Getting to Your Patrons Anonymously

De-identifying Patron Data for Analytics and Intelligence

Washington Library Association
2015 Conference
April 16, 2015

Your Presenters

Jim Loter

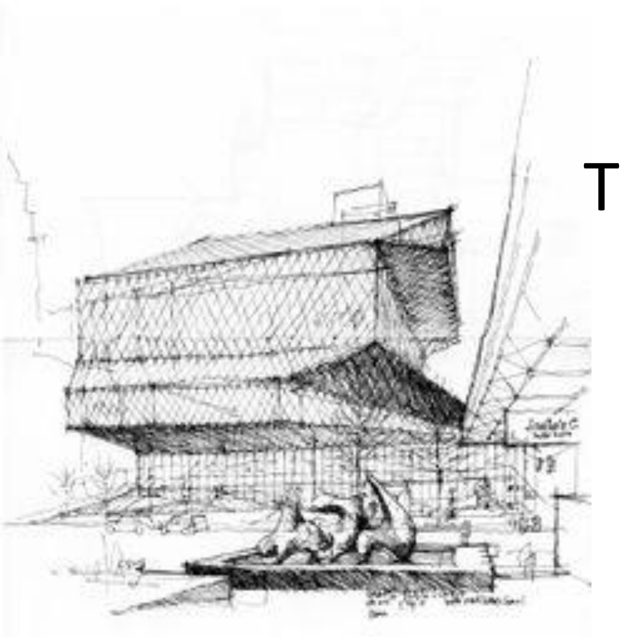
Director of Information Technology

@jimloter

Emily Morton-Owens

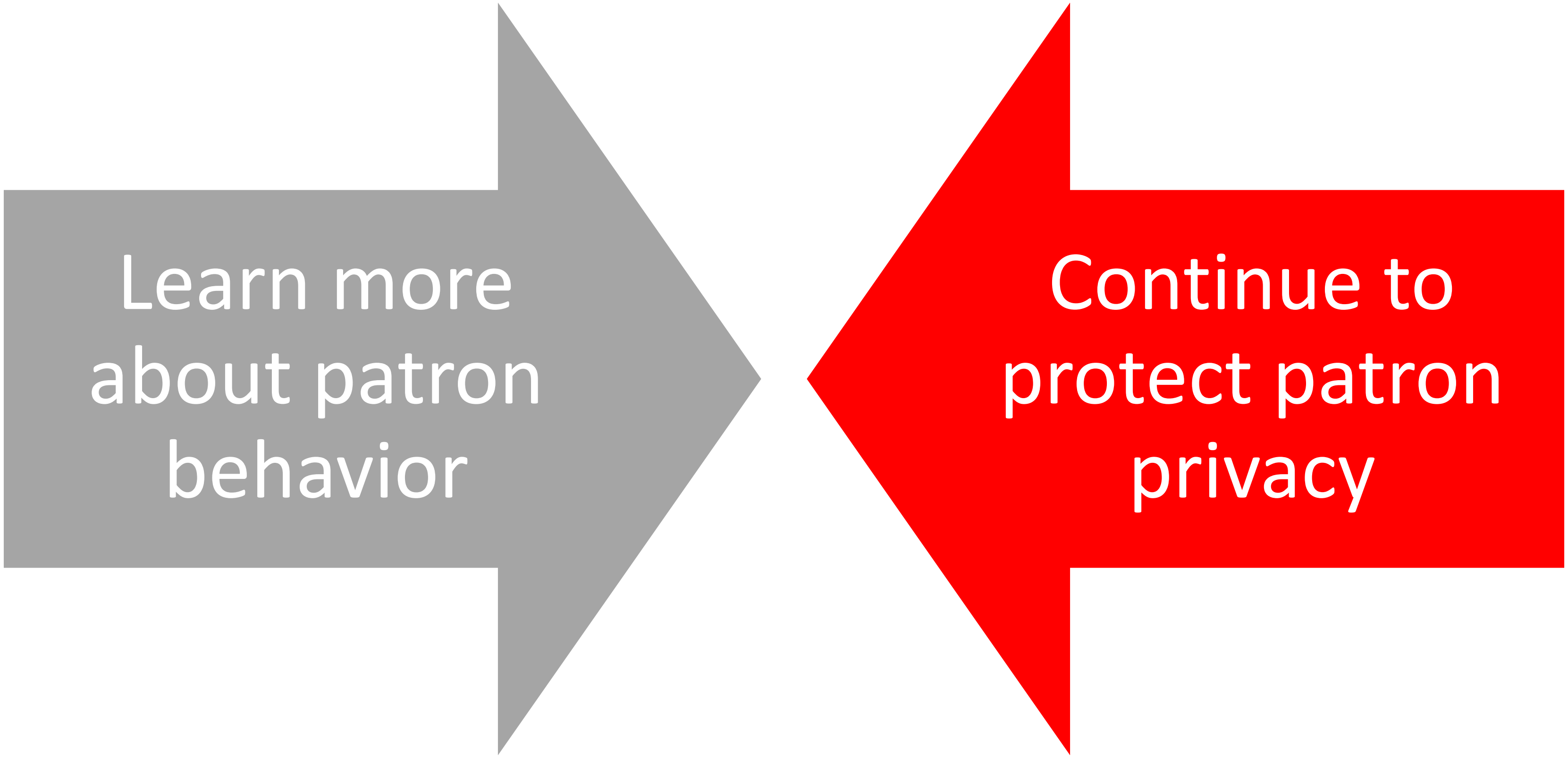
Manager of Library Applications and Systems

@bradamant



The Seattle Public Library

General Thesis



Learn more about patron behavior

Continue to protect patron privacy

Presentation Outline

- Overview of data management principles, policies, and practices
    - National
    - State
    - Library-specific
    - Data definitions
- Description of the problem
    - Unanswerable questions
    - Perceived threats and hazards
- Methods
- Examples

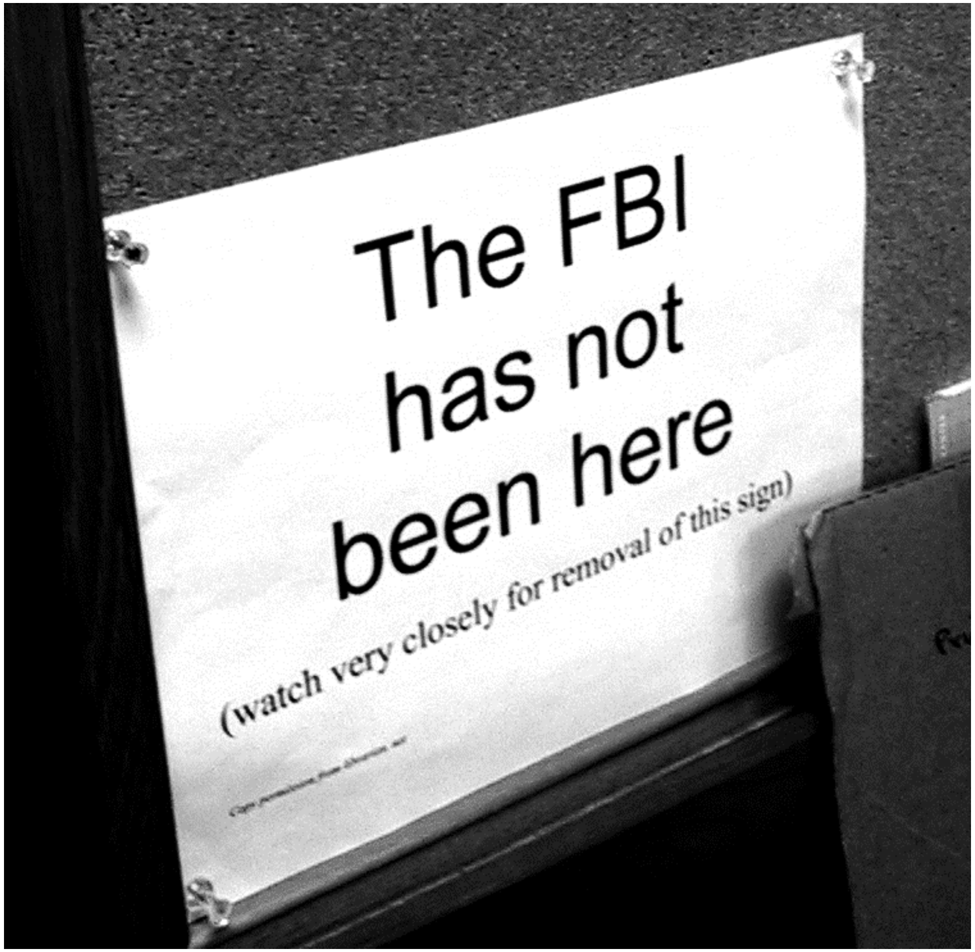
Principles, Policies, and Practices

ALA Data Management Guidelines

- Collection of personally identifiable information
  - only when necessary to fulfill the mission of the library
- Should not share personally identifiable user information with third parties, unless
  - the library has obtained user permission
  - has entered into a legal agreement with the vendor
- Make records available [to law enforcement agencies and officers] only in response to properly executed orders.

“An interpretation of the Library Bill of Rights.”
http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy

“Valid Law Enforcement Requests”



State Law: Revised Code of WA

- RCW 42.56.310: Library Records
  - Any library record, the primary purpose of which is to maintain control of library materials, or to gain access to information, that discloses or could be used to disclose the identity of a library user is exempt from disclosure under this chapter.
- RCW 19.255.010: Disclosure, notice — Definitions — Rights, remedies.
  - First & last name combined with SSN, DL #, credit/debit card number, authentication credentials, “account number”

SPL Confidentiality of Patron Data

- It is the policy of The Seattle Public Library to protect the confidentiality of borrower records as part of its commitment to intellectual freedom.
- The Library will keep patron records confidential and will not disclose this information except
  - as necessary for the proper operation of the Library
  - upon consent of the user
  - pursuant to subpoena or court order
  - as otherwise required by law.

The Seattle Public Library. “Confidentiality of Borrower Records.”
http://www.spl.org/about-the-library/library-use-policies/confidentiality-of-borrower-records

SPL Data Management Practices

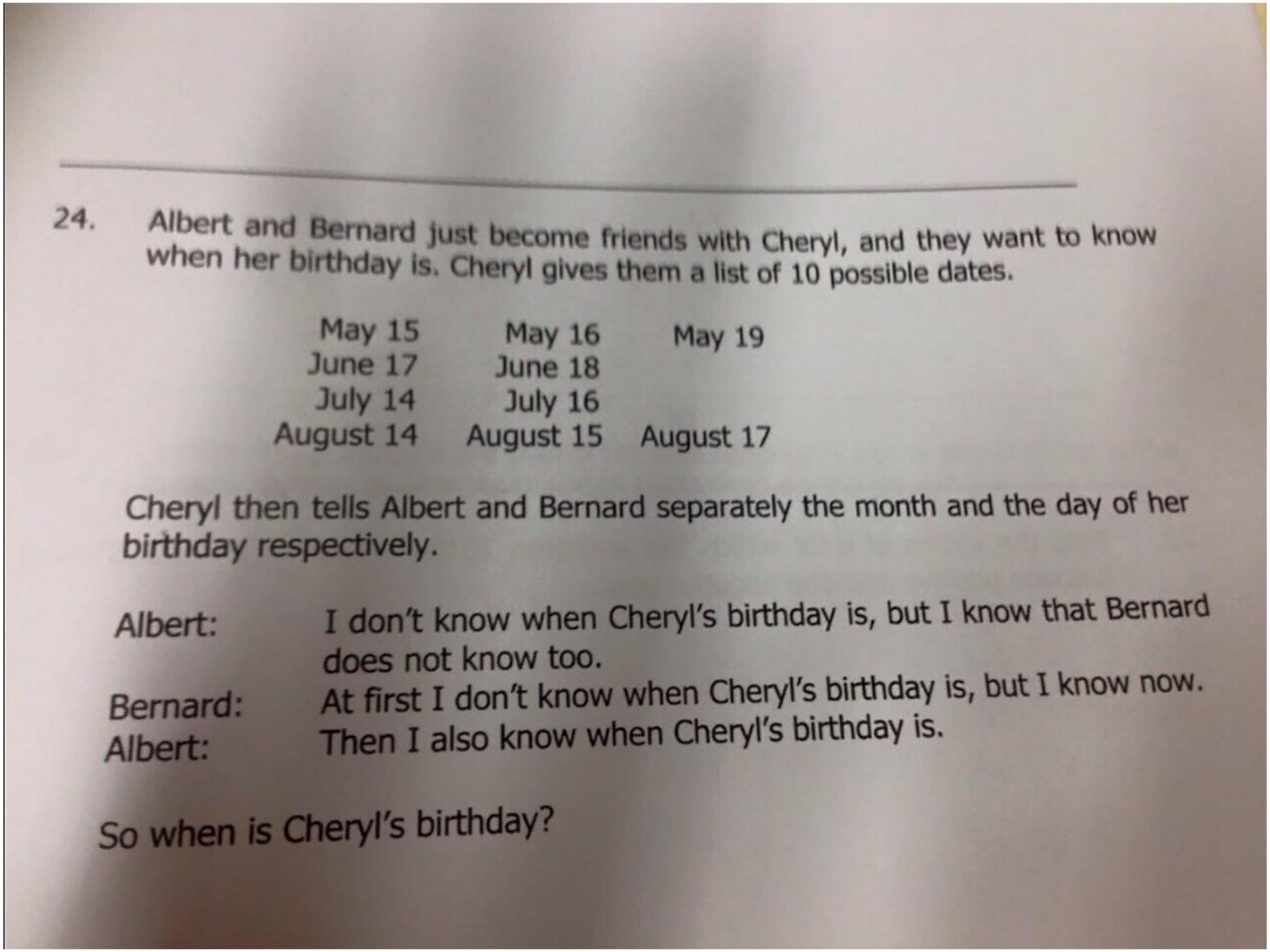
- All records connecting a patron to an item that has been held or borrowed, or to an information resource that has been accessed, are deleted upon the successful fulfillment of the transaction.
    - Circulation records
    - Public computer reservations
    - Workstation use data (log files, caches, histories)
    - Network logs

NIST: Two-part Definition of PII

1. Any information that can be used to distinguish or trace an individual‘s identity, such as name, social security number, date and place of birth, mother‘s maiden name, or biometric records
2. Any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information
   a) Libraries extend the second point by including borrowing and information seeking activity

National Institute of Standards and Technology via the Government Accounting Office expression of an amalgam of the definitions of PII from Office of Management and Budget Memorandums 07-16 and 06-19. May 2008, http://www.gao.gov/new.items/d08536.pdf

When is Cheryl’s Birthday?



24. Albert and Bernard just become friends with Cheryl, and they want to know when her birthday is. Cheryl gives them a list of 10 possible dates.

May 15	May 16	May 19
June 17	June 18	
July 14	July 16	
August 14	August 15	August 17

Cheryl then tells Albert and Bernard separately the month and the day of her birthday respectively.

Albert: I don’t know when Cheryl’s birthday is, but I know that Bernard does not know too.

Bernard: At first I don’t know when Cheryl’s birthday is, but I know now.

Albert: Then I also know when Cheryl’s birthday is.

So when is Cheryl’s birthday?

Two-part Definition of PII

Name
Email Address
Social Security Number
Date of Birth
Phone Number
Mobile Number
Address
Mother's Maiden Name

PII-1: Individual

Title
Author
URL
Web Site
Search Term
Username
Password
Call Number
Question

PII-2: Intellectual Pursuits

Data De-identification

- “Any process of removing the association between a set 261 of identifying data and the data subject.” (ISO/TS 25237-2008 [Health Informatics - Pseudonymization])
- Designed to protect individual privacy while preserving some of the dataset’s utility for other purposes.
- Make it hard or impossible to learn if an individual’s data is in a data set, or determine any attributes about an individual known to be in the data set.
- HIPAA: Data that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual

Garfinkel, Simson L. “De-Identification of Personally Identifiable Information.” April 2015. NIST. http://csrc.nist.gov/publications/drafts/nistir-8053/nistir_8053_draft.pdf

Problem statements

Current State

VEGETABLES from the SEA JILL GUSMAN
Madhur Jaffrey Eastern Vegetarian Cooking
Charmaine Solomon's COMPLETE VEGETARIAN COOKBOOK
EVERYDAY COOKBOOK vegetarian
vegie food
THE VEGETARIAN Gourmet ROZ DENNY REED EDITIONS
VEGIE FOOD LOW-FAT & DELICIOUS
DELIA'S VEGETARIAN COLLECTION
CHINESE VEGETARIAN COOKING DEH-TA HSIUNG
VEGANOMICON THE ULTIMATE VEGAN COOKBOOK
VEGAN CUPCAKES TAKE OVER THE WORLD
Fresh FROM THE Vegetarian Slow Cooker
kitchen classics gourmet vegetarian
the essential vegetarian cookbook
charmaine solomon complete vegetarian cookbook
New Asian Traditions Vegetarian Cookbook SNP•EDITIONS
PAREKH The Asian Vegan Kitchen
Vegetarian Cooking
THE LONGEVITY CHINESE VEGETARIAN COOKBOOK Gee and Goldin
vegan Over 90 mouthwatering recipes Tony Weston and Yvonne Bishop hamlyn
THE COMPLETE VEGETARIAN COOKBOOK
THE COMPLETE VEGAN KITCHEN
How it all Vegan!
Vegan Lunch Box
VEGAN WITH A VENGEANCE
DONNA KLEIN VEGAN ITALIANO MEAT-FREE, EGG-FREE, DAIRY-FREE DISHES FROM SUN-DRENCHED ITALY

penguincakes/Flickr

Unanswerable Questions

- Longitudinal questions (e.g. in medicine)
  - Long-term rather than snapshot
  - Trends, correlations, changes in behavior – not necessarily individual activity
- Questions only about the type and amount of use by demographic groups—not the content
- “Do heavy e-book users also use print materials?”
- “Do teen patrons remain active in their 20s?”
- “Do people use their neighborhood branch or use the branch where relevant materials are?” (e.g. Chinese language collection)

Privacy Requirements

- Passionate commitment to intellectual freedom
- Recognition that some patrons have no alternatives
- Intellectual content of transactions should always be purged
- Avoid keeping records that show person’s whereabouts

Serious security

AOKI

KAY

0123456789

4/16/2015

Perceived threats

Threats to patron privacy

- Law enforcement
  - Seeking intellectual pursuit data
  - Seeking patron whereabouts
- Hackers
  - Library is not an attractive target (no CC’s, SSN’s)
  - ILS data is relatively non-sensitive
- Data leak
  - Reconstruction of identity via data
  - Embarrassment/loss of trust
  - Notification costs

AOL data release (2006)

This was a screw up

There was no personally identifiable data provided by AOL with those records, **but search queries themselves can sometimes include such information.**

TechCrunch. “AOL: ‘This was a screw up’.” August 2006. http://techcrunch.com/2006/08/07/aol-this-was-a-screw-up/

AOL Example: User 4417749

n

60 single mer

c

everything

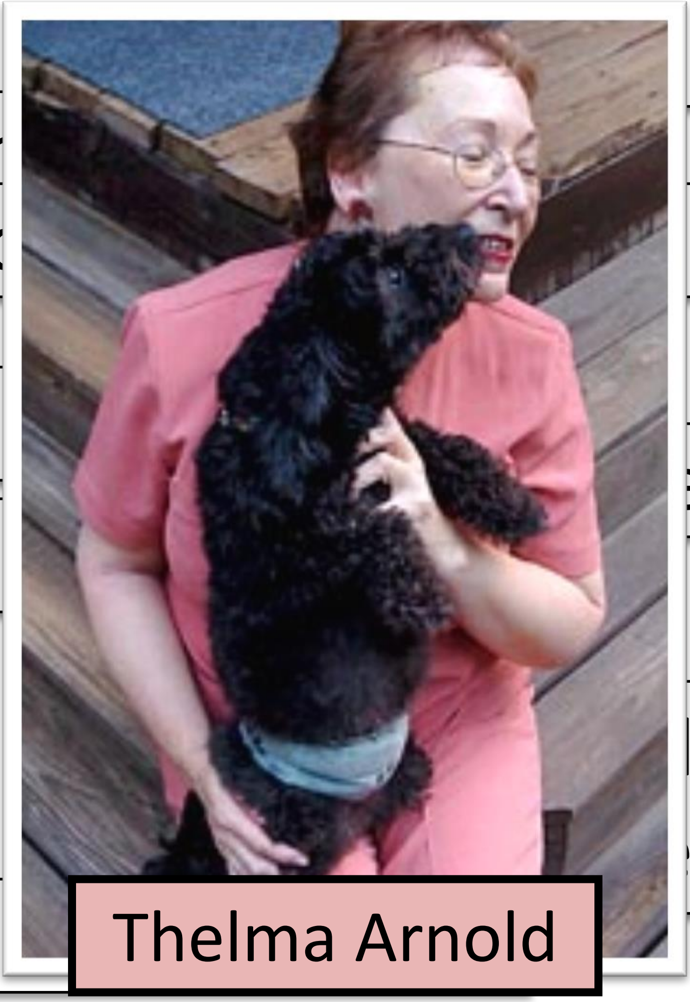
robert arnold

eorgia

marion a

low lake
t county georgia

landscapers



Thelma Arnold

Two-part Definition of PII

Name
Email Address
Address
Social Security Number
Date of Birth
Phone Number
Mobile Number
Mother's Maiden Name

PII-1: People

Title
URL
Author
Search Term
Web Site
Password
Username
Call Number
Question

PII-2: Intellectual Pursuits

Methods

Delete data

ervins_strauhmanis/Flickr

HIPAA

- Allowed to keep confidential information
- Protect against exposure and unauthorized access
- Dissimilar from law enforcement/government threat

Age vs. DOB

- DOB: 3/15/1975

- Age 40
- DOB: 3/15/1975?
- 3/16/1975?
- 3/17/1975?
- 3/18/1975?
- 3/19/1975?
- 3/20/1975?
- 3/21/1975?
- 3/22/1975?

Call numbers

- Call number: 914.30487 F683
- Format: DVD
- Collection: Beginning ESL
- Truncated call number: 91*, FIC

Timestamps vs. dates

- Timestamp
  `Sat, 11 Apr 2015 11:02:43 +0000`
- Date
  `4/11/2015, 00:44`

Extract-Transform-Load

Patrons	
Data	**PII?**
Barcode	Yes
Name	Yes
Address	Yes
Email Address	Yes
Phone Number	Yes
Date of Birth	Yes
Age	No
Gender	No
Zip Code	No
Registration Year	No

Circulation	
Data	**III?**
Barcode	Yes
Title	Yes
Author	Yes
Call Number	Yes – truncate it
Item Type	No
Branch	No
Date	No

Age	Gender	Zip	Reg Year	Item Type	Dewey 100	Branch	Date
45	Male	98117	2004	CD	700	CEN	4/1/15
45	Male	98117	2004	Book	FIC	BAL	4/3/15

Obfuscate borrower IDs

- Patron ID 12345 → `KEwHPoJpXY7K757HLmVQXHEyaEg=`
- Patron ID 98765 → `Q2se1NTE3m54zolcnS+SE19ZyTU=`
- Patron ID 12345 → `KEwHPoJpXY7K757HLmVQXHEyaEg=`

Belt & suspenders

- To identify a specific patron’s transaction, you’d need to
    - Breach ILS
    - Recreate hash algorithm
    - Breach data warehouse
    - Look up patron
- Even then
    - No intellectual content or whereabouts
    - Only the fact of types of transactions
- Strict, clear policies for staff

Data-driven + patron privacy

mladejenovic_n/Flickr

Examples

Sample Data – Workstation Use

COMPUTER	LOC	DATE	MINS	BTYPE	BSTAT	HOMEZIP	AGE	PATRONdeID	REG_YR
BALLIB08	BAL	10/1/2014	5	br	srad	98119	23	KEwHPoJpXY7K757HLmVQXHEyaEg=	2014
CENLIB5270	CEN	10/1/2014	90	br	srsen	98121	69	JeTrHceC+nwaWc/DQZ8VBfgKbL4=	1992
BALLIB08	BAL	10/1/2014	15	br	srad	98107	53	hzvXFK24blsKH9LW7Pkc5kHecto=	2014
IDCLIB12	IDC	10/1/2014	48	br	srsen	98104	63	aS4MypnjX+KV699OVM525fWB//k=	2014
UNILIB12	UNI	10/1/2014	15	br	srad	98105	36	kKdtdIrFDhQTuQwDQcqzGXkQYoc=	2010
BALLIB15	BAL	10/1/2014	25	br	srsen	98117	71	RJ0bkOnwFlmwTrFrAf/fsYJUfMo=	1992
CENLIB3009	CEN	10/1/2014	1	br	srad	98122	50	tK+QVA0PJvQk57147tU8VK08aZ8=	1996
UNILIB12	UNI	10/1/2014	15	br	srsen	98115	81	JytJE+kXHCEMpsK8lUfd4MdU/U8=	1992
CENLIB5330	CEN	10/1/2014	90	br	srad	98104	59	Q2se1NTE3m54zolcnS+SE19ZyTU=	2002
BALLIB15	BAL	10/1/2014	5	br	srad	98104	51	gS08RjQIzUGSZSuStA2Tz7MfvzE=	2013
CENLIB5401	CEN	10/1/2014	7	br	srad	98133	48	mTXkmtPG7e1Y0mMOmhxwb9RaB/c=	2011
CENLIB5330	CEN	10/1/2014	32	br	srad	98104	44	Bi1XhBLDx4Jl9yA1Y2w/tSbZrXM=	2012
CENLIB3011	CEN	10/1/2014	58	br	srad	98133	48	mTXkmtPG7e1Y0mMOmhxwb9RaB/c=	2011
CENLIB5270	CEN	10/1/2014	90	br	kcad	98035	49	T3D+yZiijOFqEuKa39/D4iURCEo=	2009
QNALIB05	QNA	10/1/2014	74	br	srsen	98109	71	y3FSFjyUfO4mc3lzUSUWMGeYLVA=	1992
CENLIB3009	CEN	10/1/2014	4	br	srad	98109	29	szB++tBCmztvhjqEx3i3/S/g2Io=	2005

Sample Situation

- Are patrons abusing 15-minute “Express” workstations?
- Old policy
  - 90 minutes per day for “Internet” workstation
  - 15 minutes per day for “Express” workstation
- New policy
  - Total of 90 minutes per day for any workstation
  - Allowed for “serial” use of Express workstations (6x per day)
  - Staff noticed (anecdotally) that “a lot” of patrons were chaining Express sessions together

What Do The Data Show?

- Before De-identification
  - Total number of sessions and minutes that Express workstations were used per day per branch
- After De-identification
  - Number of distinct (but anonymous) patrons who used Express workstations for longer than 15 minutes per day
  - For August 2014:
    - 12,770 distinct users of Express workstations
    - 3,507 used for more than 15 minutes (evidence of “chaining”)
- Over 25% of patrons used Express workstations more than 15 minutes per day

Sample Data – 2015-08-01

Patron DeID	Number of Sessions	Number of Minutes
34c4e0c201ac7f14f8eef3c14fb877ca	6	90
38b82f34e018ef6accc258e4d539cfd4	6	90
5f9511476cdda7020e6356b4a8d33419	6	90
8025078883a24a72a7f0f84077e14cef	7	90
8c2c07d77b1e8d14ffb3cb7a9489272a	6	90
b1dc081ef62a831a397623e45f9f0915	6	90
cfafdd529713f254d38dcdb480778a0e	6	90
789115f4939f7400fa8b4c3d1485b433	6	89
e3468d6731968e1081e7f4666edb5703	6	89
6af2d93348ed0c9c643cd4a74097c7f9	6	85
78668085a84eddb0869eb16a9c99ddcb	7	80
839cb6c87c4a886ab29ed9513cc008c8	5	75
83110fab344de28ea5731131ca207bdf	5	74
0ecee31153187778f8de69a41407a9bc	6	71
9b01fe4277e1153d75c35a565867986b	8	71
25ce5d330ae8cb26a17ac8798c26fb8d	5	70
eca127e85176f2392c90ff69e81cf782	5	60

Conclusions

Summary

- Store identifiable and non-identifiable information in separate locations
- Avoid storing *any* intellectually significant or identifiable data
- Build data stores that characterizes *kinds* of transactions and rough demographics
- Can be mined to analyze amount of use by demographically similar patrons
- As well as different kinds of activity done by the same people

Thank you

Jim Loter

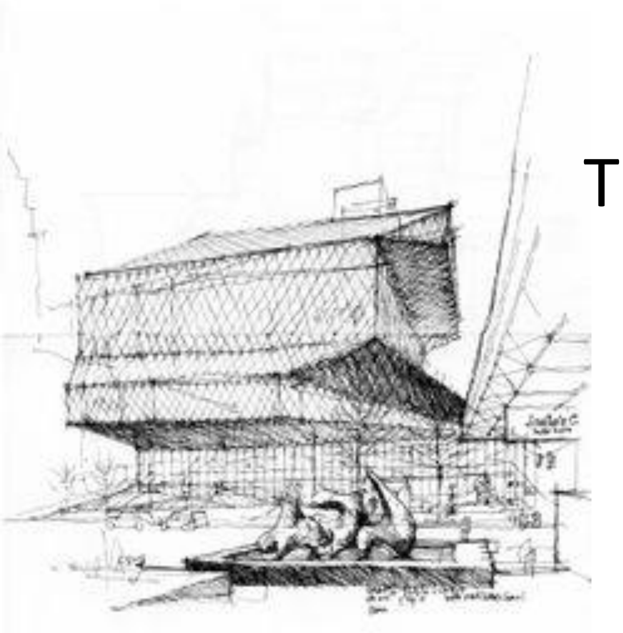
Director of Information Technology

@jimloter

Emily Morton-Owens

Manager of Library Applications and Systems

@bradamant



The Seattle Public Library