DARE UK

‘Virtual’ TREs

Service Design

Version	Author	Notes	Status
0.1	Anita Jena-Smol, Thomas Wileman & Nuno Rosa	Created 05/08/2022.	Draft
1.0	Anita Jena-Smol, Thomas Wileman & Nuno Rosa	24/08/2022	Stabilised

1. Introduction

This document describes the processes to build a technology platform that enables the use of a common framework or “service wrapper” to create on-demand Trusted Research Environments (TREs). All work streams were completed by a consortium of four partners: the Francis Crick Institute (lead), BT plc, the Institute of Cancer Research and the Rosalind Franklin Institute; and two technical partners: Infinite Lambda and Snowflake Inc. All teams contained subject-matter experts, who are focused on facilitating innovative scientific approaches, collaborations and methods for other people to work with. We believe our technology platform is a novel paradigm that has the potential to become an essential part of the technical ecosystem required to work effectively and securely with sensitive data.

We have produced a working technology platform, allowing configuration and deployment of on-demand TREs within 30 minutes. Like more typical ‘data-set-centric’ TRE architectures, our on-demand TREs are able to accommodate the needs and restrictions of individual research projects. However, our on-demand TREs also include novel features such as: i) experiment-level budget control, ii) audit and real-time management reporting, and iii) HIPAA and ISO27001 compliance. At the time of writing, our technology platform is live and being used by a research group at Francis Crick Institute; another three research consortiums are in the process of signing the Research Collaboration Agreements.

This document is supplemented by a Technical Annex which provides considerably greater detail for each section, with a view to making this work as reproducible as possible for any other interested party.

2. Stakeholders

There are a number of stakeholders whose concerns and interests need to be taken into consideration for the design of a robust common framework to create on-demand TREs for research on health data. Key stakeholders include, but are not limited to:

Stakeholder	Description
Patients	Individuals whose data will be the subject of research.
Data Custodians	Responsible for managing data access for the purpose of conducting research.
Researchers and Innovators	Individuals who require access to data in order to conduct research, and who would benefit from a more rapid access to data and improved opportunities for linking data that has until now been restricted due to the data custodians risk positions.
TRE Service Providers	Entity responsible for managing and maintaining a TRE, and keeping technical and governance systems that can protect privacy whilst providing a world class analytical experience.
Funders	Funders of research in the context of a TRE.

3. Analysis

3.1. Challenges & Opportunities

In total, 16 stakeholders were interviewed by the Business Analysis & Service Design Team at the Francis Crick Institute. Each stakeholder mapped to one of seven use cases:

NHS working with research

University working with peers

Enterprise working with healthcare providers

Research institute working with peers

Enterprise working with research

Small institute with (or without) infrastructure working with peers

Singleton lab without infrastructure working with peers

Interviewees were selected to represent as broad a set of potential users straddling multiple research councils, organisation types, industry and public healthcare to try and achieve as representative a set of challenges as possible. Feedback was collated into 12 'challenge areas':

3.1.1. Discovering Research Opportunities

As a researcher, it is difficult to find relevant, structured datasets that will help advance my research. The landscape of curated, publicly available datasets is large and complex; it is also only a tiny fraction of all research data. The norm for discovery research, given how tightly bounded and specific the research questions are, is to custom create datasets or work with known collaborators who have exactly the dataset needed to answer a specific question.

3.1.2. Setting up Research Collaboration & Data Sharing Agreements

As a researcher, in a post-SARS-CoV-2/COVID-19 pandemic world, there is much reduced tolerance for lengthy processes and bureaucratic delays when setting up Research Collaboration Agreements.

As a Legal, Governance & Compliance Team supporting research scientists, Research Collaboration Agreements are never one-size-fits-all. Unique documents because of custom created datasets and/or emerging technologies, these require input from multiple parties, so take time to get right.

3.1.3. Multi-party Collaborator Access

As a researcher, it is difficult to do truly collaborative research. This challenge is one of the fundamental administrative issues with managing a consortium, which is allocating and managing roles and responsibilities for handling, loading and transforming data. Typically, there is no pre-defined framework for these, often being dictated by whatever locally available infrastructure a consortium can negotiate to use.

3.1.4. Data Aggregation from Multiple Sources

As a researcher, it is difficult to aggregate data from local databases, public repositories and Trusted Research Environments. The nature of discovery research often means data is aggregated from many different sources (e.g. epigenomics, genomics, metabolomics, proteomics or transcriptomics). Some data may be sensitive, but a lot won't be. What is almost guaranteed is that a considerable degree of data manipulation will be needed to get a consolidated, fit for purpose final dataset.

3.1.5. Data Sovereignty

As a researcher, it is difficult to work with international datasets. Data sovereignty is the concept that data is subject to the laws of the country in which it is physically located. The legal rights of data subjects, and data protection requirements, depend on the location in which data is stored.

3.1.6. Working Safely with Patient Identifiable Data

As a researcher, it is difficult to work with Patient Identifiable Data (PID). Protection of a patient's right to confidentiality is paramount but restricts access to invaluable datasets for research - especially unstructured data. The lack of accepted platforms to work with PID places the burden of risk management almost entirely with a Principal Investigator (PI). Therefore, in addition to an excellent understanding of the ethics of research with PID, a PI must also have a good technical understanding of international data protection law, be able to guide legal professionals in the creation of robust agreements and understand the minutiae of information technology infrastructure to ensure that they are discharging their accountabilities effectively.

3.1.7. NHS Data Security

As a researcher, it is difficult to access data collected by NHS Trusts. PID and NHS Data Security are of course inextricably interlinked. The specific NHS Data Security challenge is the federated operating model of the NHS Trusts. Here, how access is requested, what constitutes acceptable release of data, and how it will be managed vary from organisation to organisation.

3.1.8. Securing Intellectual Property

As a Legal, Governance & Compliance Team supporting research scientists, negotiating Intellectual Property (IP) is never one-size-fits-all. Intellectual property refers to creations of the mind. IP is protected in law by copyright, patents and trademarks, enabling people to earn recognition for their inventions. This is especially acute with academic/industry partnerships, where the latter parties operate in a highly secure 'closed' fashion with rigorous in-house controls around IP, whereas the default position of the former is to work openly, in line with commitments to open science. This can lead to tensions in forming workable partnerships.

3.1.9. Data Loading & Preparation

As a researcher, the loading and off-loading of large or sensitive datasets into or out of collaborative spaces is complicated and time consuming. Often, a significant proportion of research time is consumed by the preparation of data, rather than its analysis. Especially, if the research project involves disparate datasets from multiple sources. In the absence of common shared environments, the 'where' we collaborate is dictated by whatever locally available infrastructure a consortium can negotiate to use and based on the rate and volume of data generation, rather than what is most robust. This often leads to huge amounts of experimental time being wasted working around the limitations of what is available, either fault-fixing or adapting sub-optimal solutions, rather than focussing on answering scientific questions.

3.1.10. Data Processing & Analysis – Running Code or Pipelines

As a researcher, it is rarely possible to analyse data all in one place. Typically, analyses are done using code pipelines, either in frameworks such as Nextflow or using custom code typically written in Python. Total flexibility to import applications, code and tools into a trusted research environment is paramount. Many popular research applications can only run on Linux, some only on Windows virtual machines. Some applications are well-suited to batch, HPC-style processing, whereas others require an interactive client-model. Often, the research environment is limiting the scope of research analyses.

3.1.11. Tracking and Auditing Compliance with Standards – HIPAA & ISO27001

As a researcher, it is difficult to know if we are able to meet audit and compliance standards. Often, the 'where' we collaborate is dictated by whatever locally available infrastructure a consortium can negotiate to use – rather than explicitly designed or managed solutions. Therefore, proving upfront compliance with data protection and security standards is extremely difficult. This causes lengthy delays to project start dates. In some cases, audit and compliance has proven an insurmountable barrier to collaboration resulting in researchers working in deliberate siloes and only sharing completed analyses rather than any combination of raw data.

3.1.12. Financial Controls – Managing Computing Resources Fairly and to Project Budget

As a researcher supported by grant-funding, financial controls and managing computing resources fairly and within project budget is an especially acute issue. The risk management implications of using cloud native solutions, where billing and cost control is entirely the responsibility of the end user, are prohibitive in environments where funding is grant limited. There is a clear need to be able to distribute resources across all

parties in a consortium according to the funding they bring. Equally, there is a clear need for cost and spend caps to be in place to prevent accidental but very financially damaging bills.

3.2. Requirements

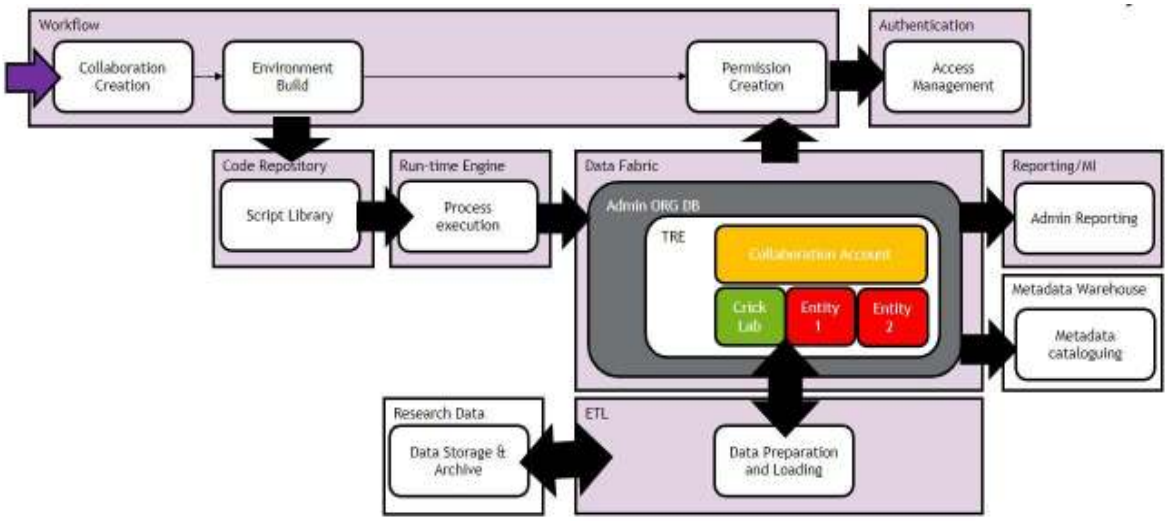
To successfully implement a technology platform that enables the use of a common framework to create on-demand TREs and address the challenges identified above, the Information Technology Office at the Francis Crick Institute propose that any solution should address the following high-level requirements:

1. Allow project-specific infrastructure configurations, but in a common administration environment
2. Enable a common set of roles and responsibilities, allowing for a standard approach to legal terms and conditions regardless of the size or complexity of the collaboration
3. Have native security controls and certifications to meet global data standards, from HIPAA in the US to GDPR.
4. Allow global deployment of infrastructure to meet international data sovereignty requirements
5. Allow access to elastic compute and data resources to cover any conceivable workload and data type, whilst also have sophisticated billing restrictions and controls, recognising that funding for collaborative research is frequently via a mosaic of different grants, each with complex restrictions on spend evidence and no contingency for overspend
6. Have tooling to allow for the import of many different data sets, and allow for the installation and running of many different analysis packages, as well as custom scripting and software
7. Be able to collect metadata to provide detailed audit and compliance reporting and monitoring capability

4. Design

4.1. Capabilities

The diagram below shows the capabilities of the solution based on the end to end journey of a research collaboration - from Inception of the collaboration through to Archive.



The **Workflow** capability is used to capture all key information input by the user. This information is fed through to the **Data Fabric** capability where the trusted research environment is built. The tool used by the Crick to enable the Data Fabric capability was the Snowflake platform.

The **Authentication** capability allows to create users, assign roles and ensure secure access to the solution based on user role. The Authentication capability ensures every user has the appropriate level of access to fulfil their role within the collaboration. User roles have been setup at the lowest level of granularity to reflect the diverse nature of any research collaboration. One individual can hold multiple roles. A detailed user role matrix can be found in the appendix at the end of this document.

A *sub-account* is a term used by Snowflake to describe a virtual space associated with a group of individuals, in which work takes place. Different parties within a research collaboration will each have their own entity sub-account that links directly to the Collaboration Sub-account. *Users* will belong to an individual entity sub-account, with varying levels of access (role dependent) to the single Collaboration Sub-account.

Code Repository and **Run-time Engine** capabilities process the information from the Workflow capability to feed through to the Data Fabric (i.e. Snowflake). Snowflake sits at the heart of the TRE and is where data research happens.

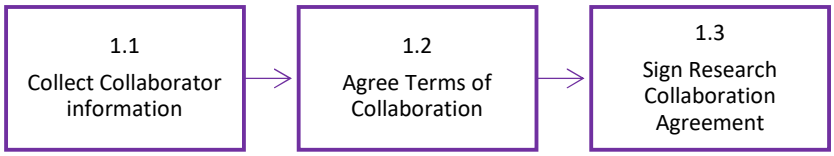
Reporting and **Meta-data Warehousing** capabilities create the management reporting layer.

ETL (Extract, Transform, Load) capabilities enable the loading and unloading of research data to and from the Snowflake Data Fabric.

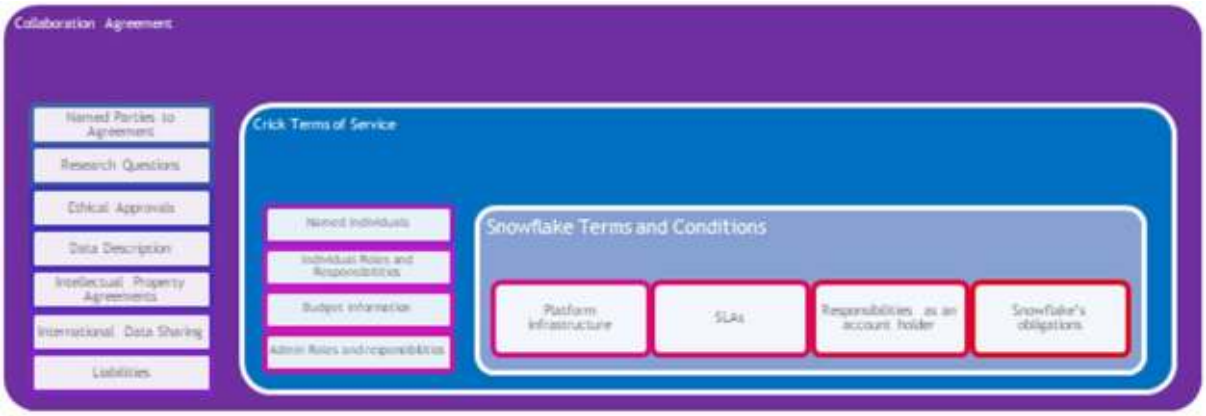
5. Business Process

The end-to-end process journey is built around the following key process stages:

5.1.1.1. Stage 1 – Create Research Collaboration Agreement



The Research **Collaboration Agreement** is arranged between members of the consortium, led by the Principal Investigator. The details agreed form the top layer of the legal agreement:



Agreement	Description
Research Collaboration Agreement	A 'Brunswick' style agreement proforma is adopted which incorporates the data sharing arrangements, and any data transfer arrangements where access across sovereign borders is unavoidable
Crick Terms of Service	Captures information about the collaboration, and sets out the obligations and roles of every party, including the Crick's role as overall administrator of the platform.
Snowflake Terms & Conditions	The Snowflake platform terms and conditions, derived from the Snowflake Master Services Agreement

Information from the Research Collaboration Agreement becomes the input data captured in the workflow tool described below. It is re-purposed to generate an agreement template for the Trusted Research Environment.

The journey starts with capturing the details of organisations and individuals playing a role in either managing or working within the environment. Details are captured from the Research Collaboration Agreement. Individuals are assigned roles and legal entities recognisable by their organisation email domain.

--- For illustrative purposes only ---

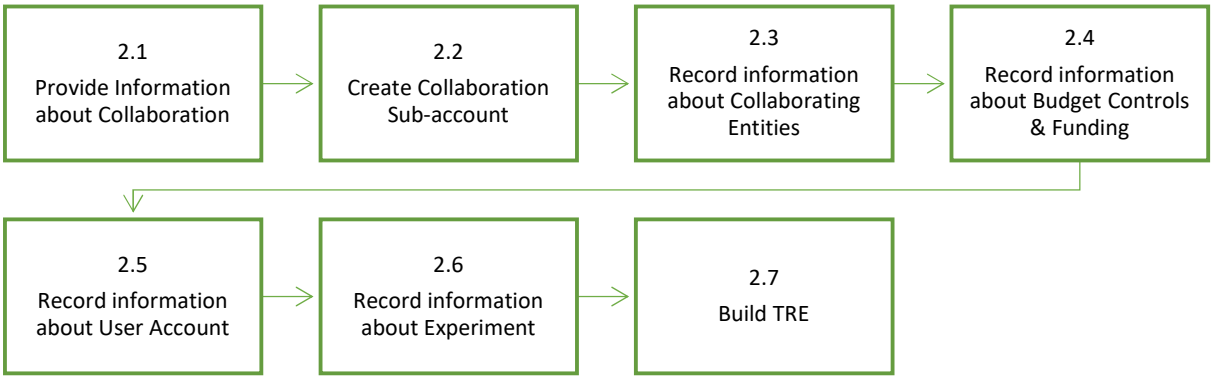
Our consortium, VISTA, involves three groups hosted at two organisations. The Principal Investigator is based at a London NHS Trust and also runs a discovery research group at the Francis Crick Institute.

The Research Collaboration Agreement has been agreed. Patient data collected by the NHS Trust is to be anonymised and shared with the Francis Crick Institute (hereafter, referred to as the 'metadata').

The NHS Trust also sends samples to the Francis Crick Institute's inhouse next-generation sequencing facility (colloquially known as a "Science Technology Platform"). Once the samples are sequenced, the genome

sequencing data and the associated metadata will be curated by a bioinformatician and analysed by a PhD Student in the discovery research group at the Francis Crick Institute.

5.1.1.2. Stage 2 – Create a New Trusted Research Environment



5.1.1.3. Provide information about the Collaboration

Details about the collaboration, also known as consortium, are entered into the workflow management tool (at the Crick, this is ServiceNow) by the Accountable Person (at the Crick, this is a Group Leader).

The following details will be captured:

Field name	In Brief	Business Rules
Working Title	A brief title or acronym	
Collaboration Description	Longer description of the collaboration, including research objectives	
Start Date	Actual start date for the collaboration	
Approximate End Date	Estimated or target end date for the research collaboration.	Captured for information only at this stage
Will Sensitive Data Be Used?	Sensitive data being used on the environment in any part of the collaboration?	A. If "YES", select "Business Critical" Account B. If "NO", select "Enterprise" Account
Will US HIPAA Data Be Used?	Are US HIPPA standards required to protect sensitive data within the environment?	A. If "YES", select "Business Critical" B. If "NO", select "Enterprise" Account
Funding Expected	Description of funding	Entries are: Core-; Grant-funded or Mixed
Comments on Funding		
Attach Collaboration Documents	Space to attach any relevant working documents about the collaboration (e.g. Research Collaboration Agreement)	

Where ***Sensitive*** or ***US HIPPA*** data has been selected, the solution will use this information to auto-select a Snowflake 'Business Critical' environment, otherwise the default is the standard 'Enterprise' environment. Note, data cannot be shared between a Business Critical (higher security) and Enterprise (lower security) environments.

5.1.1.4. *Create Collaboration Sub-account*

The request to create a consortium (collaboration) sub-account is informally 'approved' by an administrator (at the Crick this the Snowflake administrator role sits within the IT Directorate). Once approved, the details captured in the workflow tool are collated and sent to Snowflake. This triggers an event to create the consortium 'sub-account' in Snowflake.

5.1.1.5. *Record information about the Collaborating Entities*

In the workflow tool (at the Crick, this is ServiceNow), details for consortium members are entered by the Accountable person. This will take place after the consortium sub-account is created in Snowflake, so that consortium members can be linked via a unique Consortium ID.

For the VISTA consortium, three organisation-level sub-accounts will be created: one Crick Lab sub-account; one Crick STP sub-account; and one NHS Trust sub-account. All three organisation-level sub-accounts will be linked to a single Collaboration sub-account that is given a unique Consortium ID.

Other details captured at this stage include entity name, physical region, and use of sensitive data on the environment. The email domain or suffix is also recorded here; this will be used to validate individual accounts with the same email domain/suffix should be linked to the correct org entity.

Field name	In Brief	Business Rules
Consortium ID	System generated identifier to link members to the consortium	
Working Title	Brief title or nickname for the consortium	Consortium title auto-populates when the Consortium is selected
Crick Entities - Labs (internal)	Names of individual Crick labs or teams working on the project	Drop-down list is auto-populated from a directory of lab groups at the Crick. Up to six entries can be added.
Will Sensitive Data Be Used?	Will this participant bring any sensitive data to the consortium?	Drop-down entries are: HIPPA; ISO27001; Both; None A. If 'HIPPA', ISO27001' or 'Both' are selected, select "Business Critical" Account B. If "None", select standard "Enterprise" Account
Crick Entities - Non-Labs (Internal)		Drop-down list auto-populated from a directory of non-lab groups e.g. science support teams at the Crick. Up to six entries can be added
Will Sensitive Data Be Used?	Will this participant bring any sensitive data to the consortium?	Drop-down entries are: HIPPA; ISO27001; Both; None A. If 'HIPPA', ISO27001' or 'Both' are selected, select "Business Critical" Account B. If "None", select standard "Enterprise" Account
Non-Crick entities (External) - Name	Details of any external entities associated with the consortium	Free-text description of the external participant. Up to four external entries can be added.

Email suffix/domain	The email domain identifiable to this organisation	Email domains captured here will be used to associate individual user accounts with a consortium entity
Region	Enter the region from where data will originate	Organisations choose where data is geographically stored, supported regions are grouped into three global geographic segments: North/South America, Europe/Middle East, and Asia Pacific. Defining the region also determines where compute resources are provisioned. Regions do not limit user access; regions only dictate the geographic location where data is stored and compute resources are provisioned.
Will Sensitive Data Be Used?	Will this participant bring any sensitive data to the consortium?	Drop-down entries are: HIPPA; ISO27001; Both; None A. If 'HIPPA', ISO27001' or 'Both' are selected, select "Business Critical" Account B. If "None", select standard "Enterprise" Account

At the Francis Crick Institute, we chose to enable up to six internal labs, six internal teams and up to four external collaborating organisations to be added to the record. These numbers were selected based on the typical number of participants in a Francis Crick Institute research collaboration. Additional participants can be added to the consortium by creating another entry record.

When all the key information is added, the workflow tool passes the details to Snowflake to create new entity sub-accounts for each individual organisation.

Further participants can be added at any stage of the collaboration.

5.1.1.6. Record information about the Budget Controls & Funding

For each sub-account, the Accountable Person now inputs budget control details to the workflow tool (at Crick, this is ServiceNow). Details include a budget or grant codes that are identifiable back to the entity org, budget period and cost per code. These details are then used to calculate the total cost per budget control, budget percentage splits between entity sub account and collaboration subaccount, budget allocation and splits.

Field name	In Brief	Business Rules
Consortium ID	System generated identifier to link members to the consortium	
Sub Account	Select entity sub-account	System auto-populates a list of Consortium entities (entity sub-accounts) linked by Consortium ID
Notify Percent	Enter the percentage of budget spend point where Accountable person should be notified	When selected % total budget has been reached, notify the Accountable person
Project Code	Enter a project (or budget) code identifiable to the consortium, from where budget is allocated	
Grant Code	Enter a grant code identifiable to the consortium, from where budget is allocated	
Overall Budget	Enter the overall budget amount available to the consortium	This is a numerical value used for calculations later on
Contract Period	Enter the total period of the contract (the period over which the budget is allocated)	This is a numerical value used for calculations later on
Budget Control	Enter the unit of time	Dropdown entries: Days / Weeks / Months / Years
Total Cost Per Budget Control	Calculates the total cost to the consortium by budget control time period	Auto-calculate: Overall budget divided by budget control

These details reside in the budgets table linked to the entity subaccount, and are used to allocate spend to individual user roles within the entity or collaboration sub accounts.

The VISTA consortium has a budget of £30,000 to spend over a 12-month period. The Overall Budget (£30,000) is divided by the Contract Period (12) to generate a £2,500 budget per month.

Billing against each cost centre is tightly controlled. The Accountable Person can specify precise budget allocations for both the loading accounts, which are private to each entity within the consortium, and the collaboration account, which is open to all consortium members.

Field name	In Brief	Business Rules
Percentage for entity account	Select what percentage of the budget control cost is allocated to the consortium entity	Auto-calculates budget control as a percentage of total budget control cost
Percentage for the collaboration account	Select what percentage of the budget control cost is allocated to the consortium itself	Auto-calculates budget control as a percentage of total budget control cost
Recommended Profiles for your account	Select the spend profile allocation by research activity - at COLLABORATION PARTICIPANT level	Dropdown entries: Loading 50%, Processing 25% Sharing 25% Loading 25%, Processing 50% Sharing 25% Loading 25%, Processing 25% Sharing 50%
Recommended Profiles for your collaboration	Spend profile allocation by data curator / experimenter role - at COLLABORATION entity only	Dropdown entries: Curator 100% Experiment 30%, Curator 70% Experiment 50%, Curator 50% Experiment 70%, Curator 30% Experimenter 100%

The VISTA consortium decides that the £2,500 per month budget should be split 80:20 between the Collaboration sub-account (£2,000) and the three consortium organisation's sub-accounts (£500).

The Accountable person then selects the spending profile for the budget across the consortium by research activity. At the Crick we have selected this based on data loading, processing and sharing.

The VISTA consortium intends to load a relatively small amount of data to the Trusted Research Environment but conduct high-compute, multi-stage pipeline analyses. Data loading into the TRE only needs to be done by two organisations: 1. The Crick STP and 2. The NHS Trust.

The Principal Investigator (our Accountable Person) splits the budget (£500) allocated to the organisation sub-accounts equally between the Crick STP (£250) and NHS Trust sub-accounts (£250). The spend profile for the Crick STP and NHS Trust sub-accounts is then set to: Data Loading: 50% (£125); Data Processing: 0% and Data Sharing: 50% (£125). Note, no processing will be possible in the organisation sub-accounts.

Finally, the Accountable person allocates the spending profile at collaboration level between curator and experimenter. The data curator role combines datasets loaded from the consortium entities into science data sets ready for use by the experimenters. The experimenter role performs analysis on these combined science data sets, using analysis tools such as RStudio.

The Crick chose to allocate budget on a variable scale from 100% curator to 100% experimenter.

The Principal Investigator (our Accountable Person) then sets the spend profile for the budget (£2,000) allocated to the collaboration sub-account. Here, we opt for a spend profile split:
Data Curation: 10% (£200); and Experimentation:90% (£1,800).

The budget allocation process is repeated for each consortium entity, whose budget can be allocated in full or in part between the consortium and the consortium entity. Budget details are held in the workflow tool for later use.

5.1.1.7. Record information about the User Account

Aneeka Sharma is the Principal Investigator for the consortium.
Edmund Yip is the Data Manager co-ordinating the patient data at the London NHS Trust.
Steven Logan is the Senior Staff Scientist at the Francis Crick Institute's Science Technology Platform.
Deborah Maloney is the Senior Bioinformatician working in the discovery research group at the Francis Crick Institute
Erik Hanna is the computational PhD Student working in the discovery research group at the Francis Crick Institute.

The accountable person now uses the workflow tool (at the Crick, this is ServiceNow) to input user details for each consortium organisation. Individual users are identified to their organisation entity sub-account through their email domain. This is both to identify the individual to its own consortium, but also crucially for security purposes i.e. to avoid the use of public email domains e.g. @gmail.com which are typically more prone to cyber-attacks than institute email accounts.

The user's credentials will be authenticated - at the Crick, the user is sent a confirmation email which must be validated before the user account is created.

A single user can be set up with one or more roles - this is done by adding additional user records per each role. This can be added by modifying a user profile at a later stage.

Sub-account	*Role*	*Name*
Crick Lab	*Accountable Person*	*Aneeka Sharma*
Crick STP	*Data Loader*	*Steven Logan*
Crick STP	*Data Sharer*	*Steven Logan*
NHS Trust	*Data Loader*	*Edmund Yip*
NHS Trust	*Data Sharer*	*Edmund Yip*
Collaboration	*Board Member*	*Steven Logan*
Collaboration	*Board Member*	*Deborah Maloney*
Collaboration	*Data Curator*	*Deborah Maloney*
Collaboration	*Experimenter*	*Deborah Maloney*
Collaboration	*Experimenter*	*Erik Hanna*

Field name	In Brief	Business Rules
Crick Employee?	Select if user is an internal worker	System auto-populates internal user list from organisations' user directory
Consortium ID	System generated identifier to link members to the consortium	
Sub Account	Select entity sub-account the user is affiliated to	System auto-populates a list of Consortium entities (entity sub-accounts) linked by Consortium ID
Email Address	Enter individual's email address	email domain/suffix is validated against the consortium participant's org entity email domain/suffix
First Name	Enter individual's first name	
Last Name	Enter individual's last name	
Role	Enter the user's role within the entity sub account	Dropdown entries: Accountable Person/Data Loader/Data Processor/Data Sharer etc - see appendix xxx for full list
Okta Enabled	Select if user's host organisation can authenticate via the Crick's identity management solution	If selected, Crick's identity management solution [Okta] links to external identify management solution [check this]

The process is repeated for each individual user at the Consortium level by the Consortium accountable person, and again for each individual user at the Consortium entity level by the Consortium entity Accountable person.

5.1.1.8. Record Information about the Experiment

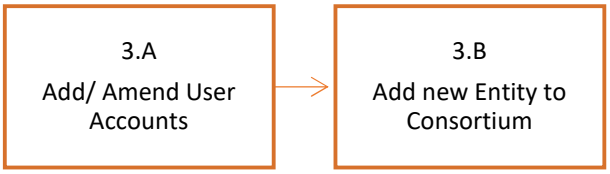
The accountable person now uses the workflow tool (at the Crick, this is ServiceNow) to input experiment details. This detail is used to ensure research activity is correctly allocated to the right budget codes. Once the user selects consortium, entity and budget descriptors, additional detail about the experiment is captured, including experiment name, description, estimated start/end dates, and budget amount. There is an option for the user to be notified if spend is close to intended budget cap. Experiment details are held in the workflow tool for later use.

Field name	In Brief	Business Rules
Crick Employee?	Select if user is an internal worker	System auto-populates internal user list from organisations' user directory
Consortium ID	System generated identifier to link members to the consortium	
Sub Account	Select entity sub-account the user is affiliated to	System auto-populates a list of Consortium participants (entity sub-accounts) linked by Consortium ID
Experiment Name	Enter short description of the experiment	
Experiment Description	Enter long description of the experiment	
Start Date	Enter start date of the experiment	
End Date	Enter end date of the experiment	
Notify Percent	Enter the percentage resource usage when user should be notified	Resource monitor will send notification (at Crick this will be notify Accountable Person) by email when this percentage resource use is reached

5.1.1.9. Build Trusted Research Environment

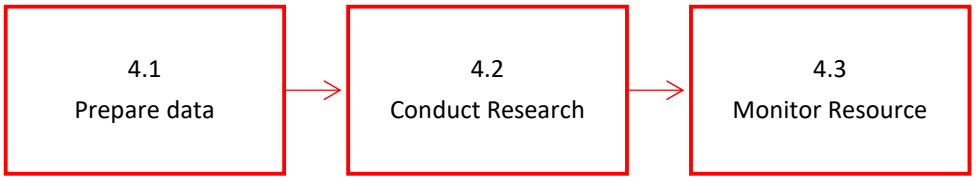
Consortium entity sub-accounts, users, budget and experiment details are now validated and approved in the workflow tool (at the Crick this is done by the Research Data Services lead). The approval step triggers the output of data passed through to Snowflake to generate entities, users, budget and activity.

5.1.1.10. Stage 3 – Edit Existing Trusted Research Environment



Consortium members are able to modify users, budgets and experiment detail throughout the research collaboration, according to their user role. Where dependencies on other entities or the collaboration exist, notification alerts are sent appropriately.

5.1.1.11. Stage 4 – Research



5.2.4.1 Prepare data

5.1.1.12. Load data

The research environment is now setup and ready to use.

Users assigned with the 'Data Loader' role can use Snowflake compatible data loading tools to load research data. This could be command line (Snowflake has its own client SnowSQL) or a number of GUI tools (at the Crick we adopted DBeaver), to establish a connection to the Snowflake environment and start loading data from the client machine or other data source.

Data is loaded to data tables on the user's entity sub-account space. The Data Loader role has privileges for off-loading data at the end of the experiment or project.

Data load activity can be ad hoc using one of the above tools, or a continuous/scheduled activity using workflow management tools and software deployment tools (at the Crick we used Apache Airflow and Kubernetes).

Consortium entity sub-accounts will always have:

A database called <<Account Name>>_DB.

This is where the loaded data resides in tables and where the processed data would also sit. This is the database that would have shares made from it to other accounts.

A database called ACCNT_ADMIN_DB.

This is an administrative database and not accessible by any of the users of the account. It is used by system resources to maintain a full copy of all the native snowflake metadata on usage and object information. This is the base source of the data that feeds the master Organisation wide equivalent database that supports Reporting.

5.1.1.13. Process data

Users can process data in the entity sub-account. To process any data in the TRE a compute facility is needed. Each role has its own. This allows to individually monitor and identify costs incurred per user.

5.1.1.14. Share Data

For every share made to the Collaboration an IN_SHARE database is required. This IN_SHARE database will be used by the Data Curator to provide access to the shared data. It acts a little like a transparent pipe enabling the Collaboration sub-account to see the shared data. It is not a copy of the data, the master of which remains in the originating entity sub-account.

The naming convention for IN_SHARE databases is:

<<ENTITY_ACCOUNT_NAME>><<SHARE_NAME>>

These IN_SHARE databases are not accessible to the Experimenter role, only the Data Curator.

There is then a further two databases:

A science database called <<ACCOUNT_NAME>>_SCIENCE_DB.

This is the database the

Experimenter role uses to develop the results of the collaboration. This database has specific schema designs, with a separate schema for each approved experiment.

ACCNT_ADMIN_DB.

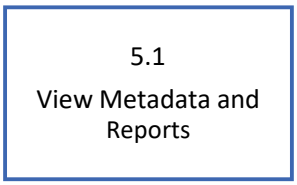
Performs the identical function as for Entity accounts, bringing together a full historical record of what was done in the Collaboration sub-account. Again, it is not accessible by people, and only used by the system.

Again, there are compute "warehouses" for each of the two roles:

DATA CURATOR: <<ACCOUNT_NAME>>_DC_XS_WH

EXPERIMENTER: <<ACCOUNT_NAME>>_EXP_XS_WH

5.1.1.15. Stage 5 – Reporting



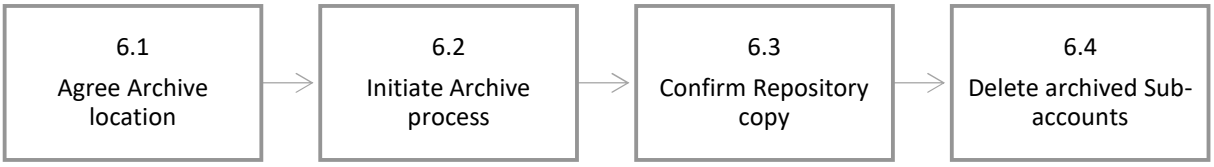
Research activity can be monitored from as soon as the environment is set up, to the point of close down and archiving. The reporting interface calls metadata within the workflow tool or within Snowflake itself.

At the Crick we used Microsoft Power BI to create a management information reporting dashboard. The dashboard includes various technical, financial, audit and compliance reports. The Crick opted to report on the following:

Field name	In Brief	Searchable by
Users	List of Users displayed by Entity, Role, Database and Compute	Individual email address
Collaborations	Displays User details across the collaboration, role, whether active user, Experiment and Share access, per each Entity	Entity (Account) name, Role
Access Failures	System level access over including successful/failed logins to the environment and to individual clients, displayed by user and over time	Entity (Account) name
Budgets	Overview of agreed budgets captured at consortium setup stage. Includes budget/grant code details and percentage allocations of research and resource activity	Entity (Account) name
Resource Monitoring	Resource usage in credits per entity, total quotas, credits used/remaining, broken down by role	Entity (Account) name
Spending	Spend profile in currency by resource service level (Snowflake Business, Snowflake Enterprise) and by usage type (e.g. compute, cloud services, data transfer)	Entity (Account) name
Spending Profile	Visualisation of Spend over time by entity	Entity (Account) name, Usage Type
Credit Usage Profile	Visualisation of resource credits billed, and storage used, over time stacked by Entity	Entity (Account) name
Table Size	Visualisation of data table sizes over time - system or science data (tbc)	Entity (Account) name
Replication Databases	Overview of replication activity per Consortium environment	Entity (Account) name

Row-level security is used to ensure users only see reporting data corresponding to their entity, unless the user has wider consortium level access.

5.1.1.16. Stage 6 – Archive & TRE Closure



At the time of writing, the archiving stage of the TRE is being built, and a summary of our intended design is described below.

Individual consortium members are responsible for archiving data loaded and processed within their own entity. Archive details are entered via the workflow tool. Input data includes a selection of objects, archive start date and details about the method and user instigating the archive. The user will be prompted to actively confirm that no metadata will be archived.

Field name	In Brief	Business Rules
Consortium ID	System generated identifier to link members to the consortium	
Sub Account	Select entity sub-account	System auto-populates a list of Consortium participants (entity sub-accounts) linked by Consortium ID
Select Objects	Select objects to archive	Dropdown list displaying objects relevant to consortium entity
Metadata	Active confirmation that no metadata will be archived	Entries are: Yes / No
Archive start date	Enter date point from which archiving should start	
Archive Method	Select archive method	Entries are: Manual / Automatic
Archived By	Select user operating archive process	If 'Archive Method' selected is 'Manual', user running the archive is selected from dropdown list
Archive Destination - filepath	Enter full path to location this data will be archived to as a fileset (.zip/.tar etc)	
Archive Destination - username	Enter username details	
Archive Destination - password	Enter password details	

On the day of the appointed archive date the following steps are followed:

- Calculate the size of the [objects selected in the archive workflow]
- Wait for any active job in Collaboration / Entity account to finish which started before midnight on ARCHIVE DATE.
- Revoke Roles from Users as of midnight on ARCHIVE DATE (to stop any new tasks being created).
- Deactivate Computes (using resource monitor) in Collaboration / Entity Account (to prevent further build of new costs / data objects).
- Disable all automated task (Snowpipe, Tasks, Metadata collection, etc) to stop further spend on active resources.
- Create ARCHIVE Role (do this now in this process rather than on subaccount creation so no confusions).
- Create ARCHIVE compute and ARCHIVE Resource Monitor (so costs are separate). Associate Archive Role to Archive Compute
- Assign Archive role to a user (or service account if automated).
- Pass JSON file to tool enacting the transfer (e.g. ELT). Or initiate the Snowpipe or PUT of-snowflake process

Upon completion of all archive processes, across all entities, Snowflake sends a request to delete the Collaboration account. Access is disabled instantly; actual deletion of the environment may take some weeks after it's disabled.

6. Further possible exploration

For the areas listed below design and build are currently in progress so are not documented here:

- Legal agreement template build from consortium data
- Tagging data (native to Snowflake)
- Archiving data back to owning consortium member
- Data analysis/pipeline tools

7. Glossary

Term	Description
Consortium	The term used for a collaboration. One consortium has many consortium members.
Consortium Member	An individual organization or team that collaborates with other members within a single consortium. Also referred to as consortium participants, consortium entity or consortium sub-account.
TRE	Trusted Research Environment: a single, secure platform in which data can be loaded, collated, analysed and processed by different research partners within a single collaboration

8. Appendix

8.1. Role Matrix – TRE Platform

						Snowflake Platform							
Role	Sub-account	Human/System	Crick	See Metadata & Reports [Power BI]	Approve Changes to the TRE [ServiceNow]	Load Data [Snowflake]	Read [Snowflake]	Write [Snowflake]	Create [Snowflake]	Update [Snowflake]	Delete [Snowflake]	Share Data [Snowflake]	Able to Incur Costs [Snowflake]
Accountable Person	Entity	Human	Group Leader	✓	✓	x	✓	x	x	x	x	x	x
Data Loader	Entity	Human	PhD / Postdoc	x	x	✓	x	✓	✓	x	x	x	✓
Data Processor	Entity	Human	PhD / Postdoc	x	x	x	✓	✓	✓	✓	✓	x	✓
Data Sharer	Entity	Human	PhD / Postdoc	x	x	x	x	x	x	x	x	✓	✓
Board Member	Collaboration	Human	Group Leader / Postdoc	✓	✓	x	✓	x	x	x	x	x	x
Data Curator	Collaboration	Human	PhD / Postdoc	x	x	x	✓	✓	✓	✓	✓	x	✓
Experimenter	Collaboration	Human	PhD / Postdoc	x	x	x	✓	✓	✓	✓	✓	x	✓
Metadata Curator	Entity & Collaboration	System	System Role	x	x	x	x	x	x	x	x	✓	x
Organisation Data Curator	Entity & Collaboration	System	System Role	x	x	x	x	x	x	x	x	x	x
Account Reporter	Entity & Collaboration	System	System Role	x	x	x	x	x	x	x	x	x	x

8.2. Role Matrix – Science and Reporting Database

Role	Description
Organisation Administrator	Role manages operations at the organisation-level. Including: • Can create sub-accounts in the organisation account. • Can view all sub-accounts in the organisation account and regions enabled. • Can view usage information across the organisation account.
Account Administrator	Role encapsulating the system-defined roles, SYSADMIN and SECURITYADMIN. This is the top-level role in the system and should be granted only to a controlled number of users.
System Administrator	The system-defined role, SYSADMIN. Role has privileges to create warehouses and databases in an account. --- If, as recommended, you create a role hierarchy that ultimately assigns all custom roles to the SYSADMIN role, this role also has the ability to grant privileges on warehouses, databases, and other objects to other roles.
Security Administrator	The system-defined role, SECURITYADMIN. Role has privileges to manage any object globally, as well as create, monitor, and manage roles and users. More specifically, this role: • Is granted the MANAGE GRANTS security privilege to be able to modify any grant, including revoking it. • Inherits the privileges of the USERADMIN role via the system role hierarchy (i.e. the USERADMIN role is granted to SECURITYADMIN).
Role Administrator	The system-defined role, USERADMIN. Role dedicated to role and user management only. More specifically, this role: • Is granted the CREATE USER and CREATE ROLE security privileges. • Can create users and roles in the account.