Using Lessons from Health Care to Protect the Privacy of Library Users: Guidelines for the De-Identification of Library Data based on HIPAA

Scott Nicholson
Syracuse University School of Information Studies, 245 Hinds Hall, Syracuse, NY 13244-4100, scott@scottnicholson.com

Catherine Arnott Smith
Syracuse University School of Information Studies, 245 Hinds Hall, Syracuse, NY 13244-4100, casmith07@syr.edu

Abstract

While libraries have employed policies to protect the data about use of their services, these policies are rarely specific or standardized. Since 1996 the U.S. healthcare system has been grappling with the Health Insurance Portability and Accountability Act (HIPAA), which is designed to provide those handling personal health information with standardized, definitive instructions as to the protection of data. In this work, the authors briefly discuss the present situation of privacy policies about library use data, outline the HIPAA guidelines to understand parallels between the two, and finally propose methods to create a de-identified library data warehouse based on HIPAA for the protection of user privacy.

Introduction

Librarians historically have been concerned with the protection of data about their patrons and the patrons' use of library services. The library is seen as a safe place to explore information, and part of that is because librarians protect the identity of their patrons. This protection has come into conflict with legal authorities in several cases. Recent conflicts have caused some librarians to turn to the deletion of large amounts of library records (Million, A. & Fisher, K., 1986; Murphy, 2003), and these deletions have long term consequences for the data-based management of and research about library services (Nicholson, 2003).

Exacerbating this privacy issue is the fact that there are no specific policies to guide librarians in dealing with their library data. Privacy of library data is currently regulated at the state level, and most of these policies are not specific in their wording about what data libraries should keep (Kennedy, 1989). The guidelines provided by American Library Association point to the state laws and thus carry the same ambiguity (American Library Association, 2004a). This lack of specificity for librarians and library system creators makes it difficult to respond quickly and in a uniform manner to threats to patron data. This was demonstrated recently when librarians' reactions earned the label of "hysteria" in their reaction to the Patriot Act (Lichtblau, 2004).

Healthcare is another field currently struggling with similar privacy issues and the public perception of its handling of these privacy issues. The users of healthcare services expect their personal health information to be kept private, yet the information is valuable to clinicians, third-party payers, and researchers; however, the philosophy of HIPAA is that such information does not need to be attached to identifiable individuals to aid in this type of research. The goal of HIPAA is not to have a chilling effect on healthcare research, but to assure that anonymized data can be provided to those who need it and the privacy of individuals protected.

The situation of library services is similar to that of medical services. People come to the library from specific demographic groups with information needs, and librarians administer appropriate treatments for those needs. Users of library services expect that their privacy will be protected. Maintaining an anonymized record of library use can help the libraries to make and justify data-based management decisions through tools such as the bibliomining process, which is data warehousing and data mining for libraries (Nicholson, 2003).. In addition, this type of record is invaluable to researchers trying to discover larger patterns of information needs and library use. However, there is no specific legislation such as HIPAA that assists library service providers by providing specific guidelines as to what to maintain and what to discard.

This work is based upon the assumption that the standards for protecting health care information are strict enough to protect the data about the information seeking of individuals. It is assumed that in most cases data about the health of a person are considered to be just as private, if not more so, than the information needs of a person; therefore, the standards for medical data protection should be appropriate for protecting data about information seeking behavior. There are several specific cases where this assumption is not true, and these are discussed at the end of this work.

We begin this article by admitting that there is an underlying difference between the application of HIPAA in healthcare and the application of similar policies in libraries. HIPAA is designed to protect information about individuals when health care information about those patients is transmitted. While the hospital that maintains medical records may keep more complete documentation, HIPAA is used to determine what is available for inspection by third parties.

In the library setting, however, these guidelines are needed to inform librarians what to maintain in their internal information systems, so that if other groups access the data, patrons' personally identifiable information will be protected. The resulting data may be useful to library management and administration, both for an individual library and for a consortial group, and to library scientists looking to better understand library processes. In addition, if other agencies demand data from the library through legal channels, the data will not incriminate individual library patrons as long as patron identity cannot be connected to library activities .

The goal of this work is to examine HIPAA, extract the basic components, and map those components onto a library domain. The result is a specific set of suggestions for those creating policies and library systems that will balance the needs to protect the privacy of patrons with the desire to maintain a data-based history for library decision-making and library science research such as bibliomining.

A Brief History of the Privacy of Library Records

While providing a safe place for patrons to get information has been a long-standing goal of most libraries, explicit concern for the records of the patrons' use of library materials has been brought to the surface in the last thirty years (American Library Association, 2004a). These concerns began long before today's electronic patron databases. Unbeknownst to most librarians, the FBI, through the Library Awareness Program, requested patron information and data on the library use of Soviet citizens in the 1970s and 1980s by appealing to the librarian's sense of patriotism (Flanders, 2002).

As librarians became aware of the Library Awareness Program in the mid-80s, they started to fight back. Across the country, librarians worked with state governments to create laws to protect the privacy of library patron data. There was no standard guiding these laws, so the result was a patchwork of state-based laws specifying different types of information that was protected, different parties that could view the data, and different instructions as to what to do with protected data. Those writing about patron data had to always include the warning "Check your state laws", as there was no common rule of thumb to guide librarians and researchers.

The terrorist activities of 2001 began a new chapter in the battle for patron privacy. In response to terrorist activities, the government launched the USA PATRIOT act ("Uniting and Strengthening America by Providing Appropriate Tools to Intercept and Obstruct Terrorism Act of 2001", Pub. L. No. 107-56, 115 Stat. 272). This act trumped the state laws, allowing agents holding appropriate legal paperwork to request information about individuals using library services. In addition, librarians were not allowed to inform patrons that their activities were being monitored. This resulted in an uproar from librarians and, in some cases, the deletion of large amounts of library data (Murphy, 2003); in reaction, researchers voiced concern with the loss of the data-based history of library services (Nicholson, 2003). With no easy method of maintaining an anonymized version of usage records, some libraries simply saw no other option but to delete the data.

This loss of data should be of concern to librarians and researchers. As resources have become scarcer, funding agencies supporting libraries demand a higher level of data-based justification. The anecdotal stories of the library's role in providing information are being replaced with the general belief that "everything is on the Internet." Libraries must be able to demonstrate their importance through documentation and data or risk being replaced with computer labs. In addition, library managers and administrators must spend shrinking budgets wisely; one lesson from the corporate sector is using available data to make justifiable decisions.

The USA PATRIOT act expired in 2003, and a new, more powerful, version is still under debate as of the writing of this article (Holland, 2004). Librarians facing a shrinking patronage and funding base cannot afford to discard valuable information about their patrons in response to a more powerful USA PATRIOT act. It is clear that guidelines are needed to aid librarians in knowing what information to collect about patrons and their use of

materials that will allow librarians to make data-based decisions while still protecting individuals' personally identifiable information.

Librarians are not alone in this struggle. Those working with medical records have always faced similar challenges in the attempt to balance capturing information about patients and treatments for research and protecting the personally identifiable information about those patients. By examining the history of this struggle and the solutions proposed to protect patients in health care, we draw inspiration for possible policy solutions for protecting personal information in library databases while still maintaining key fields for decision-making and research.

A Brief History of Censorship and Medical Records

The medical record is a feature of patient care at least as old as the Codex Hammurabi, the legal system named for an ancient king of Babylon (Spiegel & Springer, 2001). Several decades ago, pathologist Donald Lindberg, current Director of the National Library of Medicine, delineated the following typical components of the patient record (1968a, pp. 13-14).:

- "Abstract of patient's life history
- Patient's physical examination and dates
- Physical measurements not made by physician personally, e.g., height, weight
- Physical measurements made by physician
- Measurements of specimens
- History of present illness
- Diagnosis and therapy
- Logic by which decision was reached
- Measurements and estimates of the patient's thought processes and personality." (p. 14)

Little has changed since that time.

During the treatment process, information can be created that individuals may not want included in their medical records, or that health care professionals purposively do not document. Gutheil and Hilliard (2001), in an unusual article entitled "Don't Write Me Down," formulated a taxonomy of psychiatric patients' requests for noninclusion of specific items or entire sessions:

- Don't take notes.
- Don't keep records at all.
- Leave this out.
- I want you to change the record.
- I want you to destroy my record.

Rationales for these statements include legal difficulties, avoidance of stigma "for political and other reasons," concern that a permanent record will be created, "embarrassing material," and "overtly paranoid concerns"(p. 160). The authors conclude that the "clinical requirements [for records] may be visualized in terms of 'audiences' for the record"(p. 160), which audiences make up a long list including the treating clinician, other persons treating the client, other clinicians, reviewers, the legal system and, finally, the patient.

Clinical narrative text may contain combinations of information (e.g., occupation and diagnosis) that make the identity of an individual clear to a knowledgeable viewer. The problem of such surreptitious or accidental identification is one that is rarely noted in the literature, except in passing as a factor noted while in the process of other kinds of analyses. For example, Johnson and Friedman (1996) write about an NLP investigation of expressions of patient race in discharge summaries: "In cases in which race is unknown … information from the discharge summary may be available. While the percentage of cases in which this was possible was low (1.7%), analysis of more paragraphs of the summary (e.g., 'social history') are likely to yield additional information. The 'social history' paragraph in the discharge summary often indicates country of origin, language spoken, etc"(p. 540). As another example, Goodwin and Prather (2002) wrote in their case study of deidentification at Duke University: "outliers (e.g., a very young pregnant girl) were still potentially identifiable by a small number of employees who knew the patient"(p. 67). This patient's status as an outlier rendered her identifiable in clinical context.

Further, certain types of information contained in textual narrative may embarrass even those individuals who are not positively or uniquely identified by the clinical record. Consider the following topics that one legal specialist in health information suggests for inclusion in a hospital "Release of Information" policy: abortion; adoption records; blood type and donation; infectious diseases (AIDS, HIV, venereal diseases, tuberculosis); legally or clinically incompetent adults; mental health records; organ donation; sexual assault; substance abuse records; and issues relating to status as a minor, such as pregnant minors, minors who are parents, and birth control for minors (Carman, 1997).

To further complicate the problem, more individuals than simply the patient must have their personal health information removed from the medical record for deidentification to have taken place. The University of Pittsburgh's De-ID project goes a step beyond what HIPAA requires to remove all names present in medical record text—including the names of physicians and nurses involved in the patient's care. This is done because the removal of all names, regardless of whose names they are, relieves concern that a patient or family member's name could be confused with (or be identical to) that of a physician or other member of the healthcare team. Melissa Saul, a De-ID developer, makes the additional comment that in cases of quality assurance and review involving a physician's group practice, a business case can be made specifically for physician deidentification (Melissa Saul, personal communication, October 2004).

HIPAA and De-Identification of Medical Records

HIPAA defines uses and disclosures of personal health information (PHI) that must be authorized by the patient. The HIPAA Privacy Rule is a portion of the Health Insurance Portability Act that specifies the conditions under which health information may be used or disclosed for research purposes. Research is defined by the Privacy Rule as follows: "A systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" (Final Modifications, 2002). Information required for treatment, payment or operations does not require authorization; however, research is subject to HIPAA. This requires considerable changes in health information management and in research practice alike. For this reason, the American Association of Medical Colleges expressed the concerns of many researchers in writing that the Rule would "create significant obstacles to the conduct of biomedical, epidemiologic, health services, and other health-related research" (Smith, 2001).

A significant and important exception to the Privacy Rule states that health information may always be used or disclosed, for research purposes, if it has been deidentified, in other words, if it is made impossible for health information to be traced back to a particular individual. Data that has been deidentified in this way does not require IRB review.

The goal of HIPAA regulations on the de-identification of medical records is to create data that others can use while protecting the identity of individuals. Specifically, the HIPAA standard states that there should be "no reasonable basis to believe that the information can be used to identify an individual" (U.S. Health and Human Services, 2002). There are three ways this can be accomplished: through the removal of specified types of information, creating a limited data set for a specified project and research group, or to demonstrate through statistics that the chance an individual can be re-identified is "minimal" (Sweeney, 2004).

Removing Specific Identifiers

The first option to comply with the HIPAA guidelines is to remove several categories of individually identifiable health information. Individually identifiable health information is defined by US Code to be "information that identifies an individual; or with respect to which there is a reasonable basis to believe that the information can be used to identify the individual" (Definitions, 2002, 42 U.S.C. § 1320d(6)). The Privacy Rule stipulates eighteen specific data elements that must be removed for deidentification to take place. These elements are:

- "Names
- All geographic subdivision smaller than a state, including ZIP code and geocodes (except for the initial three digits of the ZIP code under certain circumstances)
- All elements of dates except year for dates directly related to an individual, including birth date, admission date, date of service, date of discharge, date of death; and all ages over 89 years, including all elements of dates including birth year indicative of such age (except that there may be a category of age 90 or older)
- Telephone numbers
- Facsimile numbers
- Electronic mail addresses

- Social Security numbers
- Medical record and prescription numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers, including serial numbers and license plate numbers
- Device identifier and serial numbers
- Web Universal Resource Locaters (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including fingerprints and voiceprints.
- Full face photographic images
- Any other unique identifying numbers, characteristic or code" (Center for Disease Control and Prevention, 2003, ¶5).

These eighteen items can be placed into these conceptual groups:

- Direct identifiers and identifiers that connect into other databases;
- Address and location information;
- Dates related to an individual;
- Contact information.

Removing these elements to create deidentified records is known as the Safe Harbor method. An important corollary to the Safe Harbor standard is that the covered entity "does not have actual knowledge that the remaining information can be used alone or in combination with other data to identify the subject" [45 CFR § 164.514(b)].

The data elements of greatest concern are those containing date and geographic information, as trends and patterns in health care data are commonly tied to these fields (Sweeney, 1997).

Limited Data Set

The second option is to create a limited data set, which is a subset of available data selected for a specific and well-documented management or research need. This is needed when a research question requires some of the information that would be removed by the Safe Harbor method. For example, researchers may believe that an outbreak of lung disease in an area may be related to an industrial plant. In order to explore this, researchers would need specific address information of a subset of individuals. The Limited Data Set option through HIPAA would allow the hospital to create a dataset of only those people with a specified set of conditions along with address information.

A Limited Data Set can contain location information and detailed date information. In order to provide researchers with a Limited Data Set, there must be a data use agreement in place. This data agreement states that the researchers will:

- "not use or disclose the information other than as permitted by the agreement or as otherwise required by law;
- use appropriate safeguards to prevent uses or disclosures of the information that are inconsistent with the data-use agreement;
- report to the covered entity any use or disclosure of the information, in violation of the agreement, of which it becomes aware;
- ensure that any agents to whom it provides the limited data set agree to the same restrictions and conditions that apply to the limited data set recipient with respect to such information; and
- not attempt to re-identify the information or contact the individual"(Center for Disease Control and Prevention, 2003, ¶ 20) .

This method allows researchers access to information that could be used to identify patients. The main complaint about this method is that it requires time and effort to set up an agreement for each research project (Sweeney, 2004).

Statistical Demonstration

The third, but currently rarely used (Sweeney, 2004) technique to determining if data is de-identified is through an application of statistical methods. Sweeney (1997), creator of the algorithms behind a privacy certification tool called Privacert, discovered that through date of birth, gender, and a five-digit zip code, 87% of the U.S population could be uniquely identified. Using the Safe Harbor guidelines, the date of birth could be turned into a year and the zip code could be reduced to a three-digit zip code to make de-identified data.

If some of this information is needed for research, other fields could be masked in their stead. For example, if the date of birth is turned into an age range, then that age range could be kept with the five-digit zip code and still mask the identity of individuals. This could be checked in a given dataset by examining how many people fall into each combination of zip code and age range to ensure that no person is uniquely identified.

The underlying concept is that a researcher uses data matching techniques to determine the chance of identifying an individual through all combinations of the demographic variables. In addition, the researcher combines the dataset with other available datasets to ensure that matching entries between datasets is not possible. The goal in this demonstration is to minimize the chance of re-identification of individuals. At this point, however, that level of minimization has not been specified and will most likely play out in court interpretations.

Problems with HIPAA

Although the goals of HIPAA are laudable, the effectiveness of its policies have been heavily debated; the heavy fines ($15,000 per individual violation) imposed on organizations have created a climate in which myths have power. Georgetown University's Health Privacy project (http://healthprivacy.org) is a resource for learning about the misinformation associated with HIPAA. The privacy and security concerns manifested in HIPAA and other policy sources have simply served to intensify public awareness of medical privacy problems. In order to accept education about confidentiality and sensitivity of data, the organization itself must have values that make these sensitivity and confidentiality needs possible. A dramatic recent illustration of terminology's potential to confuse and mislead was given by Gleason and Yates (2004) in a letter to the American Journal of Psychiatry. Their patient, a 79-year-old woman with no previous psychiatric history, attempted suicide. She had recently received a letter from her insurance company informing her of new medical information release policies under HIPAA, "misinterpreted the letter to mean that her insurance company was discontinuing coverage," and in the apparently mistaken belief that she had skin cancer, "believed she had no medical insurance to pay for treatment [and] did not want to be a burden to her husband." Gleason and Yates comment that "language in our patient's notice indicating who may have access to—and potential disclosure of—her protected health information may have contributed to her misunderstanding" (p. 374). This is a classic, and tragic, example of healthcare communication as a two-edged sword. Language that is intended to educate, delivered in the absence of context making further explanation easily available, may serve only to confuse and frighten.

Awareness of privacy issues itself has been shown to have an effect on patient care. Patients with sensitive information in their medical records have been found less likely to consent to disclosure (Merz, Spina, & Sankar, 1999), while on the clinicians' side, among 700 San Diego physicians "likely to encounter patients with significant family histories of cancer" some physicians did not even want a written notice of genetic test results to appear in patients' files, in order to maintain patient confidentiality! (Wasserman, Jones, Trombold, and Sadler, 2000).

Application of HIPAA to Library Data

Libraries need a guide as to what data fields to keep about patrons in their internal databases. In contrast, HIPAA policies were developed to determine what data should be transmitted outside the internal system to a third party.. The policies developed below for libraries will aid libraries in determining what should be passed on. In addition, this work can be used as a guide to what libraries should maintain in their internal systems to protect patrons in case a third party demands the data through legal means.

There are two stages of data storage in a library system: the time period in which a patron is interacting with a library resource, and the time after that interaction is completed. While the patron is interacting with an item (such as the circulation of a book), the library does need to connect the patron to the resource. After the work is returned, however, the patron identification can be replaced with a surrogate made up of demographic-type categorical variables for long-term storage of usage data. The HIPAA guidelines are useful in making the transition between these two data sources.

In order to provide these specific guidelines for de-identification to librarians and researchers, we will apply the concepts for creating de-identified medical data to the library setting. The most restrictive of the concepts, the safe harbor method, is easily applicable to library data. Those working in library services or creating library

systems can use this as a guideline for the data elements that should be discarded before storing data about use of library services.

One term that will be used throughout this section is a “demographic surrogate”. Just as a bibliographic surrogate contains fields that describe a book, a demographic surrogate is made of fields that describe a user. The idea behind a demographic surrogate is that some pieces of user identification are removed and replaced with a set of categorical variables. These might come from a patron application or from matching the field to another database. For example, an academic library could match a user ID to the university database to create a demographic surrogate of department and level (undergraduate, graduate, faculty, staff, etc.). A public library might match a zip code to census data in order to extract a demographic profile of someone living in that area. In each case, the set of demographics replace the personally identifiable field.

It is important when doing this to ensure that individuals are still not identified by combinations of demographics. This will require analysis during the development stage of “bin sizes” which are the number of people identified through each combination of demographic fields. If the bin size is small, then demographic categories can be combined to avoid identification of individuals.

HIPAA Safe Harbor Policies for De-Identified Library Data

The concept behind the Safe Harbor plan is that information that allows connection of library use to an individual is removed while keeping information useful for decision-making and research. Many of the HIPAA elements have counterparts in the library domain. While the specific fields in a library database may not be listed in the HIPAA guidelines , the underlying conceptual goals of protecting personally identifiable information remain the same.

The first step in de-identifying library data is to remove identifiers. There are three types of identifier elements to consider. First are direct identifiers such as name, social security number, e-mail, and telephone numbers. The second type of elements are those that provide a link into another data source that would uniquely identify someone, such as account numbers, license numbers or vehicle identifiers. The final type of personal identifier consists of combinations of variables that identify a single patron..

These variable combinations that create unique identifiers are the least obvious and may require some analysis of the library system and data to discover. For example, in an academic library setting, the combination of department (“library science”), gender(“male”), and status (“faculty”) could uniquely identify individuals or create a way to link into another database to identify someone. In order to avoid this type of identifier, small computer programs can be written to systematically explore the number of people within each variable combination. When areas of concern are discovered, one of the variable levels can be combined with another group (such as “library science” with “information science”) to prevent the identification of individuals.

The next step is to remove address and location information. HIPAA guidelines state that the smallest geographical unit that can be used is the state, with the exception that the first three digits of zip code can be kept if that zip code area contains more than 20,000 people. For a community library, this may effectively nullify any useful information from location information as the entire population of the community may be less than 20,000; in this case, the library can create a demographic surrogate from address/zip code information, append that to the item information, and discard the original address information.

The other type of location information to remove is virtual location information. In some cases, IP addresses can be used to track use of a digital service back to a particular computer. Proxy server information can be used in the same way. In a digital reference environment, there may be an audit trail available that would allow the tracking of transactions back to the computer user. These need to be cleaned in the same way that physical addresses are cleaned, and replaced with demographic surrogates when appropriate. In-library use of computers can also be a problem if there is a sign-up process that connects a user to a specific computer. Again, demographic surrogates can be used to replace both the user ID and the computer location (with a replacement of a specific IP with a more general place such as “Lower floor lab”).

Another category of information to remove is dates related to an individual, such as birth date. The age of the patron can be kept if the patron is under 89; those over 89 should be grouped into a single category of 90 and over. One date that needs to be considered carefully is the date of use of an item. This field can be very useful to library managers in considering staffing and examining patterns of material use. By knowing the days and times when items are circulated or reference questions are asked, library managers can better deploy their personnel. By observing patterns of use, librarians can better understand peak times when the collection will be heavily taxed and might choose to remove high-demand items from circulation during those periods. In addition, instead of

keeping a returned date for a circulated item, librarians can maintain a length of circulation field that will not identify a particular time.

In theory, if there is no way to connect a date of library use to an individual, then date of library use should be a safe field to include. In reality, the library may not be aware of external resources that connect an individual to a date of library use. For example, a paid parking lot in which photographs are taken of vehicles as they enter would allow connection of a car to a date of library use. If an individual is under surveillance, a third party would know a date when the individual entered a library, and could then use demographic information in conjunction with a specific date to retrieve a set of records.

Given this situation, we suggest two methods for capturing date information. In the primary data source that contains work information and demographic surrogates, only data representing the month and year of use should be kept. This will allow for the identification of larger-scale usage patterns while avoiding concerns with linking individuals to works. The second data source is similar to the HIPAA "limited data set" concept; managers will determine which fields are useful in staffing decisions, and then maintain a second data set with specific date and time information along with general surrogates for the library resource that was used (such as subject) and general demographics for the user (such as age range). The specific fields kept in this other data set should only be ones useful in decision-making, and care must be taken to ensure the metadata kept can not be matched to the larger primary de-identified data source.

Limited Data Set Concept

As described above, the limited data set articulated by the HIPAA legislation can be valuably applied within libraries. In the HIPAA context, a limited data asset allows a health care system to provide researchers with data from their databases if proper agreements are in place. In the library context, this concept does not transfer directly, as the guidelines in this work are designed to aid libraries in knowing what they should keep in their internal systems. If applied properly to library systems, there will not be any extra data to supply to researchers from the primary data source.

In the library context, the limited data set concept will be used for secondary data sources that are created alongside the primary dataset. As mentioned earlier, the library will keep full information about a patron and item during the time of use. Once that use is over, certain fields will be removed and the contents of other fields replaced with demographic surrogates to create the primary data warehouse. For a specific library situation and research need, however, there may be other data sources that are useful. These data sources will be limited to just the fields needed, and extra steps will be taken to ensure that a match cannot be made between the two data sources.

One example presented earlier is that of detailed date information for staffing. In this case, the month and year information kept in the data warehouse may not be sufficient. So, the date and time information is kept in another dataset with more general subject categories for the items used. This dataset would be useful in staffing decisions while not compromising the identity of the patrons involved.

Another useful application of the limited data set is one based upon simultaneous usage. If the unique user identifiers are removed, then the combination of items circulated at the same time or digital works examined during the same Web session is lost. This information is valuable for recommender systems. In order to maintain this type of information, a second data set can be created that contains the library resources that were used during the same session and other general demographics of use. This will allow library scientists to build recommender systems while protecting the privacy of users.

The needs for these limited data sets will depend upon the specific library setting or research question. The key consideration when creating these resources is that they can not be tied back to the primary de-identified data source. In order to ensure that these limited data sources protect privacy, the third part of de-identification, statistical demonstration, is useful.

Mathematical Demonstration

The final tool for de-identification is statistical demonstration. After creating a limited data set, demographic surrogates, or any other combination of data that may be questionable, techniques can be used to explore the data structure. The goal of this process is to ensure that, mathematically, the chance of identifying an individual is low.

There are several possible techniques useful in this exploration. If there is historical original data available, the algorithms developed to create the de-identified data warehouse can be tested to ensure individuals are not identified in any combination of the variables in the demographic surrogate. Another option is to examine the current population of the library and apply the techniques to the entire population. If the demographic surrogates uniquely identify any of the library population, then demographic variables can be combined as needed to mask individual patrons.

Future members of the library user population will not be part of this tested sample. Because of this, it is important to build in regular automated testing for new members of the library population to ensure these individuals are not uniquely identified through the demographic surrogates. If the individual is identified, either categories can be collapsed and the entire historical dataset adjusted or the individual can be flagged as someone who is not to be included in the long-term dataset.

There are more complex methods of creating datasets that are deidentified. These can involve rounding numerical data and randomly exchanging entries in demographic fields. This field, known as privacy-preserving data mining focuses on techniques that protect the privacy of the individual, but still allows for similar conclusions as those determined through more invasive methods (Clifton et al, 2004).

Involving the Users and Legal Concerns

Libraries should already have some type of privacy policy in place. This policy will need to be revised to explicitly state what information is kept about users and their library use, and might include an example of the data in the de-identified data source. In addition, the policy must include instructions to the patron as to how they can remove information about a particular use of a library service or how to set a flag on their patron record that will prevent any information kept about them at all. Neuhaus(2003) presents guidelines for this type of policy in digital reference, which is one example of a library service with significant data privacy concerns.

Another concern that libraries must consider is legal issues. Most states have privacy policies that govern library data. These restrictions from HIPAA will be sufficient to meet the standards of state legislation. Libraries should be aware of the legislation in order to know what other parties, such as researchers, may have access to the data, as some states dictate this. The ALA has collected this legislation on their Web site in the Intellectual Freedom area (2004b).

One lesson to be taken from the HIPAA research is that of taking care in the creation of policies. Librarians should work with groups of patrons to craft policy statements that inform patrons in clear and concise language about:

- What data fields are and are not collected,
- What the data will and will not be used for, and
- How patrons can remove their own data from the system.

Additional Concerns regarding the De-Identificaiton of Library Data

There are a few issues about library data that are related to the taxonomy of non-inclusion of healthcare data and the analysis of textual data presented earlier and we expect that as others work with these concepts, additional concerns will arise. The first concern is information needs that, by their nature, need to remain private. One example is researchers who ask librarians for assistance. Some researchers working on cutting-edge research, patent searching, and other classified topics may not even want the topic of their research saved in a data warehouse. In this case, the privacy concern has nothing to do with the individual; rather, it is the *topic* of the inquiry that is of concern. In digital reference archives, this is a significant problem, as the question in digital reference typically may contain context for the query as well. This concern can be dealt with by providing full disclosure to the patron about what happens with the data after a transaction and allowing the patron to remove a single transaction or to stay completely out of the archives.

Another concern comes from the richness of the archives of digital reference. While fields containing personally identifiable information can be easily removed from these archives, the user and expert may include this type of identifying data in the text of the question or the answer. These might be easy to remove, such as a signature line at the bottom of a message, or may be more difficult, if the information is used to provide context for the question. This provides a research opportunity to bring natural language processing techniques over to digital reference, to aid in the identification and removal of personal information from digital reference queries. This line

of research will also need to examine the follow-up question – how much personal context can be removed from a question and still leave a question/answer pair that has value?

Applying the HIPAA Regulations to Library Data

This article has focused on what to do, but not how to do it. The application of this concept to library data is greatly system-dependent; as there are few standards in how library systems store data, it is impossible to offer sweeping instructions as to how to apply these concepts. Most libraries use a library automation system which will not allow them to manipulate the internal data to the level required by these suggestions. There are two pathways, however, through which many libraries can apply these guidelines.

The first method is for the vendors to adopt these guidelines. Just as medical systems can be HIPAA-compliant, library automation systems can become LIPAA-compliant. This would require a larger standards group to examine the guidelines proposed in this document and adopt them for vendor use. This method would require little work by libraries, but would require library policy-makers to meet and agree upon the components that need to be removed in library systems.

The second method is by applying data warehousing techniques to separate operational data from archived data. The data kept by the library while a user is using a library service remains in the operational systems. Once the use of the library service is complete, the data is then moved from the operational system to a secondary data storage area. During this move, the fields containing personally identifiable information can be removed, leaving a safe historical data set. A library data warehouse would also allow libraries and library scientists to more easily perform analysis for evidence-based library decision making. Creating a data warehouse is the first step in the aforementioned bibliomining process (Nicholson, forthcoming).

Future Research Questions

This proposal introduces some research questions for the future:

- How much library decision-making information is lost at each stage of the de-identification? Are there algorithms that would let researchers simulate the missing data in order to draw similar conclusions?
- What standards could be presented to encourage libraries to maintain the same type of de-identified data? These standards would encourage the development of multi-library data warehouses, which would be valuable for library consortia and researchers.
- What perceptions do users have of their de-identified data? Introducing these measures will, in most cases, lead to less data captured about a user in library systems; however, the data that are captured will be more readily analyzable.

Conclusion

The goal of this work is to take the methods from health care used to create de-identified medical data and apply them to a library setting. The resulting guidelines can be used by librarians as an alternative to discarding data. This will allow libraries to maintain a data-based history of use while not compromising the privacy of their users if another party accesses the library data.

This process occurs when the user has completed their use of the library resource. Non-identifying information about the user is matched with information about the resource and stored in a de-identified data warehouse. The stages in creating the de-identified data are to first remove data fields that uniquely identify the patron, such as name and e-mail. Fields that link a patron into another database, such as a user ID, can be replaced with a set of categorical variables known as a demographic surrogate. Next, location fields, such as physical or virtual addresses, are removed and replaced by additional demographic surrogates. Specific dates are made more general to be non-identifying.

Then, secondary limited data sets are created to aid in specific research questions or management needs. These might include circulation or question counts by date and time or simultaneous uses of library resources. Mathematical techniques can then be employed to ensure the resulting data sets do not identify individual patrons.

The resulting data set will be useful to the library in making decisions based upon past data and justifying their existence to funding agencies. Consortial library groups can share data to aid in purchasing decisions. In addition, libraries can work with researchers to supply de-identified data about library use to help improve the

science of librarianship. Most importantly, if an outside agency demands parts of this database, the activities of individual users will be protected.

References

American Library Association. (2004a). *Privacy Tool Kit: 1. Introduction*. Retrieved September 22, 2004, from http://www.ala.org/ala/oif/iftoolkits/toolkitsprivacy/introduction/introduction.htm

American Library Association. (2004b). *State Privacy Laws Regarding Library Records*. Retrieved September 23, 2004, from http://www.ala.org/Template.cfm?Section=stateifcinaction&Template=/ContentManagement/ContentDisplay.cfm&ContentID=14773

Burkhardt, S. (1992). The effect on the content of mental health records when psychiatric patients are permitted access pursuant to 'patient access' laws. *Dissertation Abstracts International*, 53(6-B), 3149.

Carman, D.M. (1997) Balancing patient confidentiality and release of information. *Bulletin of the American Society for Information Science*, 23(3). 16-17.

Center for Disease Control and Prevention. (2003). *Appendix A: Research Exempt under 45CFR46.101B*. Retrieved December 3, 2004, from http://www.cdc.gov/epo/ads/section-iif.htm

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M. (2004). Tools for Privacy Preserving Distributed Data Mining *ACM SIGKDD Explorations 4*(2). 1-7.

Definitions, 42 U.S.C. § 1320d(6) (2002).

Gillette, B. (2001). To err is human--so keep patient records in an online central database. *Managed Healthcare Executive, 11*(7), 32-33.

Gleason, O.C., and Yates, W.R. (2004). Suicide attempt due to a misunderstood HIPAA notice. *American Journal of Psychiatry, 161*(2), 374.

Golodetz, A., Ruess, J., Milhous, R.L. et al. (1976). The right to know: Giving the patient his medical record. *Archives of Physical Medicine and Rehabilitation, 57*(2), 78-81.

Goodwin, L.K., and Prather, J.C. (2002). Protecting patient privacy in clinical data mining. *Journal of Healthcare Information Management, 16*(4), 62-67.

Gutheil, T. G., & Hilliard, J. T. (2001). 'Don't write me down': Legal, clinical, and risk-management aspects of patients' requests that therapists not keep notes or records. *American Journal of Psychotherapy, 55*(2), 157-165.

Holland, J. (2004). *House may revive parts of Patriot Act II*. Retrieved September 22, 2004, from http://www.guardian.co.uk/worldlatest/story/0,1280,-4508884,00.html

Johnson, S.B, and Friedman, C. (1996). Integrating data from natural language processing into a clinical information system. *Proceedings of the AMIA Fall Symposium*, 537-541.

Jones, R.B. & Hedley, A.J. (1987), Patient-held records: censoring of information by doctors. *Journal of the Royal College of Physicians of London, 21*(1), 35-8.

Kennedy, B. M. (1989). Confidentiality of library records: A survey of problems, policies, and laws. *Law Library Journal, 81*(4), 733-767.

Li, J., & Shaw, M. (2004). Protection of health information in data mining. *International Journal of Healthcare Technology and Management, 6*(2), 210-222.

Lichtblau, E. (2004, July 9, 2004). Effort to Curb Scope of Antiterrorism Law Falls Short. *New York Times,* p. 16.

Lindberg, D. A. (1968). Computers in clinical medical education. In *Conference on the Use of Computers in Medical Education.*. Oklahoma City, OK: University of Oklahoma Medical Center. 53-56.

Million, A., & Fisher, K. (1986). Library records: A review of confidentiality laws and policies. *Journal of Academic Librarianship*, *11*(6), 346-349.

Murphy, D. (2003, April 7, 2003). Some Librarians Use Shredder to Show Opposition to New F.B.I. Powers. *New York Times,* p. 12.

Neuhaus, P. (2003). Privacy and confidentiality in digital reference. *Reference & User Services Quarterly, 43*(1), 26-36.

Nicholson, S. (2003). Avoiding the Great Data-Wipe of Ought-Three. *American Libraries, 34*(9), 36.

Nicholson, S. (forthcoming). The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services. *Information Processing and Management*. Preprint retrieved June 1, 2005 from http://bibliomining.com/nicholson/nicholsonbibliointro.html.

Smith, G.R. (2001, February 8), AAMC Testimony on the Final HHS Privacy Regulations. [Testimony before the Senate Committee on Health, Education, Labor, and Pensions. Retrieved November 24, 2004 from http://www.aamc.org/advocacy/library/hipaa/testimony/2001/020801.htm (American Association of Medical Colleges, 2001).

Spiegel, A.D, & Springer, C. R. (1997). Babylonian medicine, managed care and Codex Hammurabi, circa 1700 B.C. *Journal of Community Health, 22*(1), 69-89.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics, 25*(2.3), 98-110.

Sweeney, L. (2004, July 13, 2004). *HIPAA De-idenitifcation Strategies for Hospitals.* Paper presented at the Presentation at the Easing the Burden on Research: Practical Strategies for De-Identifying Patient Data for Research and E-Health Teleconference.

U.S. Health and Human Services. (2002). Standards for Privacy of Individually Identifiable Health Information: Other Requirements Relating to Uses and Disclosures of Protected Health Information, 45 CFR Parts 160 and 164. *Federal Register, 67*(157).