

TAMU Virtual Data Library (TAMU-ViDaL): A Secure and Compliant Data Infrastructure

Hye-Chung Kum

Professor

Director TAMU ViDaL (<https://vidal.tamu.edu/>)

Director Population Informatics Lab (<https://pinformatics.org/>)

Presidential Impact Fellow

Department of Health Policy and Management, School of Public Health

Texas A&M Institute of Data Science (TAMIDS)

Texas A&M University



5/23/2021

1

1

Agenda: TAMU Virtual Data Library (TAMU ViDaL)

- Goals
- Compliance
- Basic System Pros & Cons
- Design Principles (if time permits)

5/23/2021

2

2

Do you know what a compliant data infrastructure is ?

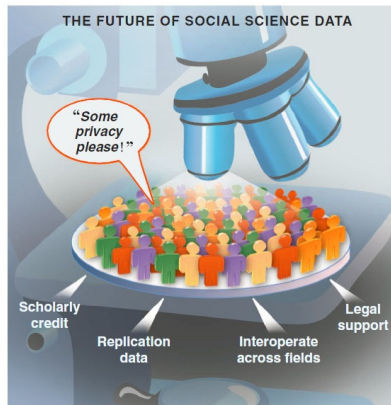


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

Gary King, *Ensuring the Data-Rich Future of the Social Sciences*, *Science*, vol 331, 2011, pp 719-721.

5/23/2021

3

- Compliant system
 - All projects must get approval
 - All activity is monitored and subject to audits
 - Data that leave the system are closely monitored
- Examples
 - VA VINCI system
 - Data enclaves
 - Federal census RDC

3

Have you used a compliance system before?

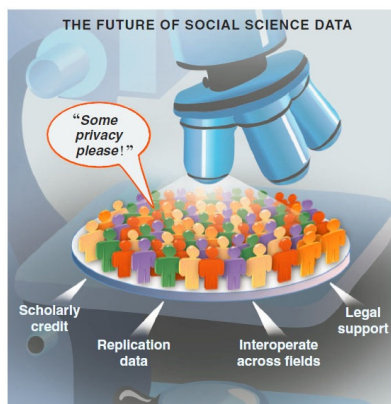


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

Gary King, *Ensuring the Data-Rich Future of the Social Sciences*, *Science*, vol 331, 2011, pp 719-721.

5/24/2021

4

- Compliant system
 - All projects must get approval
 - All activity is monitored and subject to audits
 - Data that leave the system are closely monitored
- Examples
 - VA VINCI system
 - Data enclaves
 - Federal census RDC

4

TAMU Virtual Data Library (TAMU ViDaL)

- Goal: Secure and **Legally Compliant** Data Infrastructure TAMU wide
 - Facility: West Data Center
 - Hardware: Data Intensive Cloud computing nodes (large & fast RAM nodes up to 1.5T, GPU nodes)
 - System software: VMs, Linux, and Windows
 - Basic software
 - Open source: R, python,
 - Licensed software: SAS, stata, matlab
 - Data (Data custodian):
 - Initially, Pls bring
 - Some available with approval: HCUP/Medicare/Texas Discharge PUDF data/1930 & 1940 Census data
- Support from Texas A&M's Research Development Fund (VPR) in spring 2018: \$1.4M

5/23/2021
5

5

TAMU Virtual Data Library (TAMU ViDaL)

<https://vidal.tamu.edu/>

- Management Team
 - Director (Dr. Hye-Chung Kum)
 - Associate Director of Facility & Computing (Dr. Honggao Liu)
 - Associate Director of Statistical Methods (Dr. Bobak J. Mortazavi)
 - Associate Director of User Engagement (Dr. Mark Fossett)
 - Privacy Officer (John Pryde)
- Advisory Board
 - Secure cloud computing (Dr. Dilma Da Silva)
 - Computer security (Dr. Peter Yu)
 - Legal expertise in information privacy (Cason Schmit, JD)
 - Data repositories (Dr. Bruce Herbert)
 - Human Subjects Research & IRB (Dr. Eva Shipp)
 - Statistical methods (Dr. Valen Johnson)

- User Work Group
 - [TAMU SPH] School of Public Health (Dr. Hye-Chung Kum)
 - [TEES & COE] College of Engineering & TEES (Dr. Dr. Bobak J. Mortazavi)
 - [AgriLife & COAL] Colleges of Agriculture and Life Sciences & AgriLife (Dr. Reid Stevens)
 - [TAMU COS] College of Science (Dr. Valen Johnson)
 - [TAMU CLA] College of Liberal Arts (Dr. Mark Fossett)
 - [TAMU Mays] Mays Business School (Dr. Venkatesh Shankar)
 - [TAMU CEHD] College of Education & Human Development (Dr. Whitney Garney)
 - [TAMU CON] College of Nursing (Dr. Robin Page)
 - [TTI] Texas A&M Transportation Institute (Dr. Eva Shipp)
 - [PPRI] Public Policy Research Institute (Dr. Kirby Goidel)
 - [RCHI] Rural & Community Health Institute (Dr. Peter Yu)

2/27/2019
6

6

How do we provide secure compliant computing? Privacy-by-Design



- A different perspective on privacy and research using personal data
- Personal Data is Delicate/Hazardous/Valuable
- Important to have proper systems in place that give protection but allow for continued research in a safe manner
- All hazardous material need standards
 - Safe environments to handle them in : closed computer server system lab
 - Proper handling procedures : what software are allowed to run on the data
 - Safe containers to store them : DB system
- **Minimize Information Privacy Risk**
 - **Combination of technology, governance, and transparency (operation)**



5/23/2021

7

7

Compliant Computing




- Secure Computing AND
 - Legal Requirements
 - Documentation ... on due diligence
 - Lots of logging, periodic audits
 - Training
 - Keeping current
- Requirements depends on the use case
 - HIPAA, TX HB 300, FERPA, FISMA, CUI (NIST Controlled Unclassified Information)
 - Etc...
 - Ability to work with **security expert** and **privacy expert** to work out the requirements for your project and get the required signatures about the system



5/23/2021


8

8




Usability

- Virtual Machine
- OS: Linux and Windows
- Software: R, Python, SAS, Stata, Matlab, Microsoft office
 - PI is responsible for any license fee beyond basic
 - Software fund
- Remote access
- Collaborations
 - Accounts for any collaborator as needed, with appropriate training (Sponsoring PI is responsible)



5/24/2021 9

9




Computer Specifications

- Mix of large and conventional computing
- Lots of Smaller Projects
 - Core: multi core (16+)
 - RAM capacity: up to 128 GB
 - Storage (high speed) capacity: up to 1 TB
- Large Projects
 - Core: multi core (16+)
 - RAM capacity: 128 GB - 1.5 TB
 - Storage (high speed) capacity: 1TB and up
- Backup Storage (not high speed): Automatic
 - Split space into backed up and not backed up scratch space for large computations

5/23/2021 10

10



TAMU Virtual Data Library (TAMU ViDaL)

Pros


- Immediate
 - No need to find locked office
 - Free hardware (during 3 year pilot)
 - Secure system
 - Software may be available
 - Help with getting approval on legal documents
 - Easier to collaborate
- Longer term
 - May have data (more longer term)

Cons

- MUST have an approved research plan stating
 - What is the research goals
 - What data
 - Who will have access
 - Project termination plans
- Approval process required
 - No different than before
- Must use ViDaL system
 - May not be as usable
- Shared resource
 - May get crowded
- All activity is monitored on the system with audits
- PI is responsible for all team members

5/23/2021
11

11




Compliance: Guiding Principles

- Risk-Based Compliance
 - High-Risk data will be given greater protections/controls
 - Low-Risk data will have fewer protections/controls
 - Since we cannot remove all risk (without halting research), some low risk activities will be tolerated but monitored
- Avoid Redundant Approvals and Unnecessary “Red-Tape”
 - Ex: TAMU ViDaL will not duplicate IRB approval process
- Empower PIs
 - Provide PIs with tools to manage their data and research in a compliant way
 - Ex: compliance checklists for new users, system monitoring/reports
- Check and Adjust
 - If compliance processes are not working, we will make adjustments

5/23/2021
12


12



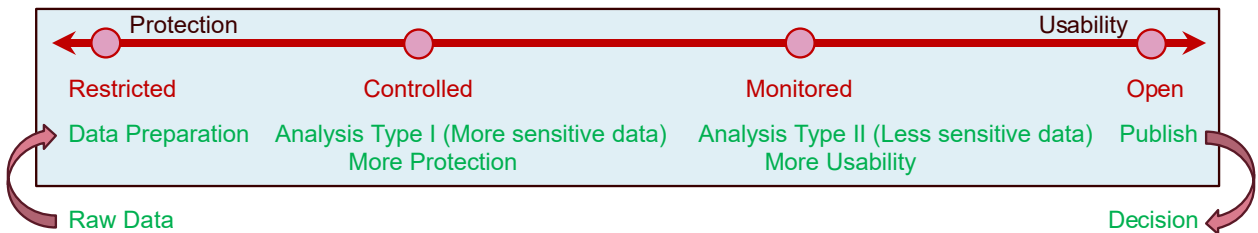
Design Principles

- First, the Minimum Necessary Standard states that maximum privacy protection is provided when the minimum information needed for the task is accessed at any given time.
- Second, the Maximum Usability Principle states that data are most usable when access to the data is least restrictive (i.e. direct remote access is most usable).
- Based on common activities in the workflow, design systems that have maximum access to the minimum amount of information required
- **Combination of technology, governance, and transparency (operation)**

13



Use Published Data for Good Decision Making (Kum & Ahalt 2013)



Deployed together the four data access models can provide a comprehensive system for privacy protection, balancing the risk and usability of secondary data in population informatics research

Kum, H.C., and Ahalt, S. (2013). Privacy by Design: Understanding Data Access Models for Secondary Data. *American Medical Informatics Association (AMIA) joint summits on translation science: clinical research informatics* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3845756/>

5/24/2021 14

14

Texas Virtual Data Library (Texas ViDaL): focus on Monitored & Controlled access

← Protection → Usability

Comparison of risk and usability

		Restricted Access	Controlled Access	Monitored Access	Open Access
Usability	U1.1: Software (SW)	Only preinstalled data integration & tabulation SW. No query capacity	Requested and approved statistical software only	Any software	Any software
	U1.2: Data	No outside data allowed But PII data	Only preapproved outside data allowed	Any data	Any data
	U2: Access	No Remote Access	Remote Access	Remote Access	Remote Access
Risk	R1: Cryptographic Attack	Very Low Risk	Low Risk. Would have to break into VM.	High Risk	NA
	R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	NA

5/24/2021
15


15

Privacy Protection Mechanism

Access	Restricted Access	Controlled Access	Monitored Access	Open Access
Protection Approach	Physical restriction to access	Lock down VM (limit what you can do on the system)	Information accountability	Disclosure Limitation
Monitoring Use	All use on & OFF the computer is monitored	All use on the computer is monitored		Trust
IRB	Full IRB approved	Expedited IRB approved	IRB Exempt (register)	Terms of Use
R1: Cryptographic Attack	Very Low Risk	Low Risk. Would have to break into VM	High Risk	NA
R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	

5/24/2021
16

16




Compliance bottom line

- SAME AS BEFORE
- Will try to facilitate and streamline process
- Compliance starts with PIs

5/24/2021 17

17



Sustainability: Business model

- Federal Census RDC Model
 - Seed Funding to get us started: RDF
 - Consortium of Institutions: infrastructure support
 - PIs from participating institutions: free for typical size projects (committee)
 - Others may pay per project
- External Funding: Ex HHSC projects

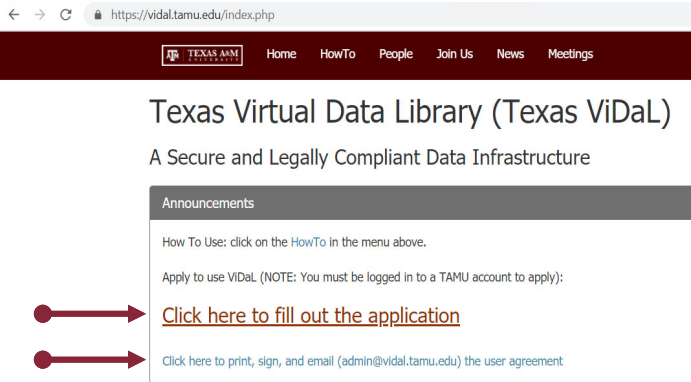
5/24/2021 18

18




How do I get access ?

- Faculty: Just apply on line - simple gform on ViDaL website
 - User Agreement
- Student: find a faculty host (similar to IRB)
 - Usually chair



2/27/2019 19

19



Texas Virtual Data Library (Texas ViDaL) User Agreement

Print as PDF, read, check appropriate role, sign, and email to admin@vidal.tamu.edu

- Data classification:
 - Your data classification level is indicated by N where ~/projN/your project folder
 - For all level 2 & 3 projects: Important - **All users must keep all project data including derived data under their root project folder.** Please be careful not to write out to your home space.
- Compliant projects are reviewed and approved by John Pryde, the privacy officer of TAMU. These are approved on a per project basis. **So you MUST apply for EACH project that requires compliance separately.** This is not a blanket approval for all projects.
- Please acknowledge vidal whenever appropriate
 - "This work was supported in part by the Texas Virtual Data Library (ViDaL) funded by the Texas A&M University Research Development Fund"
 - Please report any proposals or papers that benefit from vidal to us over email (admin@vidal.tamu.edu). This is important for tracking and reporting purposes for future funding to keep vidal affordable to us.
- Please store and share this email with all users in your project as they join. As PI, you are responsible for training your staff and making sure that they understand the requirements for using ViDaL. You can also find this information anytime on the vidal website.

- Approval is project based.
- New projects have to do new applications
- This is because each project has different requirements that must be reviewed

2/27/2019 20

20

Do you want to drive on the information highway?

- **Be prepared to get into accidents**, especially as we all figure out how to do this more safely.
- Invest in personnel to **do your due diligence** to stay up to date with the fast paced security technology and privacy regulations
- Invest in technologies (i.e., seat belts & airbags) that can **minimize the real harm when accidents do occur**
- Kum et. al. 2013.



Kum, H. C., Krishnamurthy, A., Machanavajjhala, A., & Ahalt, S. C. (2013). Social genome: Putting big data to work for population informatics. *Computer*, 47(1), 56-63.

5/24/2021

21

21

Thank You!!
<https://vidal.tamu.edu/>



Privacy is a BUDGET constrained problem

The goal is to achieve the maximum utility under a fixed privacy budget



22

22