De-Identifying Your Data

Kelley Flannery Rowan
Florida International University, krowan@fiu.edu

Follow this and additional works at: https://digitalcommons.fiu.edu/glworks

Part of the Library and Information Science Commons

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.
De-identifying your data

A Digital Scholar Studio Workshop

Kelley Rowan, Digital Archives Librarian
Agenda

- Laws, Torts, and Regulations
- Definitions, Getting Started, and Workflows
- Identifying Personal 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)
Tort Law (Precedence)

- Intrusion into seclusion
  - Improper data collection

- Public disclosure of private facts
  - Social media
  - Websites

- Intentional infliction of emotional distress
  - Cyberbullying

- Appropriation of name or likeness
- Placing a person in a false light
Fair Information Privacy Practices

Madrid Resolution, 2009

1. To define a set of principles and rights guaranteeing the protection of privacy with regard to the processing of personal data;

2. The facilitation of the international flows of personal data needed in a globalized world.
**Fair Information Privacy Practices**

- Organization for Economic Co-operation and Development (OECD Guidelines)
  - [https://mneguidelines.oecd.org/](https://mneguidelines.oecd.org/)

- Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, 1981
  - Updated to include AI and additional regulations for those storing data
  - [https://rm.coe.int/1680078b37](https://rm.coe.int/1680078b37)

- Asia-Pacific Economic Cooperation (APEC)
  - Agreed to a privacy framework in 2004
  - The Data Privacy Subgroup is working on compliance with the GDPR, 2017
  - [https://www.apec.org/](https://www.apec.org/)
State Privacy Laws

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]

Law link: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=

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

California Privacy Rights Act – 2023


Fully enforced beginning July 1, 2023
International Law

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.
The legal side of data collection, de-identification & storage

Follow the GDPR
Read the Madrid Resolution
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
Getting started with anonymization

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
Thinking about data and key storage

- **How long will the data be used?**
- **How is the data being used?**
- **How much data is there?**
- **How sensitive is the data?**
- **How much damage will be done when the data is exposed or the keys are lost?**
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.

**Other considerations:**
- Can everyone in the workplace see your computer?
- Are you using a shared computer?
- Where are you saving this data, shared drives?
## Identifying PII and personal data

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.
Tools:
fake name generator (free)
https://www.fakenamegenerator.com/order.php

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) - on
- .zip - on

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

<table>
<thead>
<tr>
<th>Name set</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>Australia</td>
</tr>
<tr>
<td>Arabic</td>
<td>Austria</td>
</tr>
<tr>
<td>Australian</td>
<td>Belgium</td>
</tr>
<tr>
<td>Brazil</td>
<td>Brazil</td>
</tr>
<tr>
<td>Chocon (Latin)</td>
<td>Canada</td>
</tr>
</tbody>
</table>

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 with your order. Use the Up/Down buttons to choose which order you want the fields in. Not all fields are available for every country. Please use the homepage to determine what information is available for the countries you have chosen.

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

Include these:
- All >>
- Up
- All <<
- Down

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)
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.

<table>
<thead>
<tr>
<th>published</th>
<th>Full text pdf/a</th>
</tr>
</thead>
</table>
Variance

Definition: varying the date and number values. Common usage is with financial data.

Technique 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

Person Information

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient No.</td>
<td>112233</td>
</tr>
<tr>
<td>Name</td>
<td>Peter Watson</td>
</tr>
<tr>
<td>Address</td>
<td>32 Elm St</td>
</tr>
<tr>
<td>City, State, Zip</td>
<td>Sunnyvale, CA, 94089</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other Info</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Card No.</td>
<td>4145 1230 0000 0062</td>
</tr>
<tr>
<td>Patient Records</td>
<td>![Image]</td>
</tr>
<tr>
<td>SSN</td>
<td>654 59 9876</td>
</tr>
</tbody>
</table>

Shuffling

Substitution

Custom Algorithm

Solix Test Data Management

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient No.</td>
<td>010101</td>
</tr>
<tr>
<td>Name</td>
<td>John Mayer</td>
</tr>
<tr>
<td>Address</td>
<td>12 Murray Ct</td>
</tr>
<tr>
<td>City, State, Zip</td>
<td>Boston, MA, 02115</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other Info</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Card No.</td>
<td>XXXX XXXX XXXX 0062</td>
</tr>
<tr>
<td>Patient Records</td>
<td>###</td>
</tr>
<tr>
<td>SSN</td>
<td>@#$ $- &amp;$#!</td>
</tr>
</tbody>
</table>
Size of the data group

A small group of data subjects undermine most forms of pseudonymization as anyone with knowledge of the subjects can de-identify the data.
Tools:

Fake data

https://www.coderstool.com/fake-test-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.
Free encryption options

1. Microsoft OneDrive (personal vault)
   Personal OneDrive only, not available for macOS
2. Folder Lock
   https://www.newsoftwares.net/folderlock/
3. AxCrypt
   https://www.axcrypt.net/pricing/
4. 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 PII
2. Do not store any PII
3. Do not share any PII
<table>
<thead>
<tr>
<th>Options</th>
<th>Prevent indexing</th>
<th>Uploaded files access</th>
<th>Anonymize responses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language, title, survey description</td>
<td>Block search engines from including your survey in their search results.</td>
<td>Indicate who should be able to view files uploaded by respondents</td>
<td>Don't record respondents’ IP Address, location data, and contact info.</td>
</tr>
<tr>
<td><strong>Responses</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey expiration, incomplete responses, back button and more</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passwords, file uploads, bot detection and more</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Post-Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thank you emails, completed survey messages, and triggers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attach point values to specific answers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Digital Scholar Studio
http://dss.fiu.edu/dss/