Data Ethics and Privacy for Researchers

Kelley F. Rowan
*Florida International University*, krowan@fiu.edu

Follow this and additional works at: [https://digitalcommons.fiu.edu/glworks](https://digitalcommons.fiu.edu/glworks)

Part of the [Data Science Commons](https://digitalcommons.fiu.edu/glworks), and the [Library and Information Science Commons](https://digitalcommons.fiu.edu/glworks)

**Recommended Citation**


This work is brought to you for free and open access by the FIU Libraries at FIU Digital Commons. It has been accepted for inclusion in Works of the FIU Libraries by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.
Data ethics and privacy for researchers

Kelley Rowan, Digital Archives Librarian
Agenda

Laws, Torts, and Regulations
Definitions, Getting Started, and Workflows
PII v. Personal (Sensitive) Data
Anonymization Techniques & Tools
Encryption Resources
Avoiding Anonymization
Data Anonymization Services
US Federal regulations

- **Privacy Act of 1974**
  - Governs federal records

- **Children's Online Privacy Protection Act of 1988, 15 U.S.C 6501-6505**

- **Gramm-Leach Bliley Act, 1999 (Financial Modernization Act)**
  - Governs banks and private financial information

- **Family Educational Rights and Privacy Act (FERPA)**
  - [https://studentprivacy.ed.gov/ferpa](https://studentprivacy.ed.gov/ferpa)

- **Health Insurance Portability & Accountability Act (HIPAA)**
  - [https://www.hhs.gov/hipaa/for-professionals/privacy/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/index.html)

4 Privacy Torts: Intrusion into Seclusion; Public disclosure of private facts and data
State Privacy Laws

https://iapp.org/resources/topics/us-state-privacy/

TITLE 1.81.5. California Consumer Privacy Act of 2018 [1798.100 - 1798.199.100]

CCPA & CPRA

link: https://oag.ca.gov/privacy/ccpa

California Privacy Rights Act – 2023


Fully enforced beginning July 1, 2023
International Law

GDPR
General Data Protection Regulation (2018)
https://gdpr.eu/

*Secure people’s data

*Make it easy for people to exercise control over their data

• **Lawfulness, fairness and transparency** — Processing must be lawful, fair, and transparent to the data subject.

• **Purpose limitation** — You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.

• **Data minimization** — You should collect and process only as much data as absolutely necessary for the purposes specified.

• **Accuracy** — You must keep personal data accurate and up to date.

• **Storage limitation** — You may only store personally identifying data for as long as necessary for the specified purpose.

• **Integrity and confidentiality** — Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).

• **Accountability** — The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.
**RTBF**

*Right to be Forgotten*
- erasure obligations

**RTBI**

*Right to be Informed*
- must provide a privacy policy
GDPR non-compliance consequences

$21,150,500 or 4% of profits, whichever is greater

Enforcement tracker: https://www.enforcementtracker.com/

Chapter 1, Article 4 (GDPR) makes no distinction between companies and individuals
• PCI DSS

Payment Card Industry Data Security Standards

https://www.pcisecuritystandards.org/

PCI Data Handling Standards
• Do not keep physical copies
• NEVER store the CVV
• Store in a secure locked place
• Limit personnel with access
• Use security cameras in area
• Assess and securely destroy every 3 months
• Provide a privacy plan for clients
Definitions

GDPR Article 4(1)
“personal data” pertains to “any information relating to an identified or identifiable natural person (‘data subject’)

- PII = personally identifiable information
- Personal (sensitive) data = any information related to an individual
- Data subject = person
Before collecting data...

- Determine your risks (impact assessment)
- Develop a security plan (VPN, data anonymization)
- Share your privacy policy
An impact assessment must always be conducted in high-risk situations such as when collecting, storing, and/or processing personal (sensitive) data.

- What are the consequences of a breach?
- How will you secure personal data?

(encryption, VPN, de-identification, key storage, 3rd party encryption, data storage and destruction)
Impact assessment

- Identify data risks
- Physical space
- Who has access (for editing and storing)
- Identify risky behaviours (accessing on various devices)
- Identify best security options
Privacy Policy

✓ What data will be collected
✓ How it is collected
✓ How it will be used and for what purpose
✓ Who has access
✓ How the data is stored and for how long
✓ How the data will be securely destroyed
✓ Contact information
Keeping data safe with anonymization techniques

**Identify**
Identify your PII and sensitive data

**Ascertain**
Ascertain whether other individuals will need working access to the data

**Determine**
Determine whether the data will be published

**Develop a key**
Secure your key
Computer Security

- Use passphrases with 2FA
- Use biometrics with 2FA
- Use GDPR compliant tools

https://gdpr.eu/compliant-services/
Workflow for teams with access to personal data

**Ideal:**

one person has access and anonymizes before granting access to others

**Good:**

2+ trusted people may work on collecting and anonymizing, including secure storage of a key.

**Weak:**

Everyone in the workplace has access, anonymization happens at the end before publishing.
Identifying PII and personal data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Last Name</td>
<td>First Name</td>
<td>FI #</td>
<td>DC# ingest</td>
<td>DC# published</td>
<td>Notes published</td>
<td>Full text pdf/a</td>
<td>embargo</td>
<td>M/PhD</td>
<td>Major</td>
</tr>
<tr>
<td>2</td>
<td>Sale</td>
<td>Tonia</td>
<td>FIDC000332</td>
<td>8431</td>
<td>3670</td>
<td>12.10.1998</td>
<td>12.10.1998</td>
<td>2 yrs.</td>
<td>MS</td>
<td>Hospitality</td>
</tr>
<tr>
<td>4</td>
<td>Bennett</td>
<td>William</td>
<td>FIDC000334</td>
<td>8433</td>
<td>3672</td>
<td>12.10.1998</td>
<td>12.10.1998</td>
<td>1 yr.</td>
<td>MS</td>
<td>speech pathology</td>
</tr>
<tr>
<td>6</td>
<td>Nelson</td>
<td>Thomas</td>
<td>FIDC000336</td>
<td>8435</td>
<td>3674</td>
<td>12.10.1998</td>
<td>12.10.1998</td>
<td>PhD</td>
<td>biomedical engineering</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Warren</td>
<td>Carmen</td>
<td>FIDC000337</td>
<td>8436</td>
<td>3675</td>
<td>12.10.1998</td>
<td>12.10.1998</td>
<td>PhD</td>
<td>electrical engineering</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Scalf</td>
<td>Raymond</td>
<td>FIDC000338</td>
<td>8437</td>
<td>3676</td>
<td>12.10.1998</td>
<td>12.10.1998</td>
<td>PhD</td>
<td>computer engineering</td>
<td></td>
</tr>
</tbody>
</table>

Look for possible points of re-identification!
Pseudonymization (masking)

The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information.

Use multiple pseudonymization techniques to achieve anonymization.

Secure the key.
Anonymized data

Anonymization is the irreversible removal of information that could lead to an individual being identified.

No longer considered PII by the GDPR.

You can achieve anonymization by combining various pseudonymization techniques.
**Substitution**

**Definition:**
Replacing one column or row of data with completely different values. These could be names or numbers.

**Effectiveness:**
Highly effective in masking data and securing privacy. Pseudonymization.

**Considerations:**
Full substitution of an entire group of data can attain full anonymization status and will no longer be considered sensitive data by the GDPR.
The Fake Name Generator believes in supporting the development community. To achieve this goal, we provide free bulk generated identity files. Please use the form below to place your order.

Step 1 - Read and agree to terms of service
- I agree to the terms of service and understand that all generated information is fake.

Step 2 - Choose output format and compression
- Output Format: Comma separated (.csv) □, Tab □
- Compression: □

Step 3 - Choose name sets, countries, gender, and age

**Name set**
- American
- Arabic
- Australian
- Brazilian
- Chinese
- Chochen (Latin)

**Country**
- Australia
- Austria
- Belgium
- Brazil
- Canada

**Gender**
- Male: 50%
- Female: 50%

**Age**
- 19 - 85 years old

Step 4 - Choose fields to include

Fields in the box on the right will be included in your order. Use the Up/Down buttons to choose which order you want the fields in.

**Don't include these:**
- Incrementing number
- Gender
- Name set
- Title
- Given name
- Middle Initial
- Surname
- Street address
- City
- State abbreviation

**Include these:**
- All

Step 5 - Enter quantity & choose delivery options

You are allowed to have three (3) orders in the queue at a time.

**Estimated wait:** 11 minutes

**Quantity:** 3000 (Maximum: 50,000, 100,000)
Tools:

Random number generator

https://numbergenerator.org/random-6-digit-number-generator#!/numbers=25&length=6&addfilters=
Nulling out

Definition:
Replacing a data field with a null value.

Effectiveness:
Highly effective. Achieves anonymization.

Considerations:
Often reduces data integrity.
Scrambling

**Definition:**
Scrambling the letters or numbers in a data field.

**Effectiveness:**
Weak form of pseudonymization. The data is susceptible to being "unscrambled" and re-identified. Can be stronger for long number sequences where the same scrambling algorithm is not used in each data field.
Tools:

word scrambler

https://www.wordunscrambler.net/word-scrambler.aspx
Shuffling

Definition:
Shuffling the values in a data field.

Effectiveness:
Weak form of pseudonymization if used alone. Susceptible to re-identification by determining the shuffling algorithm. Can achieve anonymization when used with other masking techniques.
Date Aging

Definition:
Choosing a random number of days to "age" a date.

Effectiveness:
Intermediate to strong form of pseudonymization. Somewhat susceptible to re-identification by determining the aging value.
Definition:
varying the date and number values. Common usage is with financial data.

Example:
for number values +/-10%; dates +/-200 days
Masking out

Definition:
Hiding some (not all) of the digits in a field.

e.g. xxxx-xxxx-xxxx-1079
Examples
Size of the data group

A small group of data subjects undermine most forms of pseudonymization as anyone with knowledge of the subjects can re-identify the data.
Fake Test Data Generator Tool

Generate meaningful Fake data for test purposes

Field Type
- First Name Female
- Last Name
- Company
- Job Title
- Street Address
- State
- Post Code
- Free Email

Country
- English (U.S.)

Output Rows
- 10

Output Delimiter
- Tab
**Encryption**

**Definition:**
An algorithm masks the data for you and requires a key to un-encrypt for editing and usage.

**Effectiveness:**
Highly effective unless the key or password to the encrypted folder is compromised.

**Considerations:**
Mobile versions can be less safe.

This Photo by Unknown author is licensed under CC BY-SA.
Free encryption options

1. Microsoft OneDrive (personal vault not available for macOS)
2. Proton Drive, VPN, Email: https://proton.me/drive
3. Folder Lock
   https://www.newsoftwares.net/folderlock/
4. AxCrypt
   https://www.axcrypt.net/pricing/
5. VeraCrypt
Other encryption options

https://www.techradar.com/best/best-encryption-software

For Windows
1. Secure IT 2000
2. SensiGuard
3. Renee File Protector

For macOS
1. Concealer
How to avoid anonymization

1. Do not collect any personal data
2. Do not store any personal data
3. Do not share any personal data
## Options

**General**
Language, title, survey description

**Responses**
Survey expiration, incomplete responses, back button and more

### Security

#### Passwords, file uploads, bot detection and more

**Post-Survey**
Thank you emails, completed survey messages, and triggers

### Advanced

**Scoring**
Attach point values to specific answers

---

Saved at 5:33 PM  Draft

a weasle or rug the response.

- Off

### Prevent indexing
Block search engines from including your survey in their search results.

- On

### Uploaded files access
Indicate who should be able to view files uploaded by respondents

- Only users with permission to view responses
- Anyone with the link to the file

### Anonymize responses
Don't record respondents' IP Address, location data, and contact info.

- On
Data anonymization services

1. Accelario
   https://accelario.com/

2. Anonos

3. K2View
   https://www.k2view.com/

Considerations:
Be sure services used do not receive actual data, but data already encrypted.
Kelley Rowan
Digital Archives Librarian
krowan@fiu.edu