Versatile Trusted Research Environments: An approach for Switzerland

Owen Appleton[1*], Shubham Kapoor[1*], Sergio Maffioletti[4], Gregory Shomo[3], Anna Wiegand[4], Sofia Georgakopulou[3], Sudershan Lakshmanan Thirunavukkarasu[3], Christophe Dessimoz[1], Roberto Fabbretti[5], Thomas Geiger[2], Bernd Rinn[4] and Thierry Sengstag[1,3]

1 SIB Swiss Institute of Bioinformatics, 2 Swiss Academy of Medical Sciences, 3 University of Basel, 4 ETH Zurich, 5 University of Lausanne. * Corresponding authors

Table of Contents

Executive Summary

Widespread adoption of versatile Trusted Research Environments (TREs) is essential for conducting research analysis involving sensitive personal data, which plays a pivotal role in advancing research across healthcare, social sciences, and humanities in order to meet societal challenges.

Despite the immense potential of sensitive data analysis, there are substantial risks linked to its use as well as increasing regulatory requirements and scrutiny. Unauthorised access, disclosure, or misuse of this data can have severe consequences for individuals, potentially leading to harm, discrimination, or erosion of trust in the research ecosystem. It may also lead to significant legal consequences for individuals and organisations. These concerns are further amplified by the increasing adoption of distributed computing and cloud technologies, which, while offering benefits, introduce new security challenges.

Based on experience supporting research with sensitive personal data in Switzerland, versatile TREs are the best available solution for addressing the inherent tension between facilitating research with sensitive data and mitigating the risks associated with its use. The BioMedIT TRE has been in production since 2018, bringing together ETH Zurich, the Swiss Institute of Bioinformatics, the University of Basel and the University of Lausanne to deliver a federated, versatile TRE service.

Versatile TREs cater to the diverse and evolving needs of the research community at a reasonable cost, offering flexible and adaptable environments that can accommodate a wide range of research projects. This adaptability is essential in academic research, where the specific data processing requirements may not be

fully known in advance. Relying on less secure alternatives, such as researchers processing sensitive data on personal computers or on less integrated systems poses unacceptable risks to both individuals and institutions. Based on experience delivering TREs for research, we recommend the following:

- **Promote TRE Adoption:** Organisations should actively encourage the use of TREs for research involving sensitive personal data, potentially through policy mandates.
- **Embrace Collaboration:** Building and operating TREs requires close collaboration between data providers, researchers, technology experts, legal professionals, and ethical review boards.
- **Prioritise Federation:** National efforts should focus on federating TREs to avoid duplication, leverage shared resources, and promote harmonised practices.
- **Ensure Full-Stack Management:** TRE security and compliance require a comprehensive management approach that spans from underlying infrastructure to user engagement and usability.

Experience delivering TREs in Switzerland suggests that beyond technical considerations, building and ensuring trust with the full set of stakeholders, from the public and data subjects through to hospitals, research organisations and funders is perhaps the most important factor in delivering a TRE network.

1. Building trust in the data-driven research landscape

1.1 Benefitting from sensitive personal data

The increased availability of data in recent years has empowered a massive growth in research opportunities in the fields of health sciences, social sciences and humanities. Some of the greatest opportunities come from leveraging sensitive personal data, which can both benefit the individuals involved and contribute significantly to society in general.

For the general processing of data, numerous providers offer infrastructure to allow exploitation of the increased data available, and many of these have been designed to cope with the additional demands on processing *personal- data*. This capability has been driven by the demands of GDPR[1] in Europe and New Federal Act on Data Protection[2] (nFADP) in Switzerland, which regulates personally identifiable data, such as name, email address, etc.

Beyond this, there is a class of *sensitive personal data* – including health, socioeconomic, and financial data – that is also now available for research. These go beyond personally identifiable information and carry a high risk of harm for the individuals if disclosed without consent. They require a greater level of protection, as the impact of any disclosure is so much higher for the individual, both in terms of harm to them and in terms of discouraging others from making their own sensitive personal data available for research. Elevated protection is required due to legislation in many countries, including in Switzerland, and due to ethical concerns.

Protection for sensitive personal data must operate at a different level than simple personal data. Rather than offering reasonable protection in each discrete area or level that data touches, it must be protected in a more holistic way: through a Trusted Research Environment (TRE).

1.2 The need for TREs

Sensitive personal data offers great opportunities but presents great risks at the same time. Highly sensitive data tends to originate from individuals, or through organisations which collect it, for example hospitals or

[1] http://data.europa.eu/eli/reg/2016/679/2016-05-04

[2] https://www.fedlex.admin.ch/eli/cc/2022/491/en

health services. In both cases, there has been much healthy scepticism of requests to access the data by researchers from the academic or commercial sectors. There are justifiable fears that data may be misused in ways which damage society, or directly affect the individual data subjects whose data is used. A typical example might be insurance companies basing fees to individuals on their likelihood of disease, making health insurance unaffordable and functionally unavailable for those with complex medical situations.

Early work collecting sensitive personal data typically restricted it to controlled locations, but this restricted its impact and use as well. There is now a desire to mobilise this data and combine it in ways that allow for new forms of research, with clear benefits to individuals and society, but in doing so we must still protect the individual data subjects.

TREs should exist in order to provide safe spaces where sensitive data can be securely analysed and benefited from. These safe spaces employ state of art information security measures and enhanced auditing to avoid leakage of personal data which could do harm to data subjects and thus erode their confidence in sharing personal data for research. From TREs, results of analysis can be safely exported while keeping the source data confidential and protected. Constructing such systems is complex, because they must provide good utility and ease of use for researchers while still protecting the data subjects. They typically must be remotely accessible, as they are highly expensive and not many can be built, but this brings new security risks in communicating with them. They must be supported by applicable legal agreements, though many of the stakeholders of such systems (including data subjects and researchers) will lack the legal knowledge to understand the details of the agreements.

Operating TREs will always involve complex balances between security, safety for data subjects, usability and cost, and will typically require specialist support in legal and contractual issues. They must provide a secure, controlled platform that facilitates the safe use of sensitive data for academic research. It ensures data privacy, security, and compliance with regulatory requirements, enabling researchers to conduct high-quality research while protecting sensitive information.

1.3 Whose Trust are we seeking?

Reflecting on the above, we see that the trust we build through creating and operating TREs should and must focus on three core groups. We must be:

Trusted by data providers and subjects: The need for TREs is predicated on the availability of sensitive personal data, and the preparedness of individuals or groups holding such data to share it for research. As a first priority, it is crucial that TRE operators build trust with these groups. We must be sensitive to their needs and concerns, deploy best practices to ensure the security of their data, be transparent with them on risks and in case of incidents, and work with them to encourage and enable them to make their data available.

Trusted by research organisations: While access to sensitive personal data opens new and exciting opportunities for scientific and other research advancement, risks are not limited to the data providers and subjects. Processing data received from a provider includes risks as well. Security failures and disclosure of sensitive personal data are potentially very harmful for data subjects but can also have profound consequences for data controllers and processors. These may be legal as well as financial, if they have not exercised due diligence. It is crucial to build trust with the research organisations to ensure they are confident in operating or using TREs.

Trusted by the public and wider national or regional communities: Apart from trust of the specific data subjects, broader community trust of TREs is required to ensure individuals continue to provide their data for research. Any significant data breaches would undermine public trust for all TRE operators. Equally, trust by funding agencies, regulators and government is needed to continue the provision of TRE services, and the continuing evolution of those services to match the changing needs of research communities.

2. Defining and characterising TREs

Sensitive personal data has long been collected by, for instance, medical and healthcare organisations or by tax authorities. This data has been analysed and research has been performed on it in many contexts, but the general pattern of such analysis was to perform it within an infrastructure or network which is separated from public networks. These have included for instance within the strictly controlled IT environments of a hospital, or within private high performance computing clusters operated by national authorities.

A move to distributed computing and remote services such as cloud computing is leading to a data economy where data is created, aggregated from multiple sources, analysed and leveraged in ever greater amounts, often remotely, but this breaks the existing pattern of analysis for sensitive personal data. While security has been a major concern for digital services containing personal data, platforms for research data analysis did not initially consider the elevated needs for sensitive personal data.

TRE is one of several terms used in recent years to describe systems intended to process such sensitive personal data. All terms try to describe systems able to safely extract insights from the source data without compromising the data subjects who provided it. In the absence of clear consensus definitions of TREs, we offer a working definition based on eight years of experience operating BioMedIT, the Swiss national TRE for personal health data.

2.1 Elements of a Trusted Research Environment for processing sensitive personal data

Beyond looking at specific technologies and systems, we can conceptualise TREs in various ways, both by area of activity and by organisational layer which delivers services.

At a fundamental level, processing sensitive personal data relies on four main pillars representing different areas of activity, shown in the figure below:

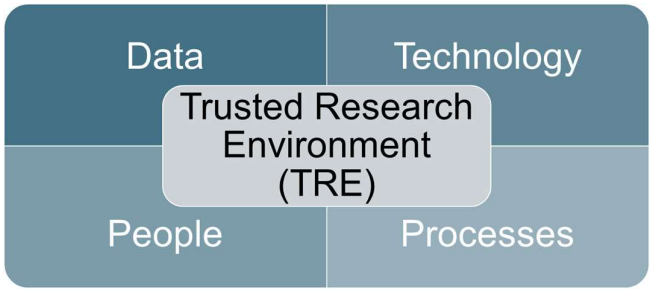

Figure 1 Pillars of a safe, effective Trusted Research Environment

Data: The platform must be able to receive, store and process sensitive personal data. Data must also be in a form that allows for its use and reuse, through data harmonisation and interoperability. This also requires significant work to build trust with data subjects, and with the initial data controllers who manage it, for example hospitals where personal medical data is collected. In many cases this also leads to deep engagement with data providers in other areas (such as working to develop their technology, people and internal processes).

Technology: The technology underlying a safe platform need not be highly specialised or unusual, instead it is more about how it is selected, designed, managed and delivered. Infrastructure components must be constructed with security and safety in mind and must be able to be adapted to the elevated needs of a safe research platform. Hence, off the shelf or generic, one-size fits all solutions will be difficult, and many technological elements will need to be customised to higher security use cases. Technologies must be integrated in such a way that their interfaces do not open security risks, so it is crucial that all technological components are managed in a coordinated way. Technology used to support TREs and how different technology components are integrated with each other, must evolve and change constantly over the course of

the lifetime of the TRE to better reflect changes in use cases, security requirements and compliance requirements.

People: Technology can help protect sensitive data, but the overall security of a general research platform also requires appropriate and security-conscious behaviour by those who use and provide services to ensure the security of the platform. This means that extensive training must be offered to ensure that all participants are aware of the legal and regulatory situation, what they are permitted or not permitted to do, how to use the relevant tools and how to otherwise manage risks to sensitive personal data. Creating this training requires a wide range of expertise in technical, legal, ethical and organisational topics.

Processes: Provision of a platform to allow research with sensitive data involves many different parties, from data providing organisations, through the organisation or federation offering the platform, through to the research groups who use the platform. Processes spanning multiple organisations are often complex and require careful design, deployment and monitoring. All stakeholders must subscribe to and use these processes, meaning that all participating organisations must actively buy into the overall platform. These are supported by structured legal agreements to ensure that processing of data is ethical, lawful and appropriate.

The above four pillars are key aspects of a TRE and are involved in all levels of services. We can also look at the elements of delivering TREs from a different perspective, as the set of layers we build up from physical infrastructure up to user support and consultancy.

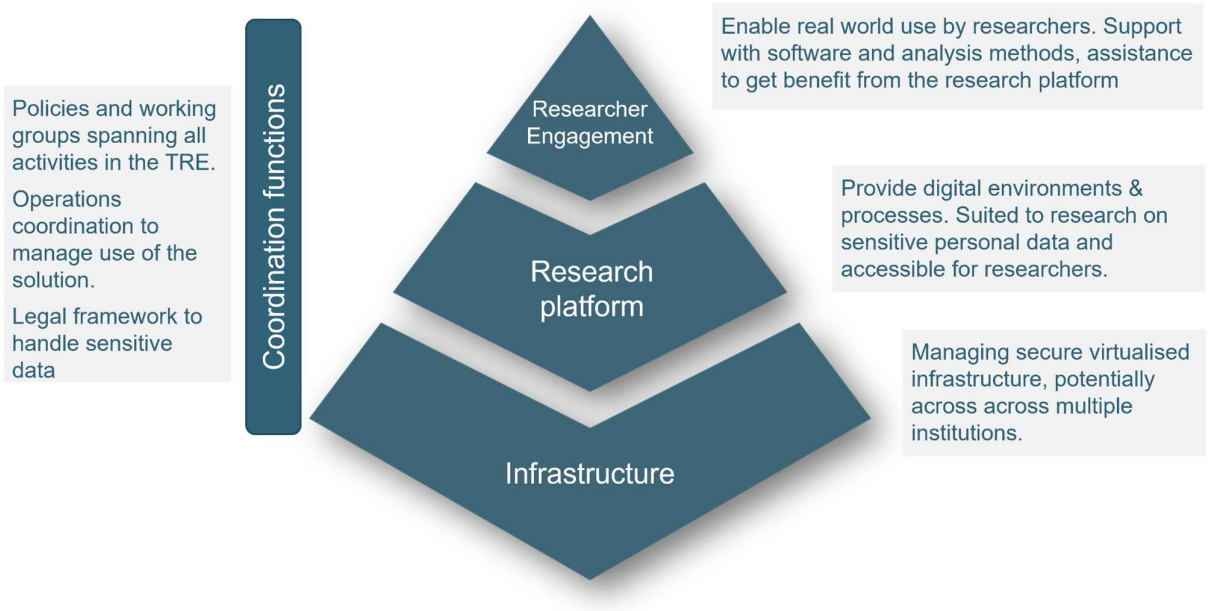

Figure 2 Layers in the delivery of TRE services

These layers can be seen as analogous to typical service delivery models such as Infrastructure as a Service, Platform as a Service and Software as a Service / Consultancy and support. Together these layers form the services needed to deliver a TRE, but in order to do so they are accompanied by coordination functions which span across the layers. This integrated coordination facilitates the interaction of the layers, delivery of high-quality services and maintaining security of the platform and the security of the sensitive personal data it hosts.

2.2 Types of platforms for handling sensitive personal data

We see four basic types of platforms for sensitive personal data based on experience deploying them in Switzerland: These are:

- **Single purpose Trusted Research Environments:** secure environments dedicated to a single use case, such as a single analysis pipeline or preset selection of tools for a specific domain. Typically, very secure as designed end-to-end, but can only perform a pre-planned set of functions which makes it difficult or expensive to support the varied usages needed for research.
- **Local processing:** Downloading the sensitive personal data to a local or personal machine for local processing. Convenient but limited power. Avoids security risks of networked services, but requires the researcher to fully secure their system to be compliant with the law and ethical standards.
- **Secure cloud infrastructure:** Virtualized remote processing and storage systems with elevated infrastructure-level security measures, such as various forms of encryption (e.g. in transit and at rest). Scalable but does not offer full stack of TRE service layers and may not be easy to manage security between secure cloud infrastructure provider and service provider.
- **Versatile Trusted Research Environments:** Secure environments comprising the infrastructure and platform layers, coordinated to provide comprehensive security. Usable for any research type depending on the software, containers or pipelines brought to it. Complex to create as it must accommodate a very wide range of use-cases, but very flexible and domain-agnostic.

Below we compare these four types of system, based on the split of responsibilities between the customer and what service provider could offer through the capabilities of the system:

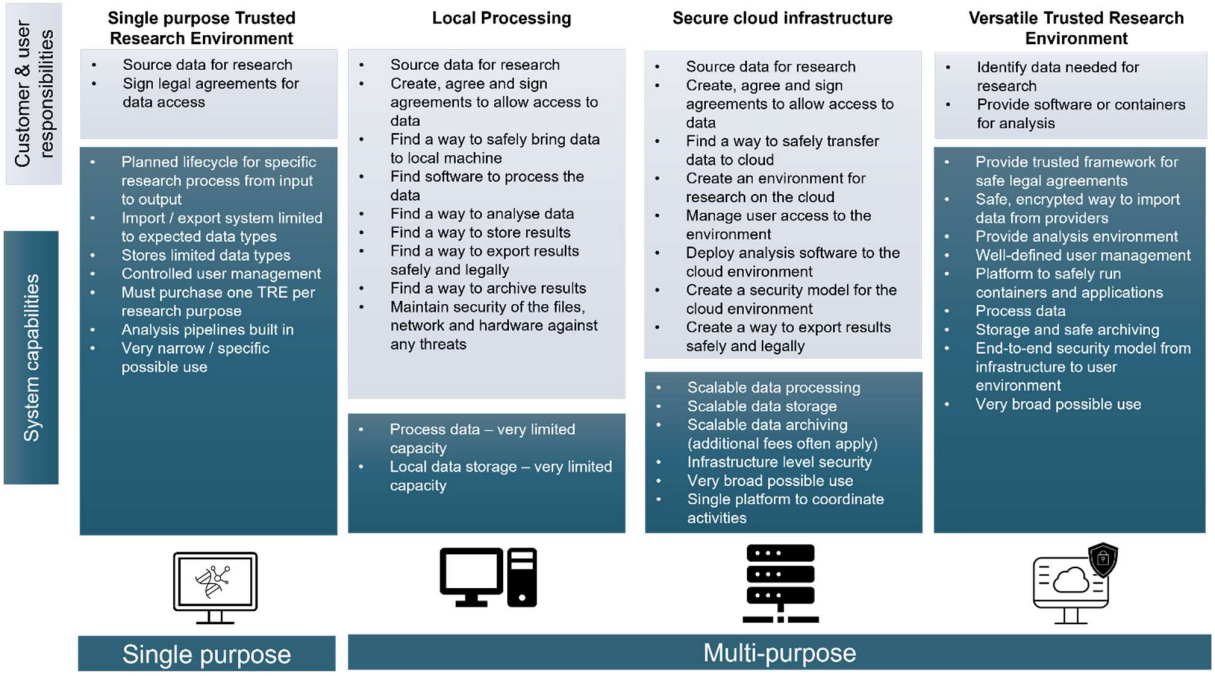

Figure 3 Typical responsibilities in different platform types for sensitive personal data handling

As can be seen, Versatile TREs offer the greatest level of flexibility, support and most comprehensive security, leaving as little as possible the responsibility of the user so they can concentrate on their research, while still allowing for a full range of different usages.

2.3 Versatile TREs as a solution for research

Versatile TREs are the type commonly built by research organisations. While it is possible to build secure online services to serve specific single predetermined purposes, where for instance a specific format of data is provided to a service, which is processed in a predetermined way, generating a known type of result. Such single purpose systems are easier to secure for multiple reasons. They tend to operate in a way that limits the interaction of users only to that needed to serve the predetermined purpose, and in theory all other avenues presenting security risks can be analysed and closed. These types of services tend to be better supported in the commercial sector, serving known and consistent needs of customers, but they do not fit research needs, nor those of the academic community.

Academic and nationally funded research communities generally have limited budgets which are needed to serve multiple types of research, and providing single purpose systems for all possible uses is unrealistic. Equally, in many forms of research one cannot know in advance exactly what sort of processing is required.

For both economic reasons and to allow for innovative research, research communities need access to Versatile TREs, which can adapt to many different types of research, offering sufficient features to support the research and balancing that with sufficient security to prevent disclosure of data and protect data subjects.

2.4 Properties of a Versatile TRE

A Versatile TRE provides a remote system for interacting with sensitive personal data with the following general properties:

TRE activity Area	Properties
Secure import of sensitive personal data from data providers	- Data should be categorised by risk-level by providers, according to a schema provided by the TRE operators. - Data from providers should be de-identified to the greatest extent possible - A sceptical approach to de-identification should be taken, assuming that pseudonymised data is not 'safe' due to both the possible existence of keys to the data or re-identification through combination of the data with other sources. - Import should include file encryption and transport encryption as well as cryptographic signing of data packages. - Data import must be based on sound agreements, including ethical approval for the given research, and legal agreements covering data transfer, use and processing. - TRE operators should engage with and closely collaborate with data providers in order to support their ability to deliver high quality, harmonised, de-identified data in a safe and secure manner. - Data import should include logging and provenance both to ensure the legitimacy of the source of data and an audit trail for legal compliance.
Secure access and user management	- Users are managed through a secure platform where they can be associated with research projects on the TRE to which they have access. - TRE operators support user management and verification, but responsibility and accountability remains with the customer - User identities are, where possible, connected to a wider AAI federation including linkage to institutional identities - Users access the TRE through a secure remote system which follows the principles of purpose limitation and least privilege, and prevent unmonitored exfiltration of data. - User access to the TRE should be encrypted and logged for auditability and legal compliance

Secure handling of data within the TRE	- The TRE provider offers security at the boundary of the TRE, letting the customers manage what goes on within it - No service should cross the perimeter of the TRE without prior security assessment, permission of the TRE operator and sufficient monitoring being in place. - Users should make all efforts to use only secure and well tested applications and containers within the TRE - Users may only handle data, including accessing, assessing, combining and processing it, in a way that is compatible with the agreements or conditions under which they were granted access to it
Secure export of permitted content	- In general, only results may be exported from the TRE. Results must be extracted in such a way through e.g. aggregation or transformation, as to make sure that data subjects are not identifiable. - Where source data is exported, it must be permitted under the agreement under which it was granted to the TRE, or based on permission and documented agreement with the initial data controller who provided it. - Export is via a managed system that provides monitoring and logging of any export, including user identity. - TRE operators support safe export, but responsibility and accountability remain with the customer
Education, training and support to ensure sufficient resources, knowledge and encouragement to act within the bounds of the law, and of ethical research	- All staff with access to the TRE (including TRE operators and customers and users) should be required to take sufficient training to understand the risks of TRE use and their responsibilities in maintaining security of data. - Provision of minimum standards or shared templates for legal and ethical agreements for the transfer processing and use of data that are compatible with the relevant legal system where the TRE operates. - Sufficient documentation should be provided to customers to let them understand and safely use the TRE
Secure & effective management and coordination of the elements of the TRE in a holistic way	- Multiple elements are required to deliver the TRE, including Secure infrastructure, the TRE service platform, and research user support. - Management of the TRE must span these elements and involve active engagement of all layers to maintain security, whether or not they are provided by the same teams or organisations. - The TRE operator should, where possible, offer specialist additional support in areas where customers may struggle to maintain local knowledge, such as legal agreements. - The TRE must support secure operations, including the platform being scanned for vulnerabilities, patches and updates being promptly applied, security incidents being handled quickly and carefully, and security reviews of key elements or customer projects being carried out.

This description is not intended to be exhaustive or perfect, but gives a strong indication of what can be expected of and required in a Versatile TRE.

3. TREs in the Swiss landscape

Work on TREs in Switzerland began in earnest with the Swiss Personalised Health Network[3] (SPHN), launched in 2017 to drive research using personal health data and led by the Swiss Academy of Medical Sciences[4] and

[3] https://sphn.ch/

[4] https://www.samw.ch/en.html

the Swiss Institute of Bioinformatics[5]. SPHN has supported more than 130 collaborative projects and has sourced data from 29 data providing organisations.

To support and enable the projects which it funded, SPHN launched BioMedIT[6], the Swiss national TRE network for sensitive personal health data. BioMedIT federates three major Swiss research universities who each have secure cloud infrastructures: the University of Basel[7], the ETH Zurich[8] and the University of Lausanne[9]. Together, and with central coordination from the Data Coordination Centre at the Swiss Institute of Bioinformatics, these form the BioMedIT TRE network. BioMedIT also engages heavily with the major national research hospitals to support and enable them to collate, harmonise, curate and export data into the TRE to support research projects.

This creates a federated and versatile TRE network, which can support a huge variety of research use cases with a good balance of usability for researchers and security and trust with data providing organisations and data subjects. Building BioMedIT has created a wealth of experience in building up multisite and federated TREs, which could be of interest to the wider European TRE community. This paper shares some of the general conclusions about the nature of TREs, but the community has also generated considerable experience in practically building TREs which it is happy to share.

4. Conclusions

This paper is intended to build on experience on building TREs in order to both provide a coherent view of what TREs are, and also support others in building up TREs in an effective way. Based on experience from BioMedIT, we summarise our experience and conclusions in the following areas:

We need TREs, and even more we need researchers to use TREs.

Sensitive personal data is a fact of modern research. We live in a data-rich society, and researchers will find a way to use this data. We need TREs to support the use of this data and allow us to benefit from it. However even more we need to ensure that researchers use the TREs to process this data type. A greater risk than not benefitting from sensitive personal data is that, not being able to or prepared to use TREs, researchers process this data on unsecure systems or their laptops. Any data breach from insecure processing of sensitive personal data would have profound negative consequences for the organisations involved. Apart from legal or financial consequences, undermining trust of data subjects and the public could lead to this type of data no longer being available to support research. To avoid this, organisations should not only support and operate TREs, but also enforce their use for researchers using sensitive personal data.

Co-develop TREs alongside research and data-providing organisations.

Developing TREs is an example of a research IT challenge. Research IT, in contrast to more corporate IT that provides stable services such as HR systems and email services, is a more dynamic environment. Research IT typically deals with a greater turnover of services which are developed and integrated in-house using new and cutting-edge technology, but with a lower level of service guarantee due to their more dynamic character. They also are more sensitive to user needs, as the user needs change more rapidly in this space, and requires close relationships with and trust of the research communities it supports.

TREs involve considerable technical development, but are more than that. They are complex collaborations between data subjects, data providing organisations, platform providers, research organisations,

[5] https://www.sib.swiss/

[6] https://www.biomedit.ch/

[7] https://www.unibas.ch/en.html

[8] https://ethz.ch/en.html

[9] https://www.unil.ch/index.html

regional/national authorities, and the general public. All of these must be appropriately engaged in order for them to be a success. Building TREs also requires, from the operators, a wide range of skills including developers and solution engineers, system administrators, customer relationship management staff, security officers, CISOs and legal officers who must work together to deliver services. It needs both the internal collaboration among providers to bring in all required capacities, and the collaboration with researchers to create trust in services and providers.

Delivering TREs in this context involves the need to track, participate in and predict needs in the evolving landscape of sensitive health data in research, and this is best achieved by becoming an active participant in the ecosystem, rather than standing outside it.

Federate, do not duplicate.

TREs are complex and expensive to develop and deliver. Having dozens of specialised TREs provides poor national value as it tends to reinvent the wheel in each case, and makes TREs over-tuned to their initial use cases. It is more effective to federate versatile TREs into a shared structure within a country. It is also more realistic in terms of cost, as it is unlikely we can support dedicated TREs for every use-case, and it is more realistic to operate versatile TREs than can adapt to changing user needs.

Federating TREs means harmonising their approaches, practices, procedures and tools, as well as the data formats they support. It means sharing best practices and solutions to reduce costs and increase effectiveness, as well as offering more consistent experience for customers across the nation. In some cases, this means a single centrally funded TRE for research, in others a relatively small set of TREs which cooperate to form a national community. This also facilitates a federation of approaches, tools and policies across national borders, as while data should generally not travel to other countries, experience and approaches can and should.

Manage the full stack, from infrastructure up to user engagement.

TREs must support processing of sensitive personal data, with the responsibilities and liabilities which go alongside this. These services are made up of many layers, from underlying infrastructure through a research platform to higher level services, support and user engagement. If management of the TRE restricts itself to a single or subset of layers, this presents significant risks through providing gaps between the layers which can be exploited. Both in order to prevent breach of sensitive personal data and to show due diligence in preventing this, delivery and security of services must be coordinated from the ground up to the most abstract services. Legal and ethical aspects must equally be managed from top to bottom to ensure all actions are legal and ethically defensible.

This is easiest in single organisation TREs, but this is not in line with the need to federate and cooperate to support communities. To achieve this in multi-organisational TREs requires a very high degree of trust and openness between participating organisations, and a preparedness to share and amend their practices to increase coherence across the network. This is difficult but possible in collaboration between e.g. research organisations who would participate in such open collaboration in order to fulfil their purposes to support research. Equally this may prove more difficult in involving commercial providers for some layers, where the need for openness is harder to justify and reduces economies of scale as it requires services to be more tailored to each customer rather than be highly commoditized.

Whether through operation by a single organisation, or through openness and tight cross-organisational coordination, TREs should be coordinated across the full stack to ensure safe, secure and trustworthy services.

Correspondence and licence

Correspondence on this paper and enquiries on the Swiss deployment of TREs should be directed to owen.appleton@sib.swiss and shubham.kapoor@sib.swiss.