2021 IEEE 17th International Conference on eScience (eScience) | 978-1-6654-0361-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/eScience51609.2021.00028

SAMPRA: Scalable Analysis, Management, Protection of Research Artifacts

Patrick G. Bridges
University of New Mexico
patrickb@unm.edu

Zeinab Akhavan
University of New Mexico
zakhavan@unm.edu

Jonathan Wheeler
University of New Mexico
jwheel01@unm.edu

Hussein Al-Azzawi
University of New Mexico
azzawi@unm.edu

Orlando Albillar
University of New Mexico
duckula@unm.edu

Grace Faustino
University of New Mexico
gfaustin@unm.edu

***Abstract*—This paper describes SAMPRA, a framework for supporting effective research on sensitive data being deployed at the University of New Mexico. SAMPRA and its associated implementation are designed to support the needs of a diverse set of use-cases from researchers across different disciplines at UNM, including from clinical neurosciences, forensic anthropology, and community health. From these use-cases, we identified a set of common unaddressed demands when handling data with privacy/protection requirements, particularly collaborative research projects, interfacing with scientific instruments, and full-lifecycle management of sensitive data. To properly address and accelerate research projects with these needs, SAMPRA a) integrates privacy-preserving storage and data transfer systems with data-centric virtual environments, and b) supports effective researcher use of the system through active collaboration between local IT personnel, campus enterprise IT service providers, and campus data librarians by defining clear roles with associated personnel. By doing so, SAMPRA seeks to meet the needs of research on sensitive data across the entire data lifecycle and avoid the pitfalls of generic "one-size-fits-all" services.**

***Index Terms*—Controlled Unclassified Information; Virtualization; Cyber-infrastructure; High-performance Computing; Research Data Management**

I. Introduction

Modern academic research is becoming more complex, more interdisciplinary, and more data intensive, but the data management, data sharing, data security, and research compliance needs associated with academic research are also growing. This is particularly true of research involving sensitive data, often termed Controlled Unclassified Information (CUI). Unfortunately, these two trends are at odds with each other—interdisciplinary, collaborative research pushes toward open, flexible, customizable computing environments while research data security and compliance needs push toward carefully managed, uniform computing environments.

Academic cyber-infrastructure systems, particularly those related to protected and sensitive data, are frequently siloed and fragmented both physically and administratively. For example, research groups with specialized data analysis needs or complex regulatory and contractual protection requirements often purchase and self-administer local or cloud-based systems. While this provides researchers a more flexible environment, it is both more expensive than centrally-provided resources, and can be fragile and provide only ad hoc support for data protection and data management.

These tensions between flexible and carefully controlled environments apply to a range of technical, operational, and policy issues. For example, requiring that all data be deidentified prior to analysis so that it can be processed on open systems forecloses research that needs to use sensitive data (e.g. sensitive location data). Similarly, using an air-gapped system in a locked room simplifies many technical and operational controls and can provide *more* flexibility in how non-networked portions of the system are configured (e.g. what software is installed) but also reduces options for preserving or sharing data and makes the system more difficult to manage and maintain. Similarly, standardizing the operating system used on these systems can make them easier to maintain, but may prevent the execution of certain applications or successfully connecting the system to some scientific instruments.

This paper describes SAMPRA (Scalable Analysis, Management, and Protection of Research Artifacts), a system for handling sensitive data being deployed at the University of New Mexico that provides the technical, operational, and policy framework for supporting diverse research demands while meeting medium NIST 800-171 compliance requirements. The SAMPRA framework enables the decentralized creation and management of virtualized protected academic computing enclaves with hardware, software, and network connectivity customized to the needs of the research. These enclaves can then be customized to researcher's needs while still complying with institutional and regulatory requirements.

SAMPRA also facilitates collaboration between varied IT personnel, including institutional Information Technology personnel, data librarians, and research computing specialists. In addition, SAMPRA includes careful integration with existing institutional resources, helping professional system administrators and data librarians collaborate with researchers to meet their data analysis, management, sharing, and protection needs. Finally, SAMPRA is designed to be incrementally managed and grown.

978-1-6654-0361-0/21/$31.00 ©2021 IEEE
DOI 10.1109/eScience51609.2021.00028

The main contributions of our work are:

1) An identification and an analysis of the needs of a set of diverse research use-cases whose needs are not completely met by previous approaches to handling academic research;
2) A general private cloud-based virtualization architecture to support custom research by academic researchers on sensitive data throughout the research lifecycle;
3) Techniques for the integration of data management services into this environment;
4) Definition of a lifecycle and collaboration framework for research on sensitive data that fully integrates support personnel from disparate campus IT and research support organizations, including identifying the the roles, responsibilities, and workflows that are part of that lifecycle and;
5) An assessment of the resulting environment's ability to meet NIST-SP-800-171 Medium compliance requirements, including an identification of gaps both institutional and within the SAMPRA environment.

In the remainder of this paper, we describe these contributions. First, Section II provides background on the needs and challenges of supporting diverse academic research on sensitive data. This section includes a description of the use cases driving the development and implementation of SAMPRA at the University of New Mexico and broadly classifies challenges and corresponding system requirements as being primarily technical, operational, or policy based. Following this, Section III maps these requirements to a high level system architecture and workflow aligned with specific stages in the research and data lifecycle, and concretely defines the roles of research support personnel for enabling research on controlled unclassified data in this architecture. Sections IV and V then discuss the implementation status of SAMPRA, including the projects which the system has thus far supported, and the overall security assessment of the system, respectively. Finally, we conclude with a brief description of related work and future research directions.

II. Background

There are a wide range of challenges in creating, deploying, and operating computing systems that support research on sensitive data by diverse research communities. In this section, we outline the challenges encountered and tradeoffs employed by such systems, as determined and illustrated by specific use-cases we examined when designing SAMPRA.

A. Current Approaches to Research on Sensitive Data

a) Custom Systems: Many institutions handle research on sensitive data using one-off systems developed on a case-by-case basis and separately for each research project. An important advantage of this approach is that it lets the researcher and/or their IT support personnel customize the system to exactly meet the needs of the research (subject to any regulatory requirements). This is particularly important when using or interfacing with specialized hardware or software. In many cases, these systems are also air-gapped to ease institutional compliance burdens and enable further customization and specialization to the needs of the research the system is supporting.

This approach also creates many problems, however. Because these systems are not systematically managed, systematic auditing, backups, and integration with other institutional infrastructure are problematic at best, and nearly impossible if air-gapped. Such systems are also often managed by local administrators, researchers, and compliance personnel, an approach which often fails to leverage and build institutional expertise. The frequent lack of systematic, scalable controls in these systems also makes them harder to maintain and more vulnerable to compromise. Finally, it can also make these systems challenging to manage and grow as needs and research requirements change. These challenges can even tempt researchers to use personal laptops, cell phones, shared printers, and commercial cloud services to store and process sensitive data, potentially violating regulatory requirements.

b) Institutional Systems: Alternatively, a number of institutions provide centralized systems that systematically meet many of the technical and operational challenges that one-off systems face in handling controlled data. This includes, for example, export-controlled compute clusters and the ResVault virtual machine system at the University of Florida [1]. Such systems more sustainably and reliably meet compliance requirements for sensitive data as they can be more easily monitored and audited, and they fully leverage central institutional computing technical, operational, and policy infrastructure and expertise.

This approach, however, often provides a limited set of system configurations that may impede research and are difficult to integrate with custom hardware and software systems. In addition, their design can, either intentionally or unintentionally, limit or eliminate customization of the system by local system administrators and other research computing and data specialists.

c) Data Segregation: Finally, many research groups use a hybrid of these two approaches based on dividing acquired data into sensitive and non-sensitive elements and processing, performing analysis primarily on the redacted non-sensitive data, and then later integrating these analyses with the original sensitive data. This approach is particularly effective when handling a data of which only a portion is personally-identifiable information. It is less effective when complex analyses needed need to include the sensitive data itself.

B. Driving Use-Cases

To further study these tradeoffs, we examined several use cases identified at the University of New Mexico and identified the challenges that they presented to the approaches described above.

a) Psychology Clinical Neurosciences Center (PCNC): PCNC at UNM utilizes computational research resources at the UNM Psychology Clinical Neurosciences Center (PCNC), the

Mind Research Network (MRN), and the UNM Center for Advanced Research Computing (CARC). This research integrates work in psychology, neurosciences, electrical engineering, and computer science to leverage sophisticated neuroimaging instruments and associated computational resources to conduct a wide range of clinical research. PCNC researchers need to be able to acquire sensitive data and associated metadata from diverse instruments, and analyze it both individually, in aggregate, and across modalities with a wide range of analytical techniques.

The current approach taken by PCNC relies on a combination of custom systems and data segregation. This is sufficient for, for example, separating EEG readings from study participant personal information. It is more challenging for other sensing modalities, however, and researchers in this group need scalable, sustainable ways of acquiring, analyzing, sharing, and preserving sensitive instrumentation data.

b) Forensic Anthropology: Research in this area at UNM includes a wide range of systems and data sources, including imaging data, decedent medical records, causes of death investigation records, and demographic information. The sensitivity of this data varies widely; some forensic data is made publicly available either directly or in redacted form, while other data is protected as personally-identifiable information either as a regulatory requirement or simply as a best practice to facilitate research. Researchers in this area generally store the original data in unredacted form, and often deidentify data using proprietary tools to share limited data products with colleagues and collaborators.

This wide range of data types has resulted in the data being stored in multiple inconsistent systems, making both data acquisition, research on the acquired data, and the systematic protection of data and generated research artifacts more challenging than is desirable. In addition, the speed and scale at which current tools can deidentify data products from the underlying archival data for collaborative research is another challenge faced by this research group. Scalable mechanisms for sharing redacted versions of the underlying data with colleagues would greatly improve their ability to conduct research.

c) Community Environmental Health Program (CEHP): These initiatives at UNM study environmental health impacts in tribal populations. There are three different centers at UNM involved in these studies, at both the UNM Health Sciences Center and on main campus, that analyze complex datasets that include both environmental studies and data from tribal participants. Managing, integrating, and analyzing these data sets, which have their own unique data use agreements, is challenging. In addition, publishing and preserving this data is challenging for similar reasons.

The current approach taken by these researchers is to either work on the underlying data on isolated systems or to redact sensitive data from the system prior to sharing with collaborators. The latter can be problematic, however, as environmental samples often include geo-spatial data essential to analysis (e.g. where a mineral assay was collected) which may need to be redacted before it can be shared. These researchers would greatly benefit from systems that provided fine-grained control of the granularity of data sharing and redaction.

C. Required Features of SAMPRA

Through consultation with the lead researchers for each of the above use cases, we identified the several challenges that academic researchers often face when performing research with sensitive data. These challenges can be classified broadly as technical, operational, and policy based.

Technical challenges result from system and architectural limitations, and require development of a flexible architecture that will support research use cases across a variety of disciplinary contexts and data types. The architecture further needs to be available to users with different permissions for secure data access, from both on and off campus. Data acquisition features need to support continuous, streaming data, large data transfers, and operate reliably at high speeds. The SAMPRA environment and formalized workflows address these technical challenges through coordination of existing, well supported networking capabilities that enable role-based access to the remote resources and support data transfer from on and off campus via a robust high-speed network. Each project will be initialized using a default VM that conforms to medium NIST 800-171 requirements, with the capability for deployment of custom applications or more stringent data security measures as needed on a per-project basis. SAMPRA will securely integrate with web interfaces to facilitate access to domain-specific and complex workflows running on the HPC clusters.

Operational challenges relate to implementation and project management concerns including human resources allocation, workflow development of documentation, and training. One key challenge is providing and verifying that users have completed training in user-facing aspects of the service, which may vary depending on project specific customizations to or extensions of the default environment. Other challenges include documenting, communicating, and enforcing research and data quality assurance workflows among a distributed group of researchers, and enabling access to deidentified data products to facilitate collaboration among research support personnel such as data librarians. SAMPRA addresses these operational challenges through project specific workflow management implemented at different stages which address data collection, annotation, analysis, publication, and preservation. To maximize efficiency, project specific customizations will integrate existing data management tools into SAMPRA infrastructure. Training materials and assessments will be developed collaboratively between SAMPRA personnel, research support personnel, and project leads. Materials will be stored in a central repository for adaptation and reuse across projects.

Policy challenges constitute a broad group of issues that may be addressed to some extent through technical or operational means, but may also relate to challenges that are institutional rather than project specific. A primary example of this is data ownership and retention. In addition to the researchers themselves, multiple stakeholders have interests and responsibilities related to data. Research administrators, for example, are accountable to sponsors for the disposition of sponsored data products. Human research subjects likewise have rights and interests relating to accessing and correcting data about themselves, as well as how data are shared and preserved. The SAMPRA environment will address policy challenges in part through the development of the training and workflow repository described above, which will be openly available to researchers and will include learning objects specific to compliance concerns relevant to the initiation and closeout phases of every project. In the initiation phase, additional assessments will verify that teams conducting human subjects research have either deidentified or properly secured data prior to uploading to SAMPRA. In the closeout phase, further compliance assessments will verify that data have been processed for publication or sharing in accordance with the corresponding IRB protocol and documented consent processes. Secure storage environments will maintain compliance with Data Use Agreements to store sensitive research data.

There is some overlap between the different classifications. For example, the integration of technical constraints and training must address regulatory requirements. Also, planning for the scalability and sustainability of the project and supported environments requires solutions that address all three classes of challenges.

III. System Design and Workflows

The challenges and requirements identified from our use cases, as discussed in the previous section, revealed that the main issues facing researchers with sensitive data were personnel and system integration issues. To address these issues, we deployed a relatively straightforward virtualization-based system architecture that can support diverse system configurations and focused on augmenting it with three different capabilities:

1) Workflows that integrate researchers, local IT staff, research computing support staff, data librarians, and central IT personnel to support a wide range of researcher needs;
2) Integrated data management facilities to enable systematic sharing and preservation of research artifacts; and
3) Customizable virtual data transfer nodes (vDTNs) to enable flexible, controlled data ingress and egress from the system.

After describing the core system architecture, the remainder of this section focuses primarily on these additional capabilities, with a particular focus on the operational workflows that support the overall system architecture and its effective usage.

A. System Virtualization Architecture

To enable diverse research configurations and support from a wide range of system personnel, we adopted a fairly standard virtualization architecture based on virtual machines, virtual desktops, and software defined networking as our core system platform. In this architecture, research projects are provided a set of virtual machines on a separate, software-defined network to which users connect using a variety of remote access technologies (e.g. ssh, VDI, RDP, etc.) This software-defined network is then connected to other networks over a virtual switch and firewall which defines appropriate external access policies. Overall, this allows both the system's virtual hardware, software, and network configuration to be customized based on the research, compliance, and data needs of the project.

Figure 1 shows the detailed system design and architecture of SAMPRA, including bare-metal edge service gateways and virtual machine hosts (vHosts). The gateways act as the gatekeeper of the SAMPRA environment, filtering all ingress and egress traffic. The gateways' northbound interfaces are connected to the UNM campus core physical network over virtual private networks (VPN) to allow end-users at research labs to access SAMPRA resources. The gateways' southbound interfaces are connected to the vHosts over a virtual distributed logical router (vRouter) deployed in virtual center. Each vHost has its own virtual network (vNet), internal virtual firewalls, virtual storage pools, and virtual machines (VMs).

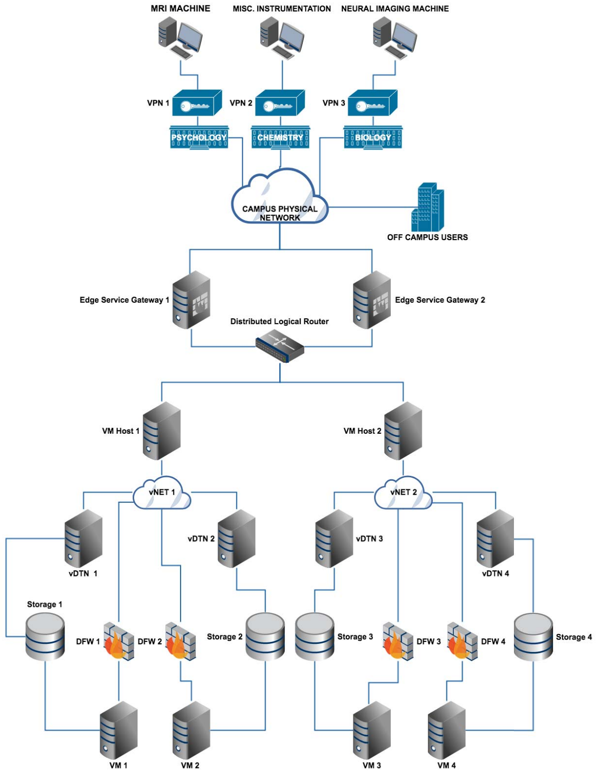


Fig. 1. Detailed SAMPRA architecture

This design makes it easy to scale out the SAMPRA physical infrastructure by simply adding new resources such as vHosts and gateways, then provisioning more vDTNS and VMs for new and/or existing research groups. Scale up of existing virtual collaborative environments depends on the available hardware resources: If the vHosts have extra free resources such as cores and memory, then scaling up is a matter of a few clicks then the extra resource will be available immediately. If there are no free hardware resources, then physical hardware scale out is required to be able to scale up the virtual resources.

B. Workflows and Supporting Personnel

Building on this architecture, we defined a set of research workflows and the roles of various University personnel in supporting them as illustrated in Figure 2.

1) Identified Workflows: The common workflows that SAMPRA supports in order to serve the research groups are as follows:

a) Project Initiation Workflow: In this phase, the research project's hardware and software environment is specified based upon the research group's requests and consultation between SAMPRA IT personnel, departmental IT personnel, and central IT personnel. In addition, operational protocols and processes particular to the research are also developed, for example the development of Technology Control Plans for export controlled data and human subjects protocols are created in collaboration with and reviewed by Research Compliance Personnel, and plans for executing the associated data management plan are developed in collaboration with data librarians from the University Libraries.

b) Project Deployment Workflow: In this phase, the University IT deploy the specified virtual infrastructure, and departmental IT staff collaborate with institutional IT personnel to provision the project's hardware and software environment including storage, network, and Virtual Desktop Interface (VDI) settings. In addition, departmental IT personnel install and configure software applications requested by the research group. Also, necessary user-facing training for the system are provided by UNM libraries, SAMPRA IT, and University IT personnel. Finally, the final system configuration is reviewed by the University IT security personnel to ensure that any changes from baseline assessed configurations still meet the data protection requirements compliance and any additional compliance requirements developed during project initiation.

c) Project Operation Workflow: During project operation, multiple related workflows happen, including training of users on the use of the system, managing research data in the system, and refining system configuration as research needs evolve. Though the roles in this phase are well-defined, the specific workflows executed are generally research-specific and are developed during project initiation.

d) Project Closeout Workflow: In the closeout phase, the data generated as the result of the research is cleaned and then published with support from the University Libraries, and transient system resources are released. The data will be available to other researchers within the department or other institutions for reproducibility or use in other research projects.

2) Roles and Responsibilities: Similarly, we identified the following roles in supporting system workflows in addition to those performed by researchers who use the system for their day-to-day work:

a) Departmental IT Staff: provide day-to-day support for the deployed system to researchers, assisting them in fully utilizing SAMPRA-provided hardware, software, and services. These staff often have intimate knowledge of the hardware and software systems with which the researchers they support work, and are the main IT coordination point for SAMPRA system and platform IT personnel (described below).

b) SAMPRA System IT Staff: provide system-specific configuration of SAMPRA based on requirements specified by researchers and Departmental IT staff. This includes creating and modifying virtual firewall rules, changing system storage configurations, and adding and removing other virtual resources from the environment for researcher use and which Departmental IT personnel directly support.

c) Platform Support Team (Central IT): provide the overall virtual machine system, virtual network platform, and storage system that enables researchers to share and transfer their data as well as perform upgrades, renewals, and maintenance of compute clusters.

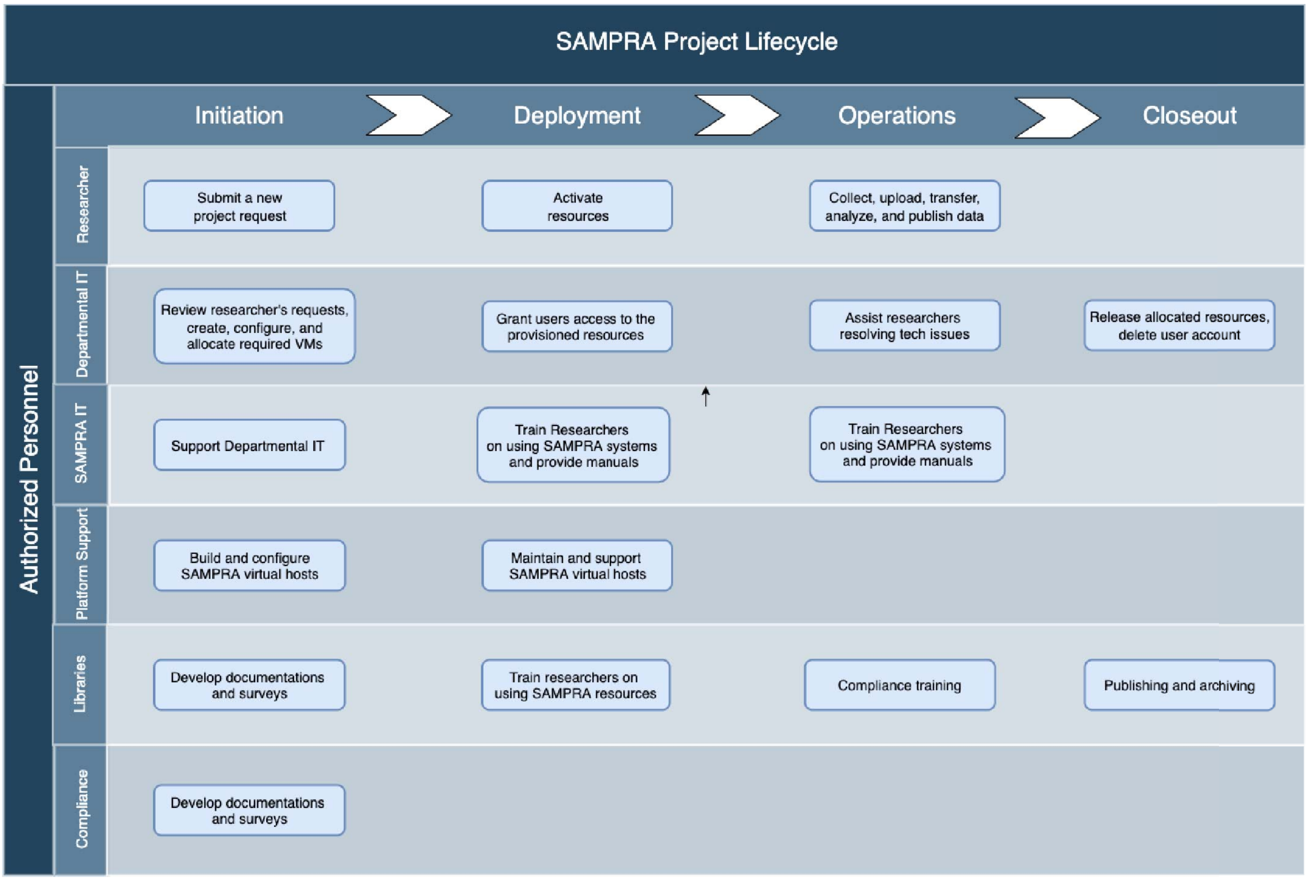
d) Data Management Support Team (Libraries IT): support the development and delivery of training materials relevant to project specific, standards-based data documentation and publication. These personnel also support closeout activities related to the preservation and archiving of deidentified data products.

e) Research compliance personnel (University Admin): audit and approve proposed system configurations, configuration changes, and operational workflows based on input from other IT personnel to ensure that the resulting system configuration meets University policy and sponsor regulatory requirements. This includes export control and industrial security personnel, as well as the University's Institutional Review Board.

C. Data Management

The increasing emphasis placed on data sharing and preservation by research funders, university administration, and publishers requires that the SAMPRA system and corresponding workflows proactively address data management through technical, operational, and policy oriented features. The SAMPRA architecture as described here is consequently informed by a set of functional requirements that are well aligned with established models of lifecycle research data management. Though there is no canonical model, DataONE [2] provides a high level conceptualization that describes intersection points between the anticipated needs of SAMPRA researchers and institutionally supported data management services.

In particular, the SAMPRA environment provides a means to dynamically support researcher needs with regard to the *collect*, *assure*, *analyze*, *describe*, and *preserve* stages of the lifecycle as illustrated in Figure 3. Each of these stages represents a fault point during which data may be lost or otherwise compromised, and SAMPRA provides a framework



SAMPRA Project Lifecycle

Authorized Personnel	Initiation	Deployment	Operations	Closeout
Researcher	Submit a new project request	Activate resources	Collect, upload, transfer, analyze, and publish data	
Departmental IT	Review researcher's requests, create, configure, and allocate required VMs	Grant users access to the provisioned resources	Assist researchers resolving tech issues	Release allocated resources, delete user account
SAMPRA IT	Support Departmental IT	Train Researchers on using SAMPRA systems and provide manuals	Train Researchers on using SAMPRA systems and provide manuals	
Platform Support	Build and configure SAMPRA virtual hosts	Maintain and support SAMPRA virtual hosts		
Libraries	Develop documentations and surveys	Train researchers on using SAMPRA resources	Compliance training	Publishing and archiving
Compliance	Develop documentations and surveys			

Fig. 2. Roles and actions throughout the research project lifecycle.

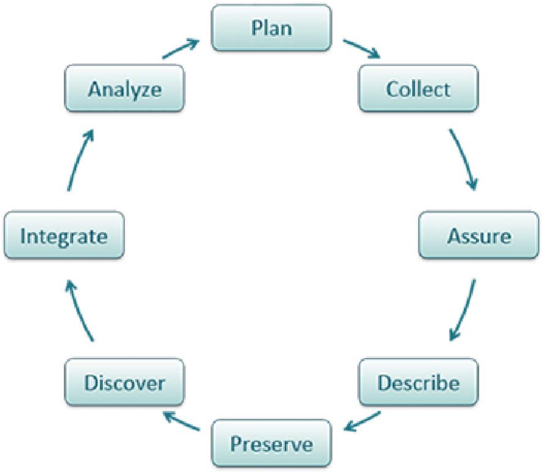


Fig. 3. A generalized research data lifecycle model. Image credit DataONE, https://www.dataone.org/best-practices.

for identifying and developing services around both general purpose and project specific needs:

- *Collection*: Systematic collection of CUI and other sensitive data within networked research environments requires the specification of role-based file upload and access privileges. SAMPRA addresses the technical aspects of this requirement through virtual machine templates that specify user permissions based on their role in a research project. Operationally, the SAMPRA team works closely with researchers to structure training opportunities around an assessment of the policy requirements that inform and constrain project specific workflows. Training content and methods are aligned with best practices as recommended by DataONE [3] and as described throughout the edited volume by Johnston (2017) [4]. Training are provided by the SAMPRA team, including data librarians from UNM Libraries' Research Data Services unit.
- *Assurance* and *analysis*: In addition to providing role-based virtual machine access and training in project specific sensitive data management topics, the SAMPRA framework includes processes for white- and blacklisting applications within and network connections to the virtual machine. Not only does this standardize the tools and applications that can be used to clean and analyze data, but system logs provide an auditable record of user access and activity.

- *Describe*: In addition to the documentation of data provenance enabled by the system logs, metadata or descriptive information about data collection and analysis methods are collected throughout the research process in coordination with UNM data librarians. Documentation requirements and best practices are included in researcher training, and collected metadata are mapped by data librarians to standards such as Dublin Core [5] that support the discoverability and preservation of published or archived datasets.
- *Preservation*: SAMPRA facilitates the transfer of CUI and other sensitive data products into institutionally supported preservation systems through automated staging of both sensitive and deidentified data products. Two specific features of the framework streamline the process of curating data for archiving: The above noted logs and metadata are co-located with the data in the virtual machines; and the data can be released for transfer through a secure network for review and curation by data librarians.

By aligning workflows both technically and operationally with specific tasks and responsibilities of research support personnel, the SAMPRA framework provides a set of structured, standardized data management capabilities that promote data sharing and preservation while reducing the compliance burden for individual researchers.

D. Data Ingress/Egress Support

Providing flexible, scalable, controlled data ingress and egress was one of the key requirements of the use-cases we studied. To support this, SAMPRA adapts the concept of data transfer nodes as part of the Science DMZ architecture pioneered by ESNet [6] to provide controlled portals for importing data from instruments and exporting sensitive data in a controlled manner. These *virtual data transfer nodes* are generally the only externally-visible part of a SAMPRA enclave, and can be configured by SAMPRA and local IT personnel to meet the data movement needs of the research in question.

The software run on a vDTN is often highly application- and research-specific, though a number of existing data processing tools can potentially ease this task. We are currently building on our experiences implementing vDTNs that provide access to deidentified data (described in the following section) to develop a general architecture for them based on the existing Globus [7] and iRODS [8] software systems.

IV. Implementation Status

We have currently implemented a prototype of SAMPRA using a 4-node VMWare ESXi/NSF cluster and deployed both the workflows and supporting infrastructure described in the previous section on this cluster. PCNC researchers are the initial system users, and are currently migrating workflows previously supported using stand-alone air-gapped systems to the SAMPRA environment. We are currently working to migrate this system and its supporting workflows into the production UNM LoboCloud environment, as well as to move additional research projects into both the current prototype and future production systems.

Our first prototype vDTN was a system responsible for extracting millions of CT images from a Picture Archiving and Communication System (PACS) used by the New Mexico Office of the Medical Investigator in support of forensic anthropology use cases. The software was primarily custom scripts and services to connect the backend imaging archive containing sensitive information to the storage and web-hosting system making redacted versions of these data publicly available. The resulting system, the New Mexico Decedent Information Database [9] is the first publicly-available large-scale archive of decedent images. We plan to re-implement this system after finalizing a general architecture for SAMPRA DTN nodes, a project that is ongoing at UNM.

V. Security Assessment

To ensure that the overall SAMPRA architecture is compliant with NIST-SP-800-171 regulations, we collaborated with UNM's Information Security and Privacy Office (ISPO) to conduct a full medium-level assessment of the overall system architecture. One goal of this overarching assessment was to simplify later assessments of researcher-customized environments. Because customizations of systems require reassessment and approval prior to the start of research, we sought to identify technical and operational controls that placed as little burden as possible on the system so as to maximize customizability and to identify any controls or policies that would potentially need to be reassessed based on changes in the underlying system.

The assessment process utilized the Department of Homeland Security (DHS) Cyber Security Evaluation Tool (CSET) to build a comprehensive site cybersecurity plan with focus on control systems. The tool was used to identify existing controls and the need for any overall system improvements. The first step in this approach was to provide answers for all of the security-control questions in the tool's template from DHS. A Plan of Action and Milestones were developed based on the identified gaps from the security assessment, resulting in relatively minor revisions to various technical and operational controls. The resulting assessment and plan serve as the primary security documentation for SAMPRA.

One key insight that we developed in the process was the importance of the use of training, operational controls, and system monitoring instead of strict technical controls when possible. Strict technical controls and limitations are necessary in some high-risk environments (e.g. when handling ITAR data), but in NIST medium environments, operational controls whose verification can be monitored through system logging, log monitoring, and alerting result in a more flexible system that can often better meet the needs of researchers.

We also found that the combination of this approach of focusing on operational controls when appropriate, a virtualized system architecture that allowed for the deployment of technical controls at precise system locations, and existing administrative reviews and processes (e.g. IRB review of handling of personal data and export control technical control plan processes), resulted in an overarching environment that supported diverse research while meeting regulatory compliance requirements. Finally, the main points identified that require careful reassessment for each system beyond the overarching system assessment are:

- Log monitoring and alerting for various OS and software configurations;
- Virtual firewall and proxy configuration; and,
- vDTN configuration.

VI. Related Work

Several systems have been deployed using virtualization to provide a platform for supporting research on controlled unclassified information, most notably the ResVault (formerly GATORVault) system at the University of Florida [1]. As in SAMPRA, virtualization in these systems increases the flexibility of the research supported in all of these systems. Most of these systems, however, have opted for supporting a small number of tightly controlled VMs with strong technical controls to ensure compliance. In SAMPRA, we have generally opted for operational controls with appropriate monitoring when possible, and used technical controls primarily at system boundaries when necessary; this increases the flexibility of research the system can support. The workflows, staff integration, and supporting technical systems described in this paper are designed with exactly this goal in mind.

In addition, a wide range of national cyberinfrastructure systems use virtualization to increase system flexibility. Both Jetstream [10] and Comet [11] are NSF XSEDE [12] systems that leverage virtualization to support research and education in more diverse areas of science and engineering. Jetstream focuses primarily on customizable virtual desktops, workstations, and big data systems using Virtual Linux Desktop similar to SAMPRA's RDP access method. Comet, on the other hand, provides virtual HPC clusters for systems that require custom software stacks.

Similarly, the Bridges [13] system uses OpenStack virtualization software to provide a wide range of hardware for supporting a broad range of research demands, primarily in the high-performance computing and big data space. Unlike the other systems described above, however, Bridges provisions and configures bare-metal systems instead of primarily leveraging virtual machines. It also provides some support for handling sensitive data, but only on specifically configured and controlled machines. Unlike SAMPRA, Bridges does not attempt to provide a flexible and configurable environment for handling CUI data.

Finally, CloudLab [14] and ChameleonCloud [15] are NSF-sponsored virtualization systems that provide infrastructure for studying a diverse range of cloud computing and networking system software issues. Both include diverse compute hardware and networking technology located at multiple sites, as well as software designed to support system software. Neither, however, primarily seeks to address issues associated with CUI research.

VII. Conclusions and Future Work

SAMPRA seeks to address technical, operational, and policy challenges facing research that handles sensitive or controlled data. It does so by deploying workflows, staff, and technical systems that support an underlying virtualized computer and network architecture. This work is increasingly important with the growing threats to academic research data by state, industrial, and rogue actors, and the increasing demands for research cybersecurity at academic institutions [16].

Key areas for future work, besides continually refining the system to meet changing regulatory and compliance requirements, are on further enhancing the data management and vDTN support in SAMPRA. Current data management and vDTN support is adequate to the demands on the system, but we have not yet integrated a number of our motivating use-cases into the system. Supporting the research demands of these systems, growing requirements for data archival and sharing, and further integration with scientific instruments will all require research on new software architectures and data and personnel workflows.

VIII. Acknowledgments

This paper is based upon work supported in part by the National Science Foundation under Grant No. OAC-1840069.

References

[1] U. of Florida Research Computing, "UFL resvault." [Online]. Available: https://www.rc.ufl.edu/services/restricted-data/researchvault/

[2] DataONE. (nd) DataONE data lifecycle. [Online]. Available: https://www.dataone.org/data-life-cycle

[3] ——. (nd) DataONE data management skillbuilding hub. [Online]. Available: https://dataoneorg.github.io/Education/

[4] (2017) Curating research data, volume 1. [Online]. Available: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

[5] D. C. M. Initiative. (2020-02-20) DCMI metadata terms. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[6] S. DMZ, 2010. [Online]. Available: https://fasterdata.es.net/science-dmz/

[7] I. Foster, "Globus online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, no. 3, p. 70–73, May 2011. [Online]. Available: https://doi.org/10.1109/MIC.2011.64

[8] K. D. Winters, M. A. Cowan, G. E. George, M. E. Gonzalez, B. Priest, O. Morris, and J. Landrum, "Analysis of ers use cases for irods," Engineer Research and Development Center (US) Vicksburg United States, Tech. Rep., 2020.

[9] H. J. H. Edgar, S. Daneshvari Berry, E. Moes, N. L. Adolphi, P. G. Bridges, and K. B. Nolte, "New Mexico Decedent Image Database," Office of the Medical Investigator, University of New Mexico, 2020. [Online]. Available: https://doi.org/10.25827/5s8c-n515

[10] J. Fischer, S. Tuecke, I. Foster, and C. A. Stewart, "Jetstream: A distributed cloud infrastructure for underresourced higher education communities," in *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models*, 2015, pp. 53–61.

[11] S. M. Strande, H. Cai, T. Cooper, K. Flammer, C. Irving, G. von Laszewski, A. Majumdar, D. Mishin, P. Papadopoulos, W. Pfeiffer *et al.*, "Comet: Tales from the long tail: Two years in and 10,000 users later," in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 2017, pp. 1–7.

[12] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "Xsede: Accelerating scientific discovery," *Computing in Science Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[13] N. A. Nystrom, M. J. Levine, R. Z. Roskies, and J. R. Scott, "Bridges: a uniquely flexible hpc resource for new communities and data analytics," in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 2015, pp. 1–8.

[14] R. Ricci, E. Eide, and C. Team, "Introducing cloudlab: Scientific infrastructure for advancing cloud architectures and applications," *; login:: the magazine of USENIX & SAGE*, vol. 39, no. 6, pp. 36–38, 2014.

[15] ChameleonCloud. (2015). [Online]. Available: https://www.chameleoncloud.org/

[16] Carnegie Mellon University and The Johns Hopkins University Applied Physics Laboratory LLC. (2020-03-18) Cybersecurity maturity model certification (CMMC). [Online]. Available: https://www.acq.osd.mil/cmmc/docs/CMMC_ModelMain_V1.02_20200318.pdf

[17] R. M. Badia, J. Ejarque, F. Lordan, D. Lezzi, J. Conejero, J. Á. Cid-Fuentes, Y. Becerra, and A. Queralt, "Workflow environments for advanced cyberinfrastructure platforms," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1720–1729.

[18] W. Hong, J. Moon, W. Seok, and J. Chung, "Enhancing data transfer performance utilizing a DTN between cloud service providers," *Symmetry*, vol. 10, no. 4, p. 110, 2018.

[19] Y. Liu, Z. Liu, R. Kettimuthu, N. S. Rao, Z. Chen, and I. Foster, "Data transfer between scientific facilities–bottleneck analysis, insights, and optimizations," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), Tech. Rep., 2019.

[20] D. Roode. (2016) A vision for research cyberinfrastructure at uci, version 4.7e. [Online]. Available: https://rcic.uci.edu/hpc3/A-Vision-for-RCI-at-UCI-Document-and-Appendices.pdf