Title

Securing Research Data: A Whitepaper for UC Berkeley

Authors
Hoffman, Chris
Christopher, Jason
Jaffe, Rick
et al.

Publication Date
2017-07-05

Peer reviewed

Securing Research Data: A Whitepaper for UC Berkeley

July 5, 2017

Authors: Chris Hoffman, Jason Christopher, Rick Jaffe (Research IT); Jon Stiles (D-Lab); Rachael Samberg (the Library); Leon Wong (Information Security and Policy)

Executive Summary

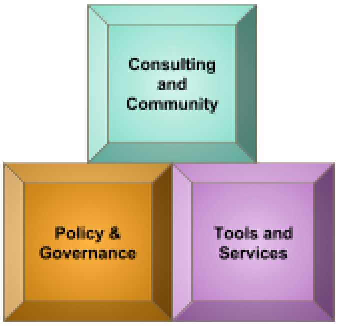
Researchers at Berkeley increasingly work with sensitive and restricted-use research data that needs to be handled securely. Most researchers, however, lack experience in managing sensitive data. They do not have training in protecting data, are unable to locate appropriate campus support for doing so, and, as a result, cobble together makeshift solutions which are often inefficient, vulnerable, or unsustainable. This places their research at risk, places the campus at risk, and threatens the opportunities of other campus researchers. Equally as important, it saps time and effort from their research, and limits the capacity for other campus researchers to leverage their investments.

In January 2017, the Research Data Management (RDM) program (a partnership between the Library and Research IT in the Office of the CIO) launched a six-month project with partners in the D-Lab and Information Security and Policy (in the Office of the CIO). The goals of this initial project were to assess the campus sensitive-data landscape from the point of view of researchers and research administrators; gauge the demand for services and guidance; benchmark services at peer institutions; and make a set of recommendations for future work.

In carrying out this project, the team created several important products:

- This whitepaper, which describes the existing campus landscape, suggests policy and practice guidelines, highlights solutions from peer institutions, and identifies some near-term and long-term strategies and solutions for campus.
- An on-going D-Lab Working Group on Securing Research Data, which pulls together the larger community of stakeholders: researchers, service providers, compliance offices, and others.
- A Research Data Classification Guideline that interprets campus Data Classification and Protection Profiles for researchers, service providers, and compliance offices. The publication of this guideline also satisfies one of the findings from a 2016 internal audit of research data management on campus.
- Box PL2 assessment is in process.

In order to both reduce risks for the campus and facilitate research that uses and produces sensitive data, a comprehensive approach must take into account: consulting and community; policy and governance; and services and tools. The recommendations outline below are discussed in more detail in the body of the whitepaper.



Introduction

Researchers at Berkeley are increasingly working with sensitive and secure research data that they lack experience in protecting, are unable to locate appropriate campus support for and, as a result, cobble together makeshift solutions which are inefficient, vulnerable, or unsustainable. This places their research at risk, places the campus at risk, and threatens the opportunities of other campus researchers. Equally importantly, it saps time and effort from their research, and limits the capacity for other campus researchers to leverage their investments.

Background

Researchers in a wide range of campus disciplines face an increasingly complex set of requirements to protect their research data, with often overlapping requirements coming from an increasing number of directions: from funders, data providers, state and federal agencies, and from the campus and university itself. This problem is further complicated by disciplinary differences across the campus, and by changes within the disciplines themselves. Social scientists have long worked with human subjects research data, but are now exploring larger data sets, new methodologies, and new research opportunities. Other fields are perhaps newer to data-driven research (e.g., the humanities); or are exploring human subjects data for the first time or in new and expanded ways (e.g., in computer science and engineering). Simultaneously, even data that do not involve human subjects are increasingly subject to data use agreements and other requirements from data providers. Not surprisingly, much of this regulatory aspect is driven by cybersecurity threats and the potential impact of a data breach, a phenomenon and trend that is rapidly and chaotically evolving. Equally important, researchers want to know how to protect the data that they are using and generating, knowing that data sets are an important and valuable part of their research and scholarship.

In the summer of 2016, a joint team from Research IT, the Library, the D-Lab, and IST evaluated UC Berkeley's service offerings for securing research data. Using the methodology developed for the RAE (Research and Academic Engagement) Redo Benchmarking project, the team defined Securing Research Data as:

> Services in support of research data sets that have restrictions set by campuses, state or federal law, or contractual agreements. Includes policy and guidance; infrastructure services for storage and computation; consulting and training; cross-campus coordination for managing restricted data; and support for coordinated development to build resources and manage relationships with data providers.

The team identified the following criteria for evaluating service offerings at UC Berkeley and at thirteen peer universities:

- Discoverable and useful policy and guidance documentation
- Secure research environments and infrastructure
- Active training and consulting

- Life-cycle process support and cross-campus coordination
- Active development of restricted data resources

Following the comparative analysis, the team ranked UC Berkeley's service offerings as a 3 (with 1 being the highest score and 4 the lowest). This three-level ranking was characterized as follows:

> Accessible policy guidance with limited services, few/inflexible research environments for secure data, environments limited in terms of compute or storage, limited service support and training, limited coordination among campus stakeholders and service providers, ad-hoc development of discipline specific resources.

In short, researchers working with data that require protections of any kind most often must perform this activity with very little campus support. See Appendix A for the RAE Redo report, "Secure Research Data and Computation".

Another campus-wide effort that demonstrated broad needs in this area was an audit of Research Data Management conducted by Audit and Advisory Services. The final report to campus, dated June 24, 2016, determined a) that the process for identifying required security protections was driven by many offices that do not necessarily work together in a coordinated way and b) that the scope of the problem and the campus-wide risk are very difficult to ascertain. That is, the campus does not know how much risk it faces, nor does it have the right policies, governance mechanisms, or even relationships with researchers that would facilitate a more accurate assessment.

Other campuses are not only doing more to help researchers, but they are also treating the ability to manage restricted data as a strategic differentiator. As data-intensive research including data science grows on campus, the importance of providing services that allow researchers to work with sensitive data of various kinds becomes strategically valuable and even critical. Indeed researchers at Berkeley are often at the vanguard of working with data that require protections of various kinds, and yet the services campus offers are not adequate. Campus data policies and enterprise services are geared towards data for campus administrative processes. At the same time, most research systems have emphasized data sharing and collaborative access. As a result, researchers often struggle to develop solutions, taking valuable time and energy away from primary research.

Project and process

In January 2017, Research IT and the Library launched a six-month project through the Research Data Management Program with partners in the D-Lab, Information Security and Policy (OCIO), the Office of the Vice Chancellor for Research, and other units within Information Services & Technology. The project charter and team members are documented in Appendix B.

The project team developed a work plan knowing that a six-month project could only be the beginning of a process to address the complex set of issues that the campus faces. At the same time, the team recognized that a comprehensive approach was needed in order to lay the groundwork for future success. Early on, the project adopted a framework crafted along the following dimensions: consulting and community; policy and governance; and services and tools.

Project Outcomes

Despite the short duration of this initial project, the team highlights these outcomes:

- This whitepaper, which describes the existing campus landscape, suggests policy and practice guidelines, highlights solutions from peer institutions, and identifies some near-term and long-term strategies and solutions for campus.
- Formation of a D-Lab Working Group (Securing Research Data) which pulls together the larger community of stakeholders: researchers, service providers, compliance offices, and others.
- A Research Data Classification Guideline for researchers, service providers, and compliance offices. The publication of this guideline also satisfies one of the findings from the internal audit of research data management on campus.
- Broader recommendations for policy specifically addressing research data
- Benchmarking and case studies from other institutions
- An inventory of existing solution designs that researchers currently use to store and analyze restricted data
- Technology service recommendations for environments for storage and analysis

These outcomes are presented below and in appendices attached to this whitepaper.

Current State

Assessment of current demand

To understand the contours of the problem, and to estimate demand for improved services, the project examined current use of sensitive and “restricted-use” data on campus. Team members aggregated information from several sources:

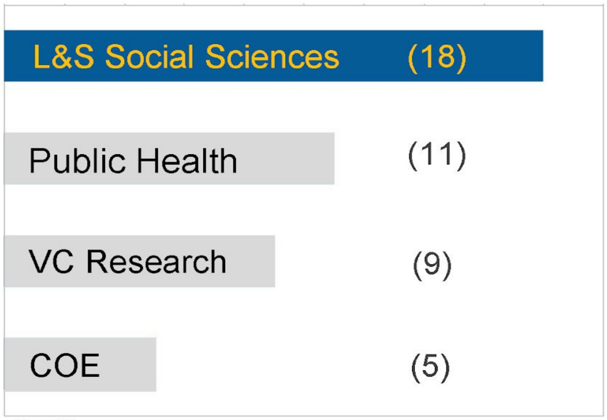
- Consultation requests to the Research Data Management program
- Data security review requests to Information Security and Policy
- Case notes from the D-Lab and from offices such as the Industry Alliances Office and the Campus Privacy Office.

These sources are all campus organizations tasked with providing consulting help, overseeing data security, or exercising regulatory or signatory authority over certain types of sensitive data.

The resultant set of 68 cases provides a glimpse of the broader need at UC Berkeley. The list comprises research projects that sought help from or engaged with these campus units. As such, it is assumed to represent only a small subset of the real demand for services on campus and likely underrepresents significant portions of the research enterprise (more on this, below). Though by no means complete, it offers rich insight into the scope and nature of that research.

At a Glance

Top schools and offices accessing help…

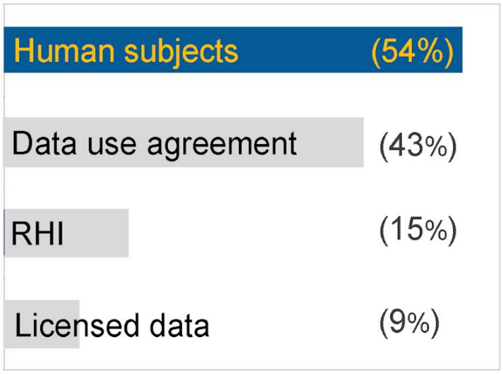


N = 68 cases

Units across campus asking three or fewer times…

Natural Resources
Social Welfare
Berkeley Law
Haas Business
Education
L&S Math & Physical Sciences
VC Undergraduate Education
RW Johnson Health Scholars
L&S College-wide
Environmental Design
VC Admin & Finance
L&S Biological Sciences

Top regulatory frameworks controlling data…

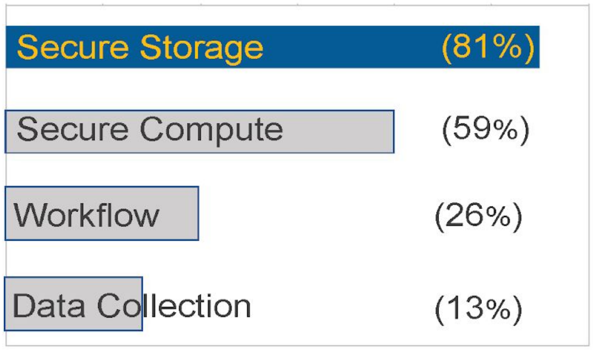


N = 68 cases, some governed by multiple frameworks
(99 regulatory framework tags, total)

Other frameworks are important, too …

HIPAA
NIST 800-171
NAGPRA
Endangered Species Act
FISMA-Moderate
Cultural Agreements
ARPA
Data Privacy
Export Controls
FERPA

Top activities researchers seek help with…



N = 68 cases, some involving multiple aspects of work
(158 activity tags, total)

Other aspects of their work need support, too…

- Planning
- Secure Sharing
- Secure Transfer
- Secure Communication
- Annual Review
- Records Retention
- Secure Access
- Re-identification

Discussion

Who is requesting help with sensitive data?

The project team was interested to know the campus affiliation of researchers requesting assistance with sensitive data. In the study sample, more than a quarter of the cases originated from a single division within the College of Letters & Science: the Social Sciences division, with the majority coming from the Departments of Sociology and Psychology. Another 16% of the requests came from the School of Public Health, and 13% from the Vice Chancellor for Research Office, which represents organized research units, centers, institutes, museums and the UC Botanical Garden. Additional requests, in smaller numbers, came from a dozen other schools, offices, and programs across campus.

In part, this distribution follows naturally from the type of research questions and data addressed by each of these disciplines. The skew in numbers towards social sciences and public health also reflects administrative efforts to increase communications among the campus's institutional review board, its Sponsored Projects and Industrial Alliances Offices, and its Research Data Management and D-Lab programs.

In contrast, a number of disciplines are probably underrepresented. Chemistry, for example, is not represented at all. Researchers working with intellectual property such as financial transaction data, market research data, and cutting-edge engineering and physical sciences technologies might rely on support resources at the college, rather than campus, level, or be more accustomed to developing local services to support data-intensive research. Researchers in disciplines new to working with human subjects data might not know about campus support resources.

The breadth of demand alludes to different needs across domain and discipline, and the complexity of managing sensitive data on campus. Greater familiarity and coordination among research support offices and programs represents a blueprint of sorts for addressing the data security needs of these types of projects.

What makes the data sensitive?

Sensitive or restricted data can be personally identifiable information (PII), confidential or time-sensitive data, or valuable intellectual property. The most protected categories of personally identifiable information ("notice-triggering" data) require that research subjects be notified (and, often, made whole to a prescribed extent) when the security of that data is breached. The monetary costs to campus of responding to such a breach can be tremendous. In addition, researchers who fail to protect the confidentiality of their data risk losing access to data in the future – both for themselves and for other researchers on campus.

The sensitivity of the data, and the conditions by which it must be handled, are regulated variously by:

- Federal and state law
- Funder or sponsor guidelines
- Data use agreements
- Licenses and contracts
- Campus policy
- Institutional review board (human subjects) approval and informed consent from the subjects themselves
- Established privacy frameworks
- Ethics.

Finally, data is kept private in some fields due to its potential value to the researcher, rather than external regulation. Much research in Chemistry (unreported here) might fit into this category.

Each of the 68 cases reviewed in this report was annotated with the regulatory framework controlling the use of data. Note that an individual case can have multiple controls; in total, 99 regulatory "tags" were assigned to these cases.

Among the projects requesting help from campus, 37 were governed by human subjects controls. Data use agreements (DUAs) from the data owner – typically a state or federal agency addressing health, welfare, and educational issues – applied to 29 projects, overlapping with human subjects controls in 21 of those cases. Personal health-related data, regulated as Research-related Health Information (RHI), constituted 10 of the cases handled, alone or in tandem with human subjects controls and DUAs. (Another three projects were ruled by the federal Health Information Portability and Accountability Act, or HIPAA; HIPAA cases are more common on campuses with a medical center.) Six projects involved data protected by a license from a publisher or other organization.

In the past year, the federal government has implemented regulations, defined by the National Institute of Standards and Technology ("NIST") 800-171 standard, for safeguarding data released by Executive Branch departments and agencies. Labeled 'Controlled Unclassified Information', this class of data appeared only a few times in the sample, but is expected to become much more common on campus. Research data governed by rules re: export controls, cultural sovereignty, endangered species, etc., were less-often used, but important nonetheless.

What are researchers seeking help with re: their sensitive data?

The project team also annotated the sample cases according to the data activities associated with the researchers' requests for support (frequently more than one per case; in total, reviewers assigned 158 activity "tags" to the 68 cases). 55 cases were tagged "secure storage" alone or in combination with other activities; 40 were tagged "secure compute" alone or in combination. Other impacted aspects of the research process include workflow design, data collection, data transfer (as part of a collaborative project or from the researcher's previous institution) and data sharing/publication. For clinical and health-related research, communications about the data were sensitive and important. Not greatly represented yet, but on the horizon, are issues related to data (records) retention, annual security reviews mandated by the data owner, and the increasing likelihood of re-identification of individual subjects as more and more personal data becomes available through the Web and through third-party aggregators.

Secure computing *and* secure storage needs frequently go together. (In the study set, 40 of the 68 cases involved the need for both.) This phenomenon holds with other aspects of the data workflow, too: collected personally-identifiable data, for example, must be transferred, stored and cleaned or analyzed securely; sensitive data stored on campus should have an appropriate retention period so it can be discarded at a specified date and the risk of data breach eliminated. A nuanced policy and support environment recognizes these connected needs across the research lifecycle.

While the cases reviewed as part of the Securing Research Data project might not represent all of campus research involving sensitive data, the lessons drawn from the analysis can be applied broadly. An effective policy and support environment requires good communications with researchers and across the many offices involved in the management of sensitive data throughout the research lifecycle; a shared understanding of the frameworks and controls involved; and a nuanced view of the processes applied to data during the full course of research. The next section looks at the current state of campus policy and support.

Campus policy, guidance, roles and responsibilities

A primary observation of the current state on campus is that there are weaknesses within policies guiding the oversight of research data. This has resulted in uncertainty among researchers and offices responsible for managing and approving agreements of various kinds.

Current UC Berkeley Research Data Policies

UC Berkeley has several campus data policies that, in combination, articulate protocols relevant to particular aspects of securing research data. Briefly, these policies include:

- *UC Berkeley Data Classification Guideline & Standard*: The UCB Data Classification Standard (https://security.berkeley.edu/data-classification-standard) applies to Berkeley "campus data," which is defined as information prepared, managed, used relating to the activities or operations of the University. No express reference is made to research data as constituting "campus data", so as currently drafted, it is not clear whether the Data Classification Standard inherently applies to all research data.

- *UCOP Records Retention Policies*: UCOP records retention guidelines exclude "research records" as a type of record subject to the retention policies (see Records Retention Schedule at http://recordsretention.ucop.edu/). This leaves the research data potentially exposed to destruction, absent countervailing contractual or funding obligations. However, *administrative* records related to the research data must be maintained, such as indicia of grants, contracts, hazardous waste usage, etc. (see Research Record Retention and Disposition at http://www.ucop.edu/research-policy-analysis-coordination/policies-guidance/record-retention/index.html).

- *UCOP Administrative Procedures Manual 020*: Notebooks and other original records of research are the property of the University. No explicit instructions are provided as to who administers, maintains, or preserves those research records, however.[1]

There are other potentially applicable campus policies, such as the Human Subjects policy, Minimum Security Standards for Electronic Information, and Minimum Security Standards for Networked Devices which apply to certain types of research data. Yet, no express reference is made in these to securing all types of research data. As such, Berkeley's campus data policies are either silent or unclear about several important aspects of a holistic or lifecycle approach to securing all forms of research data—including assignment or delineation of roles and responsibilities for research data management, storage, and retention. Addressing these aspects of securing data is critical, given that various funders and granting agencies require adherence to data management plans, and failure to properly retain research data could create problems for reproducibility or in legal disputes.

Benchmarking Policies at Other Institutions

Certain other research institutions address these needs through implementation of an express Research Data Management policy distinct from other campus data policies, which instead focus more on campus administrative data. To that end, we have benchmarked a number of other institutional research and campus policies in Appendix C.

In what we have classified as “green” tier institutions (e.g. Harvard, Stanford, University of Washington), there are separate research data policies that touch upon four key criteria relevant to securing research data throughout its lifecycle (“Four Criteria”). These “green” tier policies:

1) *Define roles and responsibilities* of researchers vs. other information security officers, and explain which campus actors and units are subject to the policy;

2) *Make applicable the policy to all storage media* types, forms of data, and data locations;

3) *Specify retention* standards and duration, and;

4) *Address intellectual property issues* pertaining to the research data, which also serve to provide baseline ownership and retention obligations.

UCLA falls into this “green” tier as it has adopted “Interim UCLA Guidance on Access to and Management of Research Data and Tangible Research Materials” that expressly provides guidance on all Four Criteria. It is not procedurally clear, however, why this interim guidance is not permanent policy, or how/whether it is adhered to in practice.

In what we have classified as “yellow” tier institutions (e.g. University of Indiana), the university’s main data policy defines “campus data” broadly enough to encompass research data. Upon closer examination, however, the policy’s other provisions and guidance remain oriented toward administrative campus data, and provide insufficient guidance on executing all Four Criteria.

Finally, the “red” tier reflects institutions, like UC Berkeley, that leave unanswered at least several of the Four Criteria relevant for a comprehensive approach to securing research data throughout its lifecycle.

Based on this analysis of research data policy readiness at UC Berkeley and peer institutions, the project team finds that the lack of policy specifically applicable to research data is a significant obstacle that must be addressed. While the broader policy or policies needed will take time to develop and will involve participation in system-wide groups already examining these issues, there are two areas where UC Berkeley can make improvements in the nearer term. First, the interim guidelines adopted at UCLA have been refined and adopted at other UC campuses, and they provide a likely model for action at UC Berkeley. Second, as outlined in the next section, enhancement of the campus Data Classification Guideline represents another opportunity that can have real impact in the near term through revisions addressing some or all of the Four Criteria.

Research Data Classification Standard

Across campus, a number of offices and individuals are involved in reviewing data protections required for research projects. Currently, they refer to the campus Data Classification Standard which is published and maintained by Information Security and Policy. The purpose of the Data Classification Standard is defined as follows:

> The Berkeley Data Classification Standard is a framework for assessing data sensitivity, measured by the adverse business impact a breach of the data would have upon the campus. This standard provides the foundation for establishing protection profile requirements for each class of data.

The defined Protection Levels range from Protection Level 0 (PL0) for public information to Protection Level 3 (PL3) for systems with an extremely adverse business impact. The examples given in the standard do not map cleanly to research. Because the Data Classification Standard was written primarily for administrative and student data, there are nuances related to research data that are not addressed. For example, the primary criterion of “adverse business impact” is one that needs to be translated to be applied meaningfully to research. Similarly, within the current standard, Protection Level 2 (PL2) is defined as data sets with a high adverse business impact but is characterised as those having a statutory requirement for notification in the case of a breach (e.g., social security numbers, financial account information). However, this notice-triggering requirement is not sufficient to identify research data sets that need a very high level of protection.

Another shortcoming of the current standard is that the actual protections required for research data sets can vary significantly (as defined by requirements from data providers). Many restricted data sets would be assigned to Protection Level 2. However the protections and controls mandated by campus policy in the Minimum Security Standards for Electronic Information (MSSEI) for PL2 data sometimes exceed what is actually required by the data provider. Sometimes these additional controls are desirable from a campus risk perspective. Other times, they place a set of constraints on research that are probably not necessary and can be very expensive and difficult to implement.

As part of this project, staff in Information Security and Policy have developed a Research Data Classification Guideline to address many of these shortcomings. See the recommendations below on the publication and dissemination of these guidelines.

Platforms, solutions, and workarounds

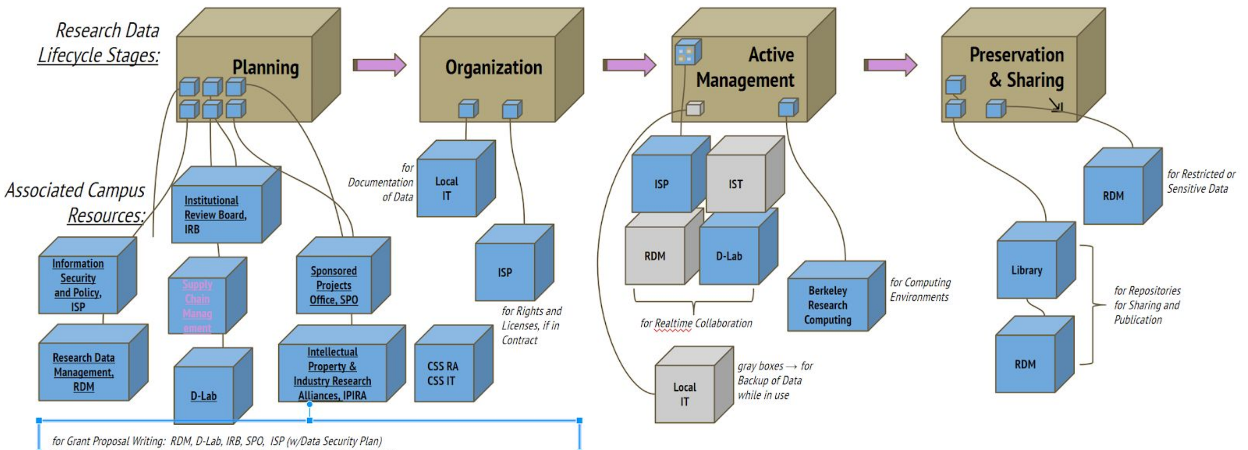
There are several services offered campus-wide that can be used by researchers working with restricted data. Box and Drive are approved for Protection Level 1 data and are licensed by campus for researchers to use at no cost. CalShare is a recharge service built on top of Microsoft Sharepoint that is certified by campus for Protection Level 2 data. IST offers a number of recharge services that can be assembled into restricted data environments of different kinds: storage, databases, virtual machines, and remote desktop (terminal) servers. These recharge services are often but not always beyond the resources of researchers, especially if they have not planned for these expenses appropriately.

Given the widespread need for services, many schools, colleges, and departments have developed solutions that work at different scales. The D-Lab is a notable example of this that offers both infrastructure, expert consulting, and training to a broad spectrum of “social science researchers (graduate students, staff, and faculty) and anyone interested in connecting with them”. Other examples that the project team worked with are found in Demography (which offers a managed virtual machine environment to affiliated researchers), the Haas School of Business, the School of Social Welfare, the Econometrics Lab, and the Institute for Research on Labor and Employment. Certainly others exist as well. However, there are many examples where individual labs or individual researchers are basically developing their own restricted data solutions. In fact, it is possible that this is the more normal pattern, and a logical response to the lack of coordination at the campus level and the significant gaps in needed services. The result of this lack of services is redundancy and inefficiency. As researchers struggle to provide their own storage and compute environments, the impact is in heightened (but invisible) risk and lost opportunities for researchers to perform research and take advantage of the results of other research.

Assessing the current state

As can be expected for a diverse University with a high profile research mission, a broad variety of offices and organizations assist with, set policy for, or have authority to regulate and approve different aspects of research data policy-setting, acquisition, management, analysis, storage, sharing, and preservation. These aspects are deeply intertwined, but organizational roles and responsibilities can be opaque, shared across organizations, amorphous, and uncoordinated. For the sake of simplicity, the organizations involved in supporting and managing research data are identified in the chart below at selected points in a lifecycle model of research data, showing points of engagement during planning, organization, active management and preservation and sharing stages. Not all organizations are identified; in particular, many ORU's and departments provide resources and support tailored to local researcher demands. [Note: Split IPIRA into IAO (incoming data) and OTL (outgoing data)] Descriptions and links to these organizations are shown in Appendix H.

Guidance - Research Data Lifecycle Stages and
Associated Campus Resources -Workflow



for Grant Proposal Writing: RDM, D-Lab, IRB, SPO, ISP (w/Data Security Plan)
for Contract Negotiation: IPIRA/IAO, SPO, ISP, D-Lab, Business Contracts, Library,

Based on aggregated case studies described above and in-depth discussions of specific research use case from the series of Secure Research Data Working Group meetings during the spring and early summer, paths through this life cycle and sets of organizations are described for some stylized personnae below. The personnae are described as if at the beginning of the process, although many existing examples are at “mid-life” or beyond, and need guidance on next steps.

Persona 1: A researcher in a social science field identifies a restricted use dataset available at established archive/repository. The repository has published specifications for proposal processes to apply for access, and expectations about how the data will be handled and secured. [Support/Liaison: CSS-RA, BCBP, CPHS, ISP, IST, D-Lab] [examples: GSS, NCES, BLS, ICPSR]

Persona 2: A faculty member has developed a long-standing relationship with a data producer/provider, and is seeking to formalize a role in hosting data provided by that entity for the campus. [Support/Liaison: SPO, CSS-RA, ISP, IST, CPHS] [Examples: Census RDC, German IAB in Econ, ARE Scanner Data, Full Count Census Data in Demography]

Persona 3: A researcher wants to access a set of sensitive qualitative data (e.g. case notes, video recordings) and is seeking a secure collaborative environment that supports coding of the data by multiple collaborators. [Support/Liaison: IRB, ISP, CSS-IT,]

Persona 4: A school, department or large center are centrally hosting and supporting sets of data for their researchers/faculty. Some or all of these data are restricted use. The unit provides researchers with IT support and liaises with other relevant units on campus. Support is strong internally. [Examples: School of Social Welfare, Fisher Center at Haas, SPH]

Persona 5: A small department, research unit, or center are supporting faculty members/affiliates/researcher’s projects. Some or all of these projects are using sensitive data. Support is heavily dependent on few staff, and vulnerable to staffing changes. [Examples: IRLE, D-Lab, GSPP?]

Benchmarking and Case Studies

As described above, benchmarking of peer universities had been conducted during the RAE Redo project (see Appendix A). However, the project team knew that a deeper examination of services and practices at other institutions would be informative. Over the course of this project, three institutions were reviewed -- Stanford, UC Davis Medical Center and UCSF. Each institution had distinguishing features and a different emphasis to their approach to securing research data. We chose two medical centers for two reasons: 1) we had reviewed other EDU environments in the RAE study described above, and were familiar with the approaches taken by those institutions (EDU), most of which offer better resources and services than we do currently, and 2) we were under the impression that because medical center research has been governed by HIPAA requirements since 1996, those institutions would be more likely have a more mature approach to managing secure data than those in EDU environments. For the most part, this assumption, proved to be true, but perhaps not to the degree we had imagined.

The recommendations we make in this whitepaper (below) were drawn, in part, from what we considered to be the more impactful aspects of these programs, in the balance of cost and feasibility of near-term implementation. These aspects are described below.

Stanford: Stanford adopted a new data classification scheme in May 2015 based on the federal data classification system (FIPS PUB 199). Stanford's classification scheme includes consideration of research data, not just administrative data, and places the majority of research data into their low risk classification (minimal requirements), with an exception for data that is regulated, such as Protected Health Information (PHI), Social Security Numbers (SSNs), and financial account numbers.

To determine whether research data poses substantive risk or not in this new scheme, they defined a data risk assessment screening process -- a series of data review and sign-off steps that are facilitated by technical staff members, who take on the burden of facilitating this process for the researcher. This facilitation service is supported by 1.5 staff FTE. The end result is a report that identifies privacy and security risks as well as recommendations for safeguards.

This facilitation work, combined with their data classification scheme and guidance are the pillars of their approach to working with regulated data.

- Stanford Data Risk Assessment webpage
- Stanford Data Risk Assessment form (Word document)

UCSF: UCSF provides researchers with two essential research tools for working with sensitive data: 1) secure file transfer, via the web, to secure storage, and 2) a Windows-based secure computing environment that includes a suite of commonly used research applications (Access, Excel, MatLab, SAS, Stata, SPSS, etc.). These tools are supported by 1.5 staff FTE. The secure computing environment at UCSF is similar to BRC's Analytics Environments on Demand service.

- MyTransfer - Secure file transfer and storage
- MyResearch - Secure data hosting and compute environment
- Applications available for research in MyResearch secure compute environment

UC Davis/UCD Medical Center: UC Davis has many (>50) federally-funded grant projects that must comply with the security requirements expressed in the NIST 800-171 report - *Protecting Controlled Unclassified Information in Nonfederal Information Systems and Organizations*. These requirements are significantly higher than those typically called for in the Data Use Agreements we surveyed. They are also higher than what is required for work with HIPAA data.[1]

[1] Lawrence Berkeley Lab (LBL) and the Stanford Linear Accelerator Center (SLAC) are open science labs and are authorized to operate at the FISMA Low standard for information security. "The **Federal Information Security Management Act** (**FISMA**) is United States legislation that defines a comprehensive framework to protect government information, operations and assets against natural or man-made threats. **FISMA** was signed into law part of the Electronic Government Act of 2002." [source]. The security standard expressed in NIST 800-171 is slightly lower than the FISMA Moderate level.

Our review of known secure data use cases on campus suggests that UCB has far fewer projects that are subject to 800-171, less than 10. However, we suspect that the number is higher and that the need for 800-171 compliant computation on campus will increase.

UC Davis has been developing a cloud-based computational environment that complies with the standard expressed in 800-171, with a planned implementation by December 2017. UC Davis and UCB Research IT staff have discussed this project, are reviewing this architecture together, and have agreed in principle to explore the possibility of co-development or co-hosting in the future.

See architectural design below:

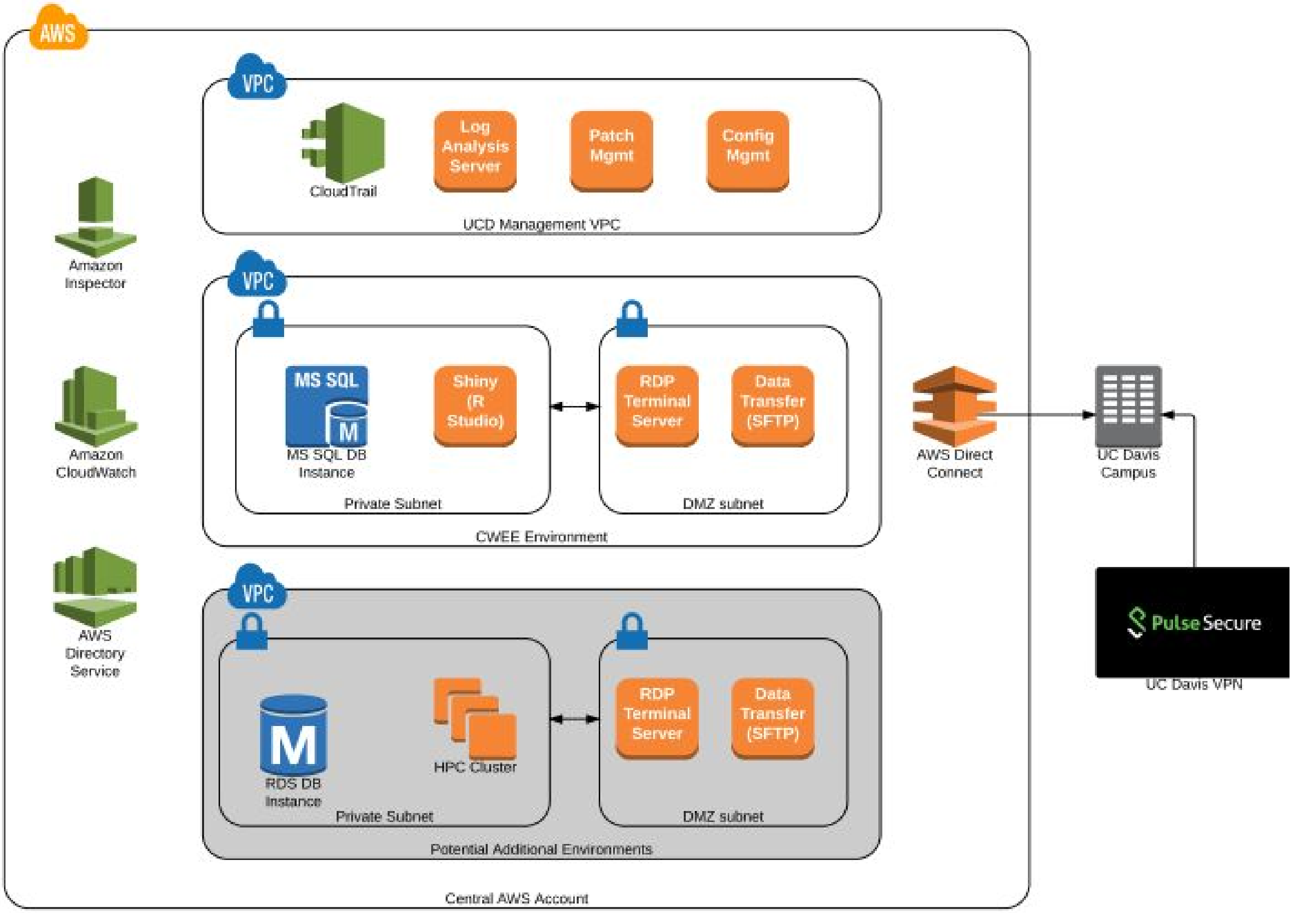
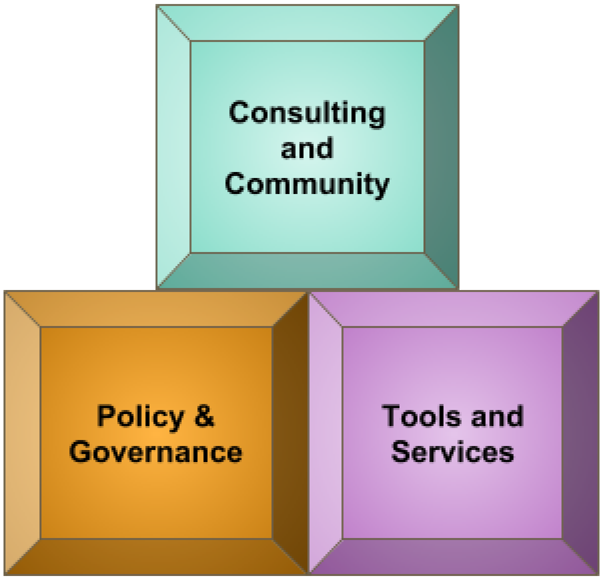


Figure 1: A cloud-based (AWS) architectural design for an 800-171 compliant computational environment, currently under development at UC Davis

The development effort for this environment is staffed by 1.0FTE, with leadership and support from the IT group in the College of Engineering. The UC Davis Medical Center research computing group is also interested in exploring partnership around supporting High Performance Computing (HPC) needs of medical researchers, as well as HPC with HIPAA data.

Recommendations

In documenting the current challenges faced by the campus and in conversations with researchers, compliance officers, and service providers, the project team has already helped the campus improve the sharing of best practices and the processes for reviewing Data Use Agreements and Data Security Plans. The current services and approaches available on campus are understood better, and service providers are building roadmaps to improve their capacity for helping researchers working with restricted data. In order to build on this progress and lay the foundation for more fundamental improvements, the project team makes the following recommendations. A comprehensive approach to restricted and sensitive data must recognize the following dimensions: consulting and community; policy and governance; and services and tools.



Consulting and Community

- Consulting and research facilitation: Improve processes for identifying protections needed for research data with the goal of facilitating research using and generating restricted data. In particular, simplify where researchers get help and clarify which offices are responsible for approval of agreements and plans. Continue to build out the partner-based consulting network in order to help researchers navigate relevant campus offices and services, complete data use agreements, and identify the required data protection controls as well as appropriate solutions and services. Ensure that environments and solutions are implemented correctly both from the researcher perspective as well as the security controls perspective. Determine what this service is called, who owns it and who participates, what else it is responsible for, what kind of support is needed in order for it to be successful, and whether policy is needed to support its role.
- Community-building: In March 2016, the D-Lab convened the Securing Research Data Working Group. Over a period of only four months, this community of researchers, compliance offices, and service providers discussed diverse perspectives related to

sensitive data and provided vital feedback to the project team. The team recommends that the working group continue to meet in order to make progress on the following goals: 1) Develop and broaden community awareness of practices, challenges, successes; 2) surface training topics to be considered by the training program; 3) review and provide feedback on guidelines, policy, and new services; and 4) perform outreach to academic domains and disciplines that are less represented (e.g., the College of Engineering, Boalt School of Law, Haas School of Business) in order to identify use cases, needs, solution patterns, and current investments in local services.

- Training: Develop a training program targeting both a) service providers especially related to the research data classification guidelines and broader policy and b) researchers as well, related to research data classification, security protections and best practices, and needed services and capabilities.

Policy and Governance

- Research Data Classification guidance: Publish and disseminate the Research Data Classification Guideline developed by Information Security and Policy. The guideline is designed to help compliance offices, service providers, and research support staff who assist researchers in identifying the data protection level and protections required for sensitive research data.
- Broader Research Data Policy: In addition, the campus should develop and adopt policy that specifically addresses the broader set of issues related to research data. As documented above, a review of other institutions demonstrated that four criteria should be met in such a policy or set of policies. UCLA's Interim Guidance (attached in Appendix D), or that of any of the "green tier" institutions, provides an example that can be implemented while broader UC systemwide issues are being discussed. The key elements of this would define roles and responsibilities of different campus actors so that there is clarity for researchers and staff about the scope of support, and so that we can better ensure a lifecycle approach to securing research data.
- Governance: Work with Research, Teaching and Learning Technologies Committee (RTLTC), other governance bodies, and stakeholders to ensure that practices and policies are aligned with campus governance, and to assess risks and opportunities for campus leadership.

Tools and Services

- New Service: Secure - Analytics Environments on Demand Service: We propose the creation of a new campus-wide service to provide secure computing based on Research IT's existing Analytics Environments on Demand (AEoD) service. The AEoD service provides researchers with scalable, interactive research computing desktop environments and is currently in active use at the Goldman School of Public Policy, Haas School of Business and the Archeological Research Facility. The service is a based on a pair of enterprise computing technologies -- VMWare and Citrix -- reconfigured for research computing. These technologies may be used to provision and distribute research desktops (virtual machines) via the web, in a secure fashion. The AEoD Service support model includes partnership with local IT staff to provide end-user support and growth, as well as assistance with documentation, planning and governance. AEoD may be used for research and instruction. In partnership with IST service teams and Information Security and Policy, this service could be built to ensure that campus standards for PL2 environments and systems are met or exceeded. An initial service proposal for Secure AEoD is attached in Appendix I.

- Box (PL2) restricted data file sharing and collaboration service: For many researchers working with restricted data, their primary need is for a service that will allow them to store data and collaborate with other researchers. For data that do not require protections, Box is the recommended file sharing service for campus by the Research Data Management program. A number of peer institutions have taken further steps to certify their instances of Box to be used for managing restricted data. Our collaboration with the bConnected team in IST:API indicates that they agree that there is demand and that Box would be an appropriate service for this purpose. They also understand what kind of work would be needed.
- Other services and tools: With the broader community on campus and beyond, continue to investigate other services and tools, especially those that have proven useful and sustainable at partners institutions. For example, UCSF and UC Davis have services developed around secure data software, data transfer and storage, and research collaboration.

Conclusions

Importantly, securing research data is not just a technical challenge. There are numerous policy and cultural issues, and the landscape of political, technological, and research challenges and opportunities is evolving rapidly. Currently experts on campus are distributed across different organizations, and while they are beginning to coordinate their work better, there is a growing demand for expertise and consulting, and for improved service offerings for storing sensitive data, for performing computations in a secure environment, for transferring data with different levels of sensitivity, for collaboration among team members both on campus and beyond, and for preserving, protecting and sharing data sets for the future.

In this six-month project, the project team has developed a fundamental understanding of the various dimensions of this challenge: consulting and community; policy and governance; and services and tools. Importantly, a community of experts from across campus has now formed, and discussions among researchers, service providers, and staff responsible for regulatory and compliance oversight have been productive. The recommendations made by the project team include both near-term actions campus can take right away as well as others that will require more discussion and eventually real resources and investment. However, the problems faced by the entire campus community are not going away. Instead, they will become only more challenging as data-intensive research continues to grow and as the regulatory environment and possible consequences increase in complexity and cost.

Appendices

A. RAE Redo one-pager: Secure Research Data and Computing (PDF)

B. Securing Research Data project charter
C. Benchmarking Policies at Other Institutions
D. Interim UCLA Guidance on Access to and Management of Research Data and Tangible Research Materials (PDF)
E. Benchmarking Questions for Secure Data and Project (PDF)
F. HIPAA - origin and background
G. Campus Units and Organizations supporting Sensitive Research Data
H. Service Proposal: Secure Analytics Environments on Demand (PDF)

References

- Leveraging Cloud Service for NIST SP 800-171

Area for Rewrite of Recommendations

Tools (Jason)
Challenge Statement
Ubiquitous data collection, unlimited data storage capacity and a steady stream of new data analysis tools have coalesced to form a Data Revolution for research. This presents an unprecedented opportunity, but the opportunity is coupled with a down-side of real and costly risk in the form of data security breaches.[2]

Recommendation
Researchers need secure computing spaces in which they can conduct research with a minimum of cyber security risk. We recommend that Research IT's Analytics Environments on Demand Service be hardened against the threat of data security breaches, and that this effort be made part of an ongoing program to address our researchers' needs for secure research computing environments and support thereof.

Policy and Governance (Chris)

How Chris tried to capture recommendations for a single RIT goal:

Secure Research Data and Computation: Implement the set of recommendations developed in the Securing Research Data whitepaper to make broad-based improvements in three areas:

A. Consulting and community: Formalize multi-departmental process for reviewing security plans and data use agreements; continue the D-Lab Working Group

[2] Berkeley News - Berkeley Alerting 80,000 individuals to cyberattack

B. Policy and guidance: Adopt interim guidelines for research data (based on work at other UCs) that address a broad set of policy issues related to research data.
C. Tools and services: Develop secure virtual machine environment (Secure-AEoD); partner with other campuses developing secure research data and computation tools and services.

Reviewer notes

From: **Steve MASOVER**

Hi Chris and team --

Thanks for a chance to look at this information-rich white paper. Given the timing, I'm going to need to limit my response to a few high level suggestions and a couple that are more detailed. I hope they prove helpful.

First, I'd like to suggest that nine is *slightly* too many recommendations to list in the *Recommendations* section. I wonder whether these can be consolidated at all; for example, can the "Training" recommendation be integrated into each of the other recommendations in the "Consulting and Community" category?

Second, I wonder if there is any way to ballpark FTE and other costs associated with each recommendation. I imagine that anyone reading this who is on the receiving end of a budget request will want to have at least a rough idea even at this possibly-early point in a process that flags need for additional campus investment. And given that the RAE document gives some (perhaps even rougher!) guesstimates of investment necessary to up Berkeley's game in this area, it would make sense to be able to map from that document's rough guesses to the (somewhat?) more refined estimates in this document.

Third, I'd like to suggest that the categories of recommendation be listed in the executive summary more directly as the "multiple aspects of this challenge" (as expressed in the final paragraph the summary). As written, it's not clear that there's a straight line that ought to be drawn between the "multiple aspects" referenced in the first sentenced, and the three elements of a "comprehensive approach" referenced in the second. I would also consider including in this brief paragraph a cost (FTE, etc.) to implement each aspect/element/category; or a sum total to implement all the recommendations listed in the more detailed *Recommendations* section. Again, I think that's a context that an executive will want to have from the outset, before diving into the meat of this white paper.

A few specifics:

- Pg. 8: "*Experience shows that though researchers often ask for help establishing a secure computing environment or accessing secure storage, secure computing and* *secure storage needs frequently go together.*" Whose experienced is referenced here? Making this explicit rather than general would both ground this assertion in specific authority, and avoid an impression of 'hand-waving' that generalized assertions such as this can evoke.

- Pg. 9-11 "Current UC Berkeley Research Data Policies" is a great section, but this sentence -- "*Finally, the "red" tier reflects institutions, like UC Berkeley, leave unanswered at least several of the Four Criteria relevant for a comprehensive approach to securing research data throughout its lifecycle.*" -- leaves one wondering -- policy addressing exactly which of the "Four Criteria" is missing at UCB? Since there are only four in question, it strikes me as odd that specificity is elided here.
- Pg. 15: "*Our review of known secure data use cases on campus suggests that UCB has far fewer projects that are subject to 800-171, less than 10. However, we suspect that the number is higher and that the need for 800-171 compliant computation on campus will increase.*" What is the basis of this "suspicion"? Is this really just extending the earlier discussion of how the team's survey of the campus is known to be incomplete? If so, I think it would be best to reference this again at this point in the paper, as grounds for the assertion, rather than make an unexplained and general reference to the team's beliefs.

Again, I hope this is helpful,

Steve

Executive Summary for Consulting and Community.

Securing research data is far from a new challenge, but the pace of change in the ubiquity of observation and the technologies to capture, store, link and preserve those observations have transformed the challenge to previously unimaginable levels. The data underlying this transformation derive from many mechanisms - new sensor technology and applications, administrative records, new tools to identify and measure biological objects, digitization of analog data (text, audio, video), traditional surveys and censuses - and are accompanied by an equally diverse set of data "owners", collectors, regulators, and interested parties. This state of affairs may be summarized in terms of five related challenges: Complexity, Confusion, Conflict, Cost and Change.

We address these challenges with four linked strategies: Communication, Coordination, Consultation, and Collaboration, which suggest the specific initial tactics summarized below and detailed in the body of this report.

Challenge: **Complexity and Confusion**. Data are obtained from an increasingly diverse set of producers who have differing concerns and desires; they are legally constrained by a variety of federal, state, and local laws and policies, and; are shepherded and managed by campus institutions at different points in the research lifecycle. This patchwork of ownership, interests, laws, management and research use creates ambiguity and confusion.

Strategy: **Communication and Coordination**. At a campus and extra-campus level, Berkeley can create coordination and communication among the strong set of internal institutions on campus and the broad network of external partners external that support researchers. While individually strong, our institutions often have overlapping missions and lack of awareness of other institutional actors, missions, and spheres of responsibility.

Tactics: **Communication and Coordination**. Develop and share a clear set of roles and responsibilities of institutional actors, delineating contacts, processes, and interactions with internal and external partners. Identify multi-unit process and paths for reviewing security plans and data use agreements. (see section X and Appendix Y for details).

Challenge: **Conflict and Cost**. Researchers and staff may fail to plan for, ameliorate risks and liabilities, or protect research data because of misunderstandings or lack of clarity in what is required or ideal, because costs of learning are high, or because the costs of fully understanding and achieving compliance conflict with more narrowly scoped research goals.

Strategy: **Communication and Collaboration**. <u>Among other strategies</u>, reduce the cost of learning and encouraging straightforward low risk communications; pro-actively create a community where researchers, policymakers, support staff, and those with regulatory roles discuss needs, and; communicate clear guidance in ways which reduce search costs and identify compatible cost-effective solutions.

Tactics: **Communication and Collaboration**. Provide for regular structured meetings between researchers, research administration, IT staff, and other support staff. Produce published working documents and discoverable guidance on specific security solutions identifying costs of implementation. (see section X and Y for details).

Challenge: **Complexity and Change.** With the number of moving parts and the pace of change in legal, campus, research and technical landscapes,

Strategy: **Consultation and Collaboration.** Aspects of working with restricted use data that are relatively straightforward or invariant should be distinguished from those in transition or where complexity requires highly tailored solutions. Consistent resources for each should be developed and made available to target audiences - researchers, graduate students, and support staff.

Tactics: **Consultation and Training.** Trainings should be created and provided for aspects that are relatively straightforward in the form of online trainings, and for moderately complex aspects and materials in the form of regular workshops. Expertise for solutions and aspects that require a great deal of tailoring should be made available through easily discoverable paths, like those now provided through RDM and D-Lab.