Best Practices for Procuring and Building a Trusted Research Environment

From federated architecture to pipeline automation and transparent pricing, Data Custodians should consider leveraging these best practices to procure and build a secure and future-proof Trusted Research Environment.

WHITE PAPER

lifebit

Introduction

The opportunities for data-driven research and innovation have never been larger. The availability of large-scale linked health data for research can accelerate our understanding of how to detect, prevent and treat disease.

When COVID-19 hit, organisations globally scrambled to set up large-scale infrastructure to make health data securely accessible for collaborative research. This presented challenges - with the vast majority of data management platforms being highly secure yet largely siloed, with little ability to combine datasets and effectively pool research resources.

Looking forward, many governments, healthcare providers, biobanks and research organisations are setting up Trusted Research Environments (TREs).[1-4] A term conceived by the UK's national institute for health data science, Health Data Research UK (HDR UK), TRE are defined as highly secure computing environments that provide remote access to health data for authorised researchers on approved studies.[5] They support the highest level of data governance by removing the need to share data physically among researchers and organisations. Data instead remains in a secure environment and is analysed *in situ* by authorised researchers with tools available in the TRE.

"In our experience with world-leading precision medicine and health research initiatives, we have seen the complex challenges faced when transitioning from traditional data platforms to a Trusted Research Environment. Incorporating these TRE best practices has enabled our clients to become the model infrastructure for health data management globally and reap the rewards in the form of collaboration and commercialisation opportunities," says Dr. Maria Chatzou Dunford, Lifebit Chief Executive Officer.

This report serves as a reference guide on best practices in establishing and procuring a TRE, to ensure your new data management platform can operate effectively and collaboratively now and into the future.

Best Practise 1: Maintain ownership

Maintain total ownership of your data and TRE to maximise security and research outputs, while minimising cost

Linked health data is of high value for research, yet its scale and sensitivity bring unique challenges for data sharing. Data custodians of population-scale biomedical cohorts have been tasked with a critical role - safeguarding their participants' data - this is the building blocks upon which all ethical approval, participant consent and public trust hinge upon. To maintain total security over the data, it must be kept exclusively in the Data Custodian's own TRE environment. To outsource data control to an external commercial company, involving the risky movement of highly sensitive participant data, is to risk participant data privacy and security.

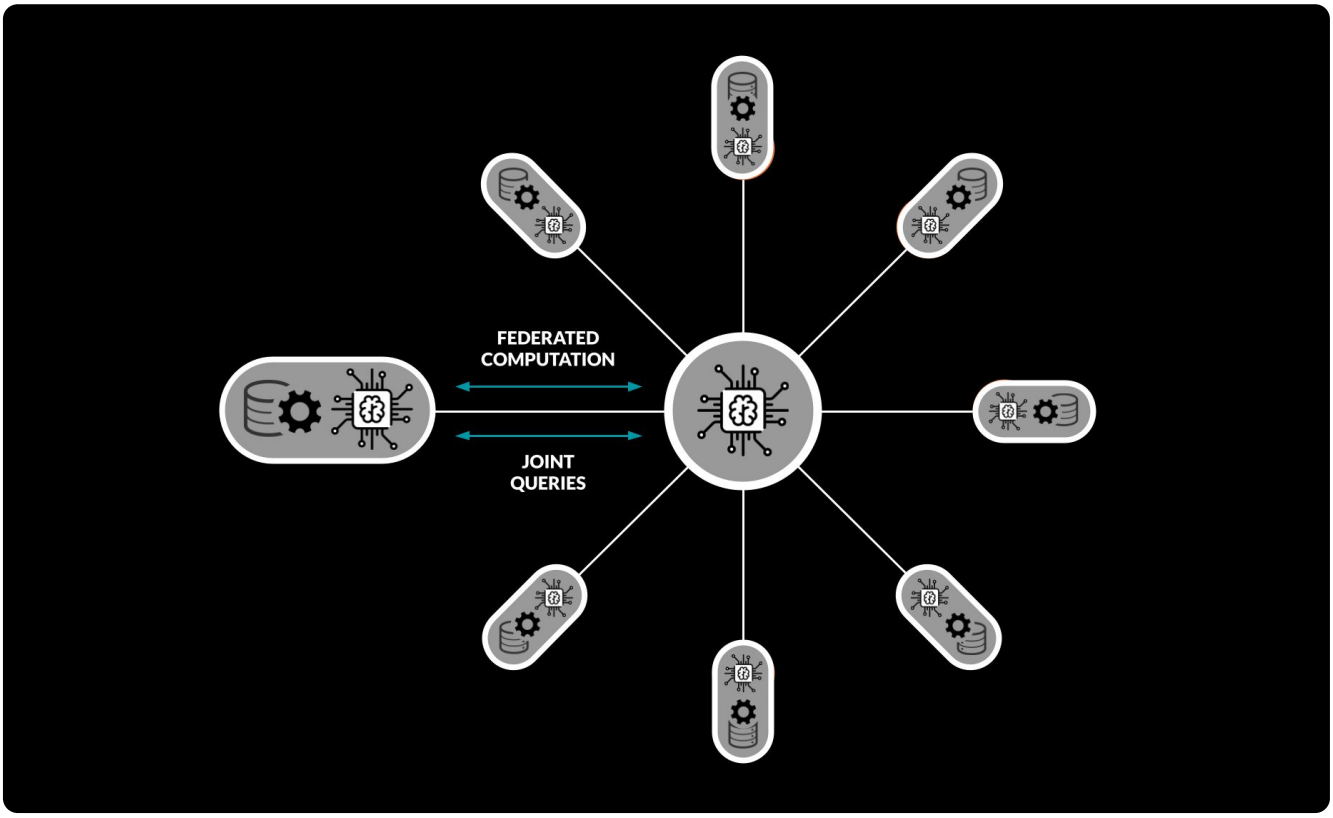
In addition, to move or copy data between TREs is to unnecessarily double costs for data storage and risk the privacy of highly sensitive participant data. In a design that Lifebit has implemented with some of the world-leading population cohorts, all data is maintained in the Data custodian's environment and a virtual file system is used to allow distributed access.[3,6] This reduces egress costs associated with data transfers, setting the TRE on a sustainable path to fast-track research.

Best Practise 2: Federate to collaborate

Adopt a federated approach to open secure collaboration and commercialisation opportunities

Federation is the future of big data analytics.[7] Federated approaches involve independent organisations hosting data in secure environments (e.g. TREs) while linking technologies (e.g. Application programming interfaces or APIs) are applied so that data can be securely analysed across multiple sites. This is an increasingly important approach for bringing together the distributed global research community and addresses the fact that data cannot be pooled for legal, regulatory or practical reasons.[8,9]

By adopting a federated approach, Data Custodians retain full security over their datasets, as all data remains securely within the bounds and security firewalls of the TRE, and only analysis and computation are taken to external datasets and cohorts. A number of organisations internationally are adopting this approach, from government and public research organisations like Genomics England and CanDIG to pharmaceutical giants like Boehringer Ingelheim.[3,10,11] With federation, researchers can gain access to larger, more diverse cohorts and organisations can gain opportunities for data collaboration and commercialisation globally.



Federated technology *enables the virtual linkage of distributed databases. This empowers researchers to run joint queries & analysis over distributed data with federated compute resources. For federated analytics, data is never replicated or moved.*

Best Practise 3: Industry-standard security and compliance

Establish industry-level compliance standards and transparent security processes to maintain public trust

TREs must be compliant with industry-wide standards that go beyond GDPR and HIPAA, for example, ISO 27001 and UK government-backed scheme, Cyber Essentials Plus.[12,13] ISO 27001 is one of the most widely recognised information security standards, defining how organisations should use security controls to manage and handle information in a secure manner. To achieve this level of certification, TREs must have a systematic approach to managing and protecting health data, including regular external audits of the full platform.

With an alarming rise in reports of large-scale data breaches and data mining activities and a long-overdue shift in public awareness towards personal data sovereignty, maintaining public trust in health data research is critical.[14-16] Conducting meaningful Patient and Public Involvement and Engagement (PPIE) and maintaining transparency on TRE compliance and data governance procedures is vital to ensure the long-term success and growth of population health initiatives.[17]

Best Practise 4: Follow the Five Safes for secure data

Implement the Five Safes to maximise security throughout the data lifecycle

The Five Safes framework, established by the Office for National Statistics in 2006, lays out a set of principles for ensuring safe access to sensitive data.[18] These span all stages of data management, from collection and processing to analysis and results reporting. The five key elements to consider are: Safe People, Safe Projects, Safe Settings, Safe Data and Safe Outputs. Staying securely within the bounds of these five pillars for secure data management is key to maintaining public trust.

A recent white paper from HDR UK built upon this framework to establish guidelines for building TREs, outlining key activities for each.[19]

- Safe people: only authorised analysts or researchers can access the data and only on approved projects. Data custodians need a process to verify the authorisation status of these individuals and need to be able to segregate data access between users. All user access and activities performed over the TRE must be recorded and logged, to enable full auditability.
- Safe projects: TREs need a transparent application process for data access, i.e. individuals need to be clear on what they are using the data for.
- Safe settings: TREs must hold data securely and have industry-standard security controls such as data encryption, no export of individual-level data and ability to track researcher/user activity.
- Safe data: data needs to be de-identified and encrypted both at rest and in transit.
- Safe outputs: TREs need a robust and transparent process to support the export of data results, also known as an Airlock. This prevents the unauthorised removal of data.

Best Practise 5: Automate the transformation to research-ready data

Automation of upstream pipelines and harmonisation processes can guarantee rapid and standardised production of research-ready data

A TRE needs to support the full data quality lifecycle, including ingestion, curation, harmonisation and quality control. TREs need automated systems within the platform to manage the large-scale raw data flowing into the environment and efficiently convert it to standardised analysis-ready data. This includes established ETL (Extract, Transform, Load) pipelines and APIs (using industry-leading standards like those of GA4GH) for interfacing between TREs and Data Custodians.[20]

TREs are frequently acquiring data from multiple data sources, calling for automated processes that harmonise disparate datasets. Transforming data to a common format, using a standard set of vocabularies, means it can be efficiently analysed using a library of standard analytic pipelines. Large-scale data harmonisation can be complex and time-consuming, we would advise selecting TRE vendors with deep experience in this process, using industry-recognised standards (e.g. HL7 FHIR) and vocabularies (e.g. OMOP common data model).[21,22] Effective harmonisation and standardisation integrates health data across organisations so that data resources can be queried more quickly and efficiently.

Best Practise 6: FAIR data

Create standardised metadata and use FAIR principles to make data findable and reusable

Metadata are data that provide information about other data, they exist to give data context. We recommend using standardised metadata and data curation standards within a TRE. FAIR (Findable, Accessible, Interoperable, and Reusable) are a set of principles that serve as guidelines for researchers wanting to enhance the reusability and discoverability of their data.[23] HDR UK's Data Utility Framework is another example of industry-level recommendations for organisations to make their data more discoverable and usable.[24]

Aligning with these industry standards brings a number of benefits to Data Custodians - making data more findable and reusable, enabling integration with other datasets or public repositories and aiding efficient data interpretation.

Best Practise 7: Multi-layered security controls

Apply trusted data controls to maximise security at each stage of the data lifecycle

Protecting data confidentiality and security within a TRE takes a multi-layered approach. Key controls that should be implemented within a TRE include:

- **De-identification:** Masking certain data to prevent a data file from being traced back to the file owner, this includes masking any potentially identifiable information with a random number or string.
- **Encryption:** The translation of data into another form, or code, so that only people with access to a key or password can read it.
- **Airlock:** A security process to manage all movement of sensitive data into and out of the TRE, whereby any movement must be approved by an authorised team.
- **Role-based Access Control:** Regulating which users can view or use resources within the TRE.
- **Tiered Access Levels:** Data access is tiered by levels according to end-user type. We provide an example approach:
    - Tier 1: Platform users can only see aggregate anonymised participant data.
    - Tier 2: Approved researchers can access anonymised individual-level data on a limited project-by-project basis, participant data access is also limited to the scope of the specific project.
    - Tier 3: Clinicians have access to identifiable, individual-level genomic and clinical data for patient care purposes.
- **Segregation:** The ability to segregate datasets and workspaces to meet compliance and restrict user access. This increasingly includes segregation of where clinical and genomic data is stored.

Best Practise 8: Procure an all-in-one solution

Procure an all-in-one TRE solution for smooth operations and mitigation of delivery risk

A common pitfall encountered during TRE procurement and build is the partitioning of services, i.e. separating the supplier contracts for billing, cloud/on-premise infrastructure and TRE. Optimally configuring the infrastructure environment for the complexity and demand of TRE systems creates a multitude of challenges during setup and if there is a disconnect between these suppliers and systems, it is likely to lead to delays and suboptimal performance.

For TRE billing and invoicing, it is key that incremental storage and computational cost are tracked across individual workspaces and that cost limits are enforced at the analysis-, user-, workspace- and organisation-level. Setting this up requires a deep understanding of platform usage, cloud provider's cost structures and the analytics workflows (e.g. bioinformatics pipelines), requirements that only experienced TRE providers are best placed to understand to ensure cost does not quickly become unsustainable.

Deploying an all-in-one solution ensures smooth operations and mitigates delivery risk, particularly essential when a TRE platform needs to be set up within the tight timeframe commonly set by public funding requirements.

Best Practise 9: Future-proof with an infrastructure-agnostic provider

An infrastructure- and cloud-agnostic TRE provider protects against vendor lock-in and project continuity risks

With fierce market competition, infrastructure vendors can make it difficult for users to migrate analysis workflows and technology stacks to a competitor's service. Likewise, infrastructure vendors can entice users in by simplifying the joining process, for example, reducing initial computing costs and then increasing costs exponentially as users need to scale. A secure exit strategy is essential to mitigate future risks and dependencies on cloud or infrastructure providers.

Selecting a TRE provider that is infrastructure- and cloud-agnostic is essential to future-proof a TRE against vendor lock-in and project continuity risk.[25] The cloud/HPC environment account should be created in your organisation's name to ensure continuity. It's also important to compare how your TRE technical requirements can be met with different vendors both now and in the future. The selected TRE provider should support all major cloud service providers and on-premise HPC environments.

An agnostic provider allows Data custodians to explore different cloud or infrastructure providers or multi-cloud environments in future, all while retaining the full ongoing operation of the platform.

This also extends to pipelines and workflows within the TRE platform, these need to be in a portable, platform-agnostic and cloud-native format, that align with open-source standards. This ensures that TRE users can continue to use them with any other service provider and retain novel IP.

Best Practise 10: An open ecosystem extends TRE functionality

Build an open ecosystem platform to seamlessly integrate with community innovations and extend platform functionality

When it comes to adopting open-source software, TRE design has a large role to play. Whether open, closed or DIY, the type of platform heavily influences how open-source software is managed and used.[26]

Closed platforms, often referred to as blackboxes, may make use of some open-source components, but the majority of the source code is proprietary and unable to be modified. The inner workings of such solutions are concealed from end-users, resulting in a lack of auditability and limited integration with third-party applications.

Some organisations choose to build their digital research platform from scratch, the DIY platform approach. By building a TRE solution using open-source components, organisations can quickly become heavily reliant on the developer community for support. In addition, open-source software frequently lacks coding and testing standards, meaning the organisation's IT team are responsible for troubleshooting implementation and maintenance, impacting the stability, security and scalability of the final DIY solution.

The open platform approach, also known as an open ecosystem approach, is an intermediate solution. These platforms are usually distributed under a licensing agreement, while also offering end-users the ability to customise their environment with additional functionalities by integrating third-party applications, tools and data via APIs. TRE end-users have diverse needs for different tools and workflows to support their analyses, it is important that they are given the opportunity to choose their preferred software and integrate it seamlessly within the TRE's existing workflows. An open platform approach mitigates future open-source risks, ensuring a sustainable and stable ecosystem.

Best Practise 11: Sustainable infrastructure has transparent pricing

Select TRE providers with transparent pricing models to ensure platform sustainability as requirements and users grow

Choose TRE providers that will provide transparent pricing for infrastructure/cloud usage, this means disclosing the actual cloud cost charged by the original cloud provider.

Cloud pricing should not be more expensive than the published pricing on the infrastructure provider's website. In addition, the cloud cost should be passing through the relevant reseller, government or public sector discounts achieved, such as the One Government Value Agreement applicable to AWS public sector clients.[27] The business model of a number of SaaS TRE providers generates funding by routinely marking-up computational and storage costs by between 50%-350%. In addition, kickbacks and hidden discounts from cloud vendors are not disclosed, meaning that both the taxpayer, as well as researchers, have to pay for higher cloud costs than necessary.

This emphasises the need to closely evaluate computational and storage costs with each provider considered and any pricing evaluation must be accompanied by the full disclosure of the official underlying cloud cost.

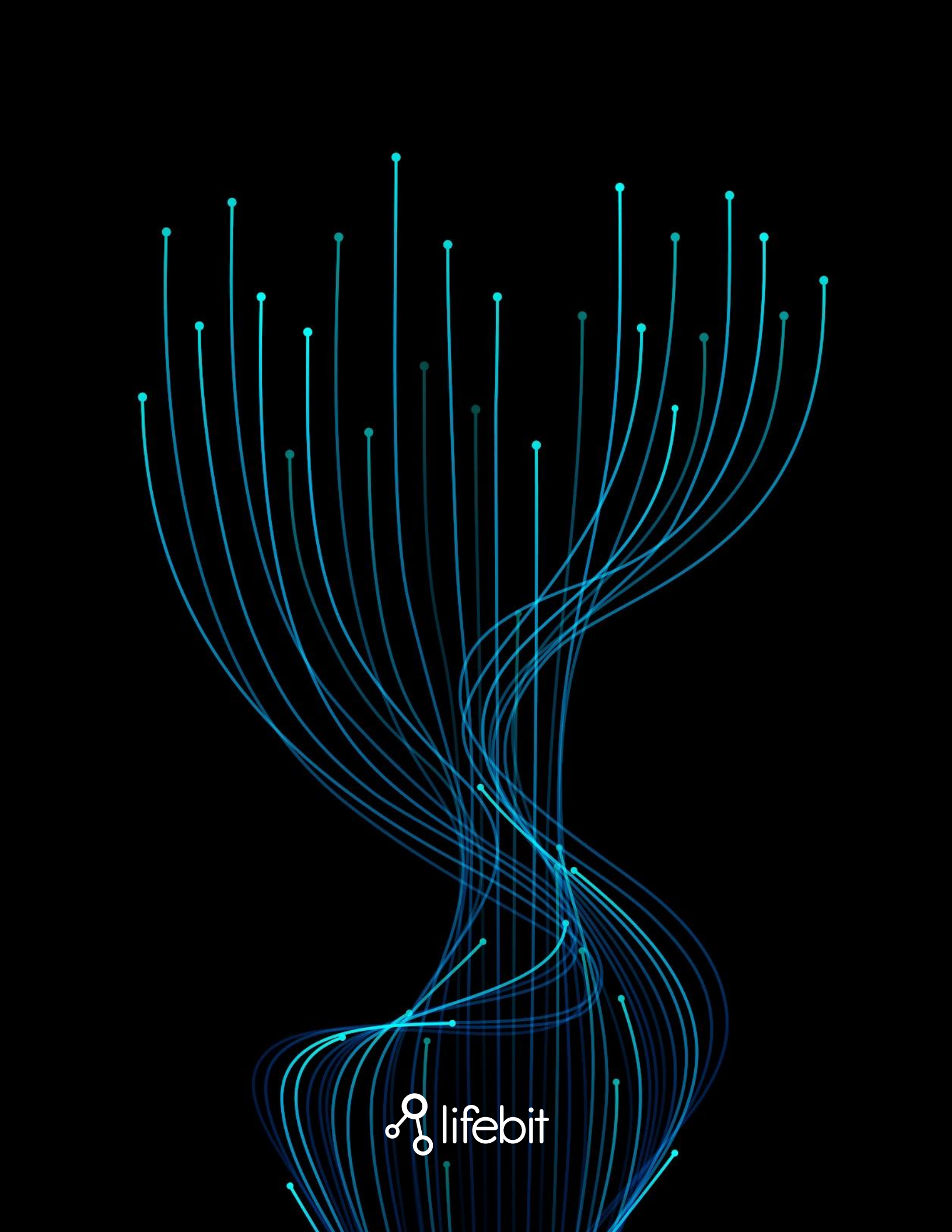
Concluding Remarks

TREs represent a sustainable and secure long-term solution for managing and making use of big data. As TREs continue to be implemented in organisations around the globe, recent policy work in the United Kingdom indicates a shift toward a system of accrediting TREs to a set of strict guidelines.[28] We recommend following these developments closely to ensure your TRE is not only meeting strict security requirements and industry standards but is also interoperable with other TREs.

Questions on procuring or building a Trusted Research Environment?
Contact us at **hello@lifebit.ai**

lifebit

1. Lifebit Awarded A Four-Year Contract for Hong Kong's Genome Project, Lifebit blog, May 2021.
2. Trusted Research Environment service for England, NHS Digital, February 2022.
3. Genomics England launches next-generation research platform central to UK COVID-19 response, Genomics England, June 2020.
4. Lifebit Counting on New UK Partnerships to Develop, Validate Federated Data Model, February 2022.
5. Trusted Research Environments, Health Data Research UK, accessed February 2022.
6. Virtual File System Overview, IBM, accessed February 2022.
7. Federated Computing Will Shape the Future of Computing, Information Week, November 2021.
8. International federation of genomic medicine databases using GA4GH standards, Cell Genomics, November 2021.
9. Building and sustaining collaborative platforms in genomics and biobanks for health innovation, OECD, March 2021.
10. CanDIG: Federated network across Canada for multi-omic and health data discovery and analysis, Cell Genomics, November 2021.
11. Boehringer Ingelheim and Lifebit announce partnership to capture transformational value of health data, Boehringer Ingelheim, March 2022.
12. ISO/IEC 27001 INFORMATION SECURITY MANAGEMENT, ISO, accessed February 2022.
13. About Cyber Essentials, National Cyber Security Centre, accessed February 2022.
14. Thousands of patients hit by NHS data breaches, Independent, July 2021.
15. Google reportedly mining millions of Americans personal health data, CBS News, November 2019.
16. The Anatomy Of Personal Data Sovereignty, Forbes, May 2021.
17. Patient and Public Involvement and Engagement, Health Data Research UK, accessed February 2022.
18. What is the Five Safes framework? UK Data Service, accessed February 2022.
19. Building Trusted Research Environments - Principles and Best Practices: Towards TRE ecosystems, UK Health Data Research Alliance, & NHSX, December 2021.
20. GA4GH: International policies and standards for data sharing across genomic research and healthcare, Cell Genomics, November 2021.
21. What Is FHIR®? The Office of the National Coordinator for Health Information Technology, accessed February 2022.
22. OMOP Common Data Model, Observational Health Data Sciences and Informatics, accessed February 2022.
23. The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, March 2016.
24. Data Utility Evaluation, Health Data Research UK, accessed February 2022.
25. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective, Journal of Cloud Computing, April 2016.
26. Open source software & its associated risks in organisational use, Lifebit, 2021.
27. One Government Value Agreement: Accelerating cloud adoption and innovation across UK government, AWS Public Sector Blog, November 2020.
28. Life Sciences Vision, Department for Business, Energy & Industrial Strategy and Office for Life Sciences, UK Government, July 2021.


lifebit