on – which is estimated to form around 70-80% of the data potentially available for any given patient. None of these data can be used for research until data controllers are comfortable that any and all personally identifiable information has been removed or redacted. For a research cohort of several hundred patients, reading everything that might be useful is infeasible, and so much that might be of value to researchers in delivering patient benefit goes unused.

Machine learning algorithms can sift huge volumes of text very quickly, but they need to be trained to identify what's risky and what's not. SARA applied natural language processing techniques to search not only for direct privacy risks (a report of a medical condition in a discharge letter, for

instance) but also indirect risks. Indirect risks can arise from commentary about a medical condition, but often elsewhere within the text in question: a discharge letter for a patient treated for an overdose might refer to suicidal intent and an incident to which the police were called. In this example, the combination of direct and indirect risks might render the text too disclosive to release for research.

SARA made good progress in characterising indirect privacy risks into broad categories, paving the way for the application of semi-automatic detection and labelling. As with SACRO's work on safe outputs (see below), the research goal here is one of creating methods and tools to support, not replace, human decision-makers. It's also important to remember that SARA's work in this area is not about assessing the risk of publishing a piece of text to the world at large but about assessing the risk of allowing an




¹⁰ A. Casey, et al. "SARA: Semi-automated Risk Assessment of Data Provenance and Clinical Free-text in Trusted Research Environments". Zenodo, 30 Oct. 2023, doi:10.5281/zenodo.10055362

approved researcher, working on an approved project to use the processed text alongside other data within a TRE.

SARA's second question looked at the next stage in the "research data ingest pipeline". Researchers working in a TRE only ever see the data approved and pre-processed for them, but it can be important to know how the data have been pre-processed to avoid further "over-processing" or the use of inappropriate techniques – and also to support the reproducibility of the subsequent research analysis. On this second point, SARA's results are an important step towards further automation and the use of reproducible analytical pipelines for research, as recommended in the Goldacre Review.

There are standard, formal ways of capturing this kind of provenance information for any given dataset, but as with free-text risk assessment, it can be a very manual process. By extending earlier work on a formal "Safe Haven ontology" and applying new tools to the ingest of research data in DaSH, the Scottish Grampian region Data Safe Haven¹¹, the SARA team were able to improve the openness and transparency of data production inside the TRE. This gives researchers a better understanding of how their data were pre-processed ahead of the research analysis and neatly complements the work done on privacy risk assessment.



An important aspect of SARA's work, particularly around the first question of privacy risk, was an ongoing dialogue between the researchers and the project's public panel. The panel provided essential feedback on what a member of the public might consider as indirect risks to privacy, for example, and these conversations directly influenced the project's direction. This aspect of the work is described more fully in the project's final report on public involvement and engagement.¹²

¹¹ Grampian DaSH. See <https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php>

¹² Stuart Dunbar, Arlene Casey, Katherine O'Sullivan, Amy Tilbrook, Elizabeth Ford, Pamela Linksted, Charlie Mayor, Jacqueline Caldwell, Milan Markovic, Ana Ciocarlan, Kathy Harrison, Nicholas Mills, & Katie Wilde. (2023). "SARA Public Involvement and Engagement Final Report". Zenodo. <https://doi.org/10.5281/zenodo.10084410>

Safe outputs: the SACRO project

Where SARA explored the safe data aspect of sensitive data research in TREs, Semi-Automated Checking of Research Outputs (SACRO)¹³ looked at the other end of the process: safe outputs, or how can we introduce efficiencies into checking research results for disclosure risk before they leave a TRE?

SACRO explored two questions related to safe outputs: is it possible to create a consolidated framework with a rigorous statistical basis that provides guidance for TREs to agree on consistent, standard processes to assist in quality assurance, and can we design and implement a semi-automated system for checks on common research outputs, with increasing levels of support for other types of output, such as trained AI (artificial intelligence) models?

Current best practice in practical disclosure checking is captured in the Statistical Disclosure Control Handbook published by the Secure Data Access Professionals working group¹⁴. A key feature is “four-eyes checking” – outputs are assessed by two independent pairs of eyes (people, of course) in sequence, with the second pair of eyes picking up anything missed by the first pair.

It is frequently the case that a large portion of the potentially disclosive statistics currently picked up by the

first pair of human eyes are both routine and statistically automatable. Using expensive and scarce human expertise to pick up mistakes in potential outputs that could be identified automatically with mathematically provable safety is both inefficient and error-prone. For non-routine cases, the second pair of human eyes is indispensable, but a statistically-based automated approach can replace the first pair without compromising overall safety.

The SACRO team captured, for the first time, the necessary rigorous statistical foundations in a guide which formalises a radical new approach to output checking. This guide views output risks as being associated not with a particular statistic but with a class of statistics, which they term a ‘stat barn’¹⁵. This taxonomic approach, which is able to classify an arbitrarily large number of individual statistics into a manageably finite number of types (the ‘stat barns’), provides the underpinning for the “ACRO engine” software.



¹³ J. Smith, et al. “SACRO: Semi-automated Checking of Research Outputs”. Zenodo, 6 Nov. 2023, doi:10.5281/zenodo.10055365

¹⁴ Welpton, Richard (2019). “SDC Handbook”. figshare. Book. <https://doi.org/10.6084/m9.figshare.9958520.v1>

¹⁵ Ritchie, F., Green, E., Smith, J., Tilbrook, A., & White, P. (2023). “The SACRO guide to statistical output checking (Version 1)”. Zenodo. <https://doi.org/10.5281/zenodo.10054629>

The ACRO engine is the result of the second part of SACRO's work, a software framework that uses the 'stat barn' rules from the formal guide to create a set of tools for researchers and output checkers.

For TREs who support the ACRO tools, researchers can utilise the ACRO engine within their analysis pipeline. By applying minimal changes to their existing analysis codes (in R, Python or Stata), they can engage the ACRO library and get an instant report on the probable safety of their current statistics. This enables the researcher to spot and fix mistakes or marginal statistics themselves, ahead of the much longer process of submitting outputs for checking and waiting for their turn in the output checker's queue. A good software development analogy is developer-driven unit testing.

Both the formal guide and ACRO engine software have been welcomed by the broader community and have already seen significant uptake. A paper describing the formal approach was presented at the September 2023 United Nations Economic Commission for Europe expert meeting on statistical data confidentiality¹⁶, and the ACRO engine has been trialled successfully within the Grampian DaSH and at the European Commission's Eurostat agency. Alongside this is the development of a new community group (Statistical Disclosure Control - Reducing Barriers to Outputs from TREs, or SDC-REBOOT¹⁷) looking at community-driven adoption of the ACRO engine and similar disclosure control tooling.

¹⁶ Derrick, B., Green, E., Richie, F., and White, P.: Towards a comprehensive theory and practice of output SDC. United Nations Economic Commission For Europe Conference Of European Statisticians, Expert Meeting on Statistical Data Confidentiality, September 2023, Wiesbaden. Online at https://unece.org/sites/default/files/2023-08/SDC2023_S5_2_UWE_Ritchie_D.pdf

¹⁷ See here: <https://dareuk.org.uk/dare-uk-community-interest-groups/dare-uk-community-interest-group-evaluation-of-automated-output-checking-and-ai-model-risk-assessment/>

Safe projects: the TRE-FX and TELEPORT projects

Taking a broad view, all five Driver Projects contribute to the “safe projects” aspect of research with sensitive data; this is the holistic nature of Five Safes. However, the TRE-FX and TELEPORT projects focused on the critical question posed by an emerging class of projects: how do we maintain a safe project when our analysis spans more than one TRE?

These are the research projects that require TRE-to-TRE federation to enable data analysis spanning distributed secure datasets and goes to the heart of the question posed in the introduction, TRE-FX and TELEPORT addressed this in different but complementary ways.

This complementarity in approaches between TRE-FX and TELEPORT is a critically important point, different kinds of data research questions requiring the capability of working safely across multiple TREs may have fundamentally different requirements in terms of their federated analysis methods or approaches. At its simplest level, this means that researchers must be able to either safely send their query to the data and safely receive a result from that query without needing to ever see the de-identified data itself, or safely and securely access distributed de-identified datasets that are presented to the researcher in a single view, or some combination of

these two approaches. Ensuring both approaches, and the underpinning technical capability and information governance acceptability, are available to researchers tackling intersectional societal challenges in the public good is critically important.

Delivering a federated network of trusted research environments to enable safe data analytics (TRE-FX)

TRE-FX¹⁸ explored the “job submission” model of querying remote TREs, in this pattern an approved researcher on an approved project interacts not with a TRE directly but with two other complementary services: a software repository service and a job submission service. The interaction of these services enables the researcher to run a scientific analysis against multiple comparable datasets held within

different participating TREs without needing to access or see the data directly.

The software repository service can be an existing, general-purpose software-hosting service such as DockerHub or WorkflowHub.eu. The researcher, knowing the nature and schema of the datasets in question, can develop a suitable analysis program, following whatever rules and formatting requirements the participating TREs stipulate. The researcher uploads the suitably packaged analysis program to the software repository and notifies the participating TREs of where to find it.

¹⁸ Thomas Giles, Stian Soiland-Reyes, Jonathan Couldridge, Stuart Wheeler, Blaise Thomson, Jillian Beggs, Suzy Gallier, Sam Cox, Daniel Lea, Justin Biddle, Rima Doal, Naaman Tammuz, Becca Wilson, Christian Cole, Elizabeth Sapey, Simon Thompson, Emily Jefferson, Phillip Quinlan, & Carole Goble. (2023). “TRE-FX: Delivering a federated network of trusted research environments to enable safe data analytics”. Zenodo. <https://doi.org/10.5281/zenodo.10055354>



TREs do not, by and large, allow arbitrary software downloaded from the Internet to execute in their environments without scrutiny and risk assessment. The TRE-FX model fully supports this qualification phase; only when the researcher’s software, downloaded from the repository service, has been approved for use is it made available to the second half of the TRE-FX solution, the job management engine.

Once the participating TREs are happy with the researcher’s software, the researcher is notified and is then able to target analysis jobs at the various instances of their code (and thus the underlying datasets) using the TRE-FX job submission service. This service receives researcher requests and forwards them to the participating TREs, where they are handled by the job management engines. These, in turn, run the pre-approved analysis code against the approved datasets, package up the results and return them individually to the submission service. The submission service assembles the partial results from the participating TREs into a final whole and passes it on for statistical disclosure output checking and ultimate release to the researcher.

TRE-FX demonstrated this approach in operation using the UK Secure e-Research Platform. Over the course of the project, the team developed a modular architecture and a number of API standards, plus an implementation of a job submission service and a job management engine (Hutch¹⁹). They also collaborated with two organisations

that provide researcher-facing workbenches – Bitfount²⁰ and DataSHIELD²¹– to enable these familiar software tools to work with the TRE-FX back-end environment. Key to this interoperability was the development and adoption of a standardised way of packaging and exchanging structured data objects between the various services involved. This standard, the “Five Safes profile” for RO-Crates²², provides a common object exchange format for a variety of data types; think of RO-Crates as an “envelope” format that provides a structured wrapper for almost arbitrary data contents.

Another important element of the federated architecture that emerged from TRE-FX was the prototype of a central registry of users, projects and data. In an environment where computational jobs are sent from a job submission service to a number of remote TREs, those TREs must have, or be able to find, enough information about who is asking for computational resources and access to data and in what context. Without this information, the TRE will be unable to make the necessary authorisation decisions, and the job will fail automatically. TRE-FX developed a service which enables TREs to look up the necessary information centrally (which users are members of which projects, and have permission to access which datasets, for example).

¹⁹ See <https://github.com/HDRUK/hutch>

²⁰ See <https://www.bitfount.com/>

²¹ See <https://www.datashield.org/>

²² See <https://trefx.uk/5s-crate/0.4/>

Connecting researchers to big data at light speed (TELEPORT)

TELEPORT²³ looked at the challenge of maintaining a safe project when the analysis spans more than one TRE from a different perspective to that of TRE-FX. That is, if a researcher working in a local TRE has permission to view two datasets, one local and one remote, can we send a query directly from the local TRE to the remote one and present the combined dataset to the researcher as a single view (a “single pane of glass”)? A second question that followed was then: if we can do so, how can the remote TRE retain governance control of its dataset, which the query to the other TRE may well have transferred?”

TELEPORT’s answer to the first question used a polystore²⁴ database connection between two TREs. A polystore is defined as “any database management system (DBMS) that is built on top of multiple, heterogeneous, integrated storage engines”, making them very promising technologies for the manifold, independent world of TRE-hosted data. TELEPORT demonstrated the feasibility of this approach through successful tests using synthetic data between national TREs in Wales and Scotland. In contrast to the TRE-FX approach, the TELEPORT approach gives the research user direct access or sight of the de-identified data to carry out their analysis, and importantly the combined dataset is presented to the research user through a single view or interface. The research users experience is that of logging into a single TRE and being able to carry out analysis working directly with de-identified data from multiple TREs.

In answer to the second question, TELEPORT demonstrated the use of a “pop-up” TRE-within-a-TRE as a mechanism to keep a running TRE project environment synchronised with an approved “known good” state. A pop-up TRE is one which is entirely software-defined and can be deployed automatically on demand. TELEPORT used the approach of creating a project-specific pop-up TRE inside a host TRE, which maintained a connection to information governance authorities at a remote TRE. The project pop-up included a polystore database combining views of local and remote datasets.

Being entirely software-defined, the pop-up TRE could be wrapped in a control environment which merges the governance requirements of both local and remote TREs. Borrowing ideas from continuous integration/continuous deployment (CI/CD) in modern DevOps software delivery, the project team demonstrated the use of off-the-shelf CI/CD management software as a “keep-alive” channel for the running pop-up TRE. If, at any point, the state of the running pop-up deviated from the approved “know good”, the CI/CD system could “automatically de-provision” or shut down the whole pop-up TRE.²⁵

²³ 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. (2023). “TELEPORT: Connecting researchers to big data at light speed”. Zenodo. <https://doi.org/10.5281/zenodo.10055358>

²⁴ See <https://wp.sigmod.org/?p=1629> for a foundational blogpost on polystores

²⁵ This is not dissimilar to “rapid unscheduled disassembly”

What might trigger this? A typical test scenario was the expiration of one researcher’s approval credentials as granted by one of the participating TREs in the course of a research project. If this were a hard requirement for any of the TREs involved in sharing data into the pop-up TRE’s polystore, credential expiry would automatically trigger a shutdown of the pop-up environment without any further human intervention required. In practice, this might be a rather drastic response, but as a demonstration of what kind of remote governance automation is possible, it is compelling.

Safe settings: the SATRE project

To date, there has been significant work around the definition of what a TRE is,^{26 27 28 29 30} with more formal ways of assessing TREs often closely aligned with broader information security standards³¹ though these are more like marking schemes than study notes.

Looking across a number of current UK TREs hosting different types of sensitive data and building upon previous work, Standardised Architecture for Trusted Research Environments (SATRE)³² set out to assimilate the essential features of these TREs into a common specification and to provide a first blueprint for new TRE builders.

Through broad community consultation and consensus-building, the project delivered version 1.0.0 of the SATRE specification³³, intending that this become a living document to be developed further in the future through community input. The specification is built on four architectural principles and four pillars, identifying 29 capabilities that a TRE should have. These capabilities are further broken down into 160 statements against which a TRE – or a TRE design – can evaluate itself. A full 75 of the 160 statements are mandatory and define the minimum set of capabilities required to be a SATRE-compliant TRE.



²⁶ UK Health Data Research Alliance. (2020). “Trusted Research Environments (TRE) Green Paper”. Zenodo. <https://zenodo.org/records/4594704#.Yd8RRdHP1zq>

²⁷ UK Health Data Research Alliance, NHSX. (2021). “Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems”. Zenodo. <https://zenodo.org/records/5767586>

²⁸ Paul R. Burton, Madeleine J. Murtagh, Andy Boyd, James B. Williams, Edward S. Dove, Susan E. Wallace, Anne-Marie Tassé, Julian Little, Rex L. Chisholm, Amadou Gaye, Kristian Hveem, Anthony J. Brookes, Pat Goodwin, Jon Fistein, Martin Bobrow, Bartha M. Knoppers. (2015). “Data Safe Havens in health research and healthcare”. <https://academic.oup.com/bioinformatics/article/31/20/3241/195451#394490978>

²⁹ Nathan Christopher Lea, Jacqueline Nicholls, Christine Dobbs, Nayha Sethi, James Cunningham, John Ainsworth, Martin Heaven, Trevor Peacock, Anthony Peacock, Kerina Jones, Graeme Laurie, Dipak Kalra. (2016). “Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research”. <https://pubmed.ncbi.nlm.nih.gov/27329087/>

³⁰ Sanaz Kavianpour, James Sutherland, Esmā Mansouri-Benssassi, Natalie Coull, Emily Jefferson. (2022). “Next-Generation Capabilities in Trusted Research Environments: Interview Study”. <https://pubmed.ncbi.nlm.nih.gov/36125859/>

³¹ See, for example, the UK Statistics Authority accreditation process for processors of UK statistical data (<https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/>) and the Scottish Government’s Safe Haven Charter (<https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/>)

³² Christian Cole, Hari Sood, Simon Li, Katie Oldfield, Matt Craddock, Nel Swanepoel, Sonya Coleman, Martin O’Reilly, Dermot Kerr, Cian O’Donovan, James Hetherington, Jim Madge, David Sarmiento-Perez, Ed Chalstrey, James Robinson, Jillian Beggs, Tim Machin, & Antony Chuter. (2023). “SATRE: Standardised Architecture for Trusted Research Environments”. Zenodo. <https://doi.org/10.5281/zenodo.10055345>

³³ See <https://satre-specification.readthedocs.io/en/v1.0.0/specification.html>



SATRE's four capability pillars make a first attempt to capture the essence of a TRE in as concise a way as possible.

They cover:

- **Information governance:** including capabilities related to quality management, risk management and training delivery.
- **Computing technology and information security:** including end-user computing, infrastructure management and information security.
- **Data management:** including identity and access management, data lifecycle management and output management.
- **Supporting capabilities:** including financial management, public engagement and project management.

A more detailed breakdown of the capability pillars can be found in Machin *et al*³⁴. The four architectural principles offer guidance on how these capabilities should be delivered.

They are:

- **Usability:** A TRE instance that works for all users minimises barriers to use, providing a productive and accessible analysis environment for research.
- **Maintaining public trust:** TREs holding public data should build and maintain the trust of data subjects and any other impacted individuals, groups, communities and organisations by protecting privacy, keeping data secure and being transparent about their work.

- **Observability:** Human-initiated and automated processes resulting in change within the TRE should be observable.
- **Standardisation:** TREs should adhere to standards or well-known patterns wherever possible.

The SATRE specification and the strong foundations behind it give us, for the first time, as a community, a set of definitions for what makes a good TRE. As the project team fully anticipate, this may not be the final word, but it is an excellent start that will continue to evolve beyond the lifespan of the project through community input.

³⁴ Machin, T., Chalstrey, E., Cole, C., Craddock, M., Hetherington, J., Li, S., Madge, J., O'Reilly, M., Robinson, J., Swanepoel, N., & Sood, H. (2023). "A Standard Architecture for Trusted Research Environments" (1.0). Zenodo. <https://doi.org/10.5281/zenodo.8411274>

Reflections

In the same way that the Five Safes should be thought of not as individual things but as aspects of a common approach to managing disclosure risk, so should the DARE UK Driver Projects be viewed as a set of complementary capabilities that collectively provide the beginnings of a common toolkit for TRE-to-TRE federation to support emerging sensitive data research in the UK.

While the Driver Projects have delivered an exciting view into what may be possible for TRE-to-TRE federation and the potential impact for sensitive data research in the UK, these projects (and the Sprint Exemplar Projects before them) have by design been exploratory. Maturing the ideas and capabilities developed through the Phase 1 project portfolios will be important looking forward, in particular the potential that exists in the interplay between these capabilities as they mature to ensure they can act as complementary parts of a common, collective toolkit. TRE-FX and TELEPORT have already provided a glimpse of just this kind of potential

for integration, equally there is potential for integrating SACRO within the workflows of both TELEPORT and TRE-FX to embed efficient statistical disclosure control at the appropriate boundaries as prescribed by information governance requirements. Or in the case of SATRE, maturing the specification to integrate not only intra-TRE capabilities (for example SARA's work on indirect privacy risk assessment within a TRE) but increasingly inter-TRE capabilities (for example TRE-FX and TELEPORT) that support the sort of routine TRE-to-TRE federation required to support emerging sensitive data research.

There have been several lessons learnt throughout Phase 1 from the Driver Projects work and the previous Sprint Exemplar Projects that will be carried forward:

- There is a strong appetite or ‘pull’ from the sensitive data research communities to both advance innovative TRE capabilities and utilise these services to deliver scientific outputs.
- Sprint-style portfolios of projects are a successful vehicle for early conceptualisation and prototyping of ideas. A one-year time window should be the minimum period of delivery. Maturing innovative infrastructure capabilities towards real-world research applications requires different mechanisms.
- Dedicated resourcing and mechanisms should continue to catalyse community-led collaborations and idea development work. The combination of broad consortia funded through the Driver Projects portfolio and direct support from the programme to enable the community to come together through the DARE UK Community Groups initiative has worked to deliver this.
- Delivering high-quality public involvement and engagement (PIE) is challenging. Coordinating all programme activities and providing focused support within specific activities is critical to delivering a coherent PIE strategy across a portfolio of projects.

In summary, DARE UK's 2023 Driver Projects have given us a glimpse of what a more standardised, more connected, and more automated federation of TREs could look like. As we look ahead to the next phase of the DARE UK programme, this is a good position to be in.



DARE UK

Get in touch

✉ enquiries@dareuk.org.uk

🌐 www.dareuk.org.uk

✂ [@DARE_UK1](https://twitter.com/DARE_UK1)

📌 [DARE UK](#)

**UK
RI** UK Research
and Innovation

HDRUK
Health Data Research UK

 **ADRUK**