Glossary and terms

Please organise alphabetically!

Access Control

The technical mechanism for controlling a known (Authenticated) user's access to the system. Sometimes referred to as Authorisation (and shorthanded as "Auth" often in concert with Authentication), as it determines what a user is authorised to do. Hutch and Bitfount use "Role Based Access Control" (RBAC) so a suitable administrator (e.g. of a project, or a TRE) can configure that certain Users have certain Roles, and those Roles have Permissions to perform authorised actions. In Hutch, Access Control checks for a user will take place after they are authenticated, by both the Submissions Layer outside a TRE and the Polling Layer inside a TRE, to ensure they are allowed to do what they are asking.

Commented [1]: I think this is excess information for a glossary definition

API

An abbreviation for Application Programming Interface, an API is a type of software interface that provides a way for two or more computer programs to communicate with each other. In contrast to a user interface, which connects a computer to a person, an application programming interface connects computers or pieces of software to each other.

Application Catalogue

A centralised inventory or repository that lists and describes all the applications or software systems used within an organisation, including their functionalities.

Application Deployment

The process of installing, configuring, and making software applications available for use within the TRE.

Application Stack

An application stack is a number of applications or tools that work in concert to form a complete software solution. Hutch is an application stack consisting of several components, some of which are developed as part of Hutch itself, and others are pre-existing and used to complete the whole solution.

Commented [2]: Would also argue this is excess for a glossary, not needed to understand the underlying concept

Asset Management Process

A systematic approach to acquiring, operating, maintaining, and disposing of assets within an organisation, aimed at maximising their value and minimising risks.

Authentication

Authentication is finding out who a user is and ensuring that they are who they say (i.e. they are authentic) via some acceptable proof. This goes hand in hand with Access Control (or “Authorisation”) and the two are both sometimes shorthanded (separately or together) as “Auth”. If there’s any doubt in the context of what “Auth” is referring to, it should be clarified!

Authentication Application

A software system that verifies and validates the identities of users or entities accessing a system through multifactor authentication.

Authentication Token

An authentication token is a piece of data that is used to authenticate the identity of a user or application. It is typically a string of characters or a digital certificate that is generated by an authentication server, and is then passed between the user or application and the server to verify their identity. Authentication tokens are commonly used in web applications, APIs, and other systems that require secure access control. When a user logs in to a system, the authentication server generates a token that is associated with the user's account and privileges. This token is then passed back to the user's browser or application, and is used to authenticate subsequent requests to the system.

Authentication tokens can be generated using a variety of methods, such as cryptographic keys, digital certificates, or session IDs. They can also be time-limited or have other restrictions to enhance security and prevent unauthorised access. The use of authentication tokens helps to ensure that only authorised users or applications can access sensitive information or resources, and can provide an additional layer of security beyond traditional username/password authentication.

Automated Disclosure Control

Disclosure control without the intervention of a human being each time. Configuring a software system to reliably detect that data it is provided with is “non-disclosive” (i.e. doesn’t disclose any information that is not allowed to be shared with the requesting person, or leave the environment where the data is held), such that a human doesn’t need to look at the data to determine its (non-)disclosive nature.

Bitfount

Bitfount is a federated data science platform that allows researchers to perform secure and privacy-preserving data analysis on distributed data. Bitfount enables researchers to analyse data held in separate, geographically distributed databases, without the need for data to be transferred or shared between databases. This is achieved through the use of a distributed computing approach, where the data analysis tasks are executed on each of the individual databases, and the results are aggregated appropriately.

Bitfount also incorporates a range of privacy-enhancing technologies, such as differential privacy and secure multi-party computation, to further protect the privacy of individuals and ensure that sensitive data remains secure. The platform also provides a web-based user interface for managing analysis tasks, accessing results and authorising access to data.

Commented [3]: If this is going to be in here I think all the terms in this sentence also need to be defined (PETs, differential privacy, multi-party computation)

Bitfount Messaging Service

Bitfount's managed service for communications across the Bitfount network. This service operates as the submission layer in the Bitfount platform.

Commented [4]: What does submission layer mean?

Bitfount Pod

A Processor of Data (pod) in the Bitfount platform is the name given to the service that operates next to data and performs approved analyses.

Bitfount Task

A Bitfount task refers to the job or set of instructions to be performed during a single piece of analysis. In TRE-FX the task will be represented as an RO-CRATE

Commented [5]: Needs defining

Certification Management

The process of managing and overseeing certifications or qualifications held by individuals or entities, including tracking expiry dates, renewals, and compliance requirements.

Code Control

The management and oversight of software code or source files, including versioning, change tracking, access control, and collaboration.

Commented [6]: Should all these be defined?

Command Line Interface

A text-based interface or environment that allows users to interact with a computer or software by entering commands or instructions using a command line interpreter.

Common Workflow Language (CWL)

Common Workflow Language is an open standard for describing how to run command line tools and connect them to create workflows. It is supported by WfExS. CWL workflow and tool descriptions are defined in YAML files.

Compliance Checking

The act of verifying and ensuring adherence to applicable laws, regulations, standards, or internal policies within the TRE organisation.

Controlled Environment

This is considered a prepared environment (e.g. a VM) for a given Project within a TRE, in which workflows will be executed against datasets. Typically in a TRE, approved users would log in to such an environment to perform their work. Available tools in the environment are controlled by the TRE admins. Any data entering or leaving the environment is checked and subject to approval. No

Commented [7]: Needs definition

communication with the public internet, or any network resources that aren't expressly approved (e.g. enforced by Firewall rules).

Within Hutch, some components of the stack sit in the controlled environment in order to execute workflows:

- the Hutch Agent
- WfExS
- (possibly approved workflow / tool definitions)
- (approved tool container images)
- Other approved services to interact with e.g. DataSHIELD components
- Target Data sources

Commented [8]: Think this should all be included in the definition of Hutch, pointing out to this definition when referring to Hutch's 'controlled environment'

Controls

Measures, safeguards, or mechanisms implemented to manage or mitigate risks and ensure the integrity, confidentiality, availability, and reliability of systems, processes, or data.

Curriculum Creation and Management

The process of designing, developing, and managing educational curricula, courses through training needs analysis for required competency.

Data Archiving *

The practice of securely storing and preserving data in a read-only format for long-term retention, typically for compliance, historical reference, or reproducibility~~future analysis purposes~~.

Data Asset Register *

A database or record that documents and manages information about the TRE organisation's data assets, including their characteristics like~~,~~ governance~~ownership~~, provenance, usage~~, access,~~ and other relevant details.

Commented [9]: My microphone doesn't seem to be working in the breakout room but I was hoping to ask a question/wondering around this, would the word 'origin' be better for general/public understanding? I'm not a researcher in this area so I may be wrong, but in terms of using accessible language could origin work without detracting from the meaning?

Commented [10]: Note this overlaps with metadata and a metadata catalogue, but they're not the same thing

Data Classification *

The categorisation or labelling of data based on its sensitivity, risk, value, or other attributes, often used to determine appropriate handling, storage, and security controls.

Data Classification Service *

A service that assists in classifying and labelling data based on predefined criteria or policies to determine sensitivity.

Data Classification Tool *

Software or tools designed to assist in the categorization, labelling, and classification of data based on predefined policies or criteria.

Data Consumers

See Researchers

Data Custodian

The person, organisation or other entity responsible for some data. They should control access to the data and protect the use of it and sharing of it (or subsets of it) to ensure regulations are followed appropriate to the type of data. This includes ensuring no private data is disclosed when (or to whom) it shouldn’t be.

Data Deletion

The process of permanently removing or erasing data from storage systems or devices, ensuring that it cannot be recovered or accessed.

Data Egress

The movement or transfer of data to infrastructure outside of TRE either through manual or automated process. Often known as data outputs.

Data Ingress

The movement or transfer of data to infrastructure inside of TRE either through manual or automated process. Often known as data inputs.

Data Lifecycle Control *

The management and oversight of data throughout its lifecycle, including storage, usage, sharing, retention, and eventual disposal~~ition~~.

Data Minimisation *

The practice of collecting, processing, and storing only the minimum amount of data necessary to fulfil a specific purpose or requirement to reduce privacy risks.

Data Pooling

In the context of Federation, but in contrast to Federated Data, Data Pooling refers to a data sharing case where different organisations or data custodians have their own data, with their own policies and autonomy of management, but subscribe to agreements which permit one or more of these individual datasets to be moved into a data handling environment different to that of their custodians, for the purpose of common linkage and onward analysis.

DataSHIELD

DataShield is an open-source software platform that allows researchers to perform secure and privacy-preserving data analysis on distributed data. DataShield enables researchers to analyze data held in separate, geographically distributed databases, without the need for data to be transferred or shared between databases. This is achieved through the use of a distributed computing approach, where the data analysis tasks are executed on each of the individual databases, and the results are aggregated appropriately.

DataShield also incorporates a range of privacy-enhancing technologies, such as differential privacy and secure multi-party computation, to further protect the privacy of individuals and ensure that sensitive data remains secure. The platform provides a range of tools and resources to support data analysis, including libraries for statistical analysis, data visualization tools, and a web-based user interface for managing analysis tasks and accessing results.

Commented [11]: - ChatGPT made this but I'm not sure this is true. Can you confirm?

Commented [12]: yep - I told you it made stuff up!

Data Sunsetting Service *

A service or process for retiring or decommissioning outdated or unnecessary data in line with the organisation and researcher project's retention policies.

Data Transfer Portal

The interface used to access a data transfer service~~An online platform or system that facilitates the secure exchange and transfer of data between individuals, organisations, or systems~~.

Data Transfer Service

A service or system that facilitates the secure and efficient transfer of data between different systems, networks, or locations.

Desktop

The graphical user interface (GUI) and environment presented to users on their computer screens, typically including icons, menus, and windows for interacting with applications and files.

Desktop Applications

Software applications designed to be installed and run on individual computers or desktop systems, often providing specific functionalities or tools.

Document & SOP Management Process

A structured approach to creating, organising, updating, and controlling documents and Standard Operating Procedures (SOPs) within the TRE organisation.

Egress/Ingress Control

The implementation of measures or controls to regulate and monitor the movement of data into and out of the TRE, to prevent sensitive data from leaving the TRE. Often known as output/input checking, or in the case of egress, disclosure control.

Electronic Quality Management System Application

A software application or platform used to manage and automate quality management processes, including document control, corrective actions, audits, and performance tracking.

ELIXIR Recommended Interoperability Resource (RIR)

An ELIXIR Recommended Interoperability Resource is a biological or biomedical data resource that has been recommended by ELIXIR (European Life-sciences Infrastructure for Biological Information) as a key data source for integration with other resources in the field. ELIXIR is an infrastructure for life-science data management that supports the integration and interoperability of biological data across Europe and beyond. Examples include ontology and persistent identifier services, metadata registries and mapping services. They have been identified as having high-quality data, robust standards, and reliable services that can be integrated with other data resources to create a more comprehensive and interoperable network of data. These resources are selected based on their scientific relevance, technical quality, and the level of support and engagement from the data providers.

ELIXIR Core Data Resource

ELIXIR (European Life-sciences Infrastructure for Biological Information) is a pan-European organization that aims to coordinate and provide access to biological data resources, tools, and services to support life science research. ELIXIR Core Data Resources (CDRs) are a subset of ELIXIR resources that provide essential data and services to the life science research community. These resources have been selected based on their scientific excellence, the breadth and depth of their coverage, their sustainability, and their strategic importance for European life science research.

External Audit

An independent assessment or review of the TRE organisation's controls, processes, or compliance conducted by external auditors or audit firms.

Federation

Federation typically refers to a fairly loose alignment of a number of entities who retain autonomy, but agree to align on certain things as a group. Ordinary real world examples of this are the United States of America and the EU, where member states are still individuals with individual laws and management, but they also subscribe to a central body's policies to enable and encourage working together.

Commented [13]: Best definition so far!!

In software terms federation typically then refers to separate organisations with their own policies and assets who agree to allow use of those assets but without the assets leaving control or ownership of the organisation.

Federated Analytics

The running of "analytics" pipelines or workflows across federated datasets. This could be as simple as having access to one or more federated datasets to allow doing some simple maths or other statistical work, but it could also be simultaneously running interconnected analytics against multiple federated datasets and then aggregating the results.

Federated Data

As per Federation above, Federated Data refers to a data sharing case where different organisations or data custodians have their own data, with their own policies and autonomy of management, but subscribe to agreements or organisations, or the use of tools which allow access to that data (e.g. for research purposes), **without the raw data ever leaving the custodian's environment**.

Federated Datasets can be summarised as follows:

- Distinct datasets
- Usually geographically separated
- With different custodians
- With strong often regulatory requirements around
    - Access
    - Anonymity
    - Security
- e.g. Health Data

Federated Identity Mapping

The process of linking or mapping user identities across multiple systems or domains to enable seamless access and authentication.

Federated Learning

The running of machine learning model training or predictions against federated datasets.

Federation Operator

An entity or role responsible for managing and coordinating federated identities and access across multiple systems or organisations.

Federated Query

Synonymous with Federated Analytics.

Firewall

A firewall is a network security system that monitors and controls incoming and outgoing network traffic based on a set of predefined rules. It acts as a barrier between a private network (such as a company's internal network) and the public internet, filtering and blocking unwanted traffic while allowing legitimate traffic to pass through. Firewalls can be hardware devices, software applications, or a combination of both. They typically use a variety of techniques to filter network traffic, such as packet filtering, stateful inspection, and application-level filtering. Firewall rules can be configured to allow or block specific types of traffic based on factors such as the source or destination IP address, port number, protocol type, and content.

Firewalls are commonly used in enterprise networks to protect against external threats such as malware, hacking attempts, and unauthorized access. They can also be used to control access to specific network resources and applications, and to enforce security policies and compliance requirements. While firewalls are an important component of network security, they are not foolproof and cannot provide complete protection against all types of threats. It is important for organizations to use a multi-layered approach to security, including regular software updates, employee training, and other security measures in addition to firewalls.

Five Safes

The Five Safes framework is a set of principles developed to guide researchers and organizations in the creation of Trusted Research Environments (TREs) for handling sensitive data. The framework was developed by the UK Data Service in collaboration with other organizations and is widely used in the research community.

The Five Safes framework consists of five key principles that should be considered when handling sensitive data:

1. Safe projects: The purpose of the project should be clearly defined, and the use of sensitive data should be necessary and proportionate to achieve the project's objectives.
2. Safe people: Access to sensitive data should be restricted to authorized individuals who have been trained in data handling and security procedures.
3. Safe settings: Data should be stored and processed in secure environments that have appropriate physical and technological safeguards in place to prevent unauthorized access or data breaches.
4. Safe data: Data should be appropriately de-identified or anonymized to protect the privacy of individuals, and access should be limited to the minimum necessary for the project.
5. Safe output: The results of the research should be appropriately disseminated to ensure the confidentiality of individuals and to avoid the disclosure of sensitive information.

The Five Safes framework provides a practical and comprehensive approach to handling sensitive data in a responsible and ethical manner, while still enabling researchers to conduct important research that can inform public policy and advance knowledge in various fields.

GA4GH TRS API

The GA4GH TRS API is a web-based application programming interface (API) developed by the Global Alliance for Genomics and Health (GA4GH) to provide a standardized way of describing, discovering, and running genomic analysis workflows.

Identity and Access Management Services

Services, systems, or processes that govern and control user identities, access privileges, authentication, and authorization within an organisation.

Identity Verification

The process of confirming or authenticating the identity of individuals or entities, often through the verification of personal information, credentials, or biometric data.

Information Asset Owner *

An individual or role accountable for managing and overseeing an information asset, including their acquisition, use, maintenance, and protection.

Integrated Development Environment (IDE)

A software application or platform that provides comprehensive tools and features for developing, testing, and debugging software applications.

Internal Audit

An independent evaluation process within the TRE organisation that assesses and improves its internal controls, risk management, and governance.

Internal Auditor

A professional responsible for evaluating and assessing an organisation's internal controls, policies, and procedures to ensure compliance, effectiveness, and efficiency.

Interoperability

Allowing mutual operation between (inter) two or more systems. Mainly letting systems work together, or communicate with each other.

If you control the development of two systems it is easy to get them to interoperate using a proprietary defined interface between those two specific systems, but good interoperability is usually achieved through the use of open standards, whereby any system can understand the standard being used and be developed to work with it, encouraging collaboration, extension etc.

Issue Management Process

A systematic approach to identifying, tracking, resolving, and managing issues or problems that arise within a TRE organisation, aiming to minimise their impact and ensure timely resolution.

IT Service Provider

A company, department, or entity that delivers information technology services or support to internal or external clients, such as network management, software development, or helpdesk support.

Learning Management System

A software platform or application that facilitates the administration, delivery, and tracking of educational or training programs, often including course materials, assessments, and learner progress tracking.

Malware Scanning Application

A software application or tool that scans and detects malicious software or malware on computer systems or networks, aiming to prevent security breaches or infections.

Management

The act of planning, organising, directing, and controlling resources, activities, and people to achieve organisational goals and objectives effectively.

Metadata

Metadata is data that describes or provides information about other data. It is used to provide context, meaning, and structure to data, and helps to make it easier to understand and use. Metadata can describe various aspects of data, such as its content, format, structure, origin, quality, and usage.

TODO: Metadata catalogue

Minimum Viable Product (MVP)

A minimal viable product (MVP) is a version of a product or service that has the minimum set of features and functionality required to meet the needs of early adopters or customers. The goal of an MVP is to quickly validate the product idea and test the market demand, while minimizing development costs and time-to-market.

Monitoring

The continuous or periodic observation, measurement, or tracking of systems, processes, activities, or events to ensure compliance, performance, or security.

Nextflow

Nextflow is a popular open-source workflow management system that allows the development and execution of data-intensive scientific workflows. Nextflow enables researchers to describe their computational workflows using a powerful scripting language that can be modified and adapted to different requirements. Workflows can be executed on a wide range of computing systems, including clusters, cloud platforms, and containers, and can be automatically scaled up or down depending on the size of the data and the available resources.

Nextflow provides a number of features to simplify the process of workflow management, including automatic data provenance tracking, built-in support for software dependencies, and a powerful error handling system. It also provides a web-based user interface that allows users to monitor the progress of their workflows in real-time and provides detailed reports of each execution.

OpenSAFELY

openSAFELY is an open-source software platform that provides a secure and efficient way to carry out analyses on large sets of linked electronic health records (EHRs). It was developed by researchers at the University of Oxford in collaboration with NHS Digital and TPP, a UK-based EHR software company. The openSAFELY platform enables researchers to conduct observational studies on real-world patient data without compromising patient confidentiality or data security. It provides a secure data processing environment that allows analyses to be carried out directly on encrypted EHRs, without the need to transfer or share data outside of the secure environment.

The platform is built using a combination of open-source technologies, including Docker, Kubernetes, and Python, and is designed to be scalable, flexible, and easy to use. It includes a range of tools and resources, such as libraries for data extraction and processing, pre-built analysis templates, and data visualization tools, which make it easy for researchers to conduct analyses on EHR data.

Policy and Process Data *

Information or data related to organisational policies, procedures, guidelines, or processes, often used for documentation, compliance, and governance purposes.

Quality Management Data *

Data collected and analysed to assess and monitor the performance and effectiveness of quality management systems, processes, and initiatives within the TRE organisation.

Quality Management Reporting

The generation and dissemination of reports or dashboards that provide insights and metrics on the performance and effectiveness of quality management processes and activities.

Quality Manager

A person responsible for managing and ensuring the quality of processes and services within an organisation, by implementing quality management systems and practices.

Registry

A centralised database, repository, or system that stores and manages information, configurations, or records related to specific entities, such as users, systems, or resources.

Commented [14]: Need to make it clear this is a general term that will normally be used in a context such as "data registry, "xxx registry", etc

Re-identification Risk Assessment Tooling *

Tools or methodologies used to assess and quantify the potential risk of re-identifying individuals from anonymized or de-identified datasets, often used in privacy-sensitive research or data sharing.

Repository Management

The management and administration of repositories, which are centralised storage locations for versioned files, code, documentation, or other digital assets.

Requirements Gathering and Monitoring *

The process of collecting, documenting, and managing the functional and non-functional requirements for the TRE based on the TRE organisation's goals and data assets.

Researchers *

Individuals or groups who utilise and analyse data for research purposes or as part of their work, such as scientists, analysts, or other professionals.

Resource Allocation

The process of assigning, distributing, and managing resources (such as personnel, finances, equipment, or time) within the TRE organisation to meet objectives and priorities effectively.

Risk Assessment

The systematic evaluation and analysis of potential risks, threats, or vulnerabilities, including their likelihood, potential impact, and the effectiveness of existing controls or mitigation measures.

Risk Ownership

The assignment of responsibility and accountability to individuals or entities for managing and mitigating specific risks within the TRE organisation.

Risk Treatment

The selection and implementation of strategies, controls, or measures to manage or mitigate identified risks, such as risk avoidance, risk transfer, risk reduction, or risk acceptance.

RO-Crate

RO-Crate is a community effort to establish a lightweight specification to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD. For details see also RO-Crate paper.

In TRE-FX we will extend several existing *profiles* of RO-Crate:

- Workflow RO-Crate profile used by *WorkflowHub* - a workflow that potentially can be executed, as consumed by WfExS. TRE-FX will register
- Workflow Run Crate profile – a record of a workflow execution, inputs, outputs and workflow. Can be produced by several workflow engines, see draft Workflow Run RO-Crate paper.

TRE-FX will add additional profiles to make a *Five safes* RO-Crate that includes permission/people.

SAIL Trusted Research Environment

The Secure Anonymised Information Linkage (SAIL) Databank is a Trusted Research Environment (TRE) located in Wales, United Kingdom. It provides researchers with a secure and controlled environment to access, link and analyze de-identified data from various health and administrative databases in Wales. The SAIL Databank provides a single point of access to a wide range of routinely collected health, social and administrative data in Wales. It allows researchers to conduct population-level studies and link data from different sources while ensuring data security and privacy protection. The SAIL Databank also offers a range of data curation and management services, including data extraction, cleaning, and harmonization.

The SAIL Databank is managed by the SAIL team at Swansea University, in collaboration with the Welsh Government and other partners. It has been approved as a TRE by the UK Information Commissioner's Office, which means that it meets the highest standards of data protection and privacy. The SAIL Databank has been used for a wide range of research projects in fields such as epidemiology, public health, and health services research. Its use has led to numerous publications, and has contributed to policy development and health service improvement in Wales and beyond.

Service Delivery *

Implementing and delivering services such as information governance, infrastructure, networks, or software solutions in the TRE.

Session Virtualization

The practice of abstracting and isolating user sessions or interactions from the underlying physical or virtual infrastructure.

Software Request Process

The procedure or workflow followed to request, review, approve, and deploy software applications or tools within an organisation.

Stack

In technology there are two uses of the term “Stack” relevant to the project:

- Technology Stack
- Application Stack

Study Closure *

The formal conclusion of a research study or project, including final data analysis, reporting, documentation, and archiving.

Study Management Portal *

An online platform or system that provides centralised access to manage research studies, including onboarding studies, control of access, and administration of compliance tasks.

Study Manager *

An individual responsible for overseeing and managing a research study or project, including planning, execution, coordination, and reporting.

Study Onboarding *

The process of onboarding or initiating a research study, including setting up necessary infrastructure, obtaining approvals, and defining protocols or methodologies.

Study Register

A centralised record or database that tracks and manages information about research studies or projects.

Supplier Management and Monitoring process

A structured approach to managing and monitoring relationships with external suppliers, vendors and contractors, including selection, contract management and compliance oversight.

Technology Stack

The "Technology Stack" (or "Tech Stack") is the different technologies (such as programming languages) that work together / are being used together for all the parts of a software solution.

In this project the following are some of the technologies used:

- .NET ("dot net")
  - C#
  - ASP.NET Core
- JavaScript
  - React
- Python
- SQL
- gRPC
- Docker

Top Management

The people within the TRE organisation accountable for ensuring the TRE is fit for purpose and fit for use.

Training Needs Analysis

The systematic assessment and identification of training requirements and gaps within an organisation, often through surveys, interviews, or performance evaluations.

Training Records

Documentation or records that track and manage information related to training activities, including attendance, completion, and certifications to demonstrate competency.

TRE (Trusted Research Environment)

A Trusted Research Environment (TRE) is a secure computing environment that is specifically designed for handling sensitive or confidential data in a way that protects privacy and ensures data security. TREs are used in research and analysis, particularly in fields such as health and social sciences, where data must be handled with care and rigour. TREs are typically used for research projects that require access to sensitive data, such as personally identifiable information, medical records, or financial data. These environments are designed to meet strict security and privacy standards, including data encryption, access controls, and monitoring.

TREs are often hosted by trusted third-party organizations, such as government agencies or universities, that have established protocols for managing and securing sensitive data. Researchers are granted access to the TRE only after meeting certain eligibility criteria and going through a rigorous application and approval process.

The use of TREs helps to mitigate the risk of data breaches or unauthorized access, while allowing researchers to work with sensitive data in a controlled and secure environment. TREs also provide a framework for ensuring that data is handled in an ethical and responsible manner, in compliance with regulatory and legal requirements.

Commented [15]: In SATRE we've run into the question of whether a "TRE" exists for a single project/dataset, or whether one TRE can host multiple independent projects/datasets with segregation enforced. Different TREs use the term differently. Do you have any thoughts on which definition we should pick?

Commented [16R15]: This is the definition that is used by HDR UK https://www.hdruk.ac.uk/wp-content/uploads/2021/09/HDRUK_TRE-One-Pager.pdf It suggests that Safe Haven and TRE are interchangeable terms and therefore a TRE could host multiple projects

TRE Implementer

A role involved in Service Delivery.

TRE Infrastructure

The set of computing resources used to implement and support a TRE. This may include desktop computers, databases, networking devices, firewalls etc. These resources may be physical (hardware owned by the TRE) or virtual (e.g. resources operated by a cloud provider).

User Documentation

Written materials, guides, manuals, or instructions designed to assist users in understanding and effectively using software applications, systems, or processes.

User Onboarding *

The process of introducing and integrating researchers and data consumers onto a TRE's systems, processes, including training, access provisioning, and orientation.

WfExS

Workflow Execution Service developed by Barcelona Supercomputing Centre. Uses *TRS* API talking to *WorkflowHub* to retrieve *RO-Crate* of a *Workflow* that is then executed in the corresponding workflow engine (Nextflow, CWL) on a compute backend, with results returned as a *Workflow Run RO-Crate*.

Workflow

In TRE-FX, *workflow* refers to **computational** workflows for data processing, handling and analysis. There are many existing workflow systems, we're focusing on Nextflow and CWL. Workflows in TRE-FX are gathered in WorkflowHub, managed by Manchester. In TRE-FX, workflows are maintained and submitted in the form of *RO-Crate* and executed within a *federated* environment of *TRE*s.

Commented [17]: This should be generalised beyond TRE-FX

WorkflowHub

WorkflowHub is a FAIR **workflow registry** sponsored by the European RI Cluster EOSC-Life and the European Research Infrastructure ELIXIR. It is workflow management system agnostic: workflows may remain in their native repositories in their native forms. Workflows are accessible through the APIs as Workflow *RO-Crate*.

(template)

Describe the term.

Terms to be added

- More stored in categories below
- Project manager
- Project management
- TRE Operators

Grouped by topic

Core terms

Controls

Measures, safeguards, or mechanisms implemented to manage or mitigate risks and ensure the integrity, confidentiality, availability, and reliability of systems, processes, or data.

Desktop (Virtual) - see virtual desktop?

The graphical user interface (GUI) and environment presented to users on their computer screens, typically including icons, menus, and windows for interacting with applications and files.

Commented [18]: Is this simple enough (or should it be defined elsewhere) - just 'what you see on your computer screen' might be simpler

Federation

Federation ~~typically~~ refers to intentional ~~a fairly loose~~ alignment of a number of entities who agree on certain things as a group ~~retain autonomy~~, but retain autonomy~~agree to align on certain things as a group~~. Ordinary real world examples of this are the United States of America and the EU, where member states are still individuals with individual laws and management, but they also subscribe to a central body’s policies to enable and encourage working together.

Commented [19]: Best definition so far!!

In software terms federation typically then refers to separate organisations with their own policies and assets who agree to allow use of those assets but without the assets leaving control or ownership of the organisation.

Interoperability

Allowing mutual operation between ~~(inter)~~ two or more systems. Interoperable systems can ~~Mainly letting systems~~ work together, or communicate with each other.

~~If you control the development of two systems it is easy to get them to interoperate using a proprietary defined interface between those two specific systems, but g~~Good interoperability is usually achieved through the use of open standards, whereby any system can understand the standard being used and be developed to work with it, encouraging collaboration, extension etc.

~~***Management***~~

~~The act of planning, organising, directing, and controlling resources, activities, and people to achieve organisational goals and objectives effectively.~~

Minimum Viable Product (MVP)

A minimal viable product (MVP) is a version of a product or service that has the minimum ~~set of features and~~ functionality required to meet the essential needs of early adopters or customers. The goal of an MVP is to quickly validate the product idea and test the market demand, while minimizing development costs and time-to-market.

Service Delivery *

Implementing and delivering services such as information governance, infrastructure, networks, or software solutions ~~in the TRE~~.

Sensitive data

Any data for which there is concern over public disclosure, and for which security controls are useful to restrict access for research purposes.

TRE (Trusted Research Environment)

A Trusted Research Environment (TRE) is a secure computing environment, or set of linked environments, with associated governance processes. TREs are used by researchers working with sensitive or confidential datasets, ~~secure computing environment that is~~ and are specifically designed ~~for handling sensitive or confidential data in a way that~~ to protect~~s~~ privacy and ensure~~s~~ data security. TREs are used in research and analysis, particularly in fields such as health and social sciences, where data must be handled with care and rigour. TREs are typically used for research

Commented [20]: FYI - can pull some language from here: https://www.researchdata.scot/our-work/data-explainers/what-are-trusted-research-environments/

Commented [21]: In SATRE we've run into the question of whether a "TRE" exists for a single project/dataset, or whether one TRE can host multiple independent projects/datasets with segregation enforced. Different TREs use the term differently. Do you have any thoughts on which definition we should pick?

Commented [22R21]: This is the definition that is used by HDR UK https://www.hdruk.ac.uk/wp-content/uploads/2021/09/HDRUK_TRE-One-Pager.pdf It suggests that Safe Haven and TRE are interchangeable terms and therefore a TRE could host multiple projects

projects that require access to sensitive data, such as personally identifiable information, medical records, or financial data. These environments are designed to meet strict security and privacy standards, including data encryption, access controls, and monitoring.

TREs are often hosted by trusted third-party organizations, such as government agencies or universities, that have established protocols for managing and securing sensitive data. Researchers are granted access to the TRE only after meeting certain eligibility criteria and going through a rigorous application and approval process.

The use of TREs helps to mitigate the risk of data breaches or unauthorized access, while allowing researchers to work with sensitive data in a controlled and secure environment. TREs also provide a framework for ensuring that data is handled in an ethical and responsible manner, in compliance with regulatory and legal requirements.

Several other terms are often used interchangeably in the literature. For example, the NHS uses the term “Secure Data Environment” (SDE), and others have adopted “Data Safe Haven” (DSH) or just “Safe Haven”.

Commented [23]: Might be good to point to a blog post outlining the nuanced differences between these (e.g. the idea of data curation (safe haven) vs data analysis (SDE/TRE))

TRE, SDE, DSH and other terms can also be used to define specific components of the overall TRE definition presented by this glossary, which will vary based on the project, initiative or institution in question.

Workspace

An individual computing environment within a TRE that a TRE user could be logged into. A workspace may have a virtual desktop interface (see above definition) or another style of interface designed with the TRE user in mind, for example a GUI that is specific to working with the research domain or the kind of data in the TRE.

TO BE ADDED

- ~~Project~~
- SDE - Secure data environments (SDEs) are data storage and access platforms, which uphold the highest standards of privacy and security of NHS health and social care data when used for research and analysis. They allow approved users to access and analyse data without the data leaving the environment.
  - https://www.england.nhs.uk/blog/investing-in-the-future-of-health-research-secure-accessible-and-life-saving/
  - https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines
- Workspaces - Project specific environments where authorised researchers/analysts are able to access the appropriate tools and data specific to their research project. An SDE/TRE may have many such environments with specific access rights. Any work carried out in a workspace should be auditable and reproducible.

Commented [24]: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines#:~:text=Secure%20data%20environments%20are%20data,the%20data%20leaving%20the%20environment.

- Repository
- Sensitive data
- Information Governance
- secure data environment
- safe haven
- Data

Security terms

Access Control

The technical mechanism for controlling a known (Authenticated) user's access to the system. Sometimes referred to as Authorisation (and shorthanded as "Auth" often in concert with Authentication), as it determines what a user is authorised to do. Hutch and Bitfount use "Role Based Access Control" (RBAC) so a suitable administrator (e.g. of a project, or a TRE) can configure that certain Users have certain Roles, and those Roles have Permissions to perform authorised actions. In Hutch, Access Control checks for a user will take place after they are authenticated, by both the Submissions Layer outside a TRE and the Polling Layer inside a TRE, to ensure they are allowed to do what they are asking.

Commented [25]: I think this is excess information for a glossary definition

Authentication

Authentication is finding out who a user is and ensuring that they are who they say (i.e. they are authentic) via some acceptable proof. This goes hand in hand with Access Control (or "Authorisation") and the two are both sometimes shorthanded (separately or together) as "Auth". If there's any doubt in the context of what "Auth" is referring to, it should be clarified!

Controlled Environment

This is considered a prepared environment (e.g. a VM) for a given Project within a TRE, in which workflows will be executed against datasets. Typically in a TRE, approved users would log in to such an environment to perform their work. Available tools in the environment are controlled by the TRE admins. Any data entering or leaving the environment is checked and subject to approval. No communication with the public internet, or any network resources that aren't expressly approved (e.g. enforced by Firewall rules).

Commented [26]: Needs definition

Within Hutch, some components of the stack sit in the controlled environment in order to execute workflows:

- the Hutch Agent
- WfExS
- (possibly approved workflow / tool definitions)
- (approved tool container images)
- Other approved services to interact with e.g. DataSHIELD components
- Target Data sources

Commented [27]: Think this should all be included in the definition of Hutch, pointing out to this definition when referring to Hutch's 'controlled environment'

Identity and Access Management Services

Services, systems, or processes that govern and control user identities, access privileges, authentication, and authorization within an organisation.

Identity Verification

The process of confirming or authenticating the identity of individuals or entities, often through the verification of personal information, credentials, or biometric data.

TO BE ADDED

- Single sign-on

Technical terms

API

An abbreviation for Application Programming Interface, an API is a type of software interface that provides a way for two or more computer programs to communicate with each other. In contrast to a user interface, which connects a computer to a person, an application programming interface connects computers or pieces of software to each other.

Application Catalogue

A centralised inventory or repository that lists and describes all the applications or software systems used within an organisation, including their functionalities.

Application Deployment

The process of installing, configuring, and making software applications available for use within the TRE.

Authentication Application

A software system that verifies and validates the identities of users or entities accessing a system through multifactor authentication.

Authentication Token

An authentication token is a piece of data that is used to authenticate the identity of a user or application. It is typically a string of characters or a digital certificate that is generated by an authentication server, and is then passed between the user or application and the server to verify their identity. Authentication tokens are commonly used in web applications, APIs, and other systems that require secure access control. When a user logs in to a system, the authentication server generates a token that is associated with the user's account and privileges. This token is then passed back to the user's browser or application, and is used to authenticate subsequent requests to the system.

Authentication tokens can be generated using a variety of methods, such as cryptographic keys, digital certificates, or session IDs. They can also be time-limited or have other restrictions to enhance security and prevent unauthorized access. The use of authentication tokens helps to ensure that only authorized users or applications can access sensitive information or resources, and can provide an additional layer of security beyond traditional username/password authentication.

Automated Disclosure Control

Disclosure control without the intervention of a human being each time. Configuring a software system to reliably detect that data it is provided with is "non-disclosive" (i.e. doesn't disclose any information that is not allowed to be shared with the requesting person, or leave the environment where the data is held), such that a human doesn't need to look at the data to determine its (non-)disclosive nature.

Command Line Interface

A text-based interface or environment that allows users to interact with a computer or software by entering commands or instructions using a command line interpreter.

Common Workflow Language (CWL)

Common Workflow Language is an open standard for describing how to run command line tools and connect them to create workflows. It is supported by WfExS. CWL workflow and tool descriptions are defined in YAML files.

Desktop Applications

Software applications designed to be installed and run on individual computers or desktop systems, often providing specific functionalities or tools.

Differential privacy

A method of analysing a dataset by looking at aggregate information without relying on individual data points. Algorithms or techniques that implement differential privacy should produce identical or close-to-identical outputs if a small number of the input data points are changed or deleted.

Federated Analytics

The running of "analytics" pipelines or workflows across federated datasets. This could be as simple as having access to one or more federated datasets to allow doing some simple maths or other statistical work, but it could also be simultaneously running interconnected analytics against multiple federated datasets and then aggregating the results.

Federated Data

As per Federation above, Federated Data refers to a data sharing case where different organisations or data custodians have their own data, with their own policies and autonomy of management, but subscribe to agreements or organisations, or the use of tools which allow access to that data (e.g. for research purposes), **without the raw data ever leaving the custodian's environment**.

Federated Datasets can be summarised as follows:

- Distinct datasets
- Usually geographically separated
- With different custodians
- With strong often regulatory requirements around
    - Access
    - Anonymity
    - Security
- e.g. Health Data

Federated Identity Mapping

The process of linking or mapping user identities across multiple systems or domains to enable seamless access and authentication.

Federated Learning

The running of machine learning model training or predictions against federated datasets.

Federated Query

Synonymous with Federated Analytics.

Firewall

A firewall is a network security system that monitors and controls incoming and outgoing network traffic based on a set of predefined rules. It acts as a barrier between a private network (such as a company's internal network) and the public internet, filtering and blocking unwanted traffic while allowing legitimate traffic to pass through. Firewalls can be hardware devices, software applications, or a combination of both. They typically use a variety of techniques to filter network traffic, such as packet filtering, stateful inspection, and application-level filtering. Firewall rules can be configured to allow or block specific types of traffic based on factors such as the source or destination IP address, port number, protocol type, and content.

Firewalls are commonly used in enterprise networks to protect against external threats such as malware, hacking attempts, and unauthorised access. They can also be used to control access to specific network resources and applications, and to enforce security policies and compliance requirements. While firewalls are an important component of network security, they are not foolproof and cannot provide complete protection against all types of threats. It is important for organisations to use a multi-layered approach to security, including regular software updates, employee training, and other security measures in addition to firewalls.

GA4GH TRS API

The GA4GH TRS API is a web-based application programming interface (API) developed by the Global Alliance for Genomics and Health (GA4GH) to provide a standardised way of describing, discovering, and running genomic analysis workflows.

High-performance computing (HPC)

Large clusters of computing resources designed to process big data and solve complex problems in the shortest possible time. This often involves multiple nodes, parallel processes and/or GPUs.

Integrated Development Environment (IDE)

A software application or platform that provides comprehensive tools and features for developing, testing, and debugging software applications.

Malware Scanning Application

A software application or tool that scans and detects malicious software or malware on computer systems or networks, aiming to prevent security breaches or infections.

Privacy enhancing technology (PET)

Any technology used to minimise the use of sensitive data as part of a data processing operation. This includes both "hard" PET such as complete anonymisation of the input data and "soft" PET such as differential privacy.

Repository Management

The management and administration of repositories, which are centralised storage locations for versioned files, code, documentation, or other digital assets.

Commented [28]: Should be defined

Secure multi-party computation

Any of a set of computational methods that allow several different parties to each independently calculate the result of a function over a set of inputs without revealing the inputs to one another.

Session Virtualization

The practice of abstracting and isolating user sessions or interactions from the underlying physical or virtual infrastructure.

Software Request Process

The procedure or workflow followed to request, review, approve, and deploy software applications or tools within an organisation.

Stack

In technology there are two uses of the term "Stack" relevant to the project:

- Technology Stack
- Application Stack

Technology Stack

The “Technology Stack” (or “Tech Stack”) is the different technologies (such as programming languages) that work together / are being used together for all the parts of a software solution.

In this project the following are some of the technologies used:

- .NET (“dot net”)
    - C#
    - ASP.NET Core
- JavaScript
    - React
- Python
- SQL
- gRPC
- Docker

TRE Infrastructure

The set of computing resources used to implement and support a TRE. This may include desktop computers, databases, networking devices, firewalls etc. These resources may be physical (hardware owned by the TRE) or virtual (e.g. resources operated by a cloud provider).

Virtual desktop / virtual desktop environment

Provision of a conventional desktop interface (*e.g.* Gnome, Windows Desktop) to a computer through a remote desktop gateway (*e.g.* Apache Guacamole), this may be through a web browser or client application.

Workflow

In TRE-FX, *workflow* refers to **computational** workflows for data processing, handling and analysis. There are many existing workflow systems, we’re focusing on Nextflow and CWL. Workflows in TRE-FX are gathered in WorkflowHub, managed by Manchester. In TRE-FX, workflows are maintained and submitted in the form of *RO-Crate* and executed within a *federated* environment of *TRE*s.

Commented [29]: This should be generalised beyond TRE-FX

TO BE ADDED

●

Information Governance

Asset Management Process

A systematic approach to acquiring, operating, maintaining, and disposing of assets within an organisation, aimed at maximising their value and minimising risks.

Certification Management

The process of managing and overseeing certifications or qualifications held by individuals or entities, including tracking expiry dates, renewals, and compliance requirements.

Code Control

The management and oversight of software code or source files, including versioning, change tracking, access control, and collaboration.

Compliance Checking

The act of verifying and ensuring adherence to applicable laws, regulations, standards, or internal policies within the TRE organisation.

Curriculum Creation and Management

The process of designing, developing, and managing educational curricula, courses through training needs analysis for required competency.

Data Loss Prevention

A set of strategies, processes, and technologies designed to prevent sensitive and confidential data from being lost, corrupted, leaked or exposed to unauthorised entities.

Document & SOP Management Process

A structured approach to creating, organising, updating, and controlling documents and Standard Operating Procedures (SOPs) within the TRE organisation.

Data leakage

Def should include how it increases the risk of re-identification as once there is a leak you know much more about the data subjects

Egress/Ingress Control

The implementation of measures or controls to regulate and monitor the movement of data into and out of the TRE.

Electronic Quality Management System Application

A software application or platform used to manage and automate quality management processes, including document control, corrective actions, audits, and performance tracking.

External Audit

An independent assessment or review of the TRE organisation's controls, processes, or compliance conducted by external auditors or audit firms.

Five Safes

The Five Safes framework is a set of principles developed to guide researchers and organizations in the creation of Trusted Research Environments (TREs) for handling sensitive data. The framework was developed by the UK Data Service in collaboration with other organizations and is widely used in the research community.

The Five Safes framework consists of five key principles that should be considered when handling sensitive data:

6. Safe projects: The purpose of the project should be clearly defined, and the use of sensitive data should be necessary and proportionate to achieve the project's objectives.
7. Safe people: Access to sensitive data should be restricted to authorized individuals who have been trained in data handling and security procedures.
8. Safe settings: Data should be stored and processed in secure environments that have appropriate physical and technological safeguards in place to prevent unauthorized access or data breaches.
9. Safe data: Data should be appropriately de-identified or anonymized to protect the privacy of individuals, and access should be limited to the minimum necessary for the project.
10. Safe output: The results of the research should be appropriately disseminated to ensure the confidentiality of individuals and to avoid the disclosure of sensitive information.

Commented [30]: This lacks that such projects should contribute to the public good

Commented [31R30]: and the project is good research: it is done correctly, has quality, has gone through ethics approval...

The Five Safes framework provides a practical and comprehensive approach to handling sensitive data in a responsible and ethical manner, while still enabling researchers to conduct important research that can inform public policy and advance knowledge in various fields.

Incident reporting

(it includes communicating a data leakage but goes beyond)

Information Governance

Information Governance (IG) is a term used to describe the principles, processes, legal and ethical responsibilities for managing and handling personal information.~~. n~~

Internal Audit

An independent evaluation process within the TRE organisation that assesses and improves its internal controls, risk management, and governance.

Issue Management Process

A systematic approach to identifying, tracking, resolving, and managing issues or problems that arise within a TRE organisation, aiming to minimise their impact and ensure timely resolution.

Learning Management System

A software platform or application that facilitates the administration, delivery, and tracking of educational or training programs, often including course materials, assessments, and learner progress tracking.

Monitoring

The continuous or periodic observation, measurement, or tracking of systems, processes, activities, or events to ensure compliance, performance, or security.

Office of the Scottish Information Commissioner~~Commisioner~~ (OSIC)

Scottish branch of the Information Commissioner's~~Commisioner's~~ Office

Privacy-Enhancing Technologies (PET)

Tools and techniques to protect the privacy of personally identifiable information.

Quality Management Reporting

The generation and dissemination of reports or dashboards that provide insights and metrics on the performance and effectiveness of quality management processes and activities.

Records management

Re-identification Risk ~~Assessment Tooling *~~

Def for Re-identification Risk: the potential risk of re-identifying individuals from anonymized or de-identified datasets, often used in privacy-sensitive research or data sharing.

Which can be assessed via Re-identification Risk Assessment Tooling: Tools or methodologies used to assess and quantify the potential risk of re-identifying individuals~~ from anonymized or de-identified datasets, often used in privacy-sensitive research or data sharing~~.

Commented [32]: Would be appropriate to separate these two as: Re-dentification, and Risk Assesment Tooling (from collab cafe)

Commented [33R32]: Actually, it looks like the term to be defined here is Re-identification Risk.

Requirements Gathering and Monitoring *

The process of collecting, documenting, and managing the functional and non-functional requirements for the TRE based on the TRE organisation's goals and data assets.

Resource Allocation

The process of assigning, distributing, and managing resources (such as personnel, finances, equipment, or time) within the TRE organisation to meet objectives and priorities effectively.

Risk Assessment

The systematic evaluation and analysis of potential risks, threats, or vulnerabilities, including their likelihood, potential impact, and the effectiveness of existing controls or mitigation measures.

Risk Ownership

The assignment of responsibility and accountability to individuals or entities for managing and mitigating specific risks within the TRE organisation.

Risk Treatment

The selection and implementation of strategies, controls, or measures to manage or mitigate identified risks, such as risk avoidance, risk transfer, risk reduction, or risk acceptance.

Statistical disclosure control

Commented [34]: Allison Noble & Madalyn could help define?

Study Closure *

The formal conclusion of a research study or project, including final data analysis, reporting, documentation, and archiving.

Study Management Portal *

An online platform or system that provides centralised access to manage research studies, including onboarding studies, control of access, and administration of compliance tasks.

Study Onboarding *

The process of onboarding or initiating a research study, including setting up necessary infrastructure, obtaining approvals, and defining protocols or methodologies.

Study Register

A centralised record or database that tracks and manages information about research studies or projects.

Supplier Management and Monitoring process

A structured approach to managing and monitoring relationships with external suppliers, vendors and contractors, including selection, contract management and compliance oversight.

Training Needs Analysis

The systematic assessment and identification of training requirements and gaps within an organisation, often through surveys, interviews, or performance evaluations.

Training Records

Documentation or records that track and manage information related to training activities, including attendance, completion, and certifications to demonstrate competency.

User agreement / terms of use

Legally binding document establishing requirements for user, terms of use

User Documentation

Written materials, guides, manuals, or instructions designed to assist users in understanding and effectively using software applications, systems, or processes.

User Onboarding *

The process of introducing and integrating researchers and data consumers onto a TRE's systems, processes, including training, access provisioning, and orientation.

Data and data management

Data Archiving *

The practice of securely storing and preserving data in a read-only format for long-term retention, typically for compliance, historical reference, or future analysis purposes.

Data Asset Register *

A database or record that documents and manages information about the TRE organisation's data assets, including their characteristics, ownership, usage, and other relevant details.

Data Classification *

The categorisation or labelling of data based on its sensitivity, value, or other attributes, often used to determine appropriate handling, storage, and security controls.

Data Classification Service *

A service that assists in classifying and labelling data based on predefined criteria or policies to determine sensitivity.

Data Classification Tool *

Software or tools designed to assist in the categorization, labelling, and classification of data based on predefined policies or criteria.

Data Deletion

The process of permanently removing or erasing data from storage systems or devices, ensuring that it cannot be recovered or accessed.

Data Egress

The movement or transfer of data to infrastructure outside of TRE either through manual or automated process.

Data Ingress

The movement or transfer of data to infrastructure inside of TRE either through manual or automated process.

Data Lifecycle Control *

The management and oversight of data throughout its lifecycle, including storage, usage, sharing, retention, and eventual disposition.

Data Minimisation *

Commented [35]: very related to IG as well

The practice of collecting, processing, and storing only the minimum amount of data necessary to fulfil a specific purpose or requirement to reduce privacy risks.

Data Pooling

In the context of Federation, but in contrast to Federated Data, Data Pooling refers to a data sharing case where different organisations or data custodians have their own data, with their own policies and autonomy of management, but subscribe to agreements which permit one or more of these individual datasets to be moved into a data handling environment different to that of their custodians, for the purpose of common linkage and onward analysis.

Data Sunsetting Service *

A service or process for retiring or decommissioning outdated or unnecessary data in line with the organisation and researcher project's retention policies.

Data Transfer Portal

An online platform or system that facilitates the secure exchange and transfer of data between individuals, organisations, or systems.

Data Transfer Service

A service or system that facilitates the secure and efficient transfer of data between different systems, networks, or locations.

Metadata

Metadata is data that describes or provides information about other data. It is used to provide context, meaning, and structure to data, and helps to make it easier to understand and use. Metadata can describe various aspects of data, such as its content, format, structure, origin, quality, and usage.

Policy and Process Data *

Information or data related to organisational policies, procedures, guidelines, or processes, often used for documentation, compliance, and governance purposes.

Quality Management Data *

Data collected and analysed to assess and monitor the performance and effectiveness of quality management systems, processes, and initiatives within the TRE organisation.

Registry

A centralised database, repository, or system that stores and manages information, configurations, or records related to specific entities, such as users, systems, or resources.

Roles

Commented [36]: Clarify whether is it a function performed a their role or whether it is a job title

Approvals Panel

Group or groups who are responsible to approve researchers and research projects to have access to the data within a TRE

Data Consumers

See Researchers

Data Custodian

The person, organisation or other entity responsible for some data. They should control access to the data and protect the use of it and sharing of it (or subsets of it) to ensure regulations are followed appropriate to the type of data. This includes ensuring no private data is disclosed when (or to whom) it shouldn't be.

Data Subjects

People whose data is held within the TRE.

Data Wrangler

Individual working within the TRE who is responsible for preparing the data for analysis which could include sanitising, linking data, anonymising, removing errors

Federation Operator

An entity or role responsible for managing and coordinating federated identities and access across multiple systems or organisations.

Information Asset Owner *

An individual or role accountable for managing and overseeing an information asset, including their acquisition, use, maintenance, and protection.

Internal Auditor

A professional responsible for evaluating and assessing an organisation's internal controls, policies, and procedures to ensure compliance, effectiveness, and efficiency.

IT Service Provider

A company, department, or entity that delivers information technology services or support to internal or external clients, such as network management, software development, or helpdesk support.

Members of the Public

This includes a range of individuals from those with no awareness of TREs, those who are interested and have some awareness as well as people have been actively involved in TREs for example in consultations or on approvals panels.

Output Checker/External Referee/Information Governance

Commented [37]: No consensus on this role term - ask SACRO their findings from stakeholders?

Quality Manager

A person responsible for managing and ensuring the quality of processes and services within an organisation, by implementing quality management systems and practices.

Researchers *

Individuals or groups who utilise and analyse data for research purposes or as part of their work, such as scientists, analysts, or other professionals.

Study Manager *

An individual responsible for overseeing and managing a research study or project, including planning, execution, coordination, and reporting. This can be internal or external to the TRE.

Senior Leadership~~Top Management~~

The people within the TRE organisation ultimately accountable for ensuring the TRE is fit for purpose and fit for use. They delegate responsibility to specialised roles within the TRE.

TRE Implementer

A role involved in Service Delivery.

Products/services/organisations

Bitfount

Bitfount is a federated data science platform that allows researchers to perform secure and privacy-preserving data analysis on distributed data. Bitfount enables researchers to analyse data held in separate, geographically distributed databases, without the need for data to be transferred or shared between databases. This is achieved through the use of a distributed computing approach, where the data analysis tasks are executed on each of the individual databases, and the results are aggregated appropriately.

Bitfount also incorporates a range of privacy-enhancing technologies, such as differential privacy and secure multi-party computation, to further protect the privacy of individuals and ensure that sensitive data remains secure. The platform also provides a web-based user interface for managing analysis tasks, accessing results and authorising access to data.

Commented [38]: If this is going to be in here I think all the terms in this sentence also need to be defined (PETs, differential privacy, multi-party computation)

Bitfount Messaging Service

Bitfount's managed service for communications across the Bitfount network. This service operates as the submission layer in the Bitfount platform.

Commented [39]: What does submission layer mean?

Bitfount Pod

A Processor of Data (pod) in the Bitfount platform is the name given to the service that operates next to data and performs approved analyses.

Bitfount Task

A Bitfount task refers to the job or set of instructions to be performed during a single piece of analysis. In TRE-FX the task will be represented as an RO-CRATE

Commented [40]: Needs defining

DataSHIELD

DataShield is an open-source software platform that allows researchers to perform secure and privacy-preserving data analysis on distributed data. DataShield enables researchers to analyze data held in separate, geographically distributed databases, without the need for data to be transferred or shared between databases. This is achieved through the use of a distributed computing approach, where the data analysis tasks are executed on each of the individual databases, and the results are aggregated appropriately.

DataShield also incorporates a range of privacy-enhancing technologies, such as differential privacy and secure multi-party computation, to further protect the privacy of individuals and ensure that sensitive data remains secure. The platform provides a range of tools and resources to support data analysis, including libraries for statistical analysis, data visualization tools, and a web-based user interface for managing analysis tasks and accessing results.

Commented [41]: - ChatGPT made this but I'm not sure this is true. Can you confirm?

Commented [42]: yep - I told you it made stuff up!

ELIXIR Recommended Interoperability Resource (RIR)

An ELIXIR Recommended Interoperability Resource is a biological or biomedical data resource that has been recommended by ELIXIR (European Life-sciences Infrastructure for Biological Information) as a key data source for integration with other resources in the field. ELIXIR is an infrastructure for life-science data management that supports the integration and interoperability of biological data across Europe and beyond. Examples include ontology and persistent identifier services, metadata registries and mapping services. They have been identified as having high-quality data, robust standards, and reliable services that can be integrated with other data resources to create a more comprehensive and interoperable network of data. These resources are selected based on their scientific relevance, technical quality, and the level of support and engagement from the data providers.

ELIXIR Core Data Resource

ELIXIR (European Life-sciences Infrastructure for Biological Information) is a pan-European organization that aims to coordinate and provide access to biological data resources, tools, and services to support life science research. ELIXIR Core Data Resources (CDRs) are a subset of ELIXIR resources that provide essential data and services to the life science research community. These resources have been selected based on their scientific excellence, the breadth and depth of their coverage, their sustainability, and their strategic importance for European life science research.

Nextflow

Nextflow is a popular open-source workflow management system that allows the development and execution of data-intensive scientific workflows. Nextflow enables researchers to describe their computational workflows using a powerful scripting language that can be modified and adapted to different requirements. Workflows can be executed on a wide range of computing systems, including clusters, cloud platforms, and containers, and can be automatically scaled up or down depending on the size of the data and the available resources.

Nextflow provides a number of features to simplify the process of workflow management, including automatic data provenance tracking, built-in support for software dependencies, and a powerful error handling system. It also provides a web-based user interface that allows users to monitor the progress of their workflows in real-time and provides detailed reports of each execution.

OpenSAFELY

openSAFELY is an open-source software platform that provides a secure and efficient way to carry out analyses on large sets of linked electronic health records (EHRs). It was developed by researchers at the University of Oxford in collaboration with NHS Digital and TPP, a UK-based EHR software company. The openSAFELY platform enables researchers to conduct observational studies on real-world patient data without compromising patient confidentiality or data security. It provides a secure data processing environment that allows analyses to be carried out directly on encrypted EHRs, without the need to transfer or share data outside of the secure environment.

The platform is built using a combination of open-source technologies, including Docker, Kubernetes, and Python, and is designed to be scalable, flexible, and easy to use. It includes a range of tools and resources, such as libraries for data extraction and processing, pre-built analysis templates, and data visualization tools, which make it easy for researchers to conduct analyses on EHR data.

RO-Crate

RO-Crate is a community effort to establish a lightweight specification to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD. For details see also RO-Crate paper.

In TRE-FX we will extend several existing *profiles* of RO-Crate:

- Workflow RO-Crate profile used by *WorkflowHub* - a workflow that potentially can be executed, as consumed by WfExS. TRE-FX will register
- Workflow Run Crate profile – a record of a workflow execution, inputs, outputs and workflow. Can be produced by several workflow engines, see draft Workflow Run RO-Crate paper.

TRE-FX will add additional profiles to make a *Five safes* RO-Crate that includes permission/people.

SAIL Trusted Research Environment

The Secure Anonymised Information Linkage (SAIL) Databank is a Trusted Research Environment (TRE) located in Wales, United Kingdom. It provides researchers with a secure and controlled environment to access, link and analyze de-identified data from various health and administrative databases in Wales. The SAIL Databank provides a single point of access to a wide range of routinely collected health, social and administrative data in Wales. It allows researchers to conduct population-level studies and link data from different sources while ensuring data security and privacy protection. The SAIL Databank also offers a range of data curation and management services, including data extraction, cleaning, and harmonization.

The SAIL Databank is managed by the SAIL team at Swansea University, in collaboration with the Welsh Government and other partners. It has been approved as a TRE by the UK Information Commissioner's Office, which means that it meets the highest standards of data protection and privacy. The SAIL Databank has been used for a wide range of research projects in fields such as epidemiology, public health, and health services research. Its use has led to numerous publications, and has contributed to policy development and health service improvement in Wales and beyond.

WorkflowHub

WorkflowHub is a FAIR **workflow registry** sponsored by the European RI Cluster EOSC-Life and the European Research Infrastructure ELIXIR. It is workflow management system agnostic: workflows may remain in their native repositories in their native forms. Workflows are accessible through the APIs as Workflow *RO-Crate*.

TO BE ADDED

- Bitfount Dataset
- Bitfount Data Source
- Hutch
- DARE
- HDR
- TRE-FX
- SATRE
- SACRO
- SARA
- TELEPORT

Commented [43]: Since these are transient should they be included in this section, could they be in a 'past projects' or make sure that in the description their short lfe is highlighted