The Washington Library Association Journal

March 2016

Volume 32 | Number 1

Alki

The Data Issue

Inside this Issue:

Gaining Insights and Protecting Privacy: De-identifying Patron Data at The Seattle Public Library

by Jim Loter

A couple of years ago, The Seattle Public Library (SPL) interviewed candidates for a director position in our marketing department. The candidates each discussed ways in which they had used customer data to analyze patterns and trends in behaviors, understand the needs of specific demographics, and generally gain insights into how people use an organization's services. In each interview, I followed up this discussion with a question: "How would you still do all that analysis for an organization that throws away all their customer data every day?"

The candidates from outside the library world were, to a person, stunned. Even those who had a sense of the public library's commitment to protecting patron privacy never thought that the library would fulfill that commitment by erasing all evidence of a patron's interactions with us. And yet that is common practice in public libraries.

There are really good reasons for libraries to manage our data the way we do. We benefit from the fact that library borrower records are exempt from public disclosure (thanks, here in Washington, to RCW 42.56.310), which permits libraries to essentially treat them as "transitory" records. By severing the connection between the borrower and the item they borrowed as soon as the item is returned, we greatly reduce the risk that a patron's borrowing activity can be revealed to others–either inadvertently through a data breach, or via official inquiry from, say, law enforcement.

But our traditional data collection practice does mean that we lose potentially valuable intelligence about our patrons–who are they, what do they like to borrow, what material is most popular among particular demographics? The knowledge that could be gleaned from all that data that we delete could help us make better collection development decisions, help us understand shifts in demographics and tastes, and, yes, help us market our services and resources more strategically.

Even as we were hiring our marketing team, however, SPL was working on a solution to this dilemma—a way that we could both preserve individual patron privacy and still expand our knowledge of who are patrons are and how they use our services. The answer to our dilemma lay in the concept of *de-identified data*.

Definitions

In October 2015, the National Institute of Standards and Technology issued a report entitled "De-Identification of Personal Information." An early draft of this report provided some of the concepts, terminology, and approaches that SPL is employing. According to the NIST report:

> De-identification removes identifying information from a dataset so that individual data cannot be linked with specific individuals. De-identification can reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing information. De-identification thus attempts to balance the contradictory goals of using and sharing personal information while protecting privacy (Garfinkel 2015).

De-identified data is distinct from *anonymous* data in that "anonymization…does not provide a means by which the information may be linked to the same person across multiple data records or information systems" (Garfinkel 2015, 2).

In other words, de-identification provides the means to link transactions to distinct individuals (i.e., "the same person") without including identifying details about the individuals. By contract, anonymization removes virtually all information about individuals so there is no way to determine if 100 activities were performed by 100 people, or if the same person performed the activity 100 times. Libraries have historically managed data via *anonymization*. This article outlines the considerations necessary to adopt a *de-identification* approach.

Another important concept in this context is *Personally Identifiable Information*, or *PII*. The NIST report makes the point that PII "is typically used to indicate information that contains identifiers specific to individuals, although there are a variety of definitions for PII in various law, regulation, and agency guidance documents." From that statement, note that what might count as PII for health care data may not be the same as what counts as PII for libraries. In the excellent article "Using Lessons From Health Care to Protect the Privacy of Library Users: Guidelines for the De-Identification of Library Data Based on HIPAA," the authors enumerate the specific data elements that are considered to constitute PII according to the Health Insurance Portability and Accountability Act (HIPAA), which can be used a general guideline for looking at library borrower data (Nicholson and Smith 2005).

continued on next page

Jim Loter was, until recently, the Director of Information Technology for The Seattle Public Library. He is now Director of Digital Engagement for the City of Seattle. Follow him on Twitter at @jimloter.

continued from previous page

Problem Statement

We first came to the de-identification approach after SPL received a grant to study the use of public libraries by members of the so-called "Millennial Generation" (those born between 1982-2000). At the time, our data collection practices included aggregating circulation and computer use data according to our three primary audiences—youth, teens, and adults. The terms of the grant, however, dictated that we collect far more granular data specifically about the Millennial demographic. But we still had a compelling interest to protect users' privacy, whether they were millennials or not. These somewhat conflicting requirement caused us to start exploring the process of de-identification.

> **"Even those who had a sense of the public library's commitment to protecting patron privacy never thought that the library would fulfill that commitment by erasing all evidence of a patron's interactions with us. And yet that is common practice in public libraries."**

For the purposes of the grant, we first established a process in our ILS (the indomitable SirsiDynix Horizon) that automatically exported circulation transaction details for "Millennial" patrons to a separate database and scrubbed the records of personally-identifiable data (name, barcode, address, etc.) For each transaction, we retained the patron's age at the time of the transaction (calculated on-the-fly from their birthdate), their home Zip code, their gender, the number of months they've been a library borrower, and their "borrower type." Note that inclusion of Zip code was one point where we deviated from the HIPAA guidelines.

This approach, however, was closer to anonymization than to de-identification. We realized that though we were now collecting and storing transaction records and demographic information for millennials, we still had no way of knowing *how many actual users* the data represented. Did we have very few prolific library users, or a great many occasional users?

Our De-identification Approach

We determined that we needed to create two datasets to meet our dual goals of learning about the borrowing habits of certain demographics and protecting individual privacy. In assembling these procedures, we adopted the persona of an "attacker"— someone who could somehow gain (or compel) access to this database–to ensure that patron privacy could still be maintained.

Both datasets include one record for each transaction (checkout or renewal) of library material. We used the same principles here also for public computer use, but for the purposes of illustration we will focus on the circulation records.

In the first dataset we include a "de-identified patron ID" in each record. The "DeID" is designed to be difficult-to-impossible to relate to a specific individual's identity but is guaranteed to belong to the same individual over time. This dataset also includes limited, non-PII demographic information about the user, and limited information about the borrowing transaction—collection code, date, and checkout location. The purposeful limiting of the transaction details further minimizes the risk of an attacker being able to reconstruct a patron's identity from the items they borrowed. An attacker might be able to see that a specific unidentified individual checked out 6 items from the Ballard library last Tuesday, but would gain no details about what those items were.

The second dataset *does not* include a "DeID" for the patron. It *does* include the same non-PII demographic information about the patron, and also includes detailed information about the items they borrowed (item record, title, time/date, checkout location, etc.). The idea behind this approach is that since there is no remnant of a patron's unique identity, the detailed item information cannot be used to reconstruct an identity. An attacker might be able to discover what titles 36-year-old men who live in Ballard like to read, but can determine nothing about specific individuals.

The "De-identified Patron ID"

The key to the system is the *de-identified patron ID* ("DeID"), which allows us to identify distinct patrons but minimize the risk of re-identification.

The DeID is a "hashed" version of the patron's numerical identifier from the Horizon database (*not* their barcode, which can change over time). A "hash" is a cryptographic algorithm that produces a unique but consistent value from a given input, but in a way that the original value cannot be determined. The SHA-256 hashing algorithm, for example, will always turn the number "123456789" into "15e2b0d3c33891ebb0f1ef609ec419420c20e320ce94c65-

continued on next page

continued from previous page

fbc8c3312448eb225." But there is (theoretically) no way to take that long output string and figure if that the original value was "123456789" or "abcdefghij" or the full text of the novel "Moby Dick."

To further secure the output, we add a "salt" to the user's Horizon ID. A salt can be a string of characters, like a password, that only we know and/or some additional element(s) from the user's record that will not change over time. With all these precautions in place, we were satisfied that even if an attacker obtained a user's SHA-256-hashed identifier there was no way for them to know (or calculate) what the input string was from that information alone.

First Use Case

The first practical use of our new datasets helped us make a policy decision about public computer use. Public computer session data is collected in a very similar manner to the methods described above.

Our libraries have both "Public Internet" workstations, which can be reserved for up to 90 minutes per day, and "Express" workstations, which can be used for 15 minutes without prior reservation. Patrons are permitted up to 90 minutes of computer use per day regardless of the type of workstation they use.

Library staff began to report that patrons were "frequently" logging on to the 15-minute Express workstations multiple times in succession, to the detriment (and annoyance) of patrons who were waiting patiently for a computer to open up. It seemed that some patrons had figured out that they could chain together up to six 15-minute sessions at the Express computers. This was obviously not the intended purpose of the Express stations. However, all we initially had to rely upon was anecdotal information from a few locations. With our traditional data collection practices, we had no way of telling if any given array of six sessions on an Express workstation were done by the same patron or by six different patrons.

The de-identified data, however, gave us our answer. Since we now had public computer session data recorded along with a patron's "DeID," we could easily tell that, indeed, many, *many* patrons at *all* library locations were logging into Express workstations 6 times per day for 15 minutes per session. We couldn't tell specifically who the patrons were, but we could tell when it was the same patron logging on 6 times versus when it was 6 distinct patrons each logging on for 15 minutes.

The new way of collecting data allowed us to determine that this problem was, indeed, rampant. We made the decision to change the way our system handled public computer sessions. Now, patrons get both 90 minutes per day on reservable public workstations but only 15 minutes per day on Express workstations. This essentially means that patrons qualify for a total of 105 minutes per day on our computers, but it was decided that this was a fairer and more equitable way to allocate time given the abuses we were seeing of the Express workstations.

Other Uses and Next Steps

As we amass de-identified data in this way, we are gaining the new ways to answer questions about how our patrons use the library while maintaining our same high standards of privacy protection. We can tell now, for example, if our heavy millennial users tend to live in particular neighborhoods and if they prefer certain kinds of materials. But we still don't know that Jane Doe checked out "A Clockwork Orange" last October, or that John Smith checked out 35 romance novels last year.

The data becomes even more powerful when cross-referenced with other data sources. If millennials in north Seattle use the library a lot but census data shows that more millennials live in south Seattle where they do not use the library, what can we learn about the patterns of usage that can influence our marketing to the south Seattle millennials?

We are only now at the point where we feel we have collected a critical mass of data that can reliably show usage patterns and trends over time. That same marketing director who was shocked to hear about our data management practices during his interview is now working to integrate our de-identified data into market segmentation studies and other research to unlock further insights into how people use our libraries. That work will directly influence how we improve our services, our collections, our facilities, and our programs.

REFERENCES

Simson L. Garfinkel. "De-Identification of Personal Information," *NIST*, last modified October 2015, accessed 15 January 2016, doi: 10.6028/NIST.IR.8053.

Scott Nicholson and Catherine Arnott Smith. 2005. "Using Lessons From Health Care to Protect The Privacy of Library Users: Guidelines For The De-identification of Library Data Based on HIPAA," *Proceedings of the American Society for Information Science and Technology* 42, no. 1: 1198–1206.