Trusted Research Environments and data management – Past, Present and Future

26 April 2021

Introduction

This paper provides the HDR UK input to the Goldacre Review request for information on past, present and planned work around Trusted Research Environments (TREs) and data management, specifically on:

- TREs (this may include commodity technology such as a server in a rack, or some cloud capacity, and the permissions structure around it; or a more complex and bespoke environment with specific analytic code and services operating within it).
- Methodological work, code or tools for data management (whether EHR data, bespoke collections of health data, or other related data).
- Methodological work, code or tools for federated analysis.

The Goldacre Review provides a timely stock-take of progress to date and consideration of the best options for future activity related to TREs and data management across England. It is important to ensure alignment between this perspective and the development of a UK-wide (and beyond) federated network of TREs. This will be further addressed in the upcoming papers from the health data science community which are currently in development or under consultation, including the data standards green paper and Trusted Research Environment white paper, which will be published in May 2021.

Public and patient involvement and engagement, open science and open code are at the heart of HDR UK's developments, and with over 150 resources available on Github, HDR UK is committed to accelerating reproducible science through open standards, data and source code.

1. Trusted Research Environments

Background

Trusted Research Environments (TREs) are secure platforms for researchers to access sensitive data. TREs are an essential way of managing risk for unauthorised access and/or re-identification of individuals from de-identified data. The re-identification risk amplifies with increased data linkage across a wider range of disparate datasets and to date, most of the focus for TREs has been to manage information security within their environments using defined processes, people and technology.

HDRUK
Health Data Research UK

TREs are fundamentally architected to manage the data lifecycle - data acquisition, cleaning, harmonisation, preparation, provisioning & deprovisioning and an analytics environment in support of research projects. Some TREs also provide data access management services as a delegated data access committee acting on behalf of Data Custodians. In many instances, TREs are co-located with the Data Custodians and often departments within the same organisation.

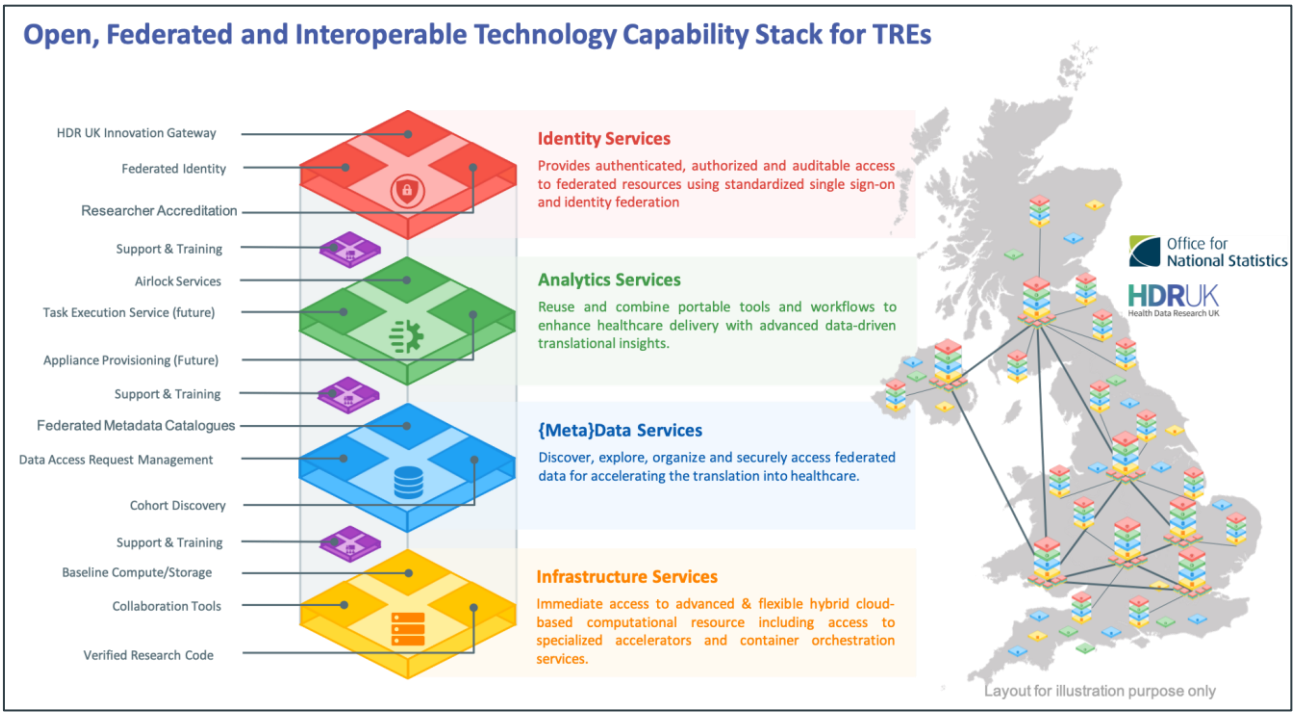
Historically, data custodians have operated on a "data release" model, where once a data access request has been granted for a research project, the required dataset is prepared (minimised and optionally linked) and released to the research project members for analysis within their own secure environment. In this model most of the risk-minimisation steps (statistical disclosure control policies) are performed upfront before releasing the data to the researchers. This adds complexity both in terms of time and effort for the data custodians to prepare the data, and often reduces the fidelity of the data for research purposes.

Shared and federated infrastructure has a long history in academia and research, and are often setup for other scientific endeavours – e.g. physics, bioinformatics where data is large and complex and the computation required is immense such that no one organisation is able to access and analyse the entirety of the data within one location. High-Performance Computing (HPC) and further Grid Computing infrastructures were specifically designed and architected to address the scale of data volume and compute required for such research projects. Handling sensitive data such as commercial and healthcare data are often bolted onto existing infrastructures to support such research projects.

The UK Health Data Research Alliance published its green paper "Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs" in July 2020 – informed by a workshop in March 2020 and stakeholder consultation during May and June 2020. It signalled a decisive pivot away from the traditional "data release" models to a model where the analysis (and analyst) accesses the data through a secure remote access service using the Five Safes framework (SAFE People accessing SAFE Data within a SAFE Settings to undertake SAFE Projects resulting in SAFE Outputs) developed by Office for National Statistics and supported by Administrative Data Research UK (ADR UK).

The pandemic highlighted the need for greater and rapid access to data for research at scale and special legislation further helped accelerate the sharing of sensitive data for research. HDR UK has been at the forefront of supporting the alignment and development of UK-wide TRE network as part of the UK Health Data Alliance and the National Core Studies, in partnership with the Office for National Statistics. The period has also highlighted the varied lineage, heterogeneity and maturity of the TREs in the UK, and even definitions of a TRE. As part of the National Core Studies HDR UK performed a survey of the technical capabilities of the TREs – the results are accessible on the Innovation Gateway as a TRE Collection.

HDR UK's focus has been leveraging existing technical capabilities where available and aligning these capabilities to be interoperable in a wider nationwide network. HDR UK does not 'own' or run any TREs. It has purposely focussed on capability development and interoperability between TREs rather than technology implementations and/or recommendations, as the latter are often decisions that reflect the operational context and maturity of the TRE.



Specific TRE investments

HDR UK has led on specific targeted investments in TRE capabilities in the Data and Connectivity National Core Study (NCS) as part of the pandemic response:

- The Data and Connectivity National Core Study (NCS) aims to make vital data discoverable via the Innovation Gateway and available in national TREs to accelerate research on COVID-19. It is led by HDR UK in partnership with the Office for National Statistics. The Data and Connectivity programme is one of six National Core Studies, that support the UK's research response to COVID-19.
- During the first six months of the Data and Connectivity National Core Study (D&C NCS) five TRE Delivery partners from the Alliance (NHS Digital, Public Health Scotland/EPCC, SAIL Databank, HSC NI Public Health Agency, Office for National Statistics) worked together to enable the development of data infrastructure and services across the UK to allow the priority research questions to be answered efficiently, in a transparent and trustworthy way. This ecosystem is being extended to include OpenSAFELY, that has been supporting the Longitudinal Health & Wellbeing NCS in particular.
- Funding for the D&C NCS is awarded by UKRI (MRC). HDR UK contracted with the five delivery partners in October 2020 for an initial six-month period (phase 0). Phase 1 funding has been awarded by UKRI for an additional 18 months.
- Key deliverables TRE deliver partners have been contracted to achieve include:
    a) Making available a core COVID-19 national data asset available in each TRE for use by all NCS
    b) Ensuring priority datasets for COVID-19 research are findable, accessible, inter-operable and reusable (FAIR) as a single "shop window" (via the Innovation Gateway)
    c) Developing a harmonised data access process for NCS related data access requests and harmonised researcher accreditation process

d) Develop technical standards and components to enable advanced analytics across TREs (federation capabilities)
e) Publishing of the Data Protection Impact Assessments (DPIA) and obtaining Digital Economy Act (DEA) accreditation

- Impacts achieved so far, include:
  - Data assets: There are now 84 NCS data assets onboarded to the TREs and discoverable on the Innovation Gateway. These include vaccine datasets, viral genomic data (COG UK), acute care data, and a joint ONS/NHS Digital joint health data asset.
  - The UK's largest linked health data research asset: By working in partnership with NHS Digital and the BHF Data Science Centre, we have created a new linked health data resource covering 54.4 million people – over 96% of the English population (BMJ 2021). This is led by the CVD-COVID-UK consortium and is available to UK researchers to collaborate in NHS Digital's new secure research environment. It is the largest research resource in the UK and is being extended to 67 million people through federated working. It is already being actively used by over 50 researchers, with all code fully open and accessible via Github.
  - Improving experience for researchers: 21 data access requests covering 83 NCS data assets have been submitted via the Innovation Gateway. Over the next 12 months we will support other NCS and urgent COVID research studies with streamlined data access requests. Remote access to Northern Ireland health and social care data for research is now available for the first time. TRE delivery partners are also supporting NCS clinical trials to enable their data flow and linkage needs.
  - Data enabled research outputs: To date, there have been 1,204 COVID pre-prints using health data and 131 published papers. Key outputs include: the BREATHE Hub study analysing effectiveness of first dose of COVID-19 vaccines against hospital admissions in Scotland, the CVD-COVID-UK consortium project which has made available linked health data for research on a cohort of more than 54 million people, and implementing a data enabled recruitment approach for the PRINCIPLE trial.
- Capability development and interoperability between TREs: Work is underway across all TREs to progress planned longer term NCS deliverables for: a) metadata catalogues with data utility framework, b) cohort discovery c) research passports and visas d) further Data Access Request harmonisation and e) federated analytics capabilities. There is still work to be done on developing security by design, remote analysis services and ingress/egress airlock service.
- Funding via the Data & Connectivity programme: This funding is intended to support a range of objectives including dataset onboarding, data provisioning, data engineering, supporting users so is not limited to the specific Goldacre Review request regarding *commodity technology such as a server in a rack, or some cloud capacity, and the permissions structure around it; or a more complex and bespoke environment with specific analytic code and services operating within it.* Investment in the Office for National Statistics TRE developments are funded separately by ADRUK.

TRE	HDR UK Investment to date		
	Data & Connectivity	**Other HDR UK**	**Total**
	Actual to date £'000	Actual to date £'000	Actual to date £'000
NHS Digital	696	420	**1,116**
Public Health Scotland/EPCC	851	460	**1,311**
SAIL Databank	1,134	2,476	**3,610**
HSC NI Public Health Agency	365	-	**365**
Total	**3,046**	**3,356**	**6,402**

Notes:

1. The wider National Core Studies Programme includes investments in ONS and other TREs. These investments are not managed through the Data & Connectivity programme so not included here.
2. HDR UK has invested £1.6m in the NHS DigiTrials Health Data Research Hub. This is not an investment in the TRE, so is not included in the table above. Further detail on the Health Data Research Hub programme is below.
3. HDR UK has other investments in compute and infrastructure that complement our TRE investments. Any investment not directly in TRE activity is excluded from the table above.
4. The award for the Data & Connectivity NCS for Phase 1 (Apr-21 to Sep-22) is £15.15m (yet to be formally announced by UKRI) covers UK-wide developments to achieve the following aims:

- Continue to respond to emerging COVID-19 research priorities, mapping key datasets required by National Core Studies, NIHR Urgent Public Health Studies and SAGE sub-groups to allow research which can inform policy and operational decision making across the UK.
- Further develop the data infrastructure and services across the UK to allow faster access to high priority health, administrative, molecular, and behavioural data assets for researchers working on the most important COVID-related studies, ensuring priority research questions can be answered efficiently, in a transparent and trustworthy way.
- Strengthen and extend the existing national Trusted Research Environments (TREs) and UK Health Data Research Innovation Gateway infrastructure through inclusive four nations approach ensuring the priority datasets for COVID-19 research are findable, accessible, inter-operable and reusable (FAIR) as a single "shop window".

2. Methodological work, code or tools for data management

As part of the Life Sciences Industrial Strategy, and a bid to UKRI, HDR UK was commissioned to unite health data as part of the £210M Industrial Strategy Challenge Fund designed to support the development of precision medicine for improved early diagnosis and treatment. A total of £37.5M was made available to the **Digital Innovation Hub Programme**, with the remaining £172.5M being awarded for genomics and five Centres for digital pathology, radiology, AI and machine learning, and enabling integrated diagnostics.

The ISCF Digital Innovation Hub Programme specifically supported the creation of seven Data Research Hubs and the UK Health Data Innovation Gateway.

UK Health Data Research Hubs: are seven centres of excellence with expertise, tools, knowledge and ways of working to maximise the insights and innovations developed from the health data. Launched in September 2019, in their first 18 months, the Hubs have made 157 datasets discoverable on the Health Data Research Innovation Gateway, have delivered 300 multi-sector projects, with over 20,000 meaningful patient and public interactions, and 2,300 training activities.

The April 2021report Improving UK Health Data: Impacts from the Health Data Research Hubs shows how the Hubs have informed UK policy decisions on the effectiveness of COVID-19 vaccines, created tools to improve clinical decision-making in the management of patients with vascular disease, and supported research in cancer, heart disease and hospital care pathways by linking routinely-collected data. Two new Hubs will join our network next month to support data research in Pain and Mental Health, funded by the MRC.

Direct investment to seven Data Research Hubs to date: £24m through to April 2022

HDR UK Data Utility Framework: This new framework shows the usefulness of data for research, it has over 100 datasets evaluated against it and is now integrated into the Gateway. Other organisations in the UK, such as NICE and NHSX, are seeking to adopt the framework, and there is considerable interest internationally from colleagues in the USA, Singapore and Israel. Key papers from the community are in development and consultation, including the data standards green paper and Trusted Research Environment white paper, which will be published in May.

Investment with MetadataWorks to develop framework: £65k (outputs here)

Investment with Inspirata to test and evaluate tools: £189k (outputs here)

Data standards

An example of HDR UK's work to enable federation across TRES has been the development of open standards for dataset metadata. HDR UK co-developed the specifications with consultation across a wide stakeholder group to inform the standard and nurtured any ecosystem of providers to support TREs provide these capabilities for research. All datasets discoverable on the Gateway are mapped to this standard, and a validation script is run against the datasets. The metadata and validation process and outputs are also

available on GitHub. In June 2020, HDR UK released the Principles for Data Standards paper, based on consultation across the community, and an updated version of this paper will be consulted on in May, along with the publication of the Trusted Research Environment white paper.

3. Methodological work, code or tools for federated analysis

Health Data Research Innovation Gateway:

The UK is home to a vast number and a rich resource of health datasets that has the potential to accelerate health data research and innovation. The purpose of Gateway was to develop a "single shop window" application to enable data discovery, data linkage, and enable health data science to take place in a safe and efficient manner across multiple TREs in the UK.

Following the development of a Minimum Viable Product (MVP), formal development of the Gateway commenced in April 2020.

As of April 2021 (12 months development), 640 datasets are available for researchers to discover and request access to, safely and securely, via the HDR Innovation Gateway, acting as the UK's unified platform for data discovery and access. There are 1,000 researchers registered on the Gateway. With over 11,500 searches each month from around the world and 640 datasets now discoverable, the Gateway is becoming the "go to" place for researchers to discover and request access to UK health datasets with the added benefit of giving much-needed transparency to the UK public on what data is available, how they are used and why. Open APIs can be found here and Open Source code for the gateway can be found here.

PA is the technology partner developing the Gateway following an open procurement process (https://www.hdruk.ac.uk/wp-content/uploads/2019/12/191016-DIH-Gateway-Phase-2-Technology-Partnership-Specification_Update-20-Nov.pdf). This is a maximum investment o £6.3m until April 2022.

Cohort Discovery:

We also achieved a major step forward in achieving the federated data ecosystem vision, with the launch of a MVP of cohort discovery research within the Gateway. This enables, for the first time, researchers to search across datasets to find cohorts of patients with specific, defined characteristics; opening up a huge potential for increased discovery, while maintaining safety and security, helping researchers get to impact faster.

Users of the Gateway can already search via keywords and can drill down further by using filters or "Collections" which group resources around specific research themes and topics. This new cohort discovery functionality co-developed by HDR UK, University of Nottingham, BC Platforms (BCP), PA Consulting and the CO-CONNECT project allows users to search by specific cohorts or demographic groups. For example, women in England between the ages of 18 and 30, with asthma that do not smoke. This adds an extra layer and dimension to data discovery provided by the Gateway, enhancing the utility of a number of these

datasets even further; and enabling new levels of analysis and insights that will ultimately feed through to the front line of improved patient care.

The Cohort Discovery tool also enables this to be done in a fast, secure, de-identified and ethical way through the continued use of TREs. For the organisations who host the datasets themselves ("data custodians"), Cohort Discovery allows them to provide access approvals much more quickly than before; but crucially to retain control to who has access to the data.

Cohort Discovery is being soft-launched in April 2021 across four core datasets, with inclusion of further datasets and developments through the year. To register for the HDR Innovation Gateway and see the new Cohort Discovery in action: https://www.healthdatagateway.org/pages/cohort-discovery-search-tool

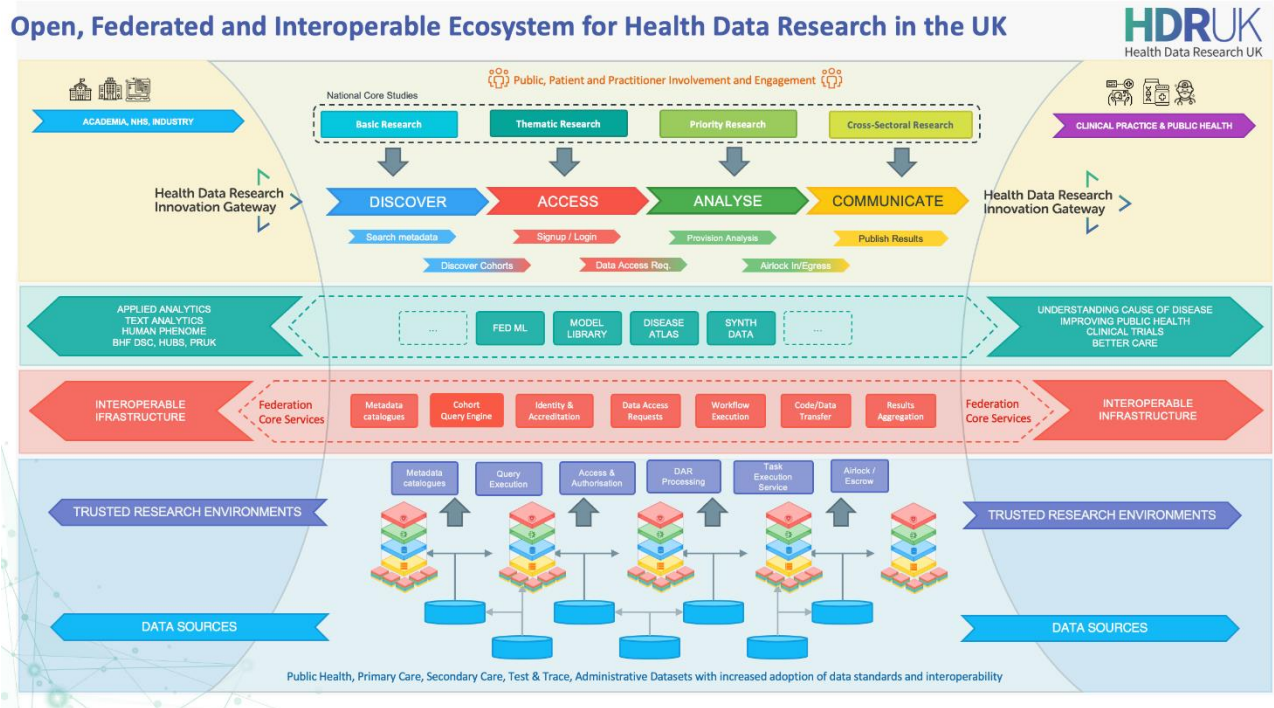
Code for Link Lite https://github.com/biobankinguk/link-lite

Investment to date: £126K (excludes investment through CO-CONNECT and Data & Connectivity)

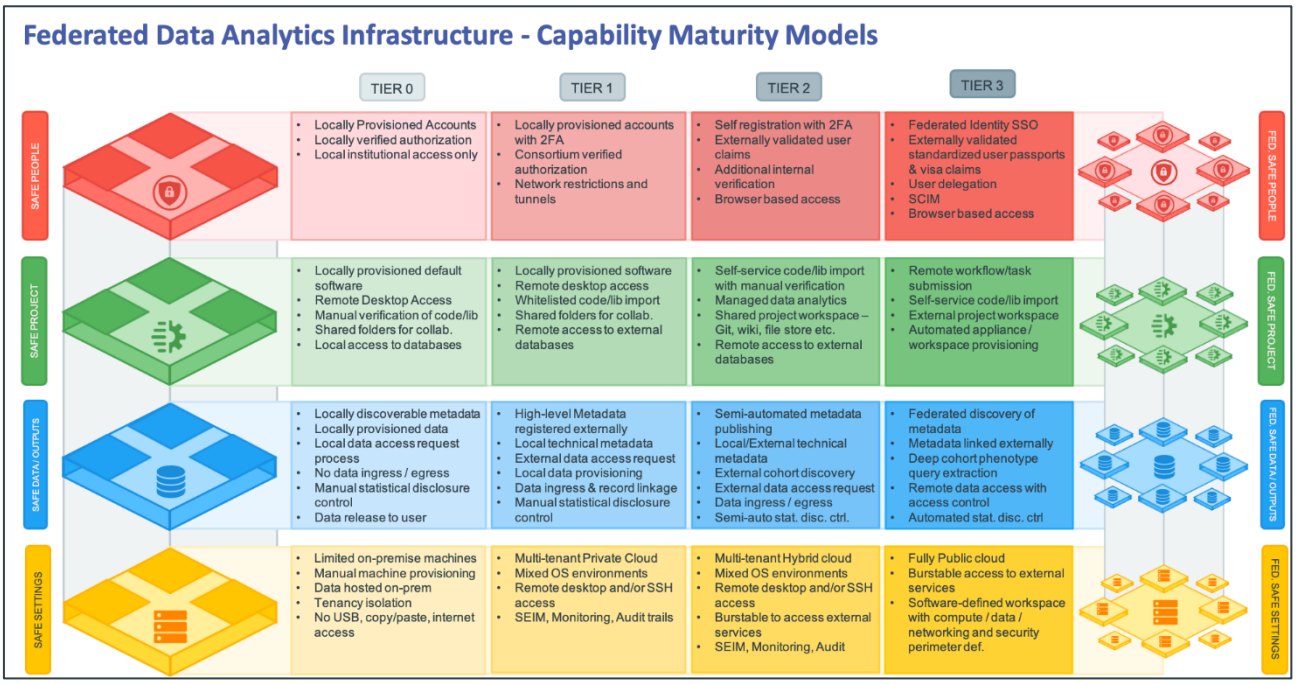
4. The Future

The pandemic has highlighted the need for an open, federated and interoperable network of TREs to support the acute need, but also the long-term vision of health data research at scale in the UK. This will empower research teams in the UK to make discoveries that improve people's lives.

The following figure provides an overview of the open, federated and interoperable health data research ecosystem developing within the UK:



Federation requires open standards, open code, open policies and shared processes. Existing TREs are at various levels of maturity:



Following the 2020 TRE Green Paper, UK Health Data Research Alliance is producing a TRE White Paper to support the federation of TREs, enabling cooperation between local, regional, national and even international TREs.

In addition to updating on progress against the six areas identified during the Green Paper consultation, the white paper will set out to:

1. Define the characteristics of a TRE node within a federated ecosystem and what the required infrastructure is to support integration.
2. Make stronger reference to the required work on data standards and highlight the work to date and priorities to be addressed.
3. Outline approaches for Safe (and cost effective) Archiving as part of Safe Outputs to ensure TREs do not limit the reproducibility of research.
4. Make recommendations on security by design
5. Address the exceptions to the TRE 'rules' and/or safe output requirements
6. Set out plans to develop supporting technical documentation