Balancing Privacy and Strategic Planning Needs

A Case Study in De-Identification of Patron Data

Becky Yoose (becky.yoose@spl.org), Library Applications and Systems Manager, Seattle Public Library

In their efforts to create and foster an evidence-based practices environment, library administrators often examine outreach efforts and collection management. Library administrators seeking to improve these areas might ask complex questions such as, "We see a gap in library use for certain age groups; for example, we see that teens and parents are active library users, but people in their twenties are not. For patrons who *are* active in their twenties, were they active users in their teens?" or, "Certain language collections see high circulation in certain branches; however, we are not sure if the patrons using those collections are traveling from other parts of the city to use those materials. Is there a way we can determine the percentage of patrons who are checking out those selected language collections outside of their home branch?"

Libraries who are looking for ways to improve outreach into their communities need information about specific patron demographic groups to provide effective targeted programs and services. Collections managers need a certain level of detail in collection data to determine if certain collections are meeting the needs of particular patron groups. Assessment and outcome-based evaluation of library programs and services cannot be effective without a specific level of detailed data. Patron data is vital for libraries to make the best use of limited resources and funding by determining what programs, services, and practices are the most effective and efficient.

The type of data needed in these analyses is also the type of data that libraries usually discard to protect the privacy of their patrons. This kind of data is considered extremely valuable by companies whose operations depend on customer data: Amazon and Facebook are two examples of businesses with various recommendation algorithms and marketing systems that are built on user behavior data. Libraries should—if not *must*—be sanctuaries from this kind of default detailed data collection, yet, because of the importance of data in evaluation and decision processes, libraries need to gain insights into their patron populations to continue to be a vital resource to their communities.

This article explores one way that libraries can be both data-informed and protectors of patron privacy with the use of a data warehouse with de-identified patron data. De-identification allows for a level of granularity that makes it possible for libraries to answer questions like the ones above, while at the same time maintaining appropriate patron privacy. After establishing a baseline understanding of library data privacy regulations, personally identifiable information, and de-identification, the article focuses on the case study of the planning and implementation of a data warehouse and de-identification plan at the Seattle Public Library (SPL). Some considerations follow for libraries investigating the options of a locally developed or vendor-hosted data warehouse solution with some more general comments about the future of data warehouses and de-identification practices in libraries at the end.

Background

Library Patron Data Privacy Regulations

Rules governing library patron data access and privacy falls within two broad areas: varying levels of special legal treatment on federal, state, and local levels and guidelines and policies provided by organizations. In the United States, the USA PATRIOT Act and, more recently, the USA Freedom Act, are the most prominent federal laws pertaining to library patron data. On the state level, each state has a different approach to defining the privacy of library patron data. The state of Washington, for example, does not have laws that explicitly protect patron data; however, state law does call out library records as an exemption from public disclosure under RCW 42.56.310. Other states have stronger patron privacy laws, including laws regarding parental access to their child's account information and when patron data can be disclosed outside of the library.[i] Finally, for libraries tied to local governments, there are additional records management and privacy regulations to follow in addition to the state and federal laws and regulations. While Seattle does not have any specific regulations regarding library patron records, for example, there are more general privacy and record management regulations by which city departments must abide.

Outside of legislation and regulations, various organizations provide guidelines and best practices regarding the privacy of patron data. The American Library Association's (ALA) Library Bill of Rights and interpretations thereof serve as one of the major sources for US libraries to reference for their approach to managing patron data (ALA 1996). In ALA's Policy concerning Confidentiality of Personally Identifiable Information about Library Users, ALA specifies that confidentiality of patron data extends to a variety of different data sets—including database use, use of library services and facilities, and information from reference/research inquiries—and that this information must be protected from unauthorized access by government agents outside of a warrant (ALA 1991). Beyond ALA, the International Federation of Library Associations (IFLA) also provides more general guidance for libraries in terms of how to approach handling patron data (IFLA 2016). IFLA recommends libraries to abstain from the collection of patron data that would compromise the privacy of said patrons, limit the data collected from patrons, and to educate both patrons and staff about how to protect their privacy, be it online or in the physical world.

Personally Identifiable Information

The National Institute of Standards and Technology (NIST) divides Personally Identifiable Information (PII) into two categories. The first category, PII-1, is information that can directly identify a person, including name, birthdate, address, and Social Security Number. The second category, PII-2, pertains to an individual's activities that can be linked back to that individual. NIST lists several examples of such information, including medical, educational, financial, and employment information (United States 2008). In the context of libraries, the second category of PII includes the intellectual pursuits of the patron, including reference interactions, search queries, and circulation history. This kind of data, in sufficient enough quantities, can be used in certain circumstances to reverse engineer an identity. A famous example of re-identification using PII-2 data is the America Online release of search data in 2006. Even though the data was edited to remove some PII, the amount of PII-2 data present in the dataset enabled researchers to identify searchers by specific search patterns and queries (Techcrunch 2006).

De-identification

Since library patron data contains both categories of PII, libraries must consider the various risks regarding what data should be stored and used for operational use, along with the additional risks of having PII stored with third party vendors. If a library wants to have some ability for longitudinal analysis with regards to library collections and services, then they need to construct a way to track unique data points without identifying unique individuals

i. For a list of state laws regarding library record privacy, please visit http://www.ala.org/advocacy/privacyconfidentiality/privacy/stateprivacy

through PII disclosure (intentional or accidental). Anonymizing the data does not allow for this type of analysis, making it difficult to use the otherwise rich context that historical data would have provided.

Outside of anonymization, another approach to consider for long-term analysis of unique data points is de-identification. The de-identification process focuses on scrubbing particular PII data in a data set while at the same time keeping the data in a state where one can still track unique data points (Garfinkel 2015). With the removal or obfuscation of several PII-1 and PII-2 data points, one's ability to identify a particular individual in a data set is severely hampered, if not made impossible to do.

De-identification is a viable option for protecting the privacy of individuals in particular datasets that are used to track behavior or trends on an individual level. In practice, library patron data de-identification has its unique challenges and considerations. The following case study shows how the SPL approached these challenges with the construction of their data warehouse.

Case Study–The Seattle Public Library

The First Iteration: Targeted Population Market Analysis

The SPL, consisting of twenty-seven physical libraries as well as mobile library services, serves the Seattle community. The Library, as part of its efforts to better serve its community, applied and received a grant in 2013 and 2014 to conduct a marketing research project regarding patrons in the "Millennial" generational age range. The goal of the project was to increase the use of Library services and resources by Millennials over a specific time period. At this time, the only data sets available to conduct this research were from the data sources themselves, primarily from the Library's SirsiDynix Horizon integrated library system (ILS). Because the ILS has both PII-1 and PII-2, the Library was faced with the problem of needing to manipulate the data in a way that would protect patron privacy within policies and regulations but at the same time ensure that the data can provide both the insight desired to gain more traction with the targeted population as well as ways to measure the effectiveness of any actions informed by the analysis. In short, the Library needed a way to track individuals without identifying who they are to get a more granular picture of current Library usage by the target population instead of the more aggregated view that traditionally has been the default in market analysis.

In an attempt to meet the needs of the project, the Library decided to create a separate internal database with exported circulation transactions from the ILS. The transactions had most PII-1 scrubbed or manipulated in a manner to obfuscate identity; for example, the age of the patron at the time of the transaction was entered into the database instead of importing the date of birth attached to the patron record. In the end, the data from the database was used to create a persona for the marketing department to use in developing services and programs for the targeted population. Regular snapshots of circulation transactions from the ILS were imported into the database to help measure the success of the above programs. At the end of the grant-funded project, the data gleamed from this database did play a major role in meeting the goals of the project (increased library usage by the target population by 15 percent over the course of a summer) (Yoose and Halsey 2016).

The Transition to a Data Warehouse

Nevertheless, the analysis could not answer some questions regarding the type of activity being seen in the circulation transactions. The database, while storing individual circulation transactions, was not set up in a way to track circulation transactions by unique individuals. The data was mostly anonymized and the Library could not tell what percentage of the transactions came from particular patrons. For marketing, knowing the type of library usage by patrons can shape outreach and events. Does the usage indicate a core of dedicated library patrons who make extensive use of the library services and resources, or does the usage show a group of patrons who make a couple of transactions, but in greater numbers? Knowing the pattern of use on the individual level gives marketing and outreach a sense as to where to spend resources in their programs and events.

Another consideration for the Library was the ability to use the database for the market study for other projects. The data collected in the database primarily served one purpose—to track individual level circulation transactions of a certain age group. Unfortunately, this focused approach in building the database left little room for other uses of the data by other departments who, for example, might want to see circulation transactions across multiple age groups or branches. The database had a positive, real-world impact, and the desire was to find a way to bring that success into other parts of the Library.

To address the above issues and needs, the Library began work on a data warehouse, the successor to the database used in the grant-funded project above. The data warehouse would incorporate multiple sources of data in a central location, giving different departments in the Library the ability to report on the same data instead of the previous practice of performing multiple exports of raw

data from the data sources themselves, which opens up a variety of problems regarding consistency of reporting as well as privacy and security of data containing patron PII. The data warehouse also provides the opportunity for the Library to balance the needs of data analysis and patron privacy through various de-identification techniques and approaches in the warehouse architecture.

Data Warehouse Architecture—Approaches to Security and Privacy

PII-1 AND PII-2

The approach as to what to include in the data warehouse is guided by the NIST definition of PII-1 and PII-2. Between the two categories, PII-1 tends to be clearer in terms of what needs to be excluded from the warehouse: full name, home address, library barcode, patron record number, and so on. There are a few pieces of PII-1, though, that can be obfuscated to keep some level of granularity in the warehouse for data analysis. For example, many people might be familiar with the case of replacing the date of birth with age. For reporting purposes, the age is just as useful as having the date of birth; for privacy purposes, listing the age instead of the date of birth makes it more difficult to re-identify a person through the warehouse data.

Another way to obfuscate individual data while not tying PII-2 data back to individuals is data aggregation. In the case of title usage statistics from a major digital resources vendor, several staff needed the ability to report on title usage by certain demographic characteristics, such as home branch, age, and council district. Instead of having the demographic information all in one table tied to a specific title, multiple tables were created with each one having a different demographic indicator. For example, one table has title information tied to age group, another table has the same title information tied to the borrower type code from the ILS, and so on. Data stored in these tables were also analyzed against the existing data in the data warehouse, resulting in the adjustment of an existing title circulation table for the same vendor to minimize the overlap of data points between the table and the newer aggregated tables.

EXTRACT-TRANSFORM-LOAD

In a data warehouse, the data goes through a three-step process called extract, transform, and load, or ETL. The ETL process is key to ensuring that no raw PII data enters the data warehouse proper. The following example of importing circulation transactions illustrates the general ETL process of importing data into the warehouse:

1. A script exports the non-PII patron data from the patron record and the item record from the ILS and imports the data into a staging database outside of the data warehouse. During this process the script also transforms the full call number into a truncated call number to obfuscate the PII-2 data point.
2. In the staging area, scripts then prepare and pull together the two separate datasets, matching the de-identified patron information with each transaction.
3. A script then loads the transformed data from staging into the appropriate data warehouse table.

By using the ETL process, the Library has more control as to what data to export from various systems and what data is imported into the database and in what state that data is at the point of import. An ETL process reduces the risk of accidental inclusion of unobfuscated PII or other data that could be used to identify an individual.

PATRON DE-IDENTIFICATION

As mentioned above, the Library needed a way to perform longitudinal analysis without identifying specific individuals. To research use of a particular resource or service over a period of time, however, a way to track distinct data points was needed. One approach is to record all of the transactions with the age and home branch of the patron. The problem with this approach, though, is that it restricts the ability to answer questions such as "do people who check out ebooks still use print?" The essential key to answer questions such as the one listed is that we need to know that Person A is Person A and Person B is Person B, and nothing more.

The solution to tracking distinct data points for the data warehouse is a de-identified patron ID, or De-ID. The De-ID consists of the borrower record number from the ILS, plus a few other key pieces of patron information, run through a SHA-256 hashing algorithm.[ii] In addition, we add a salt to the ID for added security.[iii] The creation of the De-ID happens outside of the data warehouse.

ii. SHA stands for Security Hash Algorithm. SHA-256 refers to a specific set of cryptographic hashing algorithms designed to create strings of text that cannot be reverse engineered back to the original data fed into the algorithm.
iii. "Salt" refers to random data that is inserted during the hashing process, making it more difficult for potential attackers to reverse engineer the algorithm used to create the hashed value.

ACCESS

Access to the database portion of the data warehouse is tightly controlled. Only the IT staff who maintain the warehouse have full read-write access to the database. Select library staff have a read-only direct connection to the database. This mitigates the risk of unintentional (or intentional) changes to the data in the database.

The reporting portion of the data warehouse includes a section on the staff SharePoint intranet where staff can access "canned" reports created by IT, such as collection usage by collection code or circulation numbers of items by branch. Staff cannot access the full database and all the tables from the site, though staff can access select tables of raw data (again, access provided by IT). The data warehouse as a whole is covered under existing policies and procedures regarding access to patron information. The data in the warehouse is treated like data in the ILS—staff already have strict, clear policies about how, when, and why they can access patron data.

RISK MITIGATION

Data warehouses and de-identification cannot be fully free from risk of re-identification of individuals; nevertheless, the warehouse's structure is as such that said risk of re-identification is low. Some risk mitigation strategies are mentioned above: de-identification, obfuscation of PII data, data aggregation, and controlled access of raw data. Another mitigation is the overall architecture surrounding the warehouse. To identify an individual's transactions in the data warehouse, one has to do the following:

1. Breach the ILS database and locate the patron record.
2. Recreate the hash algorithm used in creating that patron's De-ID, including figuring out the salt and the pieces of information used for the De-ID before they are hashed.
3. Breach the data warehouse database and query the table.

The risk for each step varies, depending on various circumstances surrounding each step. Risk mitigations for breaching the two databases above include following best practices and security standards for server and network security, as well as creating and enforcing appropriate access and permissions for user accounts for each system. Nonetheless, given enough resources and time, a potential attacker could execute a successful breach of either database. Recreating the hash algorithm, on the other hand, would be the most difficult out of the three steps above, provided that those who created the algorithm do not fall victim to a social engineering attack or unintentional release of information, such as revealing what pieces of information are included in the creation of the De-ID.

There are other risks beyond someone breaking into the Library's systems, including government requests and data leaks. The de-identification and PII obfuscation guidelines for the data warehouse only leaves the fact that certain kinds of transactions happened, and no specifics, including specific websites visited, titles borrowed by individuals, and so on.

The Data Warehouse's Effect and Considerations

The data warehouse proved useful early in its inception. In the first iteration of the data warehouse, the Library included usage statistics from the library computer reservation system. The data in the warehouse was obfuscated to only include the date and length of time for each session, tied to a De-ID. Because the data was structured in a way that staff can track unique and repeat computer sessions within a period of time, the Library was able to analyze the existing public computer usage policy and adjust the policy to minimize the misuse of the Express workstations (Yoose and Halsey 2016; Loter 2016).

Currently, the data warehouse has reached a critical milestone in housing several types of circulation data by title, aggregated with obfuscated demographic information, such as age range and Census Tract information. The reporting features of the data warehouse have reached a milestone with the launch of a SharePoint site where staff can run "canned" reports, including circulation by branch in a specific timeframe, off of the database.

The future of the data warehouse at the Library will only see growth in the data it houses and the reporting features for staff. Nonetheless, with the increase in data and reporting, the data warehouse's future will be guided by a governance structure. While the IT department is the business owner for the majority of data that resides in the warehouse, the data warehouse ultimately serves the organization's reporting and statistical needs. For the warehouse to be viable in the long term, the warehouse must reflect the business needs of the organization. Other departments in the organization—including Technical and Collection Services, Public Services, and Administrative Services—therefore have a key stake in the warehouse, particularly what data is stored, establishing the authoritativeness of data stored in the warehouse and how it is reported out to both internal and external audiences. Including the stakeholders in the governance of the data warehouse gives the opportunity for the warehouse

to meet organizational needs while it provides the chance for education about the abilities and limitations of current data collection and management practices at the Library with the overarching theme of balancing patron data privacy with reporting needs.

Practical Implications for Libraries

Libraries are asked to provide data for making mission-critical decisions surrounding the allocation of resources. A data warehouse can be a valuable asset for a library in making these decisions without creating major risks in using patron data in the decision-making process. Libraries considering their own data warehouse should consider several factors and risks in deciding to either create their own data warehouse or contract a vendor in creating/hosting a similar product.

> IT [IS] IMPORTANT THAT LIBRARIES USE THE DATA COLLECTED BY THEIR LOCAL SYSTEMS AS WELL AS REMOTE SERVICES IN A RESPONSIBLE MANNER THAT PROTECTS THE PATRONS BUT AT THE SAME TIME DOES NOT NEGLECT THE ORGANIZATIONAL NEEDS FOR EVALUATION AND INFORMED DECISION MAKING

Service Population Size

One reason why the SPL's data warehouse can be effective is the size of the service population that the Library serves. Smaller library systems would run a greater risk of identification, even with de-identification methods. Smaller library systems run a greater chance of having distinct data points tied to specific individual outliers. For example, if a patron lives in a zip code with a small population and does not belong to the majority demographic groups of that zip code, that patron would become easier to identify in a database even with a De-ID and obfuscated PII-1 information.

Available Resources

The SPL has the resources to build and support an in-house data warehouse, including server space, software, and the technical skills of several staff. Some of these skills include knowledge of database architecture, hashing algorithms, obfuscation and aggregation approaches, ETL procedures, and SQL. If libraries wish to secure the information in a data warehouse, a base level of skills, knowledge, and resources are needed to mitigate risks of unintentional disclosure of PII-1 and PII-2.

Data Ownership and Liability

For libraries who wish to contract with a vendor to create a data warehouse or something similar, it is vital for the library to retain ownership of the data they send to the vendor. On a foundational level, libraries differ from vendors in the sense that vendors do not have a commonly held standard of ethics and principles that libraries hold surrounding patron privacy. While libraries are bound to uphold the ethics and principles held by the profession, vendors are not under any professional obligation to do so. Not owning the data in the vendor system increases risk, including exposure of data in a wider data breach, accidental or intentional data leaks, and so on. In addition, the library puts itself in greater risk if there is no liability clause in the vendor contract in case there is a breach or leak. Finally, if the library decides to leave a vendor and does not own the data in the vendor system, the vendor is under no obligation to delete the data if there is no clause in the contract for deletion upon cancellation of services. One way to mitigate the risks mentioned above is to include a data liability clause in the vendor contract, such as the one developed by the SPL in the appendix.

Security and Privacy

The approach to security and privacy for both locally hosted and vendor hosted data warehouses differ in the level of control a library has over the environment. A locally hosted warehouse offers more control over the level of security and privacy a library can build into the data warehouse; the tradeoff, though, is that there needs to be enough resources and skillsets on hand to implement and maintain the desired level of security and privacy. A vendor should have the resources and skillsets, but then the tradeoff is less control over the security and privacy practices applied to the data warehouse.

Future Considerations

Data warehouses, when combined with de-identification of patron data, can be a valuable tool for libraries

needing data for assessment and strategic planning. The level of granularity provided by de-identification enables libraries to conduct longitudinal research and analysis that can lead to more effective distribution of limited library resources. Going back to the questions asked in the introduction of the article, by analyzing the de-identified data, a library can create targeted programs and services focused on retaining active teen patrons when they cross over to the next age group if the data shows that active patrons in their twenties were active in their teens. If the data shows that a sizable number of patrons from one home branch are traveling across the city to use another branch library's language collection, then collection and branch managers can plan ways to grow that language collection's footprint in the home branch in question for easier access.

Not every library can cleanly implement a de-identified data warehouse, partly because of limitations of current de-identification practices (particularly for small data sets) and partly because of resource limitations, be it staff or budget. As de-identification methods evolve, risk of re-identification in small datasets might decrease. There are ways for these libraries to gain similar insights without a full data warehouse implementation, but the risks of potential exposure or identification of unique patrons tied to their activity are still considerable given current practices. In addition, libraries who do reach out to vendor solutions, such as the case of St. Paul Public Library in 2015, face increased scrutiny from other libraries as well as the community that they serve (Gilbert 2015).

Given the rise of evidence-based practices and assessment in libraries in recent years, combined with tying outcomes to future funding and resource allotments, it becomes only more important that libraries use the data collected by their local systems as well as remote services in a responsible manner that protects the patrons but at the same time does not neglect the organizational needs for evaluation and informed decision making. De-identification—and, to a larger extent, anonymization—of data is one of many tools that libraries have at their disposal in conducting responsible data assessment. Unfortunately, this tool is out of reach for some libraries to implement in-house. These libraries, under pressure to produce data for both internal and external audiences (and funding), look outward to vendor products to meet those needs. Libraries have several products to choose from, but the matter of libraries consolidating all patron activity data with a third-party vendor cannot be left unaddressed by the library community. Some of this conversation is already taking place in the form of the ALA Privacy Guidelines and upcoming checklists,[iv] but there is room for the conversation to grow. The community will need to test and to solidify ways to hold both parties—libraries and vendors alike—accountable for protecting patron privacy.

iv. A current list of the guidelines, as well as the upcoming checklist, can be access through http://www.ala.org/advocacy/privacyconfidentiality.

References

American Library Association (ALA). 1996. "Library Bill of Rights." http://www.ala.org/advocacy/intfreedom/librarybill.

———. 1991. "Policy concerning Confidentiality of Personally Identifiable Information about Library Users." Ammended June 30, 2004. http://www.ala.org/advocacy/intfreedom/statementspols/otherpolicies/policyconcerning.

Garfinkel, Simson L. 2015. "De-identification of Personally Identifiable Information." *NIST*. http://csrc.nist.gov/publications/drafts/nistir-8053/nistir_8053_draft.pdf.

Gilbert, Chris. 2015. "Coming soon to your St. Paul library: Data tracking." *Minnesota Public Radio News*. June 30. Accessed January 12, 2017. http://www.mprnews.org/story/2015/06/30/library-analytics.

International Federation of Library Associations (IFLA). 2016. "IFLA Statement on Privacy in the Library Environment." April 5. http://www.ifla.org/publications/node/10056.

Loter, Jim. 2016. "Gaining Insights and Protecting Privacy: De-identifying Patron Data at The Seattle Public Library." *Alki* 32(1): 11-13. Accessed January 12, 2017. https://wala.memberclicks.net/assets/Alki/alki_mar2016_v32-1-v3.pdf.

TechCrunch. 2006. "AOL: 'This Was a Screw Up'." http://techcrunch.com/2006/08/07/aol-this-was-a-screw-up/.

United States. 2008. *Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information: Report to Congressional Requesters*. Washington, DC: US Govt. Accountability Office. http://purl.access.gpo.gov/GPO/LPS111810.

Yoose, Becky, and Stephen Halsey. 2016. "De-identifying Patron Data to Balance Privacy and Insight." Presented at Public Library Association Conference, Denver, Colorado, April 7. http://2016.placonference.org/program/de-identifying-patron-data-to-balance-privacy-and-insight/.

Appendix. The Seattle Public Library Data Liability Addendum for Vendor Contracts

ADDENDUM

CONFIDENTIALITY OF SEATTLE PUBLIC LIBRARY RECORDS AND DATA

The Seattle Public Library (SPL) collects and manages records and data which require confidentiality under one or more federal or state laws, or under recognized industry standards, including but not limited to, the following:

- Health Insurance Portability and Accountability Act of 1996
- Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009
- Children's Online Privacy Protection Act of 1998 (COPPA)
- The Privacy Act 1974 (as specified in the National Institute of Standards and Technology (NIST) SP 800-122)
- Washington State RCW 42.56.310
- Family Educational Rights and Privacy Act of 1974
- The American Library Association Library Bill of Rights
- United States Constitution, including the first and fourteenth amendment

Specifically, a provider of services to SPL will not reveal or disclose any data or records, either physical or electronic, which are designated as confidential by the Library or which pertain to SPL patrons when such data or records could be used in any manner to identify a Library patron or any references or materials that a specific Library patron accesses.

A provider of services to SPL must treat all the designated or individually identifiable SPL records as confidential and protected. Encryption of such data while in motion or at rest, and restricting access to confidential data, are typical methods of data protection. No SPL records or data shall be released by the provider to any third party without the prior written consent of the SPL.

In the event that the provider violates this addendum, then said provider agrees to indemnify, defend and hold harmless SPL and its employees from and against any losses, costs, expenses, liabilities (including attorney's fees), penalties and sanctions arising out of or relating to such violation. This addendum does not limit the provider's liability as specifically established under law.

The Parties hereto agree that this amendment modifies, changes, amends and has precedence over any contradictory language in the contract between the Parties.

Provider_____ Date___

Seattle Public Library_____ Date___