Security Options for Restricted-use Research Data

John E Marcotte, PhD

ICPSR

University of Michigan

May 2015

Security Options for Restricted-use Research Data

Research Data often have special security requirements. Laws and regulations compel compliance.

The purpose of security is to prevent disclosure (or at least make disclosure difficult!)

My Perspective

- Research Data Provider
- Security Professional
- Researcher

Security Options for Restricted-use Research Data

I. Data and Disclosure
II. Requirements
III. Compliance
IV. Security Options
V. Questions and Discussion

I. Data and Disclosure

- Data
- Restricted-use Data
- Sensitive Data
- Disclosure Risk

Data

- The term “Data” can convey different ideas to researchers and computing professionals.
- For researchers, Data refer to the information to be analyzed
- For computing professionals, Data refer to all information.

Restricted-use Data

Restricted-use Data contain information that is not publicly available

- Restricted-use Data have security requirements
- Data may be restricted-use because they are sensitive, disclosive or proprietary

Sensitive Data

Information that can cause harm or legal jeopardy; damage reputation

- Some examples are:
- Health information
- Drug use
- Criminal record
- School record

Disclosure Risk

Chances of re-identification of research subjects (individuals or organizations)

Disclosure is the identification of subjects

- Personally Identifiable Information (PII)
- Indirect or inferential risk based on combination of variables
- Disclosive data may lead to re-identification

Disclosure

- Unauthorized people obtain access to data or summary of data

- Explicit identifiers

  Laws requiring notification and remediation

II. Requirements

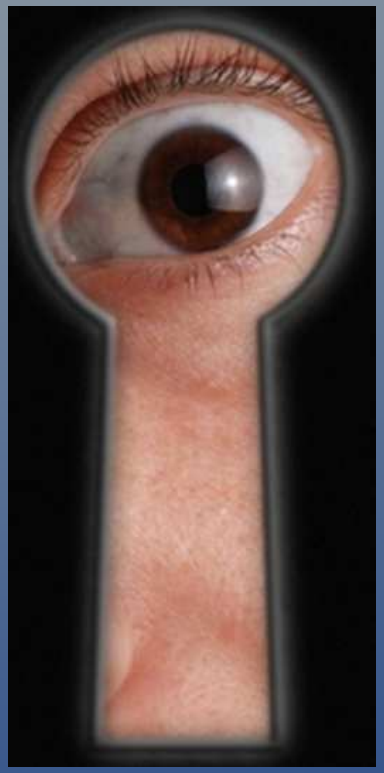
For researcher to analyze restricted-use data, they must submit:

- Data Security Plan
- IRB approval
- Data Use Agreement between institutions
- Confidentiality pledges

Data Security Plan

Data Security Plan describes how researcher and institution will prevent misappropriation of data and inadvertent disclosure

Security Risks



- Unauthorized access
- Break-ins
- Hijacking of the system by malware or botware
- Interception of network traffic
- Loss
- Theft
- Eavesdropping
- Paper output
- Human error

Special Security

Special security to prevent disclosure:

- Encrypting information at rest and transport
- Blocking unencrypted files and information from being copied to the Internet
- Vetting of results for disclosiveness
- Monitoring of processing to prevent the unauthorized transcribing of disclosive material

Data Leaks

II. Compliance

- How to ensure compliance:
  - Rely on researcher's agreement to follow protocols
  - Implement technology
  - External review
- Regulations

Regulations

- FISMA/NIST
- FIPS 140-2
- Laws:
  - Confidential Information Protection and Statistical Efficiency Act, (CIPSEA)
  - Family Educational Rights and Privacy Act (FERPA)
  - Health Insurance Portability and Accountability Act (HIPAA)

Professional Staff

Professional staff are needed to:

- Document and implement FISMA/NIST security controls
- Implement special security
- Administer systems
- Vet output

Loss versus Disclosure

- For secondary data analysis, loss is often preferable to backups and redundancies that might increase the risk of disclosure.
- Loss of encryption keys will render data unreadable but is preferable to increasing disclosure risk

Consequences

may occur even if disclosure is inadvertent

- Costs of remediation
- Loss of reputation
- Suspension of research
- Suspension of funding
- Barred from future projects

III. Security Options

- Locked room
- Guarded room
- Vetted output
- Terminal Server
- Non-networked computer
- Private network
- Encryption

Concerns

- Where are data stored
- Where are data viewed
- Collaboration
- Simultaneous access to data from different sources
- Costs and usability

Secure Server *and* Client

- Client computer is as important as server
- Client can still disclose

Collaboration

Data protection requirements often impede collaborations

Simultaneous Access

- Researcher can have access to data from multiple sources but not at the same time

- Data become unacceptably disclosive when sources are combined

Security levels

Depending on the sensitivity and the disclosiveness, restricted-use data for research can be accommodated in one three levels

1) **Low to Moderate**
2) **Moderate to High**
3) **High to Very High**

Security Levels

	Encryption	**Internet**	**Output**	**Processing**
Restricted-use 1	Encrypted	Internet blocked	Self-vetted	Self-monitored
Restricted-use 2	Encrypted	Internet blocked	Vetted	Self-monitored
Restricted-use 3	Encrypted	Internet blocked	Vetted	Monitored

Restricted-use 1

Option 1

- Non-networked computer
- Locked office
- Server and client are same machine
  - Pro: Relatively cheap to setup; two computers with KVM switch
  - Con: Impedes collaboration

Restricted-use 1

Option 2

- Private network without Internet
- Locked room for client and server
- Server and client are different
  - ❖ Pro: Better collaboration
  - o Con: May need two client computers

Restricted-use 2

Option 1

- Terminal Server or Virtual Appliance that allows incoming connections only; files cannot be copied out
- Only authorized personnel can transfer files
  - ❖ Pro: Allows external vetting
  - ❖ Pro: Data never leave server
  - ❖ Pro: Collaboration space
  - o Con: Expensive to setup for only one or two projects
  - o Con: Still need office for client

Restricted-use 2

Option 2

- "Cold" room with secure access
- Only keyboard, mouse and monitor are accessible
- Only authorized personnel can transfer files
  - ❖ Pro: Allows external vetting
  - ❖ Pro: Data never leave "cold" room
  - o Con: Researchers must go to special room
  - o Con: Impedes research because of inconvenience

Restricted-use 3

Option 1

- Batch server
- Results are vetted before being returned
- No access to original data
  - ❖ Pro: Allows external vetting
  - ❖ Pro: No travel required
  - o Con: Slow process for getting results that may only be intermediate or for debugging

➢ Extra: Synthetic data with same structure as original data for interactive testing before batch runs

Restricted-use 3

Option 2

- "Cold" room with guard
- Only keyboard, mouse and monitor are accessible
- Only authorized personnel can transfer files
  - ❖ Pro: Allows external vetting and monitoring
  - ❖ Pro: Data and notes never leave "cold" room; guard inspects everything
  - o Con: Researchers must travel to special location
  - o Con: Impedes research because of inconvenience
  - o Expensive because of personnel

Costs and Usability

- Restricted-use 3 is the most expensive

  and presents the most barriers to research

- Restricted-use 2 is expensive for one project and requires researchers to submit output for vetting.

Two-factor Authentication

Increasing requirement for research data

- Something you know
  password
- Something you have
  fob
- Biometric authentication:

fingerprint, retinal scanners

Questions and Discussion