Protecting Privacy in the Library Patron Data Lifecycle

Becky Yoose
Library Data Privacy Consultant, LDH Consulting Services

Pacific Library Partnership, January 2020

This project was supported in whole or in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. The opinions expressed herein do not necessarily reflect the position or policy of the U.S. Institute of Museum and Library Services or the California State Library, and no official endorsement by the U.S. Institute of Museum and Library Services or the California State Library should be inferred.

Workshop Housekeeping – Guidelines

- All responses and questions are valid.
- Assume good intent.
- When you disagree, challenge or criticize the idea, not the person.

- Be mindful of the time.
- One speaker at a time.
- Speak from your own perspective.
- Help protect others’ privacy by observing the Chatham House Rule.

Workshop Housekeeping - Logistics

IANAL; Consult legal staff for legal advice

Exercises and Discussions - what to expect

General Overview vs Deep Dive

Privacy measures are only as strong as the least-knowledgeable person working with patron data

Workshop Schedule

9:00 – 9:20: Welcome and housekeeping

9:20 – 10:00: Training and exercises

10:00 – 10:10: Break #1

10:10 – 11:00: Training and exercises

11:00 – 11:10: Break #2

11:10 – 11:45: Training and exercises

11:45 – 12:00: Wrap up

Introduce Yourself!

1. Name
2. Job title and where you work
3. List one thing that you do to protect your personal privacy

Section One:
Library Patron Data Overview

Libraries and Patron Data - Where It Is In Your Library

- ILS
- Database backups
- Print management systems
- Server logs
- Reference chat/desk logs
- Public computer/wireless traffic logs
- ILL requests
- Security camera footage
- Card reader logs
- Meeting room reservations
- Authentication system logs
- Library programs
    - Attendance logs
    - Feedback responses

Libraries and Patron Data – The “Big Three”

Vendor and third party application data

Paper forms

Staff email

Discussion –
Where does patron data live at your library?

Personally Identifiable Information [PII] and Library Data

PII 1 - Data about a patron

- Name
- Physical/email address
- Phone number
- Date of birth
- Patron record number
- Library barcode

PII 2 - Activity that can be tied back to a patron

- Search & circulation histories
- Computer/wifi sessions
- Reference questions
- Electronic resource access
- IP Address
- Program attendance

Exercise – Identifying Patron PII

Exercise Answer Key - Identifying PII

PII 1

- Rosalind Franklin [Name]
- roseHelix@gmail.com [Email]
- 12345678910 [Barcode]
- 1111 [PIN]

PII 2

- THE ONLY WOMAN IN THE ROOM [Checked out book title that needs renewing]
- January 14, 2020 4:27 PM [Timestamp]

Section Two:
Library Patron Data And Regulations

“What happens if my library has…” (Federal Laws)

Medical-related data? HIPAA

Student data? FERPA

Data from minors (<13 years)? COPPA

Library Patron Data in California

California Gov Code § 6267
“All patron use records of any library … shall remain confidential and shall not be disclosed” except when a person is acting within scope of their duties, written authorization from patron, or by court order.

Patron use records definition

- Written or electronic record identifying patron
- Written or electronic transaction of patron’s use of library resources

California Gov Code § 6254
Disclosure exemption for library circulation records that identify “borrower of items available in libraries” (not including records of fines imposed on borrower)

Do libraries need to worry about CCPA?

- California Consumer Privacy Act of 2018 (CCPA)
- Regulates the sale of personal information by covered businesses
- Gives California residents:
  - Right to access what personal information is collected and shared with service providers and other third parties
  - Right to request deletion of information
  - Right to opt out of sale of personal information
- Areas of concern for libraries
  - Special requirements for minors
  - Household information as part of personal information definition

What about GDPR?

- Most US libraries are not under scope
- Exceptions – academic libraries whose institutions have EU presence and collect/process EU resident personal data
- GDPR rights to individuals to access, change, limit processing, and delete data
- GDPR best practices to adopt – Privacy by Design (PbD)

Section Three:
The Library Patron Data Lifecycle

Library Patron Data Lifecycle

Collection

What data is our library collecting?

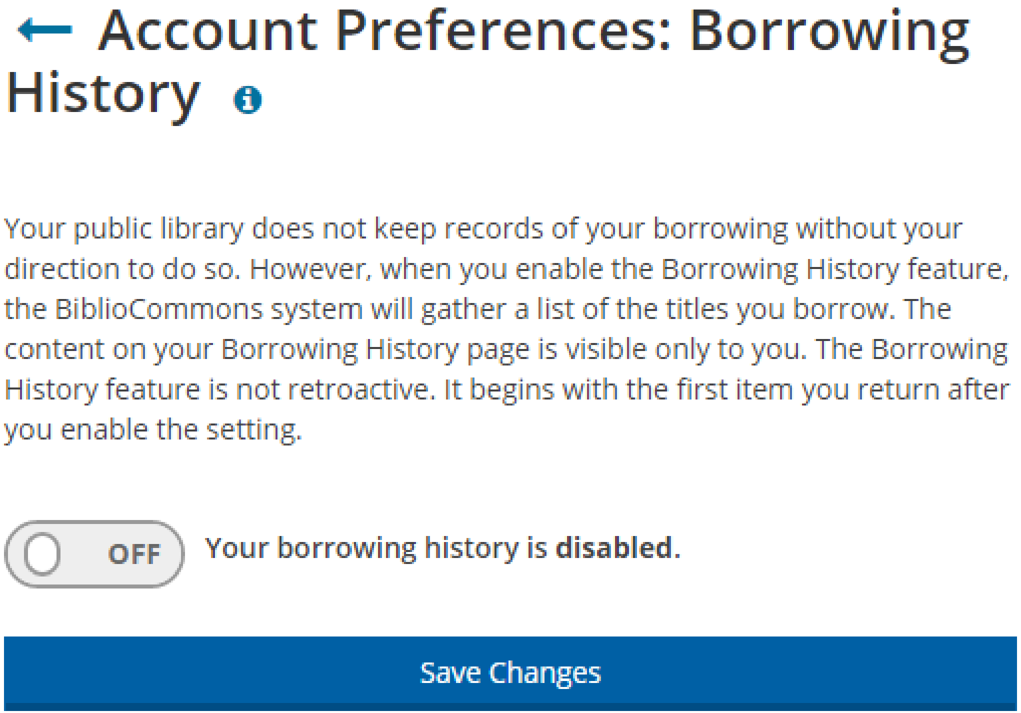
Example systems:

- Integrated Library Systems
- Server/Application logs
- Web analytic software
- Social media pages
- Survey software

Data collection includes:

- Patron & circulation information
- IP address/UUID
- Timestamps
- Search history
- Link clicks on sites and in marketing emails

Vendors and patron data collection – defaults



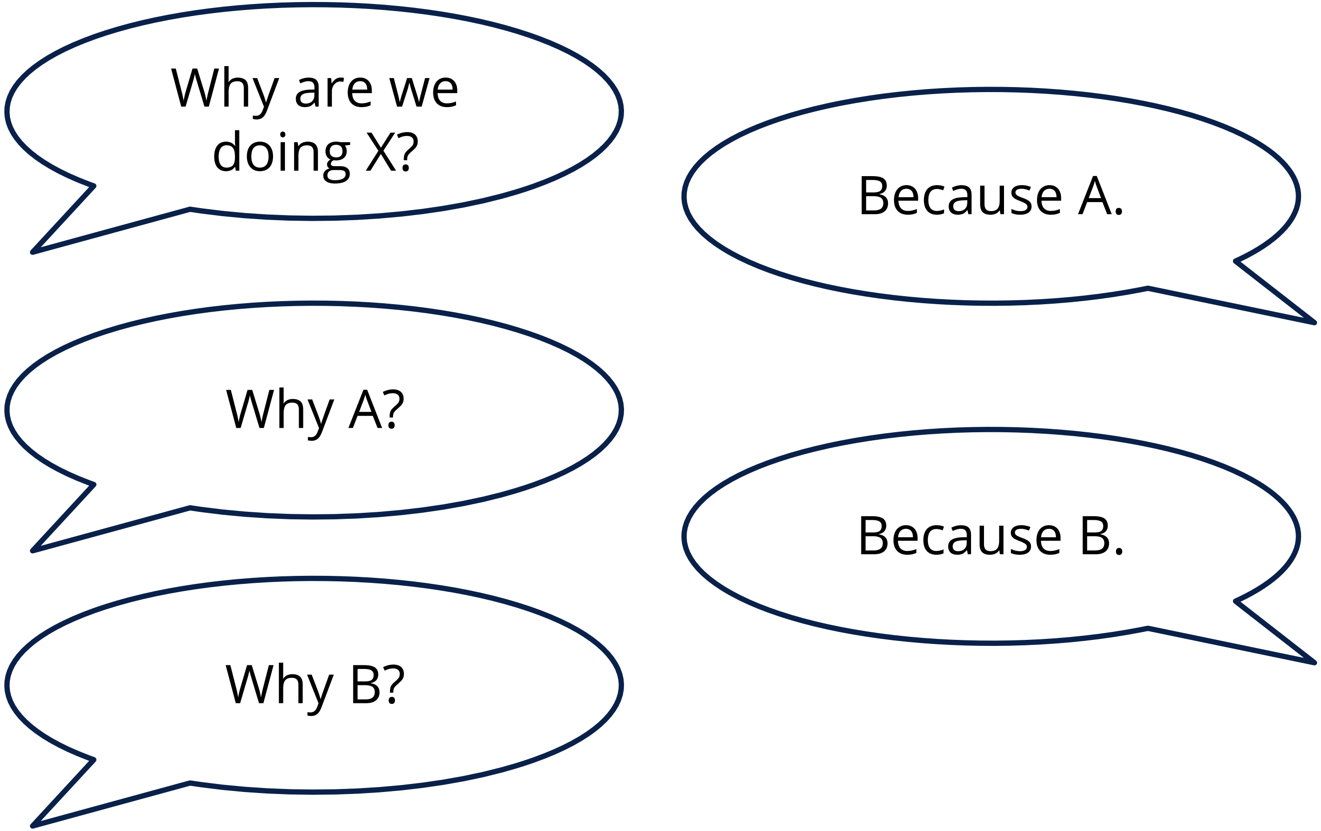
← Account Preferences: Borrowing History

Your public library does not keep records of your borrowing without your direction to do so. However, when you enable the Borrowing History feature, the BiblioCommons system will gather a list of the titles you borrow. The content on your Borrowing History page is visible only to you. The Borrowing History feature is not retroactive. It begins with the first item you return after you enable the setting.

OFF Your borrowing history is **disabled.**

If you do not have a *demonstrated* business need to explain why you are collecting a data point, *then you should not collect that data*.

The Five Whys Method



Why are we doing X?

Because A.

Why A?

Because B.

Why B?

Exercise –
“Why’s that, again?”

Library Patron Data Lifecycle

Storage

Where are we storing data?

Native systems and applications

Data extracted, exported, or otherwise taken from the native system and applications

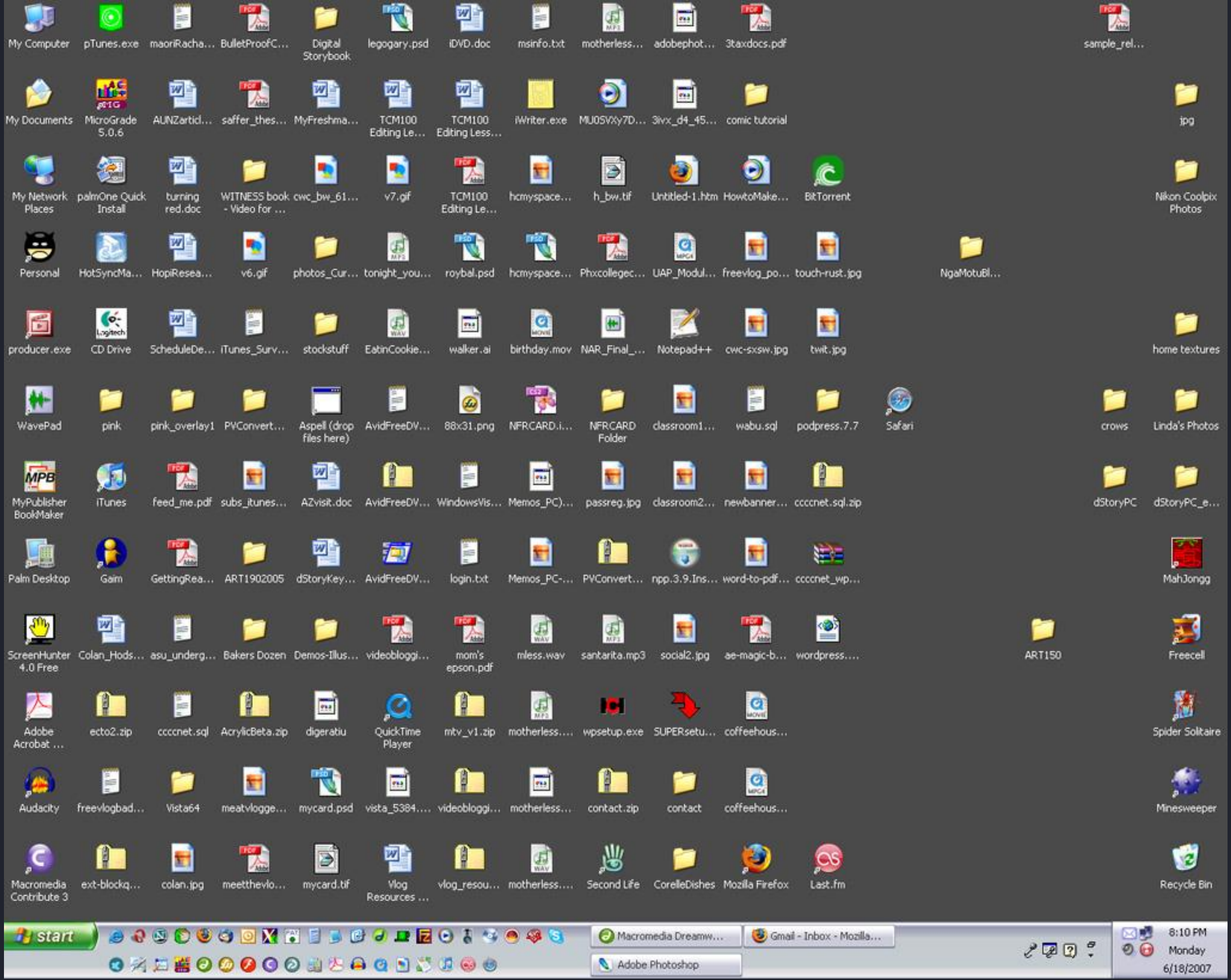
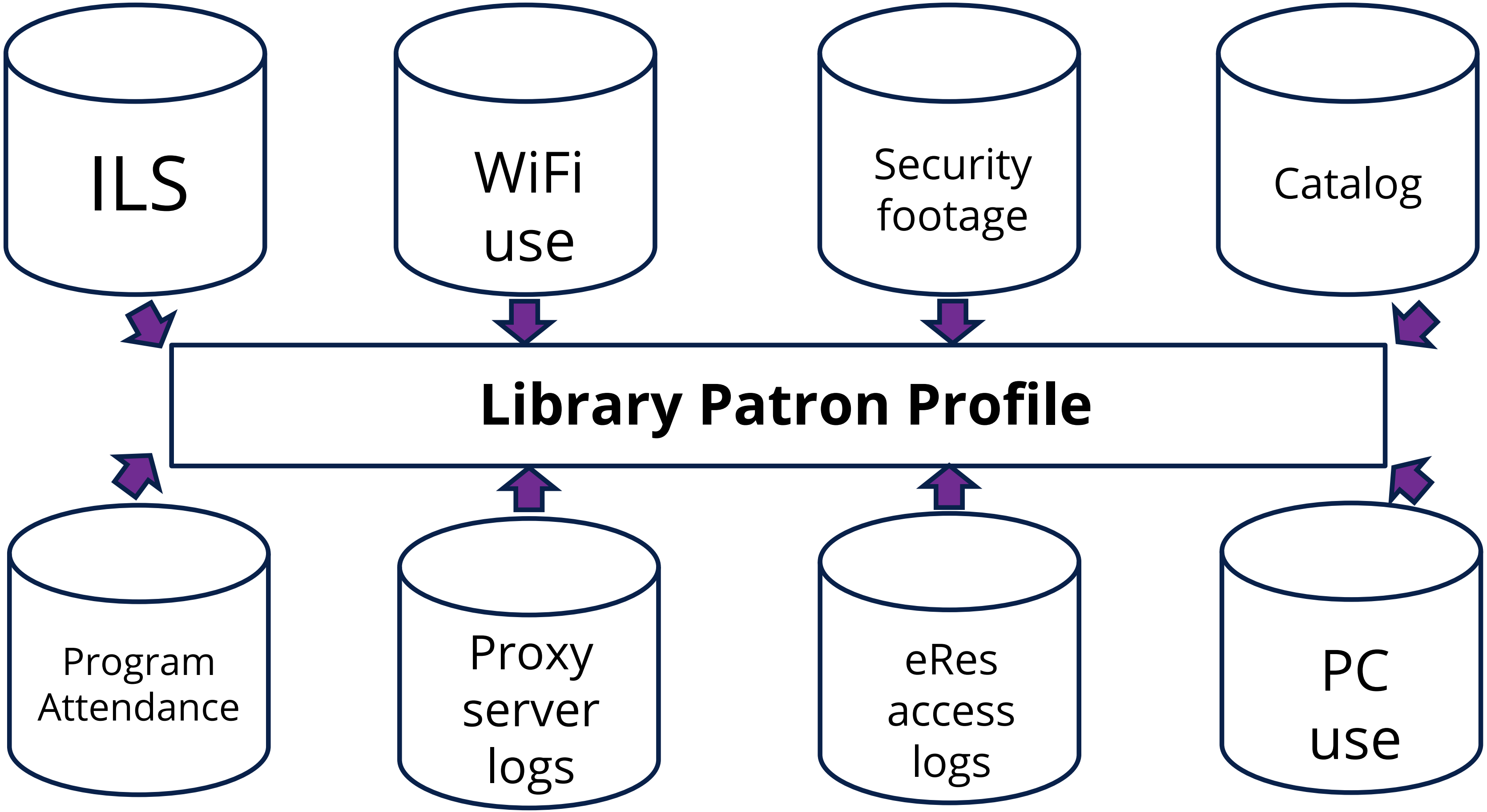


Image source – Before by Cheryl Colan, CC BY-NC 2.0, https://www.flickr.com/photos/hummingcrow/576989434/

Image source – wtf-am-i-supposed-to-do-wit by kris krüg, CC BY-SA 2.0, https://www.flickr.com/photos/kk/18768224/

Office 365



ILS

WiFi use

Security footage

Catalog

Library Patron Profile

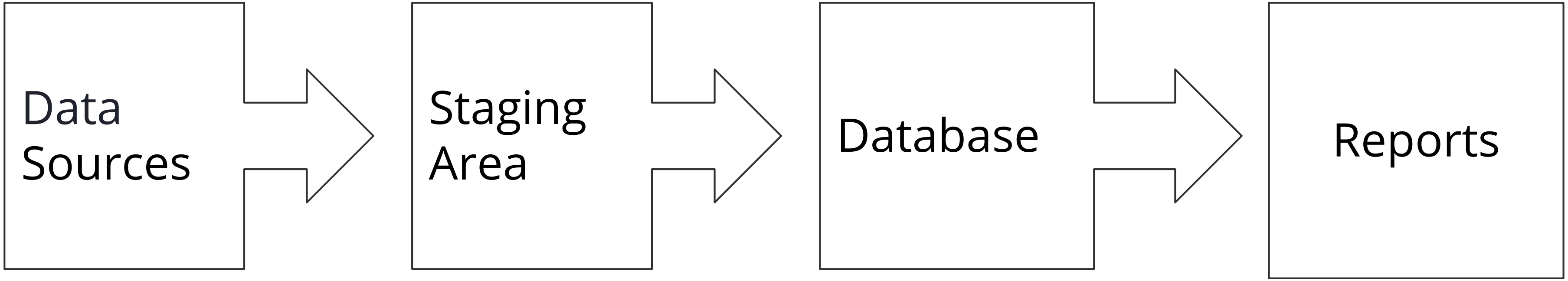
Program Attendance

Proxy server logs

eRes access logs

PC use

Example of Storage - Data Warehousing



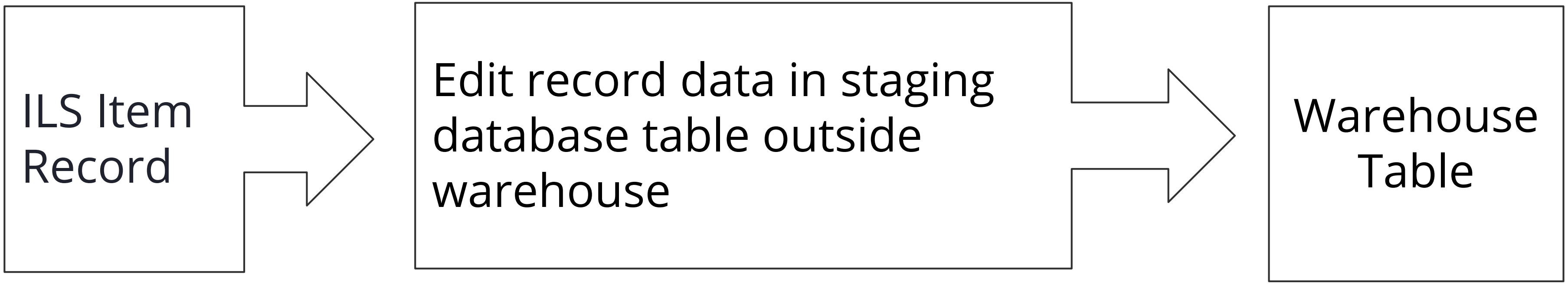
Data Sources

Staging Area

Database

Reports

ETL - Extract, Transform, Load



ILS Item Record

Edit record data in staging database table outside warehouse

Warehouse Table

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM
FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

De-identification of Library PII Data

Obfuscation

- PII 1
  - Date of birth vs age

Truncation

- PII 1
  - Full address vs zip code
- PII 2
  - Call numbers

Aggregation

- PII 1
  - Age vs age ranges
- PII 2
  - Very high level call number ranges

De-identification of Library PII Data

Pseudonymization

Some considerations:

- Algorithms
- Hashing and salt

Differential Privacy

All the mathematical equations!

Exercise –
De-identifying Data

Exercise - Raw vs De-identified Data

Raw data

- Date of birth – 10/24/1977
- 27 43rd St, Town, WA 92471
- NX180.I57 M275 2015
- FIC HARRIS 2018
- 11 Apr 2020 9:24 - 10:24

De-identified – one way

- 42 years old
- 92471
- NX180
- FIC
- 4/11/2020, 01:00

Obligatory Disclaimer De-identification of Data (Part One)

Data de-identification methods do not provide adequate privacy protection for these types of data:

- Outliers in service population
- Small overall service population or subset (degree program, etc.)

Vendors and Third Party Apps Considerations

CRMS and data analytics products areas for privacy negotiations during the acquisition process:

- Request for Proposals
- Functional Requirements
- Contract Negotiations
- Contract Addendums

Data storage: What, Where, Who Has Access

Data sharing: What, Where, Who, Raw vs De-identified vs “Anonymized” vs Aggregated

Data privacy policy: what to negotiate, theirs (if they have one) or yours

Library Patron Data Lifecycle

Access

Who has access to what data?

Physical Access

- Desktop computers
- Laptops
- Mobile devices
- Server room/data center
- Offices and desks
- Flash drives
- File cabinets
- Security camera terminal and tapes

Electronic Access

- User permissions
- Administrator account information
- Vendor access to local systems
- System log and database access
- Administrator site access

Exercise –
“What can possibly go wrong?”

Some Data Access Best Practices

- Lowest/most restrictive level of access to meet operational needs
- Lock everything - restrict access to hardware through physical barriers
- “Lock” everything
    - Require login for systems and physical devices
        - Enable multi-factor authentication (MFA) if possible
    - Encryption of hard drives and mobile devices
    - Remote wipe for mobile devices
- Regular audits of...
    - Account access to systems
    - Keys (physical and electronic)

Library Patron Data Lifecycle

Reporting

Do staff really need access to all the data for reports?

Consider giving staff the following:

- Database views for report queries and building
- Connections to data through Data BI tools (Tableau, Power BI, etc.)
  - Related - restrict access to data used in the report when published
- Dashboards
- “Canned” reports with some parameters for customization (date, location, etc.)

How much data to provide in the report?

Report on the highest level of data that will meet the core need of the report audience…

… except when release of entire data sets is expected.

Obligatory Disclaimer De-identification of Data (Part Two)

Data de-identification methods are subject to varying re-identification risks, primarily through PII 2 data:

- Identifying patterns
  - Example - AOL
- Fuzzy matching
  - Example - Taxi Cab Data

Example of fuzzy matching – Library edition

Data set #1

- Patron physical address
- Patron barcode
- Patron age

Data set #2

- Call number
- Subject headings and topics
- Patron barcode

Patron barcode match between two data sets = possible reidentification and tracking of patron activity tied to unique real world individual

Accuracy of reidentification increases with additional data sets (including Open Data and data broker data sets)

Sometimes the only response…

… is to not release data.*

*except when required by law

Exercise – Data Patron Request

Library Patron Data Lifecycle

Retention

How long are we storing data?

… when no longer needed operationally?

… 30 days?

… 1 year? Rolling year?

… in perpetuity?

… what about backups?

… what about vendor systems?

Library Patron Data Lifecycle

Deletion

Deleting data

Electronic data

Backups and logs – when and how are they deleted?

Wipe the drive after electronic data deletion

Look out for data living “outside” local and vendor systems

Physical data

Shred paper and dispose of shredded paper properly

Properly dispose of disks, drives, other hardware, including degaussing or otherwise physical destruction of the disk or drive

Vendor considerations

- How does the vendor delete physical and electronic copies of your patron data?
- When you leave, can you take your data with you?
- Can your patrons take their data with them?
- Can your patrons request their data to be deleted with a vendor?

Library Patron Data Lifecycle Roundup

Exercise –
What Would You Do?

Section Four:
A Data Inventory Starter Kit

Tips for your first data inventory and beyond

- Start small – select one system to focus on
- Ask around – sometimes you learn things that you wouldn’t otherwise
- Vendor hosted systems – focus on what you can control when recommending actions based on inventory
- Privacy and security audit integrations with data inventory
- Setting up a schedule for inventories

Pick one piece of technology (locally or vendor hosted) that your place of work is using or has used in the past.

Possible questions to ask:

- What patron data are you collecting?
- How is the patron data being used?
- Where is the patron data being stored? Don’t forget backups, log files, etc.
- How long are you keeping patron data?
- How are you deleting patron data when it’s no longer needed?
- Who has electronic and physical access to the patron data?

Section Five:
Wrap up

What is one thing from this workshop that you can put into practice or discussion at your library when you return?

Thank you

:-)

LDH Consulting Services

Becky Yoose
Library Data Privacy Consultant
LDH Consulting Services

Email: becky@ldhconsultingservices.com