SaxInstitute

Secure Data Access Platform Review

Summary of requirements, assessments and enhancements

October 2022

Enquiries regarding this report may be directed to the:

Sax Institute
www.saxinstitute.org.au
sure-admin@saxinstitute.org.au
Phone: +61 2 91889500

SURE is operated by the Sax Institute and is supported by the Population Health Research Network which is an initiative of the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS), and by the NSW Government.

Contents

Background

The Sax Institute was funded by the PHRN to review the current state of secure data access platforms for linked data research, including a review of user requirements and an assessment of available platform technologies.

The results of the work and the enhancements made to the Secure Unified Research environment as a result of the review are detailed below.

Report on system requirements, use cases and capabilities

The *Report on system requirements use cases and capabilities* was informed by consultations with 50 data custodians, ethicists and users of SURE. The consultations were assisted by Dr Michael Frommer and conducted through semi-structured interviews with key stakeholders, as well as small focus groups within organisations. Scripted questions were tailored to each organisation and informant for relevance and likely interests. Informants were advised when invited to participate, and again at the start of each consultation, that the subject of discussion would be secure remote-access computing environments generally, rather than just SURE.

The key findings from the Report were:

- That secure remote access computing environments should offer services that are flexible to reflect the research life cycle, giving users greater flexibility through shorter minimum periods of access, and self-service access for 'burst' computing as a high priority for any system
- That secure remote access computing environments need pricing to be realistic, acceptable and flexible to better align with the diverse nature of research funding arrangements and organisational structures
- That secure remote access computing environments should consider implementing new privacy and security controls and reporting to assist data custodians and ethics committees to approve trusted-user models of access to enduring linkage collections in the future, and to respond to new regulatory requirements such as the Data Availability and Transparency Act.
- Several data service providers emphasised the desirability of having a choice of multiple environments. Multiple environments were seen as beneficial to cater to the diverse needs of researchers, data custodians and data service providers. As the use of multiple environments increases, it will be important to ensure that each environment is complementary to the others, rather than developing redundant capabilities that are already available elsewhere.

Assessment of platform technologies to deliver enhanced services

The Architecture Practice (TAP) was engaged by the Sax Institute to produce an independent assessment of existing secure e-research platforms that could meet researcher and data custodian needs. The existing secure e-research platforms that were assessed included:

- Secure Unified Research Environment (SURE) – Sax Institute
- E-Research Institutional Cloud Architecture (ERICA) – University of NSW
- Secure e-Research Platform (SeRP UK) – University of Swansea, UK.

Platform assessment

TAP assessed the capacity of each of the three platforms to meet the key requirements of data custodians and the research community. The requirements against which the platforms were assessed were:

- Realistic and acceptable pricing
- Flexible computing and storage capability
- Ability to support high-performance computing and large data sets
- Flexibility with supporting new and innovative approaches to research.

Secure Unified Research Environment (SURE) – Sax Institute

- **Realistic and Acceptable Pricing –** SURE is priced realistically and at an acceptable level and provides certainty over fees charged to researchers and protection against “unplanned expenses” in its current approach.
- **Flexible Computing and Storage Capability –** Researchers are able to request additional computing power and storage via a service request, which is typically acted on quickly. There is currently no provision for this to be user controlled with automated scale up and scale down.
- **Support for HPC and Large Date Sets –** SURE does not provide support for HPC and large data sets that requires distributed and / or parallel processing to manage. However, this is mitigated to a certain extent by the introduction of the hybrid-cloud approach, which is still early days and will continue to evolve to support this key requirement. In addition, GPU nodes are being procured.
- **Flexibility and Support for Innovation –** Currently, the ability to work with new and innovative research approaches is not available in SURE as the access to tools such as AI and ML is not supported. However, this is mitigated by the introduction of the hybrid-cloud approach, allowing

for researchers to leverage capability within AWS. Note that this is still early days and is very much a live project and will rapidly evolve to support this key requirement.

Key findings

In considering the strengths of the current SURE offering, and taking into consideration the move to support a hybrid-cloud model with a linked AWS tenancy as well as the planned procurement of GPU nodes to enhance the existing SURE ecosystem, the following are the research and data custodian use-cases best addressed using SURE:

- Where there is a preference for a full-service approach to delivery of the research environment, where all aspects are maintained by the Sax Institute with support provide by the SURE Admin help desk
- Where strong security and data governance by an independent third-party is required, supported by fully maintained accreditations and processes
- Where there is a desire for certainty and consistency in pricing that is also realistic
- Where rigid and tight control over the availability of and access to data is required by data custodians
- Where neutrality and lack of alignment to one or more specific research institutions is considered of benefit.

E-Research Institutional Computing Architecture (ERICA) – UNSW

- **Realistic and Acceptable Pricing –** The pricing for ERICA “bundles” is realistic and acceptable. However, as there is an element of “pay for what you use” within the ERICA model, and no support for predictive modelling of usage patterns, there is scope for “unplanned expenses”.
- **Flexible Compute and Storage –** Researcher on-demand access to scale up and scale down of compute and storage is available within an ERICA “bundle” in a flexible manner. However, there is a limitation in that scaling up or down outside of the capacity of the bundle is not supported and will require a service request to be raised.
- **Support for HPC and Large Date Sets –** ERICA provides full support for HPC and large data sets that requires distributed and / or parallel processing to manage.
- **Flexibility and Support for Innovation –** The ability to work with new and innovative research approaches can be taken advantage of in ERICA as the access to tools such as AI and ML is available in the AWS tenancy. However, this capability is not provided in any of the standard ERICA bundles by default and is only available on request and for an additional cost.

Key findings

In considering the strengths of the current ERICA offering, the following are the areas best addressed using ERICA:

- Where easily managed access to HPC and GPUs is required
- Where strong security is required, supported by fully maintained accreditations and processes

- Where the ability to transfer, and share code is required
- Where there is a requirement for managing large data sets by parallel processing and distributed computing
- Where flexibility and user managed choice as to what software is included in a workspace is required
- Where access to advanced analytics, such as AI and ML is required in a flexible and manageable way
- Where the ability to take advantage of commercial public cloud benefits is required.

UK Secure e-Research Platform (SeRP UK) – University of Swansea

- **Realistic and Acceptable Pricing –** There has been no pricing information provided for SeRP UK at the time of completion of the assessment.
- **Flexible Compute and Storage –** On demand access to scale up and scale down of compute and storage with SeRP is available in some circumstances, depending on the method used to access the trusted research environment.
- **Support for HPC and Large Date Sets –** SeRP provides full support for HPC and large data sets that requires distributed and / or parallel processing to manage.
- **Flexibility and Support for Innovation –** The ability to work with new and innovative research approaches can be taken advantage of in SeRP as the access to tools such as AI and ML is available via the use of Jupyter Notebooks. In addition, the capability to integrate with imaging and genomics platforms supports this key requirement.

Key findings

In considering the strengths of the current SeRP offering, the following are the areas best addressed using SeRP:

- Where this a need for enduring linked data assets
- Where there is a requirement for managing large data sets by parallel processing and distributed computing
- Where access to advanced analytics, such as AI and ML is required in a flexible and manageable way
- Where the ability to transfer, and share code is required
- Where strong security is required, supported by fully maintained accreditations and processes
- Where easily managed access to HPC and GPUs is required
- Where the concept of a federated model for data access is under consideration and considered to be of value
- Where flexibility in access to the required level of compute and storage, only for the time required is important.

SURE enhancements delivered

In response to the feedback received in consultations for the *Report on system requirements, use cases and capabilities*, **t**he Sax Institute made enhancements to SURE in 2021 and 2022 to improve the experience of users.

Flexible computing resources

In October 2021, SURE launched new, higher-powered virtual machines at reduced fees in response to user requirements for greater flexibility. The new virtual machines are delivered within 24 hours of request and are available for as little as one month, giving users the flexibility to apply and pay for computing power only when needed.

The new virtual machine specifications are

- 2x Power: 32GB RAM, 4 CPU cores at $412/month
- 4x Power: 64GB RAM, 8 CPU cores at $652/month
- 8x Power: 128GB RAM, 16 CPU cores at $1,132/month.

At the same time, the Workspace setup fee which was previously $2,674, was abolished to lower the barrier to entry for users.

Deployment of AI Capability

In June 2022, SURE installed GPUs in its hardware stack to support the application of artificial intelligence and machine learning techniques by health and human services researchers. There are two projects that will begin using the GPUs for research in late 2022.

Secure Sandbox

SURE launched a Secure Sandbox capability in 2022 to make synthetic versions of actual datasets available to users without the need for lengthy ethics and data custodian approvals. The Secure Sandbox is being piloted between August and November 2022 with synthetic National Integrated Health Services Information (NIHSI) data, with other synthetic datasets to follow. It is expected to benefit researchers by giving them experience with the data prior to making an application to ethics committees and data custodians, and by letting them trial the development of code to be applied to the actual datasets once all approvals have been granted, significantly improving the timeliness and efficiency of research.

File access logging

While SURE has always kept logs of files entering or leaving the environment, it can now log each user's actions inside the environment. The logs contain information on users' access and actions with each file and provide an added layer of information security control in response to requirements identified by data custodians.

Visit: saxinstitute.org.au

Email: communications@saxinstitute.org.au

Phone: (02) 9188 91500