Data Privacy Practices and Challenges In The Era of Big Data

Becky Yoose
b.yoose@gmail.com
@yo_bj
SeaSPIN 6/5/2018

Housekeeping

- Current Position- Library Applications and Systems Manager at The Seattle Public Library
- Slides will be available, along with speaker notes and resources
- The views expressed within this presentation belong to me and not the opinions of my current employer.
- IANAL

Part One:
Data and You - a Brief Primer

Data is everywhere

Evidence based practices

Market analysis/segmentation

Open data programs

Operation assessments

Learning analytics

Data warehouses

Customer relation management systems

Business Intelligence/data visualization tools

Where does user data live?

- Databases
- Database backups
- Server logs
- Customer support or help desk chat logs
- Email
- Clickthrough tracking from emails
- Authentication system logs
- Survey and feedback responses
- Digital fingerprint tracking (browser, OS information)
- Contractors/subcontractor systems

Personally Identifiable Information [PII]

PII-1 - Data about a person

- Name
- Mailing/email address
- DOB
- Gender
- Username/password
- Credit card number
- Social Security Number

PII-2 - Data about a person’s activities

- Search history
- Transaction history
- Website visit sessions
- Customer support questions
- Linked accounts
- Geolocation

Part Two:
Why Should You Care About Privacy

Users care about privacy…
or do they?

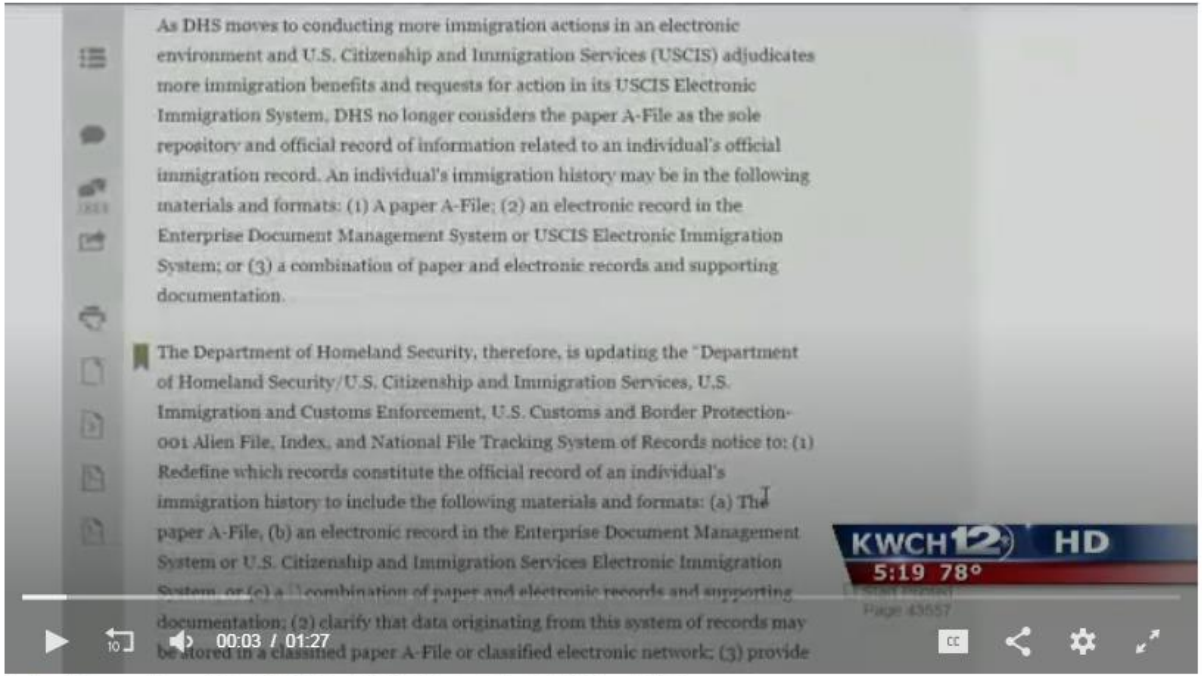
Privacy protects your most vulnerable users

Photo credit - https://twitter.com/SPLBuzz/status/917459655269474305

“It is up to people with HIV to decide to whom they talk about their status, and on what terms… It may be a commercial app, but as an LGBTQ app Grindr has responsibilities to the wider communities. That does not include sharing something as profoundly personal (and still stigmatised) as HIV status... Having an app that wraps itself in the rainbow flag passing on that status to third parties without their consent is a betrayal.”

- Owen Jones, April 2018

Federal authorities to begin checking immigrants' social media



By Brenda Carrasco | Posted: Thu 5:00 PM, Oct 19, 2017 | Updated: Thu 5:47 PM, Oct 19, 2017

Wichita, Kan. (KWCH) It's an amendment to the Privacy Act of 1974, allowing the Department of Homeland Security to collect social media handles, aliases and online search results from someone who is or has gone through the legalization process, including US residents and naturalized citizens.

If you don’t care about privacy, the lawyers will make sure that you do.

(Just) Four types of regulations surrounding data

- HIPAA/HITECH
  - Medical data
- FERPA
  - Student data
- COPPA
  - Users under 13

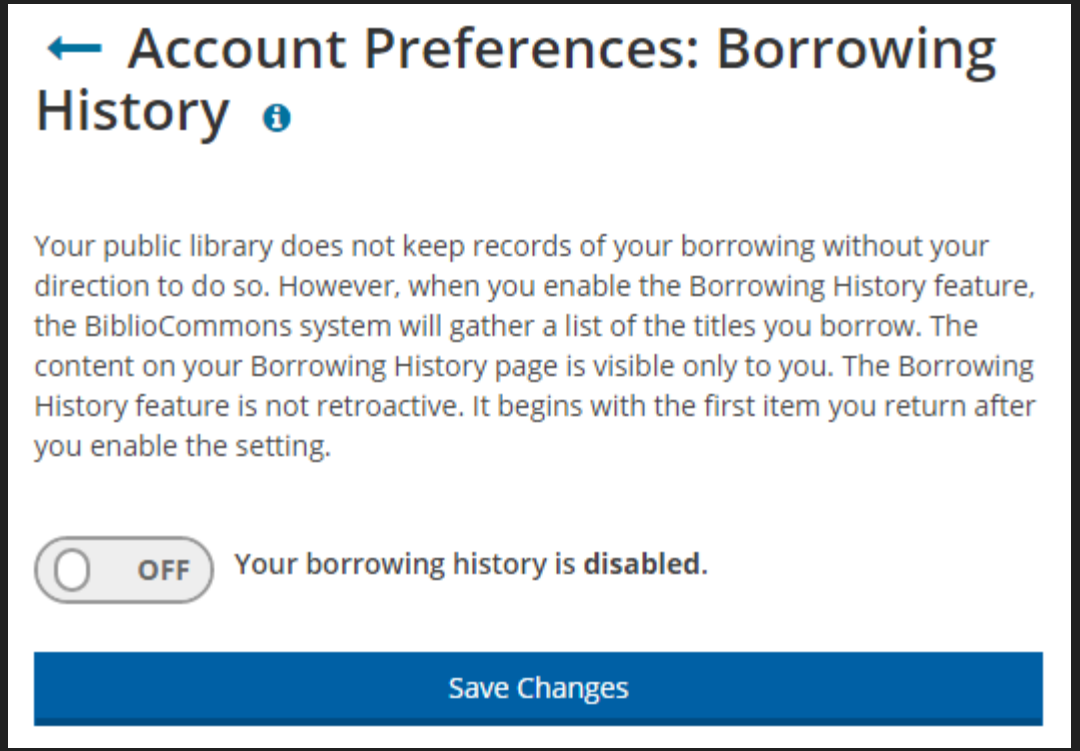
Obligatory GDPR mention

Part Three:
How do you balance privacy with operational need for data?

Privacy By Design

- Proactive not reactive; preventative not remedial
- Privacy as the default setting
- Privacy embedded into design
- Full functionality – positive-sum, not zero-sum
- End-to-end security – full lifecycle protection
- Visibility and transparency – keep it open
- Respect for user privacy – keep it user-centric

Opt-in Vs. Opt-out



Data life cycle

Collection

What data is being collected?

WHY are we collecting it?

AKA “The fight against #dataFOMO”

Storage and retention

Where is data being stored?

What possible other versions are being stored and where?

How long are we keeping $DATA_FIELD?

Data life cycle

Access and Reporting

Who has access to the physical hardware/space?

Who has what permissions to our systems, servers, databases?

What happens to access when staff change jobs, leave, etc.?

Who can see what reports?

Deletion

Physical media destruction

Electronic media destruction

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)

Example - 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

Example - De-identification of Library PII Data

Pseudonymization

Some considerations:

- Algorithms
- Hashing and salt

Differential Privacy

All the mathematical equations!

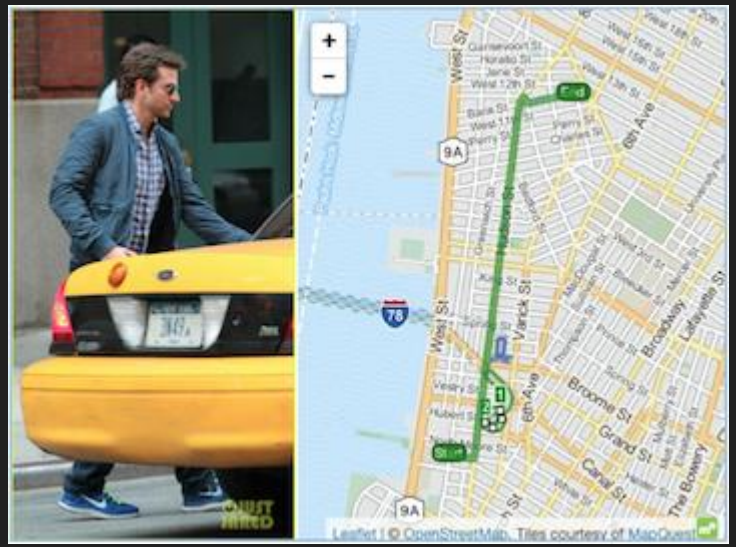
PII-2, I see you

Re-identification through search patterns

User 4417749's Search Queries:

- "numb fingers"
- "60 single men"
- "dog that urinates on everything"
- "Arnold"
- "landscapers in Lilburn, Ga"

Re-identification through fuzzy matching



Data Life Cycle Starter Kit

- What data are you collecting?
- What was the process of deciding what data to collect?
- What PII are you collecting? What anonymization/de-identification methods are you using, if any?
- How long are you keeping that data? Where is that data being kept? Don’t forget backups, log files, etc.
- How are you deleting that data when it’s no longer needed?
- Who has access to that data? To the physical system running the software?

Discussion Questions

- What privacy policies and procedures do you have in your organization?
  - How effective are those policies? Are they enforced consistently?
  - If you don’t have a policy - why? What would it take to get a policy in place?
- What codes of ethics or professional standards regarding privacy are you aware of?
  - What are some of the strengths and weaknesses of those codes and standards?

Q & A

b.yoose@gmail.com
@yo_bj - Twitter

Libraries and Privacy Resources

ALA Privacy Checklists
http://www.ala.org/advocacy/privacyconfidentiality/library-privacy-checklists

Library Freedom Project
https://libraryfreedomproject.org/

San José Public Library Virtual Privacy Lab
https://www.sjpl.org/privacy

Digital Privacy and Data Literacy Project
https://dataprivacyproject.org/

Does any of this look familiar to you?

HTTP by default/no support for HTTPS

Unsecured physical server access

Unencrypted data and backups

No backups or backups stored in perpetuity

No standardised record retention policy

No database access restrictions or policy

Improper or incomplete de-identification/anonymization of data

No strategies for data deletion when customer leaves vendor

Collecting more data than needed (vacuuming every datapoint possible)

Providing user PII data to other companies

Tracking user activity, location, etc. without consent

Sharing user information to third parties without consent or notification

No public privacy policy