Sweeney, L. *Risk Assessments of Personal Identification Technologies for Domestic Violence Homeless Shelters*. Carnegie Mellon University, School of Computer Science. Technical Report CMU-ISRI-05-133. Pittsburgh: November 2005.

Risk Assessments of Personal Identification Technologies for Domestic Violence Homeless Shelters

Latanya Sweeney, PhD

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890
latanya@privacy.cs.cmu.edu

November 2005

Abstract

In an attempt to perform a national unduplicated accounting of visit patterns across domestic violence homeless shelters, while respecting the confidentiality of the clients who are the subjects of that accounting, the United States Department of Housing and Urban Development ("HUD") has sponsored locally administered Homeless Management Information Systems ("HMIS"). These are computerized data collection and processing systems designed to capture person-specific information over time from homeless persons being serviced by local shelters. In order to maintain client safety and to insure high degrees of compliance, HUD agreed that the name and Social Security number of each client of a domestic violence homeless shelter are not to be forwarded to HMIS. Instead, a newly created identifier termed a "unique identification number ("UID") can be used. A question posed is, "*how do shelters construct UIDs with minimal risk of re-identification while still achieving an accurate unduplicated accounting*?"

The work reported herein provides a framework for reasoning about and assessing proposed technical solutions that may answer this question. Eight categories of technologies (encoding, hashing, encryption, scan cards/RFID, biometrics, consent, inconsistent hash, and distributed query) are examined and a set of recommendations provided. Results suggest that inconsistent hashing, distributed query and (regular) hashing may be easier to bundle with policies and best practices to create an effective solution. Scan cards, encryption, and biometrics create new kinds of risks to consider. Consent and encoding are technically the simplest to implement but harbor serious dangers that are difficult for any particular implementation to overcome. Biometrics is the only technology that authenticates clients; all the other technologies tend to rely on non-verified information from clients. While significant differences and trade-offs exist in the use of these technologies, there is no magic technology as much as practices that must be bundled with any chosen technology in order to demonstrate minimal risk of client re-identification and maximum correctness in computing an unduplicated accounting.

Keywords: *identity management, personally identifying information, unique identifier*

Table of Contents

Index of Figures

1. Executive Summary

In an attempt to perform a national unduplicated accounting of visit patterns across domestic violence homeless shelters, while respecting the confidentiality of the clients who are the subjects of that accounting, the United States Department of Housing and Urban Development ("HUD") has modified the kind of information it recommends these shelters share with HUD-sponsored locally administered Homeless Management Information Systems ("HMIS") [1]. A HMIS is a computerized data collection and processing system designed to capture person-specific information over time from homeless persons being serviced by any homeless program, including domestic violence homeless shelters. Information gathered from all homeless service programs that are geographically co-located is compiled by a HMIS operated by a planning office (called a "Continuum of Care" or "CoC" in HUD documents) that is local to those programs. Information collected at homeless programs is not directly forwarded to HUD. Instead, de-duplication is to be performed by the local planning office and the resulting de-identified, unduplicated aggregate information is then forwarded to HUD.

Special privacy considerations are given to the clients of domestic violence homeless shelters so that client information provided by a domestic violence homeless shelter to a HMIS cannot be reasonably re-identified to the clients who are the subjects of the shared information. HMIS are to gather information from local domestic violence homeless shelters in such a way that client confidentiality is maintained yet an accurate unduplicated accounting of visit patterns can still be achieved across homeless programs by planning offices. The overarching question posed is, "*how do domestic violence homeless shelters help achieve an overall accurate unduplicated accounting across homeless programs with minimal risk of re-identification to their clients*?"

HUD recognized that HMIS must accept less identifiable information on domestic violence homeless clients in order to maintain client safety and to insure high degrees of compliance. Along these lines, HUD agreed that the name and Social Security number of each client of a domestic violence homeless shelter is not to be forwarded to HMIS [2]. In order to associate individual service information across shelters without these explicit identifiers, HUD introduced the notion of a newly created unique identifier ("UID") to uniquely identify clients. But then, "*how do shelters construct UIDs with minimal risk of re-identification while still achieving an accurate unduplicated accounting*?"

Constructing UIDs with minimal risk of re-identification while still achieving these kinds of accounting tasks has utility beyond HMIS. The ability to track individual utilization while maintaining privacy is important to numerous public health and administrative simplification efforts at the local, state and federal government levels. Examples include gathering incidence and prevalence information (e.g., the numbers of people having disease X), computing utilization across combinations of social services (e.g., how many of those receiving service Y also receive service Z), and evaluating social service performance over time (e.g., how many of those who received service Y at one time still used service Z at a later time). The goal in these kinds of efforts is for policymakers to be able to measure counts, utilizations, and outcomes without necessarily knowing the identities of the people who are the subjects of these measures.

This paper reports on an examination of 8 categories of technologies for constructing UIDs and reports general findings in terms of the utility and privacy protection afforded by each. The 8 categories of technologies are listed in Figure 1 and they include the kinds of technologies that have either been considered or are being considered in a variety of service tracking scenarios. A framework is provided for reasoning about proposed technical solutions for generate and match UIDs. (For more information, see section 5.)

Encoding
Hashing (regular)
Encryption
Scan Cards / RFID
Biometrics
Consent
Inconsistent Hashing
Distributed Query

Figure 1. Categories of technologies for constructing UIDs.

While many factors must be considered in determining which technology is most appropriate for shelters and a planning office in a particular region, the technology assessments provided herein suggest that inconsistent hashing, (regular) hashing, and distributed query may be easier to bundle with policies and best practices to get an effective solution. Scan cards, encryption, and biometrics create new kinds of risks to consider. Consent and encoding are technically the simplest to implement but may harbor serious dangers that are difficult for any particular implementation to overcome. Biometrics is the only technology that authenticates clients; while the other technologies tend to rely on non-verified information from clients. (For more information, see section 6.)

Both general and technology-specific recommendations appear below. These recommendations concern information collected from clients of domestic violence homeless shelters and are not necessarily intended to be more generally applied to other homeless populations whose information may also be captured in HMIS.

1.1 General recommendations

Recommendation #1: Care to maintain an unduplicated count across planning office jurisdictions is needed to account for situations in which a single client has information appearing in the data of more than one planning office. Coordination of privacy protection schemes is necessary across planning offices that service a geographical region in which shelters within the region report to different planning offices but service some of the same clients. (For more information, see section 3.3.)

Recommendation #2: Given a "Proposed Solution" (i.e., a UID technology bundled with policies and practices for the construction, maintenance and use of a UID technology for clients of domestic violence homeless shelters), a person skilled in statistical, computational and/or legal principles, as appropriate, should certify in writing that the Proposed Solution has a minimal risk of re-identification when the solution is considered with other publicly and readily available information and techniques. Such writing should address vulnerabilities for inappropriate re-identifications by various categories of insiders. This is termed a "compliance statement" and should be made available for inspection. (For more information, see section 5.5.)

Recommendation #3: Given a Proposed Solution, a person skilled in statistical and/or computational principles, as appropriate, should certify in writing that the Proposed Solution provides a reasonably accurate unduplicated accounting of client visit patterns to shelters within the regional setting it is to be deployed. Such writing should include possible false match and missed match rates. This statement is termed a "warranty" and should be made available for inspection. (For more information, see section 5.5.)

Recommendation #4: The fields *date of birth* and *ZIP code of last residence*, which are among the data elements HUD recommends HMIS collect, should contain information less specific than the month, day, and year of birth and all 5 digits of the ZIP (or postal) code. (For more information, see section 4.3.)

Recommendation #5: UID values assigned to Clients of domestic violence homeless shelters should not be used (i.e., stored or referenced) by any non-HMIS program to which the Clients may participate. (For more information, see section 4)

Recommendation #6: A shelter should consider releasing information to the planning office on a client some time after the client has left the shelter. (For more information, see section 3.4.2.)

Recommendation #7: Shelters and planning offices should train personnel on the responsibilities and accepted practices for collecting, storing and sharing client information. (For more information, see section 3.4.2.)

Recommendation #8: Shelters and planning offices are already required to issue and post privacy notices to clients about the data collection, sharing, and linking practices of the shelters and planning offices in which the client's data will be part [1]. Beyond the role this requirement plays as a Fair Information Practice, this requirement is also important to help ensure the integrity of the information a client provides in forming the client's UID. (For more information, see section 4.)

1.2 UID technology-specific recommendations

Recommendation #9: If the technology for constructing UIDs uses non-verifiable information from the client, then instruments that instill client trust in the overall system should be deployed; otherwise, the UID should use verifiable source input from clients.

Recommendation #10: If the technology for constructing UIDs uses encoding, then the fields {*part of name, date of birth, gender, ZIP code of last residence* } should be avoided as the basis for encoding the UID.

Recommendation #11: If the technology for constructing UIDs involves encryption or hashing, then "strong" cryptographic methods should be used and the description of the method should be included in the warranty or compliance statement.

Recommendation #12: If the technology for constructing UIDs involves encryption or hashing, then accompanying practice should control access to and document an audit trail of specific uses of the encryption/hashing function. A description of these practices related to the capture and auditing of uses of the encryption/hashing function should be included in the warranty or compliance statement.

Recommendation #13: If the technology for constructing UIDs involves scan cards, then accompanying practices are needed to avoid issuing multiple cards to the same client and to prevent card sharing and swapping among clients. A description of practices related to avoiding these unwanted activities should be included in the warranty or compliance statement.

Recommendation #14: In cases where consistent UIDs are assigned to Clients over time, once Planning Offices link and de-duplicate Client visits, stored copies of the linked information should have all UIDs removed.

Recommendation #15: HUD should consider selecting one or two of these technologies (e.g. inconsistent hashing and/or regular hashing) and providing an example of a complete UID solution as an example for shelters and planning offices. A complete UID solution consists of technology instantiations and accompanying model policies and practices and warranty and compliance statements (as previously discussed).

2. Introduction

The number of homeless Americans appears to have dramatically increased in recent years, but no one actually knows the current number of homeless persons and counting them may not be as easy as it may first seem. At stake are resource allocations, program evaluations, and billions of dollars necessary for managing and resolving what may be one of the most serious social and economic crises of our time.

2.1 Examples of increases in the numbers of homeless Americans

Numerous anecdotal examples illustrate that the numbers of homeless Americans seem to be increasing over time and that related spending has reached dramatic heights.

HUD's Emergency Shelter Grants program funds resources for basic shelter and essential supportive services by awarding grants to state governments, large cities, urban counties, and U.S. territories. These awards totaled $10 million in 1987 and had grown to $115 million by 1997, with continued increases thereafter [3].

A report from the Northeast Ohio Coalition for the Homeless in 2005 that addressed the overflow of shelters in Cleveland Ohio, asserted that shelter costs in 2004 was 5.6 times the cost 10 years earlier for men and 9.4 times the cost 10 years earlier for women [4]. They predicted further increases over the next 10 years due to increased demand and warned that at the current rate of increased demand, county and city public sector funding will be exhausted.

A 2001 study of 27 U.S. cities reported that 37% of all requests for emergency shelters and 52% of all requests for emergency shelters from families were unmet in that year due to a lack of resources [5].

In April 2002, over 33,000 homeless people were provided emergency shelter each night by the New York City Department of Homeless Services [6]. This was the highest number they had recorded, and the cost of homelessness rose to record heights as well. According to a report by the New York City Independent Budget Office, New York City agencies spent almost $1 billion on homelessness in Fiscal Year 2001 [7].

Congress appropriated over $1 billion dollars to homeless assistance programs in the Fiscal Year 2002 HUD Appropriations Act [8].

2.2 Congress directs HUD to report on homeless service utilization

In response to noted increases in homelessness, which seem to reflect a growing social and economic crisis, Congress deemed it critical for the United States Department of Housing and Urban Development ("HUD") to work with local jurisdictions to develop an unduplicated accounting of homeless service utilization. Congress directed HUD to perform an unduplicated count[1] of homeless persons sufficient to provide annual reports to the Committee on Appropriations documenting the demographics and utilization patterns of homeless persons based on collected count data [8][10].

In the Fiscal Year 2002 HUD Appropriations Act, Congress allocated $2 million dollars specifically to continue work on a homeless data collection and analysis project that had begun the year before in the Fiscal Year 2001 HUD Appropriations Act [9]. This project seeks to document the demographics of homelessness, identify patterns in service utilization, and record the effectiveness of assistance programs. The work reported herein addresses *ways to* achieve (and not to achieve) the unduplicated accounting within this data collection and analysis project.

2.3 Earlier attempts to count the number of homeless Americans

There have been previous attempts to count the number of homeless Americans by counting the number of people who are in shelters or on the streets at a given point in time.

On March 20, 1990, federal employees of the U.S. Bureau of the Census, in satisfaction of their duties as set forth in the U.S. Constitution, attempted to determine the exact number of Americans in the U.S. population by physically verifying the existence of each person, including an attempt to count every homeless person and gather related demographics [11]. Under this effort, termed Shelter-and Street night, thousands of federal employees visited homeless shelters, inexpensive hotels, all-night eating establishments, bus stations, street corners and various urban places identified by local jurisdictions as places where homeless people are likely to be found. Employees were instructed not to ask who was homeless and not to awaken any persons found sleeping. Instead, they were told to count all visible persons (including children) found in these places and record demographics as either provided or as they appeared to the census taker. These efforts were able to add 240,140 homeless people to the official census count.

A more comprehensive estimate was provided by the Urban Institute using the 1996 National Survey of Homeless Assistance Providers and Clients [12]. The survey was designed to provide information about the providers of homeless assistance and the characteristics of homeless persons who used services by sampling 76 metropolitan and non-metropolitan areas, including small cities and rural areas at two points in the year. On a given night in February, 842,000 in 637,000 households were found homeless. On a given night in October, 444,000 people in 346,000 households were found homeless. Converting these point counts into a national annual projection, researchers at the Urban Institute estimated that between 2.3 and 3.5 million people were homeless in that year [13].

[1] The term “unduplicated count” is misleading. In ordinary language it tends to imply that the answer is a single number. In terms of the Congressional directive, it is actually an unduplicated accounting of shelter visits –i.e., the distinct visit patterns of each client across shelters.

2.4 Limits of point-in-time counts

Point-in-time studies, like those mentioned above, give a limited static picture by only counting those who are homeless at specific places during a narrow slice of time. No explicitly-identifying person-specific information is necessarily collected, so double-counting can occur when clients use more than one service (i.e., appear at more than one point) during the capture period. An example is a client receiving meals at one facility and lodging at another during the same night; such a person may be counted once, twice, or not at all. Seasonal and climate variation may be missed altogether. Important differences in client circumstances may not be captured. For example, the frequency and lengths of time in which particular clients are in and out of homelessness is typically not captured by a point-in-time count. Prolonged unemployment, sudden loss of a job, lack of affordable housing, and domestic violence contribute to episodes of homelessness, while severe mental illness and addiction disorders often account for chronic homelessness. For these reasons, point-in-time studies may misrepresent the magnitude and nature of homelessness.

3. The HMIS Approach

In response to Congress' directive, HUD elected not to use the traditional point-in-time approach, but opted instead to develop and introduce national data and technical standards for locally situated computer systems that collect, process and share details of each client's utilization of service related to homelessness. These are termed Homeless Management Information Systems ("HMIS"), which are described in terms of the parties to and from which data flows and the data elements that constitute information flow. At this writing, the initial data elements had already been altered to protect the privacy of domestic violence shelter clients from intimate abusers, but other privacy concerns remain which are addressed herein.

3.1 Data flow in HMIS

Using HMIS, information does not flow directly from a homeless service provider to HUD. Instead, HMIS introduces an intermediary (termed a "planning office" in this writing and referred to as a "continuum of care" or "CoC" in HUD documents)[2] that is local to a group of homeless service providers (e.g., shelters). The purpose of the planning office is to establish an HMIS for a group of service providers. Information flows from clients to service providers, who in turn, provide visit information to the local planning office. Because clients are expected to consume services from multiple providers, the planning office can then associate visits across providers over time to provide an unduplicated accounting to HUD.

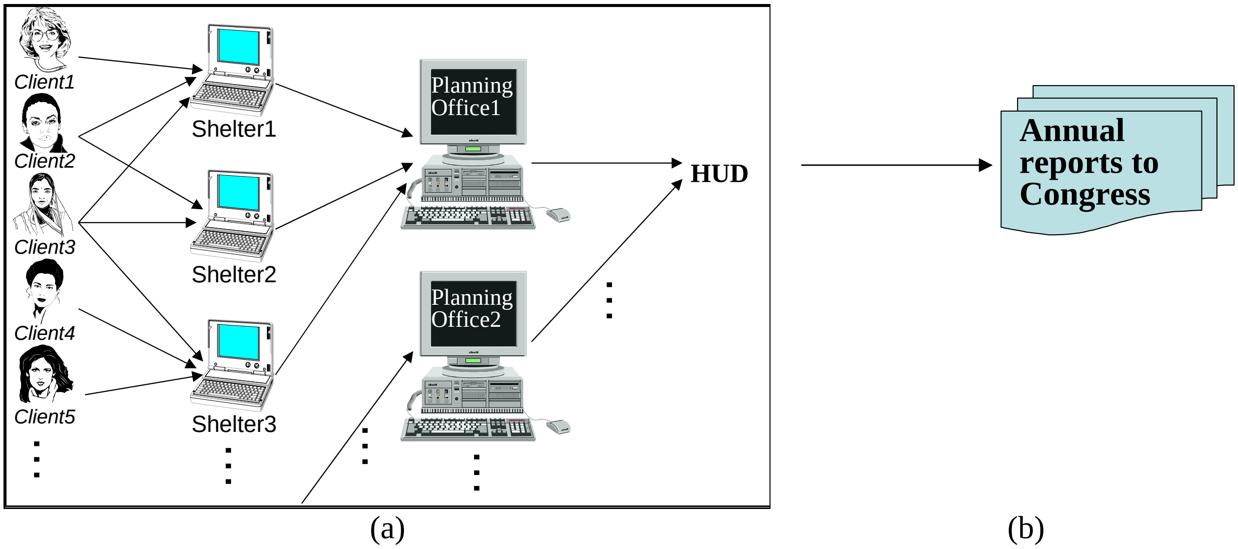


Figure 2. Flow of information from Clients to HUD: (a) Clients give information to Shelters, which report information to Planning Offices, which in turn provide non-identifiable unduplicated count information to HUD, (b) which becomes the source data for annual reports to Congress.

[2] The purpose of a planning office is broader than HMIS, but for the purposes of this writing, planning offices are examined narrowly in their HMIS context.

While HMIS includes services beyond providing shelter, the work reported herein is specifically focused on clients who visit domestic violence shelters. Hereinafter, unless otherwise noted, references to "Shelters" are exclusively domestic violence shelters and may generally apply to a suite of homeless service providers. Similarly, references to "Clients" are homeless persons serviced by Shelters and to "Planning Offices" are the CoCs servicing Shelters.

Figure 2(a) depicts the flow of information from Clients to Shelters through Planning Offices to HUD. A Client visits one or more Shelters. Each Shelter provides information to one Planning Office. HUD uses non-identifiable information from Planning Offices to provide annual reports on the utilization patterns of homeless people to Congress; see Figure 2(b).

3.2 Comparing HMIS to point-in-time approaches

Because Client demographics and specific visit data are captured on each visit, many of the shortcomings found with point-in-time studies may potentially be resolved by the HMIS approach.[3]

For example, HMIS seeks to record sufficient information to allow the same Client to be identified on subsequent visits to the same or other Shelters, thereby thwarting the potential for double counting. Associated date and length of stay information may be recorded to identify seasonal, climate and temporal visit patterns. Recording the reason given for each visit may help identify utilization characteristics related to different kinds of homelessness, and tracking Clients across the same and different shelters can provide recurrence and duration rates.

3.3 Concern about selecting planning offices

It is understood that a Client may visit one or more Shelters, which is why de-duplication across Shelters is necessary, but if the same Client visits Shelters reporting to different Planning Offices, then the de-duplication effort can be thwarted.

For example, consider the case of Boston and Cambridge Massachusetts. These are two cities between which people regularly walk and ride multiple times a day. If each of these cities has their own Planning Office, then a single Client being serviced by a Shelter in Cambridge and by another Shelter in Boston, would be counted twice –once by the Planning Office for Cambridge and again by the Planning Office for Boston. Similar situations can exist with Planning Offices located in close proximity to one another irregardless of city, county, or state lines. To combat this problem, the following recommendation is made.

[3] One shortcoming of both the survey used by the Urban Institute and the HMIS approach is the sole reliance on service providers. Homeless people who are not using shelters or covered services are not captured. These include homeless people who may live in automobiles, make-shift housing (such as cardboard boxes or tents), or doubled-up situations.

<u>Recommendation:</u> *Care to maintain an unduplicated count across Planning Office jurisdictions is needed to account for situations in which a single Client has information appearing in the data of more than one Planning Office. Coordination of privacy protection schemes is necessary across Planning Offices that service a geographical region in which Shelters within the region report to different Planning Offices but service some of the same Clients.*

At present, HUD funds about 400 Planning Offices. This funding extends beyond HMIS to the coordination and funding of homeless services at the local level. A Planning Office defines its own geographical service area and competes to receive HUD funds for homeless programs. Because geographical service areas are not dictated by HUD, cooperative coordination of privacy protection schemes in overlapping areas allows a Client's utilization pattern to be determined without compromising the identity of the Client.

3.4 The intimate stalker threat

Almost as soon as the first HMIS standards were announced, privacy concerns emerged over the need for protections for Clients of domestic violence shelters [14][15]. Tracking victims of intimate domestic violence who seek refuge in Shelters may be necessary for HMIS accounting, but many feared HMIS data collection and sharing might become a vehicle to further endanger victims whose information would appear in HMIS data because of attempts to remove themselves from harmful situations.

3.4.1 Concerns are real

Domestic violence shelters have historically had to protect Clients from intimate and aggressive abusers and concerns are well founded. Over 31% of all women[4] murdered in the United States are murdered by husbands, boyfriends , or exes – the majority killed after attempting to leave an abusive relationship [16][17]. The National Institute of Justice estimated that 73% of domestic violent assaults go unreported largely because of women's lack of faith in the system [17].

Personal stories are quite chilling. As an example, consider a case from Los Angeles, California [18]. In 2001, a woman's husband was unemployed and had been drinking heavily. When she refused to have sex with him, he attacked her, prevented her from calling for help, and held her captive in her home. Various other incidents recurred. Eventually she was able to get a spot in a family shelter for herself and her two children. After leaving the shelter, the husband quickly tracked her down and strangled her to death with a belt.

[4] While the wording used has a bias that women are victims and men are abusers, it is important to note that men are also victims and that abusers can be male or female.

3.4.2 The intimate stalker

The "intimate stalker" (an name given in this writing to an intimate abuser who stalks a Client) challenges computer systems that record and share Client visit information in several ways. First, the intimate stalker typically has knowledge of various personal facts about the Client that may be recorded in collected data by the Shelter in which the victim resides. For example, an intimate stalker is likely to know the victim's name, date of birth and Social Security number, which may not be readily known by the general population. Second, the intimate stalker tends to be highly motivated to locate a targeted Client. For example, repeated violations of court orders and police reports of escalating incidents of death threats, stalking and harassment are common. Finally, an intimate stalker may use insider access (either his own or compromising an insider who has access to the data) to gain location information on a targeted Client. For example, an intimate stalker may persuade a family member or friend to assist in revealing a Client's Shelter location by expressing a desire to reconcile for the sake of the children or because situations (such as obtaining a new job) have changed.

No one solution addresses all these concerns, however some recommendations can be made immediately and others will be made in subsequent sections.

One recommendation, as stated below, is to thwart the intimate stalker's ability to locate the Client by making sure visit information shared with the Planning Office is no longer current. This protection is not a first line of defense against an intimate stalker and should not be the only protective action taken. It merely offers supplementary protection. Stronger protections, which will be examined later in this writing, guarantee that the location of any Shelter in which the Client has historically visited cannot be learned by the intimate stalker. Stronger protection is important because some Clients tend to re-visit the same Shelters and an intimate stalker's knowledge of a historic visit can pose future problems.

Recommendation: *A Shelter should consider releasing information to the Planning Office on a Client some time after the Client has left the shelter.*

Another recommendation, as stated below, is aimed at helping thwart the stalker's ability to recruit or compromise those with insider access to Client information. This protection only provides supplemental protection. Stronger protections, which are examined later in this writing, guarantee that the Client's information cannot be found in shared or stored information.

Recommendation: *Shelters and planning offices should train personnel on the responsibilities and accepted practices for collecting, storing and sharing client information.*

3.4.3 HUD's removal of explicit identifiers

A privacy protective action taken by HUD involved changing HMIS standards to allow Shelters to provide Client information without making reference to any client explicit identifiers (e.g., name and Social Security number). Instead, an approved proxy, coded, encrypted, hashed, or other alternative (termed a "unique identification number" or "UID") is to be used by Shelters to provide client information to Planning Offices, provided each Planning Office has the ability to recognize the occurrence of the same clients in the same and different shelters (including shelters that are not domestic violence provider shelters) over time.

3.5 Details of HUD's data elements

HUD requires certain data elements be sent from Shelters to Planning Offices. The data elements that HUD requires Shelters to provide to Planning Offices are termed the "Universal Data Elements," and consists of a record for each Client's visit to a Shelter and includes the Client's UID. The original data elements were modified to use UIDs, in lieu of explicit identifiers, as shown in Figure 4. Shelters participating in HMIS must collect the Universal Data Elements and share them with the Planning Office at least once a year in a privacy-preserving manner that includes replacing name and Social Security number with UIDs.

"Program-Specific Data Elements" are additional fields of information that Shelters may be required to provide on each Client visit. All McKinney Vento funded Shelters that are required to complete an Annual Progress Report are required to collect and share certain Program-Specific Data Elements with the Planning Office[5]. Figure 5 lists the Program-Specific Data Elements and identifies which data elements are required for the Annual Progress Report.

HUD places no further restriction on the information collected between Clients and Shelters. Beyond the noted data elements, Shelters may elect to collect additional information for their own purposes. A Unique Person Identification Number ("PIN") is included among the Universal Data Elements. This field allows a Shelter to store its internal reference number for a Client. However, care must be taken to share only when the PIN is sufficiently privacy-protecting, as noted in the following recommendation.

<u>Recommendation:</u> *A Shelter should only use a PIN that does not include the Client's name, Social Security number, or other characteristic that may be used alone or in combination with other data elements to re-identify a Client. If a characteristic is not allowed as part of a UID, then it should not be used as a PIN, because PINS must satisfy the same privacy requirements as UIDs.*

[5] See http://www.hud.gov/offices/cpd/homeless/apr.doc.

In summary, Figure 3 shows the flow of information from a Client through the Planning Office to HUD using the Universal and the Program-Specific Data Elements.

Hereafter, the information transmitted from a Shelter to a Planning Office is collectively termed the "Dataset" in this writing and refers to the Universal Data Elements unless otherwise stated.

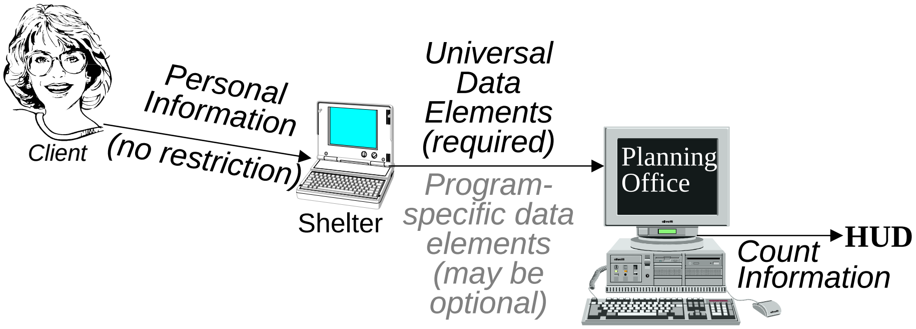
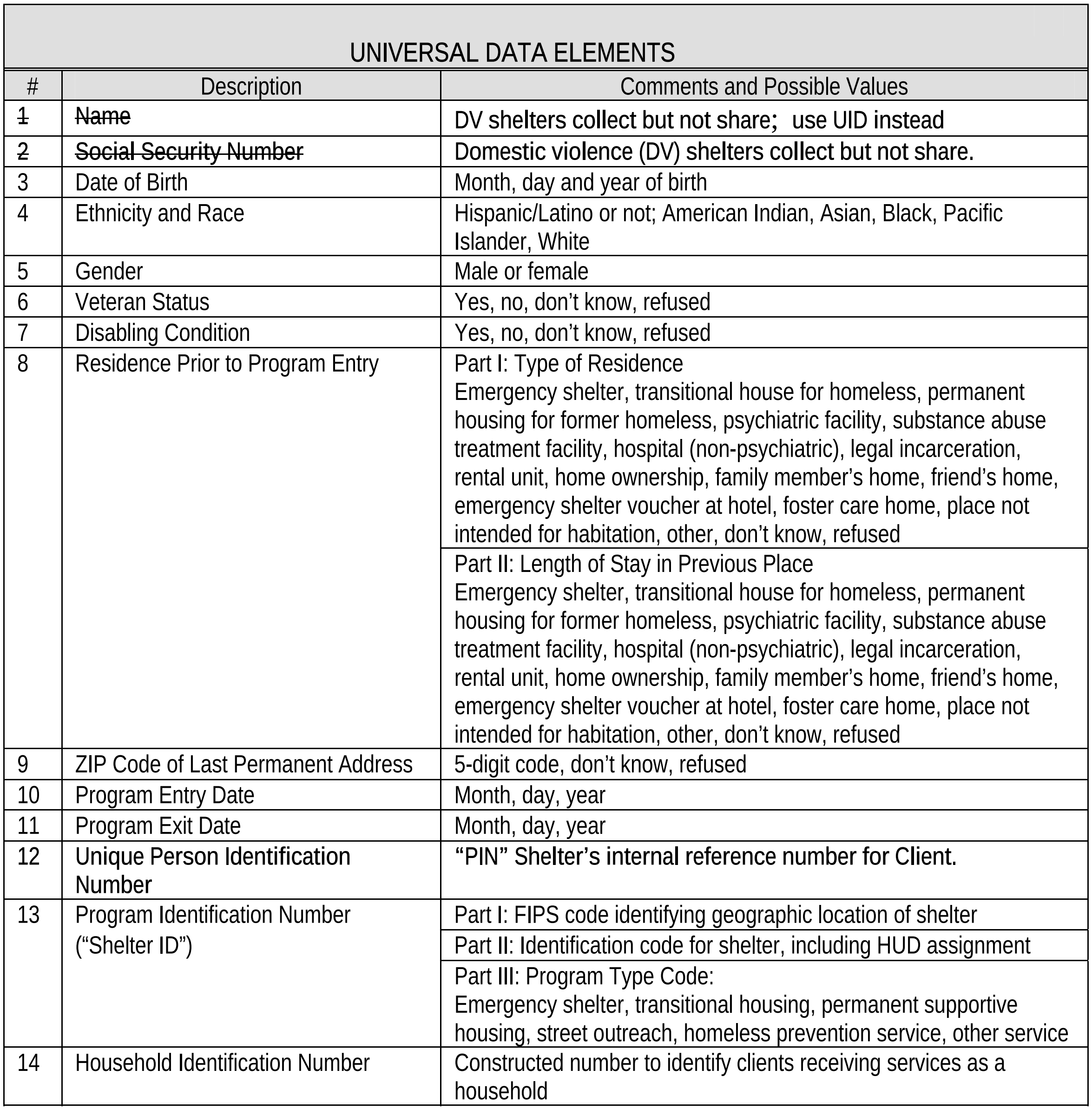


Figure 3. Flow of information: Client gives explicit personally identifying information to the Shelter, which provides the Universal Data Elements and Program-Specific Data Elements to the Planning Office, which in turn provides to HUD, non-identifiable, unduplicated count information of Client visits across all Shelters in the Planning Office's region.



UNIVERSAL DATA ELEMENTS		
#	Description	Comments and Possible Values
~~1~~	~~Name~~	DV shelters collect but not share; use UID instead
~~2~~	~~Social Security Number~~	Domestic violence (DV) shelters collect but not share.
3	Date of Birth	Month, day and year of birth
4	Ethnicity and Race	Hispanic/Latino or not; American Indian, Asian, Black, Pacific Islander, White
5	Gender	Male or female
6	Veteran Status	Yes, no, don't know, refused
7	Disabling Condition	Yes, no, don't know, refused
8	Residence Prior to Program Entry	Part I: Type of Residence Emergency shelter, transitional house for homeless, permanent housing for former homeless, psychiatric facility, substance abuse treatment facility, hospital (non-psychiatric), legal incarceration, rental unit, home ownership, family member's home, friend's home, emergency shelter voucher at hotel, foster care home, place not intended for habitation, other, don't know, refused
		Part II: Length of Stay in Previous Place Emergency shelter, transitional house for homeless, permanent housing for former homeless, psychiatric facility, substance abuse treatment facility, hospital (non-psychiatric), legal incarceration, rental unit, home ownership, family member's home, friend's home, emergency shelter voucher at hotel, foster care home, place not intended for habitation, other, don't know, refused
9	ZIP Code of Last Permanent Address	5-digit code, don't know, refused
10	Program Entry Date	Month, day, year
11	Program Exit Date	Month, day, year
12	Unique Person Identification Number	"PIN" Shelter's internal reference number for Client.
13	Program Identification Number ("Shelter ID")	Part I: FIPS code identifying geographic location of shelter
		Part II: Identification code for shelter, including HUD assignment
		Part III: Program Type Code: Emergency shelter, transitional housing, permanent supportive housing, street outreach, homeless prevention service, other service
14	Household Identification Number	Constructed number to identify clients receiving services as a household

Figure 4. HMIS Universal Data Elements includes the generated unique identification number (UID).

PROGRAM-SPECIFIC DATA ELEMENTS			
#	Description	Need for Annual Progress Report	Comments and Possible Values
1	Income and Sources	Yes	Part I: Source of Income Earned income, unemployment insurance, supplemental security income (SSI), Social Security disability (SSDI), veteran's disability, private disability insurance, worker's compensation, temporary assistance for needy families (TANF), general assistance program (GA), Social Security retirement income, veteran's pension, former job pension, child support, alimony, other source, no financial resources.
			Part II: Total monthly income in dollars
2	Non-cash benefits	Yes	Food stamps, MEDICAID health insurance, MEDICARE health insurance, state children's health insurance, women-infants-children program (WIC), veteran's medical services (VA), TANF child care, TANF transportation services, other TANF services, public housing, other source.
3	Physical Disability	Yes	No, yes
4	Developmental Disability	Yes	No, yes
5	HIV/AIDS	Yes	No, yes
6	Mental Health	Yes	Part I: Mental health problem – no, yes
			Part II: Expected indefinite duration – no, yes
7	Substance Abuse	Yes	Part I: Problem: none, alcohol, drug, dully diagnosed
			Part II: Expected indefinite duration – no, yes
8	Domestic Violence	Yes	Part I: Experience –no, yes
			Part II: Time of experience past 3 months, 3-6 months ago, 6 to 12 months ago, more than a year ago, don't know, refused.
9	Services Received	Yes	Part I: Date of service – month, day, year
			Part II: Type of Service Food, housing, material goods, financial aid, transportation, consumer assistance, legal services, education, health care, HIV/AIDS services, mental health care, substance abuse services, employment, case management, day care, personal enrichment, outreach, other.
10	Destination	Yes	Part I: Destination Emergency shelter, transitional housing, permanent housing for formerly homeless, psychiatric facility, substance abuse treatment center, hospital (non-psychiatric), legal incarceration, rental unit, home own, family home, friend's home, hotel paid by shelter voucher, foster care, place not meant for habitation, other, don't know.
			Part II: Tenure Refused, permanent, transitional, don't know, refused
			Part III: Subsidy Type None, public housing, Section 8, S+C, HOME program, HOPWA program, other housing subsidy, don't know, refused.

PROGRAM-SPECIFIC DATA ELEMENTS			
#	Description	Need for Annual Progress Report	Comments and Possible Values
11	Reasons for Leaving	Yes	Housing opportunity, completed program, non-payment of rent, non-compliance with project, criminal activity, reached maximum time allowed, needs could not be met, disagreement with rules or people, death, disappeared, other
12	Employment	No	Part I: Employed – no, yes
			Part II: If employed, number of hours worked past week
			Part III: If employed, tenure --permanent, temporary, seasonal
			Part IV: If not employed ,looking for work – no, yes
13	Education	No	Part I: In school – no, yes
			Part II: Received vocational training – no, yes
			Part III: Highest Level of School Completed No schooling, nursery school to 4th grade, 5th or 6th grade, 7th or 8th grade, 9th grade, 10th grade, 11th grade, 12th grade with no diploma, high school diploma, GED, post-secondary school.
			Part IV: Post-Secondary Education If high school diploma or equivalent, earned Associated Degree, Bachelor's, Masters, Doctorate, other graduate/professional degree.
14	General Health Status	No	Excellent, very good, good, fair, poor, don't know
15	Pregnancy Status	No	no, yes
16	Veterans Information	No	Part I: Military Service Era Persian Gulf, post Vietnam, Vietnam era, between Korean and Vietnam wars, Korean war, between WWII and Korean war, World War II, between WWI and WWII, World War I.
			Part II: Duration of active duty in months
			Part III: Served in a war zone – no, yes
			Part IV: If served in War Zone, Specify Zone Europe, North Africa, Vietnam, Laos and Cambodia, South China Sea, China-Burma-India, Korea, South Pacific, Persian Gulf, other.
			Part V: If served in war zone, number of months served
			Part VI: Received hostile or friendly fire –no, yes
			Part VII: Branch of the Military Army, Air Force, Navy, Marines, other.
			Part VIII: Discharge Status Honorable, general, medical, bad conduct, dishonorable, other.
17	Children's Education	No	Part I: Current enrollment status – no, yes
			Part II: Name of School (explicitly stated)
			Part III: Type of School – public, parochial-private
			Part IV: Last date of enrollment –month, day, year
			Part V: If not enrolled, Identify Problem Residency requirements, availability of school records, birth certificate, legal guardian requirements, transportation, lack of preschool program, immunization requirements, physical examination requirements, other.

Figure 5. Program Specific Data Elements are supplemental information that may be made available to planning offices.

3.6 The unduplicated accounting

The motivating end product for HMIS data collection and sharing are the annual reports HUD will provide to Congress, which will report on homeless demographics, utilization patterns, and service availability. These reports are termed the "Annual Homeless Assessment Report" ("AHAR"). To produce the AHAR, Planning Offices use HMIS data to provide aggregate count information to HUD.

HUD will provide the first AHAR to Congress in 2006 using HMIS data collected in 2005. An initial draft of the data analysis for the 2006 AHAR shows how HMIS data elements contribute to the AHAR [19]. Basic questions addressed by the AHAR focus on emergency shelters and transitional housing for individuals and for families. Figure 6 has a sample of the kinds of questions answered by the AHAR using HMIS data elements. The sample questions pertain to individuals at emergency shelters, but similar questions exist for transitional housing and for families. Notice that all the data elements are used except UID and PIN (recall name and Social Security number had already been removed). A Planning Office provides HUD with answers to these questions, which are aggregated counts and not the raw data used to compute the counts.

A Planning Office can generate a "De-identified Dataset"[6] to perform the de-duplication and compute the unduplicated count information needed for the AHAR by linking Client demographics to Shelter utilizations using Client UIDs. The resulting data, which does not itself have to further include Client UIDs and PINs, is de-identified.

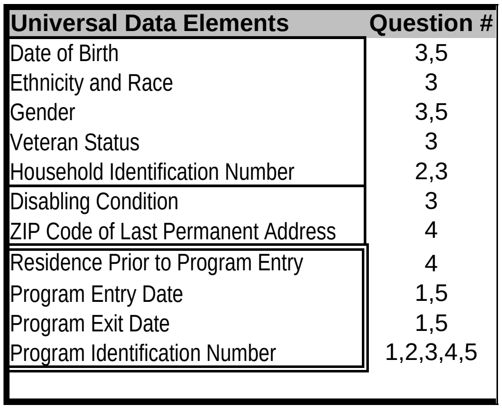
The UID is used to identify data relating to the same Client. Once the visit records are grouped by Client, the UIDs are no longer needed. A sequentially assigned Client number from 1 to the total number of distinct Clients appearing in the dataset can be used to reference Clients in the De-identified Dataset.

PINS are not needed in the De-identified Dataset. If a data problem occurs, the Planning Office has the originally received data for communicating with a Shelter using the Shelter's PIN.

Similar to UIDs, once Clients belonging to the same households are linked together, the Household Identification Number can be replaced with a sequentially assigned number from 1 to the total number of distinct households appearing in the dataset.

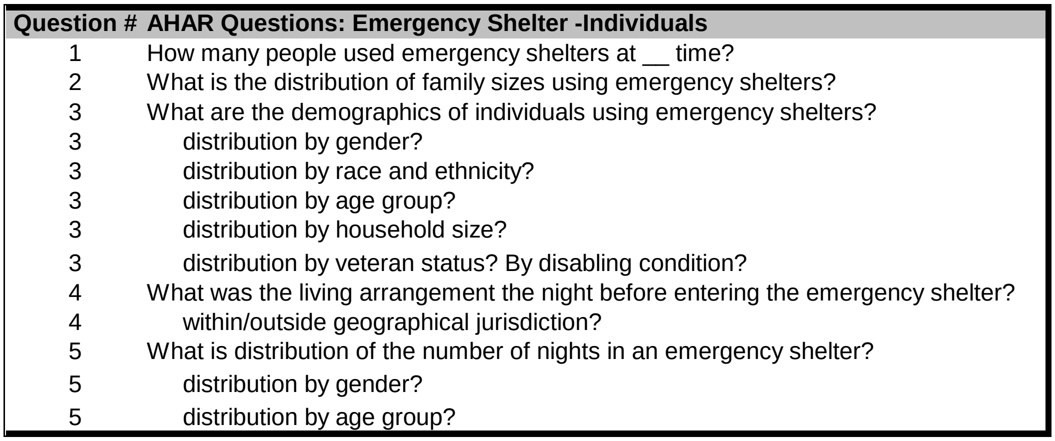
Figure 7 shows an example for a single Client. The Client's utilizations relate to her demographics but not to her explicit identity. Clients belonging to the same household are linked by sharing the same Household Identification number. Figure 8 provides an example of four clients, two of which are in the same household. The de-identified data can be used to compute values necessary to forward to HUD for the AHAR. Removing PINs and replacing UIDs and Household Identification numbers adds privacy protection to the De-identified Dataset, though more privacy protections are needed, as discussed in the remainder of this writing.

[6] While the De-identified Dataset is sufficient for computing the aggregate unduplicated count information that is forwarded to HUD, Planning Offices are not required to use the exact de-identified dataset described above.



Universal Data Elements	Question #
Date of Birth	3,5
Ethnicity and Race	3
Gender	3,5
Veteran Status	3
Household Identification Number	2,3
Disabling Condition	3
ZIP Code of Last Permanent Address	4
Residence Prior to Program Entry	4
Program Entry Date	1,5
Program Exit Date	1,5
Program Identification Number	1,2,3,4,5

(a)



Question #	AHAR Questions: Emergency Shelter -Individuals
1	How many people used emergency shelters at __ time?
2	What is the distribution of family sizes using emergency shelters?
3	What are the demographics of individuals using emergency shelters?
3	distribution by gender?
3	distribution by race and ethnicity?
3	distribution by age group?
3	distribution by household size?
3	distribution by veteran status? By disabling condition?
4	What was the living arrangement the night before entering the emergency shelter?
4	within/outside geographical jurisdiction?
5	What is distribution of the number of nights in an emergency shelter?
5	distribution by gender?
5	distribution by age group?

(b)

Figure 6. Data elements from Figure 4 above (a) associated with sample questions answered by the AHAR (b). Planning Offices provide HUD with aggregated unduplicated count information as answers to the questions.

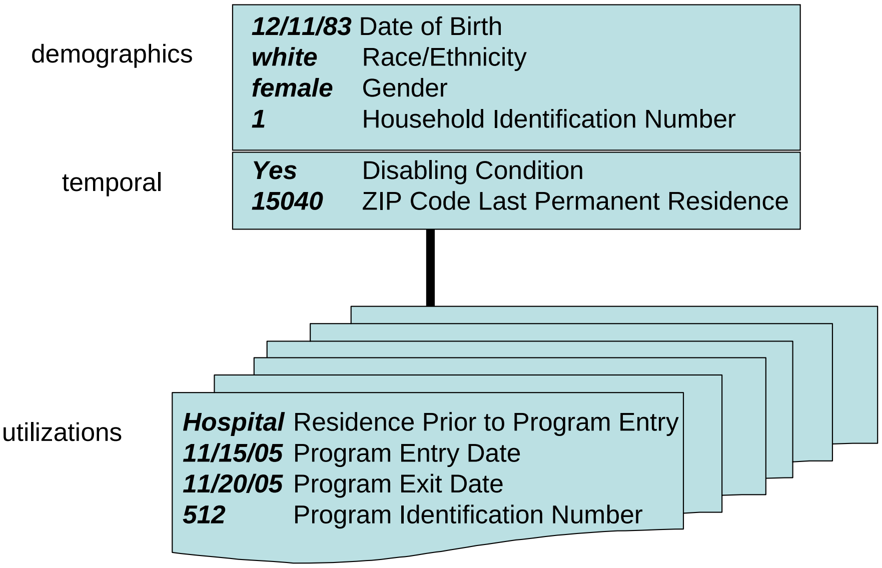

demographics
12/11/83 Date of Birth
white Race/Ethnicity
female Gender
1 Household Identification Number
temporal
Yes Disabling Condition
15040 ZIP Code Last Permanent Residence
utilizations
Hospital Residence Prior to Program Entry
11/15/05 Program Entry Date
11/20/05 Program Exit Date
512 Program Identification Number

Figure 7. De-identified data for a Client includes demographics, some information that may change over time (disabling condition and ZIP of last residence), and program utilizations.

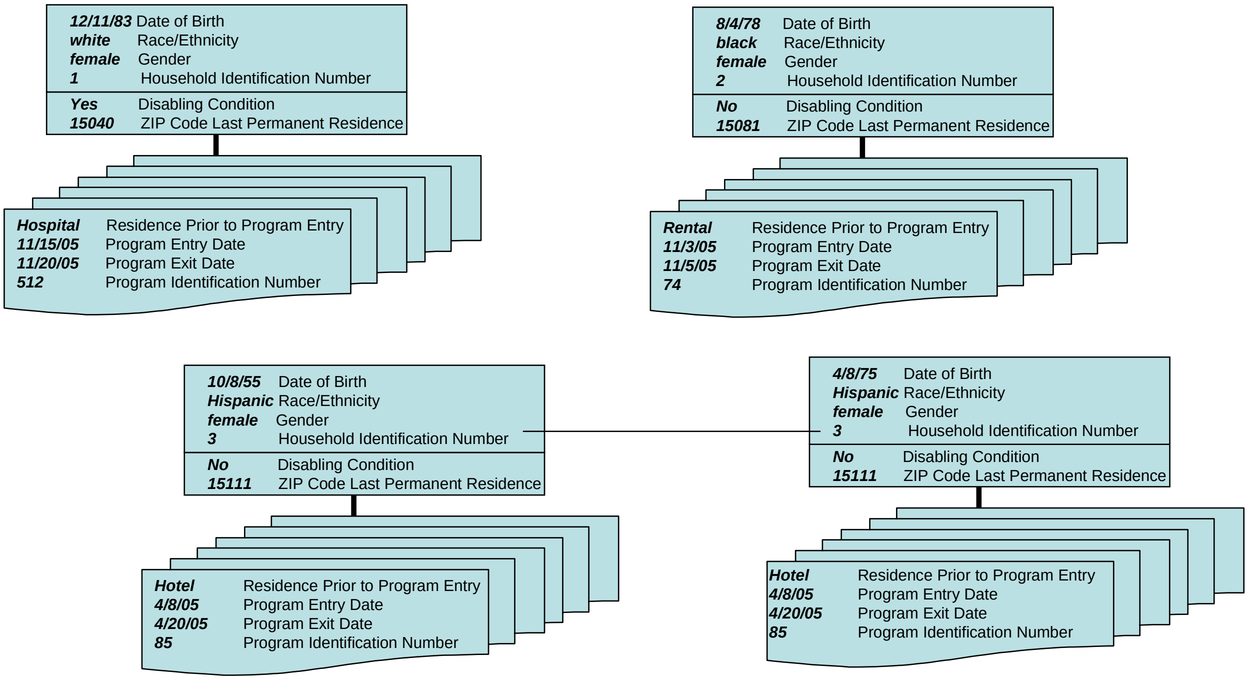

12/11/83 Date of Birth
white Race/Ethnicity
female Gender
1 Household Identification Number
Yes Disabling Condition
15040 ZIP Code Last Permanent Residence
Hospital Residence Prior to Program Entry
11/15/05 Program Entry Date
11/20/05 Program Exit Date
512 Program Identification Number
8/4/78 Date of Birth
black Race/Ethnicity
female Gender
2 Household Identification Number
No Disabling Condition
15081 ZIP Code Last Permanent Residence
Rental Residence Prior to Program Entry
11/3/05 Program Entry Date
11/5/05 Program Exit Date
74 Program Identification Number
10/8/55 Date of Birth
Hispanic Race/Ethnicity
female Gender
3 Household Identification Number
No Disabling Condition
15111 ZIP Code Last Permanent Residence
Hotel Residence Prior to Program Entry
4/8/05 Program Entry Date
4/20/05 Program Exit Date
85 Program Identification Number
4/8/75 Date of Birth
Hispanic Race/Ethnicity
female Gender
3 Household Identification Number
No Disabling Condition
15111 ZIP Code Last Permanent Residence
Hotel Residence Prior to Program Entry
4/8/05 Program Entry Date
4/20/05 Program Exit Date
85 Program Identification Number

Figure 8. De-identified data for Clients includes utilization patterns. Some Clients are linked together by sharing the same Household Identification Number (depicted by the link between the bottom Clients).

4. Data Linkage Threat

Before introducing methods to assess UID technologies (Section 5) and assessing 8 UID technologies (Section 6), some background on data privacy threats specific to Clients in Shelters is needed.

Beyond the intimate stalker threat in which information about a single Client is sought (see Section 3.4), the data linkage threat involves learning information about most, if not all, Clients by matching the information to other available data in order to use HMIS data inappropriately. This kind of activity is most likely to occur at Planning Offices where linking can be used to learn information about a larger number of Clients than those at just one Shelter. Protecting privacy in this setting cannot involve thwarting all linking, because the HMIS de-duplication task the Planning Office performs on the data requires linking records that belong to the same Client across Shelters. Instead of thwarting all linking, privacy protection in the HMIS setting involves thwarting linking attempts that may re-identify Clients.

Figure 9 provides an example in which the Dataset is linked to publicly available voter information on {*ZIP*, *date of birth*, *sex*} to re-identify the records in the Dataset by *name*. The more uniquely occurring {*ZIP*, *date of birth*, *gender*}, the more fruitful the re-identifications.

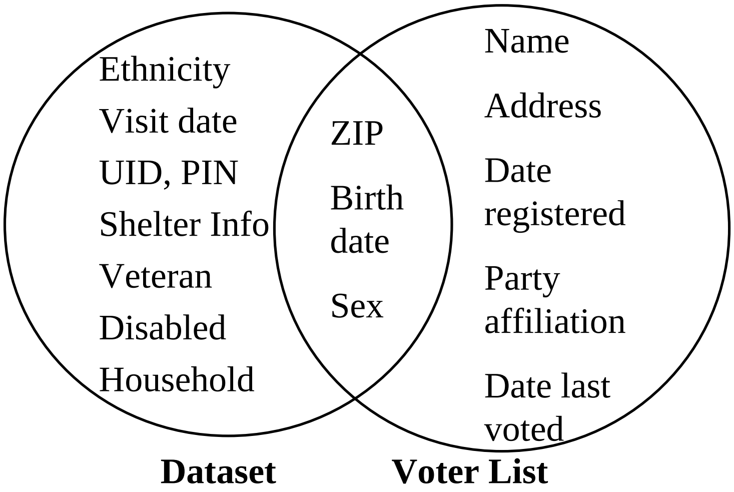


Figure 9. Example of linking Dataset to a publicly available population register, such as voter list, to re-identify the names of Clients appearing in Dataset.

Most UIDs are designed to be uniquely assigned to Clients, so as a result, UIDs can also be used as the basis for linking datasets. That is not surprising given that HUD introduced UIDs into HMIS in order to link Client visits. However, if the same UIDs are also used with non-HMIS data, then they become the basis for linking HMIS data beyond the HMIS context. The following recommendation is aimed at thwarting secondary uses of HMIS data using UIDs.

<u>Recommendation:</u> *UID values assigned to Clients of domestic violence shelters should not be used (i.e., stored or referenced) by any non-HMIS program to which the Clients may participate.*

In some cases, Planning Offices may decide to use HMIS data outside the HMIS context and in so doing, may purposefully link HMIS data to other non-HMIS data, even though this is unnecessary to achieve HMIS objectives. UID technologies can be constructed to thwart this behavior, as discussed later in this writing, but if this activity is desired, then Clients and Shelters should be made aware of this practice and any increased risk that may result. This is the motivation behind the following recommendation.

Recommendation: *Shelters and Planning Offices should make sure privacy notices to Clients explain the data collection, sharing, and linking practices of the Shelters and Planning Offices in which the Client's data will be part.*

4.1 Re-identification

A "re-identification" results when a record in Dataset can reasonably be related to the Client who is the subject of the record in such a way that direct and rather specific communication with the Client is possible. Figure 10 provides a depiction of a re-identification in which external information is linked on month and year of birth (9/1960), gender (F), and ZIP code (37213) to identify the visit information as belonging to Ann. The re-identification is sufficient to send a letter to Ann's home.

For another example, consider Figure 9 in which Dataset is linked to a voter list to re-identify Client visits by name, even though Client names had been omitted from the visits in an attempt to protect privacy (recall Section 3.4.3).

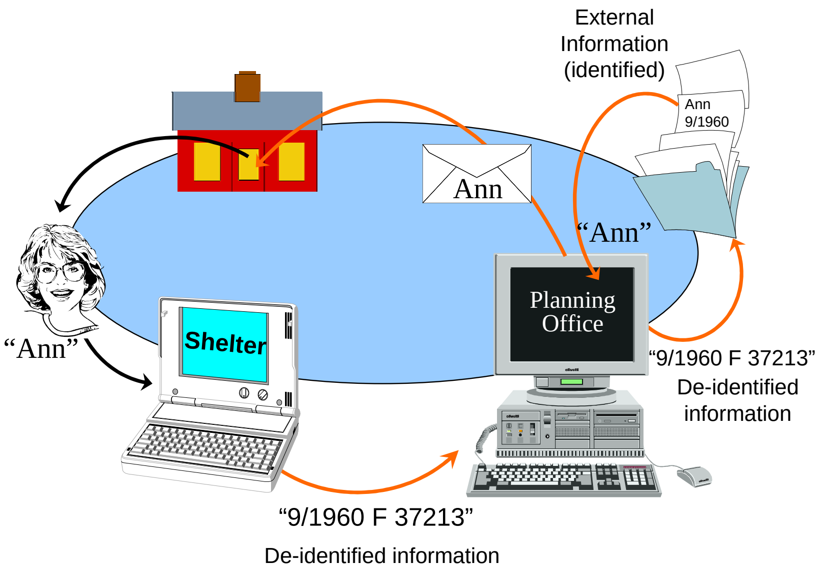


Figure 10. Depiction of re-identification. Ann leaves her home and gives her explicitly identified information to the Shelter. De-identified information about Ann is provided to the Planning Office, but in this depiction, the information can be used with external information (or personal knowledge) to re-identify the information as belonging to Ann. A re-identification occurs if there is sufficient information to directly communicate with Ann (not limited to mail), shown in the diagram as mailing an envelope to her original residence (or alternatively, sending the letter to Ann at the Shelter in which she resides).

4.2 Identifiability

One way to report the risk of re-identification is to determine the number of people to whom a record could refer. This is termed "identifiability." Figure 11 shows two examples in which information is released and compared against a known population. On the left, Figure 11 (a), each of the released profiles are ambiguous in terms of head shape and shading. Neither can be uniquely identified. The top released profile matches Hal and Len indistinguishably and the bottom profile ambiguously matches Jim and Mel. The release shown on the upper right of Figure 11 (b) is different. There is only one person in the known population (Hal) having the same color and head shape. In this case, the record referring to Hal is uniquely re-identified even though many of Hal's details had been removed.

While unique re-identifications obviously pose a privacy problem, so do situations in which a record maps ambiguously to a few known people. In Figure 11(a), both released profiles map to two individuals, but these people are both explicitly known, so they can both be contacted with little effort. Of course, the larger the number of people to whom a record refers, even if all of the people are known, the greater the effort usually needed to contact so many or make use of the information.

Counting the number of possible re-identifications for a record is a useful measure of privacy risk, but what is needed is a way to estimate the number of people to whom a record might refer.

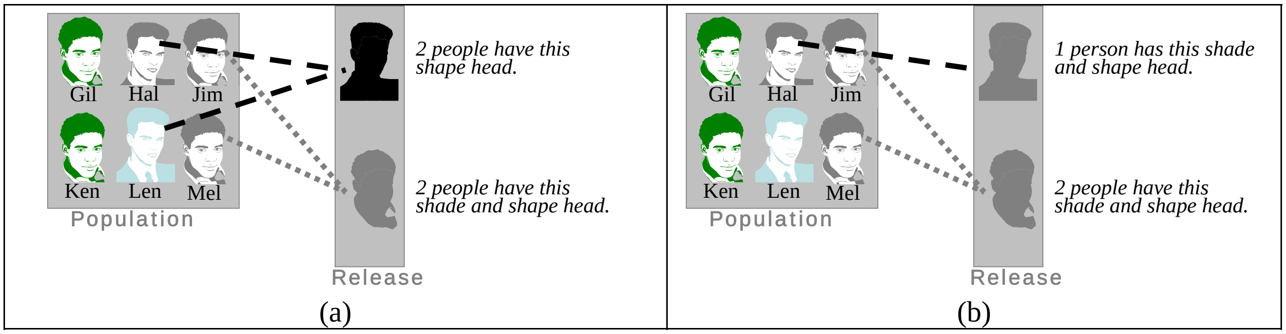

Gil
Hal
Jim
Ken
Len
Mel
Population
2 people have this shape head.
2 people have this shade and shape head.
Release
(a)
Gil
Hal
Jim
Ken
Len
Mel
Population
1 person has this shade and shape head.
2 people have this shade and shape head.
Release
(b)

Figure 11. The identifiability of the profiles released in (a) are each ambiguously re-identified to two named persons. The top profile released in (b) is uniquely re-identified to Hal.

4.3 Identifiability of Dataset

The Risk Assessment Server is a commercially available system that reports re-identification risks by estimating the number of named persons to which each record could relate given its model of the U.S. population and its knowledge of publicly available datasets [20]. The output of the Risk Assessment Server is a plot of identifiability estimates, in graduated size groupings, that report the number of people to which a released record is apt to refer.

Figure 12 shows the results from the Risk Assessment Server based on {*date of birth*, *gender*, *5-digit ZIP*} from Dataset. The lower left plot shows that 87% of the population are uniquely identified by these characteristics. As age information is generalized and as geographical reference to the Client's prior residence is made less specific, uniqueness deteriorates and privacy protection increases. For example, {*year of birth*, *gender*, *5-digit ZIP*} drops the unique identifiability to 0.04% (see the lower right plot in Figure 12).

Dataset currently requires Shelters to provide the full month, day and year of birth and all 5 digits of the Client's last residential ZIP code, yet the AHAR uses only gross age values and geography relative to Shelter's service area (refer to Section 3.6). The following recommendation is aimed at increasing privacy protection by changing the level of specificity in these fields.

Recommendation: *The fields* date of birth *and* ZIP code of last residence, *which are among the data elements HUD recommends HMIS collect, should contain information less specific than the month, day, and year of birth and all 5 digits of the ZIP (or postal) code.*

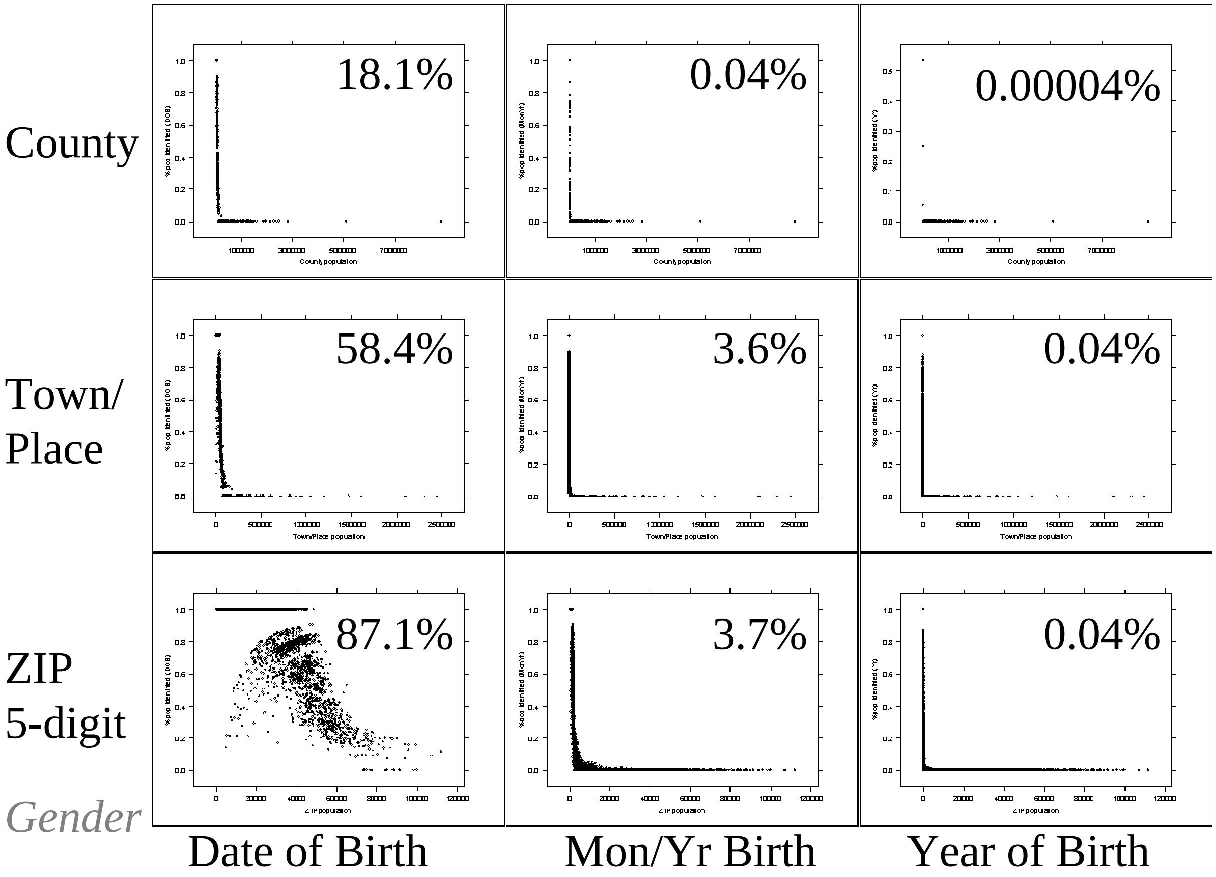


Figure 12. {*date of birth, gender, 5-digit ZIP*} uniquely identifies 87.1% of USA population, but as ZIP is made less specific, the identifiability drops to 18.1% (bottom to top). Similarly, as the age of the client is made less specific, the identifiability drops to 0.04% (left to right). All values include gender. The horizontal axis of each sub-plot is the number of people who reside in the geographical area and the vertical axis is the percentage of the population uniquely identified by the noted combination of demographics noted. As the demographics are aggregated, the points move towards 0% identifiable.

4.4 Privacy concerns in Program-Specific Data Elements

Planning Offices that receive Program-Specific Data Elements (Figure 5) have some additional privacy concerns to consider to best protect Client data.[7] Program-Specific Data Elements may be linked to other available programmatic information to re-identify Clients. This vulnerability differs among municipalities and states as different kinds of secondary data from related programs are available.

A Planning Office is assumed to have multiple versions of data available, each having different re-identification risks and therefore different access policies. Figure 13 provides an overview. In terms of re-identification risk, the most sensitive data is that which first arrives at the Planning Office from the Shelter. These data may be separated into the Dataset used for the unduplicated accounting (the Universal Data) and the Program-Specific Data. No UIDs should appear in the Program-Specific Data. The De-identified Dataset is of least risk. A Planning Office may make internal access policies commensurate with these levels of risk. This advice regarding the maintenance of various versions of data is for consideration by Planning Offices and is not required.

Different versions of the data have different purposes. The originally received data should be maintained intact for quality control of Client information with Shelters (using PINs). The De-identified Dataset (modified to have less specific values of ZIP and date of birth) offers the least risk of re-identification and can be used to compute the unduplicated count information. In cases where the Shelter does not provide Program-Specific Data, the Dataset and the Originally Received Data are the same.

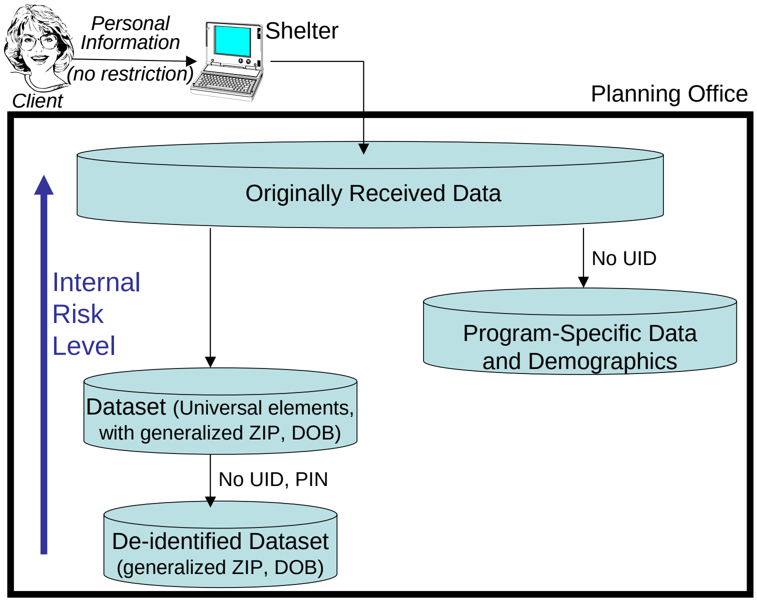

Personal Information
(no restriction)
Client
Shelter
Planning Office
Originally Received Data
No UID
Internal Risk Level
Program-Specific Data and Demographics
Dataset (Universal elements, with generalized ZIP, DOB)
No UID, PIN
De-identified Dataset (generalized ZIP, DOB)

Figure 13. Versions of data maintained by a Planning Office with relative internal risk of re-identification. The originally received data has the most internal risk and the De-identified Dataset has the least.

[7] The requirements of the Program-Specific Data elements reside outside the scope of this work. However, some relative re-identification risk is noted.

5. Methods for Assessing UID Technologies

Since HUD's introduction of a UID in Dataset, many Planning Offices and Shelters have already explored technologies they might deploy to construct, maintain, and use UIDs. Other Shelters and Planning Offices are just getting started in this process. The goal of this writing is to describe how to assess plans and technologies in terms of their ability to perform an unduplicated accounting while protecting privacy. This document itemizes what should be the content of the assessment and what problems it should address.

In this writing, a "Proposed Solution" is a UID technology bundled with an accompanying set of policies and practices for the construction, maintenance and use of a UID technology for Clients of Shelters in HMIS. The entire package, UID technology, policies and practices, bundled together, is the subject of the assessment.

The overall problem for which UIDs have been introduced is easy to understand. It is termed the "HMIS Unduplicated Count Problem" and is stated below.

<u>The HMIS Unduplicated Count Problem.</u>
Given a set of Clients, a set of Shelters, and a Planning Office, where Clients visit Shelters, and Shelters report Dataset to the Planning Office on the Clients that visit, how should information about Clients be reported to Shelters and to the Planning Office such that the Planning Office can identify distinct visits of Clients across Shelters but not the identities of the Clients?

In order to determine whether a Proposed Solution is a sufficient solution to the HMIS Unduplicated Count Problem, an assessment must be done that demonstrates that the Proposed Solution remains useful for HMIS purposes while still being minimally invasive to privacy. Framed this way, the HMIS Unduplicated Count Problem is an optimization problem. On the one hand, a Proposed Solution should provide an accurate accounting of distinct Client visits. On the other hand, a Proposed Solution should protect the privacy of Clients. The sufficiency of a Proposed Solution is based on performance guarantees that can be made. Specifically, a performance guarantee that the Proposed Solution has a minimal risk of re-identification when the solution is considered with other publicly and readily available information and techniques is termed a "Compliance Statement" in this writing. Similarly, a performance guarantee that the Proposed Solution provides a reasonably accurate unduplicated accounting of client visit patterns to shelters within the regional setting it is to be deployed is a "Warranty" in this writing. An assessment of a Proposed Solution is done by providing Compliance and Warranty statements.

In the next subsections, more information about Warranty and Compliance statements will be provided. But first, the notion of "source information" and "de-duplication instrument" are introduced.

5.1 Basic terms

A UID technology involves transforming some source information collected from the Client at a Shelter, into a UID. The ideal is to have a UID uniquely associated with a Client such that no two UIDs relate to the same Client, and a Client has only one UID. Resulting UIDs are used by the Planning Office to identify the same Clients across Shelter visits by matching UIDs or by using a "de-duplication" instrument. These terms are further described in the next subsections.

5.1.1. Source information

Source information is something a Client holds or knows that forms the basis of the Client's UID. Common examples of source information are name, date of birth, and Social Security number. The source information is not the same as the UID, but instead is used as the basis for a method (or algorithm) that computes a UID from it. For example, an algorithm for constructing a UID could involve concatenating the Client's date of birth with the first 4 letters of the Client's first name. For example, Alice with birthdate 9/12/1960 would have UID "09121960ALIC."

In some cases, the source information may rely solely on volunteered verbal information from the Client. This is termed "non-verifiable" source information. Client information is just accepted as stated and is not checked against other credentials.

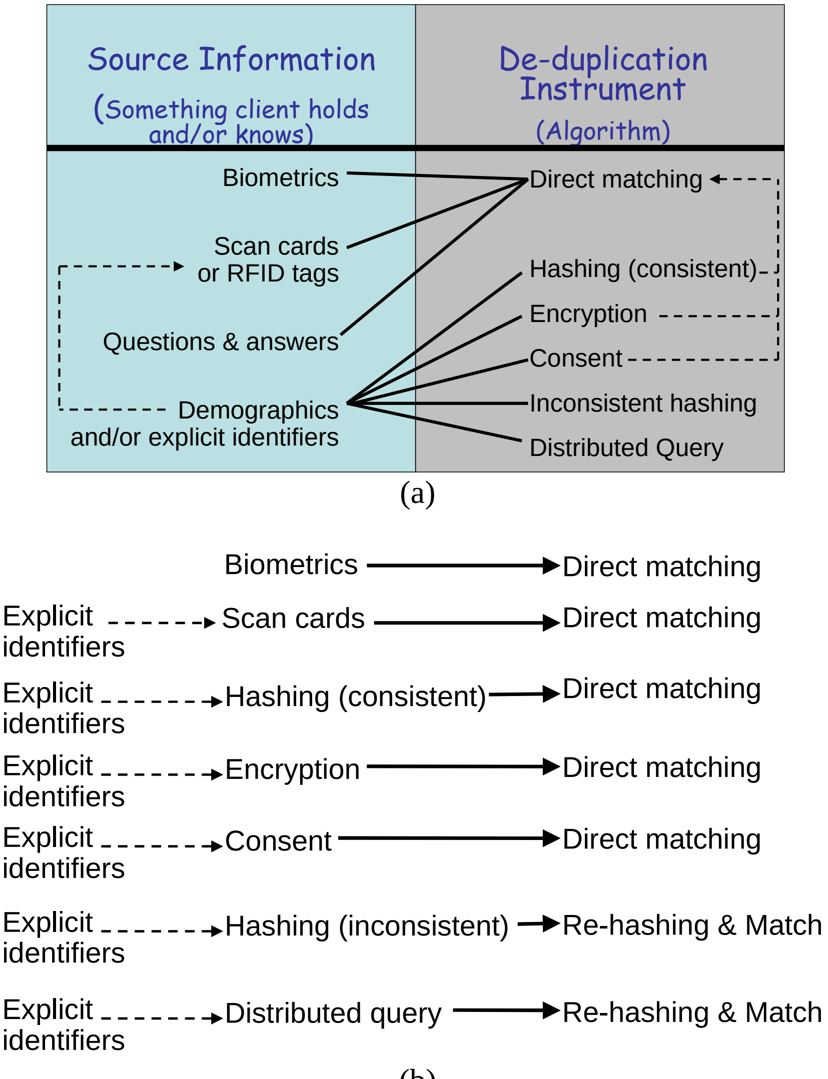
An interesting example of non-verifiable source information for UIDs is realized by allowing a Client to makeup her own UID (e.g., "100678") or by constructing a UID based on Client answers to simple questions like "your favorite color, song, and ice cream" or "which picture most resembles your first love." As long as the Client answers consistently across multiple Shelter visits, the UID will be associated with the Client. As long as the questions tend to evoke unique answers from each Client and Clients answer the same way on each visit, then the resulting UID will be uniquely associated with a Client.

"Verifiable" source information is something provided by the Client that can be confirmed. Examples include a driver's license or a fingerprint.

5.1.2. De-duplication instrument

A set of algorithms that describe how to construct a UID from source information and how to use UIDs to match Clients are collectively termed a "UID technology." Algorithms that construct UIDs may be as simple as concatenating parts of Client demographics, as demonstrated above, or more complicated as computing a unique value for a Client. Algorithms that match (or "de-duplicate") UIDs can be as simple as comparing two numbers, or as complicated as computing probabilistic matches.

Figure 14 (a) includes UID technologies already being considered. Source information includes biometrics, scan cards, question-and-answer, and the use of demographics and explicit identifiers. De-duplication instruments include directly matching (or linking) assigned, hashed, or encrypted values. Inconsistent hashing and distributed query are de-duplication instruments that do not simply match constructed UIDs. The UID technology termed "consent" merely checks whether Client permission was given. Each of these categories of UID technologies will be further described when they are assessed in Section 6. Figure 14 (b) shows some sample ways these are combined.



(a)

(b)

Figure 14. UID Technologies, assessed in Section 6, are broken down by source information and de-duplication instrument (a). Sample ways source and de-duplication instruments combine are shown in (b).

In Figure 14 (a), the solid line linkages between source information and de-duplication instruments show combinations of source information currently under consideration by some Planning Offices. Notice that biometrics, scan cards, question-and-answer, and demographic source information use direct matching to determine whether two UIDs match. Hashing, encryption, consent, inconsistent hashing, and distributed query all use demographics and/or explicit identifiers (e.g., Social Security number) as source information. The dashed lines in Figure 14 show secondary relationships. Demographics and explicit identifiers may be stored on scan cards. Hashed and encrypted values use direct matching for de-duplication. Consent also uses direct matching on demographics and/or explicit identifiers for de-duplication.

Assessing a UID technology involves producing Warranty and Compliance statements. Each of these are further described below.

5.2 Warranty statement (utility)

Given a Proposed Solution to the HMIS Unduplicated Count Problem, a Warranty shows that a reasonably accurate unduplicated accounting of records from Shelter Datasets is possible by the Planning Office. Below are fundamental issues that should be addressed by a Warranty.

The Warranty should demonstrate how de-duplication is done in the general case and identify the Proposed Solution's overall performance. Measures of accuracy should be included and cases that inflate or deflate the overall accounting should be addressed.

The behavior of the Proposed Solution using non-verifiable source information and verifiable source information should be examined. Particular attention should be given to the behavior of the Proposed Solution if Clients provide bad source information, such as purposeful name misspellings, wrong information, plausible differences in the information, or no information in part. Finally, consider the extent that the Proposed Solution can instill client confidence. This is particularly important when using non-verifiable source information because in these cases the system relies significantly on the cooperation of the Client.

Figure 15 lists considerations for Warranty statements.

WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?*
Verifiable source information	*Can problems occur if the UID is based on verifiable source information? What if the information is not correct?*
Client confidence and trustworthiness	*The more Clients (and those who regularly intake Clients) trust the overall system and are encouraged to provide truthful information, the more likely Clients will actually provide truthful information. How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)? How would a lack of trust effect overall performance?*
Inflated accounting	*What are the circumstances under which de-duplication is likely to inflate the accounting? What are the circumstances in which a known Client is is not recognized (even if this does not actually inflate the count)? Explain the circumstances that generates these false negatives.*
Deflated accounting	*What are the circumstances under which de-duplication is likely to deflate the accounting? What are the circumstances in which a known Client is considered to be a different Client (even if this does not actually deflate the count)? Explain the circumstances that generates these false positives.*
Handling bad or missing input	*What is the effect of bad, incomplete, or missing source information on performance? How are these cases handled? (Note: "bad" information refers to accidental typing or other input mistakes.)*

Figure 15. Warranty Statements should seek to answer these questions.

5.3 Compliance statement (privacy)

Given a Proposed Solution to the HMIS Unduplicated Count Problem and publicly and readily available data and techniques, a Compliance Statement shows that the number of Clients who may be re-identified from the records in a Shelter Dataset is minimal. Below are fundamental issues that should be addressed by a Compliance Statement.

Consider any vulnerabilities the intimate stalker may exploit. Refer to Section 3.4.

Consider the ability to link Datasets, which include UIDs, to other available information in an attempt to re-identify Clients. Refer to Section 4.

“Dictionary attacks” should be considered. The idea of a dictionary attack is to generate UIDs for all possible source values and then see which results match UIDs stored in Dataset. Because the source that produced the UID is known, the information about the Client becomes known. Dictionary attacks assume the attacker has access to the UID technology and knowledge of what source information is used.

Here is an example of a dictionary attack. Assume a UID technology uses encryption to compute a number from the Client’s Social Security number. Even without knowing how encryption works (which will be discussed in Section 6.3), one can use a dictionary attack to learn the source information that generates a UID. Figure 16 shows an encryption method that when given a Social Security number produces a UID.

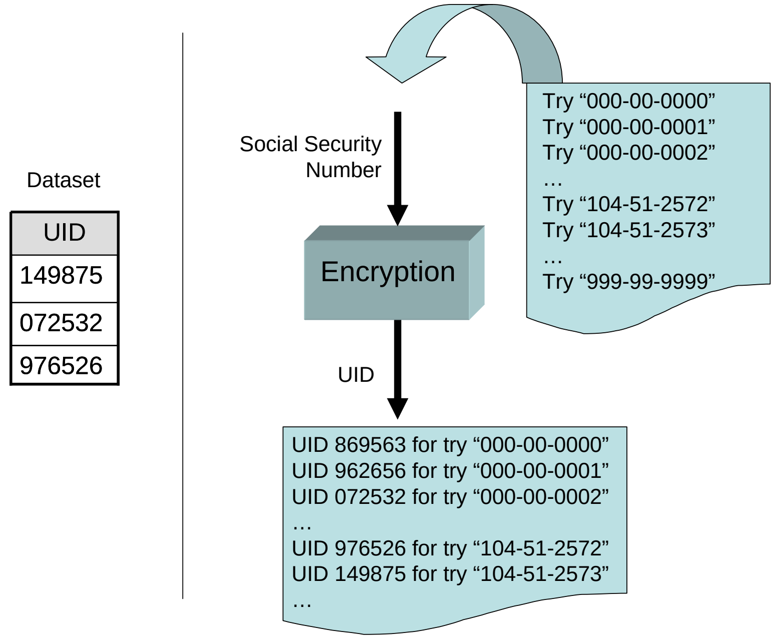


Figure 16. Example of a dictionary attack. Given a Dataset having UIDs 149875, 072532, and 976526, the knowledge that UIDs are encryptions of Social Security numbers, and access to the encryption function, a dictionary attack allows the UIDs to be learned by trying all possible Social Security numbers and seeing which Social Security numbers encrypt to the observed UIDs. In the example above, the Social Security number 104-51-2573 encrypts to 149875.

Suppose the Dataset contains the UIDs: 149875 and 072532. We can use a dictionary attack to learn the Clients' Social Security numbers that produced those UIDs by trying all possible 9-digit values and seeing which 9-digit Social Security numbers produce the UIDs that appear in the Dataset. As noted in Figure 16, the Social Security number "104-51-2572" produced UID 149875 and the Social Security number "000-00-0002" produced the UID 072532, so the Clients have the Social Security numbers "104-51-2572" and "000-00-0002," respectively.

A dictionary attack can be combined with linking to re-identify Clients by name. Assume a UID technology encrypts a combination of a Client's date of birth and gender to produce a UID. These same values also appear in the voter list (see Figure 9). So, computing UIDs for every voter in the voter list allows us to match UIDs in the Voter list to UIDs in the Dataset to re-identify Clients by name.

Beyond the intimate stalker threat, linking attacks, and dictionary attacks, an assessment should also examine to what extent the algorithm for producing the UID can be "reverse engineered." For example, given the following list of UIDs: 09121960ALIC, 10251974JANE, …, one can conclude that the UID is constructed by concatenating the month, day and year of birth with the first 4 letters of the first name. In this example, observing the UIDs revealed the method for constructing the UIDs. Given a Client's name and date of birth, the Client can be found in the Dataset.

A Compliance Statement should also identify any new legal or technical privacy risks that may be introduced based on the existence of the Proposed Solution's UID. This is considered "exposure."

Here is an example. If a Proposed Solution uses fingerprints as the source information for UIDs in such a way that a UID database is a fingerprint database, then the existence of the resulting database of Client fingerprints may be useful to law-enforcement. The existence of the database's usefulness to third parties poses new privacy concerns for Clients and thereby, increases exposure.

Figure 17 lists considerations for Compliance statements.

COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	*What vulnerabilities exist for the intimate stalker?*
Re-identification: Linking	*What vulnerabilities exist for re-identification of UIDs (and Dataset) using data linkage? What is the identifiability of the Dataset?*
Re-identification: Dictionary Attack	*What vulnerabilities exist for re-identification of UIDs (and Dataset) using a dictionary attack?*
Re-identification: Reversal	*What is involved in reverse engineering the UID construction method?*
Exposure	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?*

System Trust

The overall system consists of intakers, who enter Client information, insiders, who have access to Client information for a variety of meritorious reasons, and the Shelters and Planning Offices themselves. Which parties are heavily trusted?

Figure 17. Compliance Statements should seek to answer these questions.

5.4 Other factors

There are many other factors that may contribute to a decision of which UID technology to use that are not part of the assessment. Among these are trust and economics. Where trust is placed differs among Proposed Solutions. Some solutions put more trust in the Shelters (e.g. distributed query), in the Clients (e.g., UID technologies using non-verifiable source information), or in the Planning Offices (e.g. consent).

Another key factor can be the economics of constructing, installing and maintaining the system. Some states are constructing systems for administrative oversight of social programs, so weaving HMIS requirements into those systems can be cost-effective, but doing so, may dictate the use of a particular UID technology.

Another factor can be available technical expertise.

While all these kinds of factors are important to the decision-making process, they are excluded from demonstrating the worthiness of the Proposed Solution. Warranty and Compliance statements demonstrate utility and privacy protection independent of these concerns.

5.5 Putting the pieces together

The goal of this section is to provide guidance on what should constitutes an assessment. The goal of the next section is to provide some overall assessments of 8 categories of technologies.

In summary, an assessment is a thorough review and analysis of a Proposed Solution that should be completed before a Proposed Solution is put to real-world use. Assessing a Proposed Technology requires a confluence of technology, policy, and sometimes law. Groups of people lacking the proper expertise or not focused on key issues pertinent to Warranty and Compliance issues can lead to poor assessments. The goal of this section and the next is to help those engaged in this process to ask themselves the right questions and to identify the right kinds of expertise needed. Along these lines, the following two recommendations are made.

Recommendation: *Given a Proposed Solution, a person skilled in statistical, computational and/or legal principles, as appropriate, should certify in writing that the Proposed Solution has a minimal risk of re-identification when the solution is considered with other publicly and readily available information and techniques. Such writing should address vulnerabilities for inappropriate re-identifications by various categories of insiders. This is termed a "Compliance Statement" and should be made available for inspection.*

Recommendation: *Given a Proposed Solution, a person skilled in statistical and/or computational principles, as appropriate, should certify in writing that the Proposed Solution provides a reasonably accurate unduplicated accounting of client visit patterns to shelters within the regional setting it is to be deployed. Such writing should include possible false match and missed match rates. This statement is termed a "Warranty" and should be made available for inspection.*

5.6 Privacy, not computer security

One word about computer security before continuing. This writing relates to data privacy concerns and not to computer security issues. It is assumed that any Proposed Solution operates in a computational environment having adequate computer security to authenticate users, limit access, combat intrusions and prevent eavesdropping. This writing does not address computer hacking, break-ins, viruses, or unauthorized computer users, because such issues appear to be adequately addressed with commercial computer security solutions. For general reference, see Pfleeger [21].

Instead, this writing addresses ways to limit authorized users from doing unauthorized tasks with available data. For example, the intimate stalker either has access to the Dataset already or obtains assistance from someone with access. Linking Dataset to other available information in order to re-identify Clients can only be done by someone with access to the Dataset. If someone does break-into the computer system and gains access to Dataset to attempt these things, the safeguards described in this writing will thwart their efforts. Described in this manner, these safeguards provide some privacy protection even in the face of a computer security breach. But more generally, these safeguards thwart unwanted activities by most of those who work with Dataset regularly.

6. Gross Assessments of UID Technologies

Overall assessments of 8 categories of UID technologies are presented in this section using the assessment criteria stated for Warranty and Compliance statements in the previous section. A summary of results appears in Section 7. This section provides details by UID technology.

The assessments presented in this section are not complete assessments. They examine only the UID technologies and not the accompanying policies or practices that may address noted concerns. Nonetheless, these assessments are useful in comparing UID technologies and in identifying the kinds of issues that accompanying policies and best practices need to address prior to deployment.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 18. Level of severity or difficulty of a problem is determined by shading.

For each of the UID technologies, the answers to the questions posed for Warranties (see Figure 15) and for Compliance statements (see Figure 17) are addressed with respect to that technology in the absence of accompanying policies or best practices. If a "problem" is described in answering the question, it should be addressed by accompanying policy or practice or by modification of the UID technology from the generally assumed form. A shaded code is assigned to denote the severity or difficulty of the problem: the darkest shading denotes a "serious problem," a dark hash pattern denotes a moderate problem, a light hash pattern denotes the existence of a "problem," a light shade with no pattern denotes a situation that "may be a problem," and no shading signals that there is not likely to be a problem. Figure 18 shows the shadings and patterns. Comments related to System Trust have no associated shading because these comments merely reflect where trust is placed.

The following categories of UID technologies are examined in the noted subsections.

6.1. Encoding
6.2. Hashing
6.3. Encryption
6.4. Scan Cards / RFID
6.5. Biometrics
6.6. Consent
6.7. Inconsistent hashing
6.8. Distributed query

6.1 Encoding

Using "encoding" to produce UIDs simply involves concatenating parts of source information to form a UID. De-duplication is then performed by simply matching resulting UID values.

Figure 19 provides an example of a UID constructed by encoding the fileds {*date of birth*, *gender*, *ZIP*}. Specifically:

encode(9/12/1960, F, 37213) = "09121960F37213"

In this example, the digits of the date of birth, a letter for gender, and the 5-digits residential ZIP code are merely concatenated. While this example uses all characters in the source information, encoding sometimes uses only some characters, such as using the first 5 letters of a person's last name.

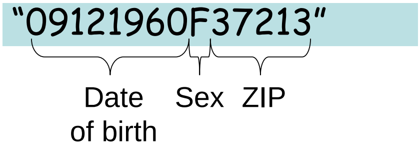


Figure 19. Example of making a UID by encoding {date of birth, gender, ZIP}.

An obvious problem with encoding is that given a series of UIDs and some source information, an attacker can often deduce what parts of which source information appears in the UID and where in the UID it appears.

Figure 20 and Figure 21 provide a gross assessment of encoding as a UID technology. Issues related to utility and the warranty statement appear in Figure 20. Issues related to privacy and the compliance statement appear in Figure 21. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

ENCODING --WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information	░	*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information is a reliably captured biometric, but biometrics seem unlikely source information for encoding (refer to hashing, encryption, or inconsistent hashing). So, determining what would constitute verifiable Client information for encoding would be important.
Client confidence and trustworthiness	▓	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Encoded UIDs tend to be transparent, which can limit Client and intaker confidence by exposing information. Accompanying practices should seek to build Client and intaker trust. An example of a transparent code that would still maintain trust would be to allow Clients to make up their own UID or to use answers to simple questions as source information (see Section 5.1.1).
Inflated accounting	░	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Count inflation can occur in cases where a Client provides different source information on different visits. In these cases, different UIDs are generated and therefore will not match to each other even though they are assigned to the same Client. This relates to the comment above on non-verifiable information.
Deflated accounting	■	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation can occur in cases in which a Client provides incomplete or missing information or different source information on different visits, or in which a bad method is used for generating UIDs. In these cases, the same UID is generated for different Clients and therefore visit information will combine inappropriately, generating serious accounting problems. Deflation is more likely than inflation.

…continued on next page …

Handling bad or missing input	■	*What is the effect of bad, incomplete, or missing source information on performance?* Typing mistakes that go uncorrected, as well as incomplete or missing information, can generate different UIDs for a Client than would have been generated with complete and properly entered information. This tends to inflate accounting by generating spurious UIDs for Clients having multiple visits. On the other hand, having the same incomplete and missing information across Clients will deflate accounting because different Clients would have the same UID. See comments on inflated and deflated accounting above.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 20. Gross Warranty assessment of encoding as a UID technology.

ENCODING –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	■	*What vulnerabilities exist for the intimate stalker?* In typical cases where demographics are the source information encoded, serious problems may exist. Demographics tend to be visible within the encoding, making identification more transparent to an intimate stalker.
Re-identification: Linking	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because demographics tend to be the source information used with encoding and demographics appear in other available data, linking tends to be a serious problem. Analysis of specific risk should be based on the re-identification of demographics over the actual population from which Clients are drawn.
Re-identification: Dictionary Attack	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* A dictionary attack can be done by executing the encoding function over all legal combinations of source information. For any generated UID that matches a UID in the Dataset, the Client's source information is learned. This may pose a serious problem depending on the source information and encoding method used. A combination dictionary-attack and linking attack can also be a problem. For example, suppose some other data (Other Data) is to be linked to a Dataset in which UIDs are encoded using source information and the same source information appears in Other Data. UIDs can be produced for the source information in Other Data, and then, UIDs in Dataset are matched to UIDs in Other Data to link Client data.
Re-identification: Reversal	■	*What is involved in reverse engineering the UID construction method?* Because encodings tend to be transparent, casual (or visual) inspection can often be used to describe the encoding algorithm. Even in cases where the encoding appears more cryptic, inspecting known cases can often reveal the encoding method.
Exposure	▩	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of encodings enable risks of linking described above and can make demographics on Clients transparent which can increase re-identification risks beyond the HMIS context.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
□	No problem likely, or not applicable

…continued on next page …

System Trust
Which parties are heavily trusted?

All insiders are heavily trusted not to decode UIDs or exploit the knowledge they may learn about the encoding scheme. If the encoding scheme is obscure, then the scheme itself is heavily trusted in the belief that no one, no matter how heavily motivated, will learn or share the scheme. Additionally, if the encoding scheme is obscure, insiders with access to the encoding method are heavily trusted.

Figure 21. Gross Compliance assessment of encoding as a UID technology.

6.2 Hashing

Using "hashing" to produce UIDs involves computing a number from source information. De-duplication is then performed by simply matching UID values.

Figure 22 provides an example of making a UID by hashing the fields {*date of birth*, *gender*, *ZIP*}. Specifically:

hash(9/12/1960, F, 37213) = "8126r1329ws"

Unlike encoding, the hashed value is not transparent, as it was with encoding (Section 6.1).

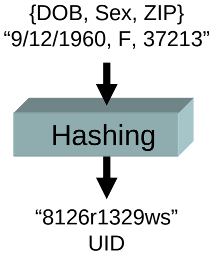


Figure 22. Example of making a UID by hashing {date of birth, gender, ZIP}.

Hashed UIDs are consistently produced. That is, each time the hash function is given the same input, it produces the same UID.

A vendor can create their own hash function, but it has been shown that these "ad hoc" approaches can be reversed, especially if someone is highly motivated to do so. Protection using an ad hoc hash function is good only as long as no one learns the actual hash function used. Rather than using ad hoc hash functions, cryptographically "strong" hash methods are highly recommended. With a strong hash function, everyone can examine the method being used, but even with intense inspection, it has been proven that no one can reverse the process without performing more computation than can be reasonably performed [22].

Hash functions have the property that they do not preserve the natural ordering typically found in source values. Two consecutive values (e.g. ZIP codes 37212 and 37213) tend to have radically different hashed values (e.g., "x41768" and "z1Rx5G"). This is good for privacy, but can be bad for utility.

Figure 23 and Figure 24 provide a gross assessment of hashing as a UID technology. Issues related to utility and the warranty statement appear in Figure 23. Issues related to privacy and the compliance statement appear in Figure 23. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

HASHING –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently because similar source values have radically different hashed values. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information	□	*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information is a reliably captured biometric.
Client confidence and trustworthiness	▒	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Hashed UIDs tend to appear cryptic, which can instill Client and intaker confidence. However, problems can emerge in cases where the requested source information is sensitive, notwithstanding the cryptic appearance of the UID itself. Educating Clients and those who perform intake regularly and/or issuing privacy notices may help.
Inflated accounting	▒	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Count inflation can occur in cases where a Client provides different source information on different visits (see comments for non-verifiable source information above). In these cases, different UIDs are generated and therefore will not match to each other even though they are assigned to the same Client.
Deflated accounting	■	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation can occur in cases in which a Client provides incomplete or missing information or different source information on different visits, or in which a bad method is used for generating UIDs. In these cases, the same UID is generated for different Clients and therefore visit information will combine inappropriately, generating serious accounting problems. Deflation is more likely than inflation.

…continued on next page …

Handling bad or missing input	■	*What is the effect of bad, incomplete, or missing source information on performance?* Typing mistakes that are go uncorrected, as well as incomplete or missing information, can generate different UIDs for a Client than would have been generated with complete and properly entered information. This tends to inflate accounting by generating spurious UIDs for Clients having multiple visits. On the other hand, incomplete and missing information is likely to deflate accounting because different Clients whose entries are missing the same information may have the same UID.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
□	No problem likely, or not applicable

Figure 23. Gross Warranty assessment of hashing as a UID technology.

HASHING –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	■	*What vulnerabilities exist for the intimate stalker?* In typical cases where demographics is the source information used with hashing, serious problems may exist. Access to the hash function can allow the intimate stalker (working with a compromised insider) to generate a Client's UID, and then to use the UID to identify the Client's Shelter location in the Dataset. Control and auditing of hash function use is important to thwarting this problem.
Re-identification: Linking	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because demographics tend to be the source information used with hashing and demographics appear in other available data, linking tends to be a problem if access to the hash function is not controlled and audited. Practices should limit and account for hash function use. Risk analysis should be based on the re-identification of demographics over the actual population from which Clients are drawn.
Re-identification: Dictionary Attack	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* A dictionary attack can be done by executing the hash function over all legal combinations of source information. For any generated UID that matches a UID in Dataset, the Client's source information is learned. This may pose a serious problem depending on source information and hash method used. A combination dictionary-attack and linking attack can also be a problem. For example, suppose some other data (Other Data) is to be linked to a Dataset in which UIDs are hashed using source information and the same source information appears in Other Data. UIDs can be produced for the source information in Other Data, and then, the UIDs in Dataset are matched to the UIDs in Other Data to link Client data to Other Data. Practices should limit and account for uses of the hash function.
Re-identification: Reversal	▨	*What is involved in reverse engineering the UID construction method?* If a "strong" hash function is used, then it is highly unlikely that the method will be reversed. For this reason, strong rather than ad hoc hash functions should be used. If strong methods are not used, then attention must be paid to the ability to reverse the method.
Exposure	□	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of hashed UIDs used only in the HMIS-context is not likely to expose Clients to additional risks beyond those mentioned above.

■	Most severe/difficult problem
▩	Moderate problem
▨	A problem
░	May be a problem
□	No problem likely, or not applicable

…continued on next page …

System Trust
Which parties are heavily trusted?

If the hash function is ad hoc (not strong), then the function itself is heavily trusted in the belief that no one, no matter how heavily motivated, will reverse the function. It also requires trusting the developer of the ad hoc hash function.

Additionally, no matter whether the hash function is ad hoc or strong, insiders with access to the hash function are heavily trusted.

Figure 24. Gross Compliance assessment of hashing as a UID technology.

6.3 Encryption

Using encryption to produce a UID involves computing a number from source information. De-duplication is then performed by simply matching UID values. This is the same as hashing (Section 6.2), except with encryption there exists a "key" such that whoever has the key can reverse the process to take a UID and reveal some (or all) of the source information that produced it.

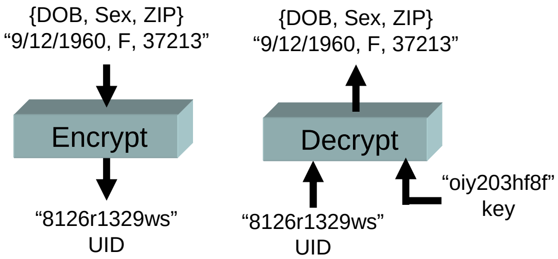


Figure 25. Example of making a UID by encrypting {date of birth, gender, ZIP}. With the key, the process is reversed to reveal the original source information.

Figure 25 provides an example of making a UID by encryption the fields {*date of birth*, *gender*, *ZIP*}. Specifically:

encrypt(9/12/1960, F, 37213) = "8126r1329ws"

Then,

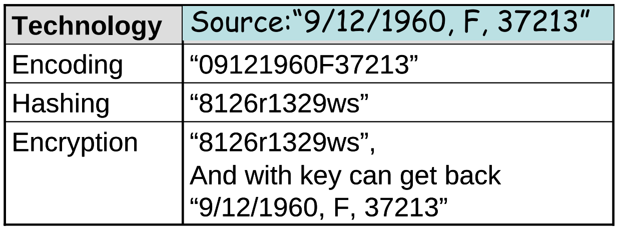
decrypt(*key*, "8126r1329ws") = "9/12/1960, F, 37213"

Encrypted UIDs, as with hashing, are consistently produced. Each time the encryption function is given the same input, it produces the same UID.

A vendor can create their own encryption function, but it has been shown that these "ad hoc" approaches can be reversed, especially if someone is highly motivated to do so. [This is the same as was discussed with hashing in Section 6.2.] Protection using an ad hoc encryption function is good only as long as no one learns the actual encryption function used. Rather than using ad hoc encryption functions, cryptographically "strong" encryption methods are highly recommended. With a strong encryption function, everyone can examine the method being used, but even with intense inspection, it has been proven that no one can reverse the process without the key [22].

Encryption functions have the property that they do not preserve the natural ordering typically found in source values. [This is the same as was discussed with hashing in Section 6.2.] Two consecutive values (e.g. ZIP codes 37212 and 37213) tend to have radically different encrypted values (e.g., "x41768" and "z1Rx5G"). This is good for privacy, but can be bad for utility.

Encoding, hashing and encryption are very similar, as shown in Figure 26. However, encoding tends to visibly reveal source information where as hashing and encryption values do not. Encryption, in comparison to hashing, has a key that can reverse the process.



Technology	Source:"9/12/1960, F, 37213"
Encoding	"09121960F37213"
Hashing	"8126r1329ws"
Encryption	"8126r1329ws", And with key can get back "9/12/1960, F, 37213"

Figure 26. Comparison of encoding, hashing, and encryption. Encoding tends to transparently reveals the original source values. Encryption has a key that can reverse the process.

See Figure 27 and Figure 28 for a gross assessment of encryption as a UID technology. Issues related to utility and the warranty statement appear in Figure 27. Issues related to privacy and the compliance statement appear in Figure 28. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

ENCRYPTION –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information		*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information is a reliably captured biometric.
Client confidence and trustworthiness	▦	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Encrypted UIDs tend to appear cryptic, which can instill Client and intaker confidence. However, problems can emerge in cases where the requested source information is sensitive, notwithstanding the cryptic appearance of the UID itself. Educating Clients and those who perform intake regularly and/or issuing privacy notices may help. The existence of a key that can unlock Client information may also reduce Client confidence.
Inflated accounting	░	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Count inflation can occur in cases where a Client provides different source information on different visits (see comments for non-verifiable source information above). In these cases, different UIDs are generated and therefore will not match to each other even though they are assigned to the same Client.
Deflated accounting	■	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation can occur in cases in which a Client provides incomplete or missing information or different source information on different visits, or in which a bad method is used for generating UIDs. In these cases, the same UID is generated for different Clients and therefore visit information will combine inappropriately, generating serious accounting problems. Deflation is more likely than inflation.

…continued on next page …

Handling bad or missing input	■	*What is the effect of bad, incomplete, or missing source information on performance?* Typing mistakes that are go uncorrected, as well as incomplete or missing information, can generate different UIDs for a Client than would have been generated with complete and properly entered information. This tends to inflate accounting by generating spurious UIDs for Clients having multiple visits. On the other hand, incomplete and missing information is likely to deflate accounting because different Clients whose entries are missing the same information may have the same UID.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
□	No problem likely, or not applicable

Figure 27. Gross Warranty assessment of encryption as a UID technology.

ENCRYPTION –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	■	*What vulnerabilities exist for the intimate stalker?* In typical cases where demographics is the source information used with encryption, serious problems may exist. Access to the encryption function, or the key with the decryption function, can allow the intimate stalker (working with a compromised insider) to generate a Client's UID, and then to use the UID to identify the Client's Shelter location in the Dataset. Control and auditing of the encryption and decryption functions are important to thwarting this problem.
Re-identification: Linking	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because demographics tend to be the source information used with encryption and demographics appear in other available data, linking tends to be a problem if access to the encryption and decryption functions are not controlled and audited. Practices should limit and account for encryption and decryption use. Risk analysis should be based on the re-identification of demographics over the actual population from which Clients are drawn.
Re-identification: Dictionary Attack	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* A dictionary attack can be done by executing the hash function over all legal combinations of source information. For any generated UID that matches a UID in Dataset, the Client's source information is learned. This may pose a serious problem depending on source information and encryption method used. A combination dictionary-attack and linking attack can also be a problem. For example, suppose some other data (Other Data) is to be linked to a Dataset in which UIDs are encrypted using source information and the same source information appears in Other Data. UIDs can be produced for the source information in Other Data, and then, the UIDs in Dataset are matched to the UIDs in Other Data to link Client data to Other Data. Practices should limit and account for uses of the encryption function and also for key use.
Re-identification: Reversal	▒	*What is involved in reverse engineering the UID construction method?* If a "strong" encryption function is used, then it is highly unlikely that the method will be reversed. For this reason, strong rather than ad hoc encryption functions should be used.

…continued on next page …

Exposure	▩	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of encrypted UIDs means there exists a key that can unlock the UIDs without permission, thereby increasing Client risks beyond the HMIS context.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

System Trust
Which parties are heavily trusted?

If the encryption function is ad hoc (not strong), then the function itself is heavily trusted in the belief that no one, no matter how heavily motivated, will reverse the function. It also requires trusting the developer of the ad hoc encryption function.

Any party that has access to the decryption key is heavily trusted.

Additionally, no matter whether the encryption function is ad hoc or strong, insiders with access to the encryption function are heavily trusted.

Figure 28. Gross Compliance assessment of encryption as a UID technology.

6.4 Scan Cards/RFID

Using Scan Cards as a UID technology involves issuing a card containing a UID to each Client who presents for service. The card can store a photo, serial#, randomly assigned number, and/or demographics. Figure 29 shows a depiction of a scan card in which only a serial number and picture appear.

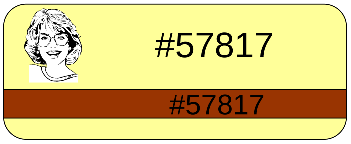


Figure 29. Depiction of a scan card with a serial number and photograph visible. The magnetic strip stores the serial number, but the serial number stored on the strip is not visible to the naked eye.

Scan cards that have a magnetic strip on one side resemble credit cards. Information is stored on the magnetic strip that can be read by a card reader even though the information is not visible to the human eye. In fact, these magnetic strips are typically readable by most card readers, and therefore, the ability to read scan cards is not limited to card authorized readers. Card readers outside those located at Shelters could read the cards.

Radio frequency identification (RFID) cards have no magnetic strip. Information is still stored within the card and can be read by an RFID reader. But unlike magnetic strip cards, RFID content intended for one reader is not as easily read by other readers. In fact, expensive RFID cards and readers offer exclusive protection. Only authorized readers are easily able to read specific kinds of cards. Finally, RFID cards come in a variety of sizes, some smaller than a dime (and many cost less than a dime too).

The decision of what appears printed on the card is important in assessing its use as a UID technology. If Shelter information appears, others may learn information about the Client from merely viewing the card.

The information stored on the card is the UID. The source information can be a randomly assigned number, demographics, or some other value. If a serial or random number is assigned, the Planning Office will most likely have to coordinate issuances of numbers across Shelters.

See Figure 30 and Figure 31 for a gross assessment of using scan cards as a UID technology. Issues related to utility and the warranty statement appear in Figure 30. Issues related to privacy and the compliance statement appear in Figure 31. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

SCAN CARDS / RFIDs –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	▓	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Assume non-verifiable information is the basis for a UID stored on a card. Then, if the Client consistently uses the card, no problem is likely. But if cards are borrowed or swapped, or if Clients have multiple cards issued with different UIDs (e.g., with card replacement), problems are likely.
Verifiable source information	░	*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information that can be stored on a scan card is a reliably captured biometric (see Section 6.5). Printing photographs on the card may be considered a means to verify identity, but intake personnel must be trained to actually verify appearance.
Client confidence and trustworthiness	█	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Scan cards may pose serious problems based on the existence of the card and on information appearing on the card. Assume a Client was issued a card and subsequently returned home to the abuser. The card, if found, can instigate trouble. Further, if information about the location of the Shelter or the UID itself are actually printed on the card, the intimate stalker may gain sensitive information.
Inflated accounting	█	*What are the circumstances under which de-duplication is likely to inflate the accounting?* The issuance of additional scan cards to the same person can inflate the count if new cards have different UIDs. Accompanying practices should address how registration of cards is done and how lost cards are handled. This is likely to be a common problem. Swapping cards among Clients does not actually inflate the count, but it does generate false visit patterns in which visits of one Client are incorrectly associated with another.
Deflated accounting	░	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation is not likely to occur with scan cards unless the information used to generate the UID associated with the card is badly chosen. Most ways in which UIDs stored on scan cards are likely to be generated pose no problem. For example, randomly generated UIDs would not pose a problem. But if source information produces the same UIDs for different people (i.e., different cards assigned to different Clients but having the same UIDs), then visit information would combine inappropriately, generating accounting problems.

…continued on next page …

Handling bad or missing input	░	*What is the effect of bad, incomplete, or missing source information on performance?* Bad or missing information is not likely to effect the performance with scan cards unless the information used to generate the UID associated with the card is badly chosen. Most ways in which UIDs stored on scan cards are likely to be generated pose no problem. For example, randomly generated UIDs would not pose a problem. But if the method relied on source information that could have bad or missing information, then deflated accounting is possible because different Clients whose entries are missing the same information may have the same UID.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 30. Gross Warranty assessment of using scan cards as a UID technology.

SCAN CARDS / RFIDs –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	▩	*What vulnerabilities exist for the intimate stalker?* In cases where printable information appearing on the card itself includes Shelter location or the UID itself, viewing the card may reveal sensitive information. Practices should address information appearing on the card and its possible use by the stalker. Care nust also be taken that the UID dies not reveal or use information available to the intimate stalker.
Re-identification: Linking	▩	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* If demographics are stored or printed on the card, linking will be a problem. Risk analysis should be based on demographics over the actual population from which Clients are drawn. However, other possibilities, beyond demographics, exist as the basis for providing UIDs for scan cards.
Re-identification: Dictionary Attack	░	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* If the UID associated with a Scan Card is just a random number, then a dictionary attack is not likely. However, if the UID associated with a Scan Card uses demographics or biometrics, then vulnerabilities may exist (see Section 5.3 and Section 6.5).
Re-identification: Reversal	░	*What is involved in reverse engineering the UID construction method?* If the UID associated with a Scan Card is just a random number, then reversal is not likely. However, if the UID associated with a Scan Card uses encoding or hashing, then vulnerabilities may exist (see Section 6.1 and Section 6.2)).
Exposure	■	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of the Scan Card in the Client's possession and any information printed on the card can expose a Client's consumption of Shelter services to an intimate abuser, for example. Care should be taken about the information printed on the card. The severity of this problem can be easily resolved by avoiding such printing on the card.

■	Most severe/difficult problem
▩	Moderate problem
▨	A problem
░	May be a problem
	No problem likely, or not applicable

System Trust
Which parties are heavily trusted?

Assuming a scan card stores only a randomly assigned number and no printed information is visible, then scan cards place trust in Clients in the belief that Clients will use the same card on recurring visits, will not swap cards and will provide the same source information on card replacement or re-issuance.

Figure 31. Gross Warranty assessment of using scan cards as a UID technology.

6.5 Biometrics

Using a biometric as source information for a UID technology has the advantage that the biometric is something always present with the Client and that typically does not change. The most common biometric is a fingerprint. Figure 32 shows how a fingerprint is used as source information. A fingerprint can be used as source information to a hash or encryption function or the fingerprint itself can be the UID.

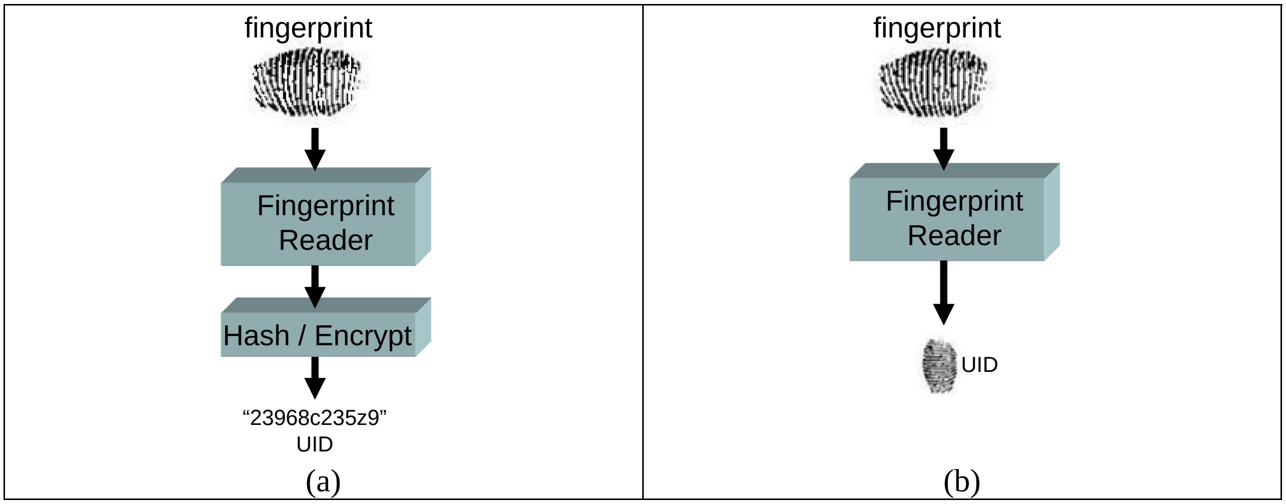


Figure 32. Fingerprint as source information to a hash or encryption function to generate a UID (a); or, used as the UID iteself (b).

Fingerprint readers have become inexpensive and as a result, fingerprint reading is becoming popular for all kinds of new uses, such as a way to gain access to a car or a refrigerator or to use a computer keyboard. Of course, inexpensive capture devices tend to be horribly inaccurate, but reasonably priced devices perform reasonably well. It is important to test the accuracy of a fingerprint system on the population with which it will be used. The combination of a particular fingerprint system with a specific population should be checked for consistency and accuracy. Check that the same person is recognized to be the same person (and not someone else). Also confirm that a person who has been in the system continues to be recognized (and not considered a new person).

For some explained and unexplained reasons, there are some people whose fingerprints cannot be reliably captured [23]. Finger cuts, scars, amputations, disease, infection, and overall disabilities and abnormalities can pose fingerprint capture problems. Hands having excessive moisture or dryness can frustrate fingerprint capture. Unofficial FBI statements claim that persons involved with certain drugs and persons who regularly scrape their fingertips on abrasive surfaces, such as concrete, cannot be reliably fingerprinted. If so, some homeless people who spend significant time on concrete sidewalks may be difficult to fingerprint.

If fingerprint images are captured and used as UIDs, Shelters and Planning Offices would maintain a de facto fingerprint database of Clients. The existence of such a database may invite linking requests (unofficial and official), especially from law enforcement. Whether matching latent prints to a crime scene or confirming identity, law enforcement requests serviced by Shelters may alter how some Shelters and Clients have historically viewed the homeless service environment. An increase in court orders demanding copies of Client prints, the UID construction method, and all Client UIDs is a likely possibility.

See Figure 33 and Figure 34 for a gross assessment of using biometrics in UID technologies. Issues related to utility and the warranty statement appear in Figure 33. Issues related to privacy and the compliance statement appear in Figure 34. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

BIOMETRICS (fingerprints) –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information		*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Does not require non-verifiable source information from Clients.
Verifiable source information		*Can problems occur if the UID is based on verifiable source information?* A biometric that can be consistently and reliably captured can provide independent, invariant Client information that is not likely to be bad or to cause problems.
Client confidence and trustworthiness	░	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* UIDs based on biometrics are generally invariant to Client trust though some attention should be given to establishing Client acceptance of what may be perceived as an invasive process. Otherwise, Clients may purposefully try to generate bad captures, if possible, in an attempt to thwart the system.
Inflated accounting	▓	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Inexpensive technology or poor quality biometrics can inflate counts when the same person generates different UIDs. In most cases, Clients are likely to undergo a registration process to generate a database of known Clients. Then, when a Client appears on a subsequent visit, if the presenting biometric is not found, the count is not inflated, but administering the process is slowed by having to repeat captures until a matching biometric is found. Attention should be spent on testing the accuracy of the biometric capture on the specific Client population. Sometimes, using multiple captures can improve results. Another possible remedy is to use better technology.
Deflated accounting	▓	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Inexpensive technology or poor quality biometrics can deflate counts when multiple people map to the same UID. Attention should be spent on testing the accuracy of the biometric capture on the specific Client population. Sometimes, using multiple captures can improve results. Another possible remedy is to use better technology.

…continued on next page …

Handling bad or missing input	░	*What is the effect of bad, incomplete, or missing source information on performance?* If the biometric is presented, the information provided is not typically bad or missing, even though the provided information may not necessarily be properly captured. Care must be taken to test the accuracy and consistency of the biometric system on the specific Client population. Procedures should address how misses and mismatches are handled (see discussion above on inflated and deflated accounting).

█	Most severe/difficult problem
▓	Moderate problem
▒	A problem
░	May be a problem
□	No problem likely, or not applicable

Figure 33. Gross Warranty assessment of using biometrics in UID technology.

BIOMETRICS (fingerprints) –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	░	*What vulnerabilities exist for the intimate stalker?* In cases where the biometric capture program can be made to work with artificial or previously captured images, rather than live capture, a problem may exist. For example, a stalker having access to a fingerprint image of a Client and the fingerprint capture program could generate a UID. The risk of such an occurrence is increasing as the number of fingerprint capture devices become more commonly used in daily life. Ways that non-live prints may be used with the biometric system should be understood and addressed.
Re-identification: Linking	▓	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* As the use of biometrics becomes increasingly popular in society, the ability to link other data to biometric data increases. For example, as more people are fingerprinted and inexpensive fingerprint capture devices become increasingly common, many more databases to which to link fingerprints will exist. A UID that uses a fingerprint as source information may not necessarily store an image of the fingerprint sufficient for linking to other fingerprint databases; this depends on the specifics of the method used for constructing the UID from the fingerprint. Care should be taken to understand this method and related risks. The fingerprint databases maintained by law-enforcement require particular consideration. For example, one cannot simply refuse to obey a court order demanding copies of captured Client fingerprints, the UID construction method, and all associated UIDs for the purpose of matching Client prints against a criminal database. On the other hand, if the database did not exist, no such request could be made. A privacy policy and notice informing Clients of potential risks should be considered.
Re-identification: Dictionary Attack	░	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* In the general case, exhaustive search is not likely though this should be confirmed in any particular solution proposed. However, a dictionary attack using a large biometric population database (e.g., law-enforcement fingerprint database) may re-identify Clients whose fingerprints are already captured there. Risks associated with linking prints with law-enforcement data should be assessed, and consideration given to the possibility of receiving a court order for such. In these cases, the method that related prints to UIDs would be used with image not live-scan data, a difference which may matter to some proposed solutions. A privacy policy and notice informing Clients of potential risks should be considered.
Re-identification: Reversal	░	*What is involved in reverse engineering the UID construction method?* Reverse-engineering a method that converts a biometric to a UID is not necessarily as fruitful as just using the method to make the associations (see linking and dictionary attack above). However, if the UID method requires live scan capture, motivation exists to perform the reversal.

…continued on next page …

Exposure	■	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of captured biometrics on Clients can expose Client information to be the subject of court orders and search by law-enforcement and others.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

System Trust
Which parties are heavily trusted?

Shelters and Planning Offices are heavily trusted to design systems in such a way that either linkages to law-enforcement databases are highly unlikely, or the Client is clearly informed.

Figure 34. Gross Compliance assessment of using biometrics in UID technology.

6.6 Consent (permission technology)

"Consent" as a UID method refers to a permission technology. The database technology that stores Client information at Shelters includes a permission flag which records whether a Client has granted permission to have her data forwarded to a Planning Office. Only the information of Clients who have granted permission is forwarded. The information of all other Clients is not forwarded. Figure 35 provides an example in which Ann and Claire have granted permission, and therefore their information is forwarded, but Betty and Donna have not granted permission, so their information is not forwarded.

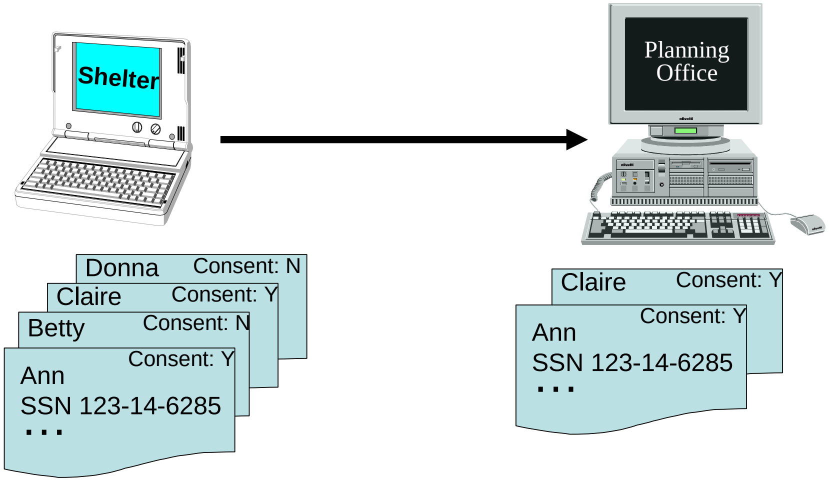


Figure 35. Consent used as the basis for deciding which Client information is forwarded to the Planning Office. Information provided to the Planning Office is explicitly identified by name and Social Security number.

Information provided to the Planning Office when consent is used typically has explicitly identified UIDs, such as name and Social Security numbers. Of course, some other UID could be used, but such cases are covered in those sections of this writing. This section addresses the situation in which the basis of de-duplication is matching explicitly identified information (e.g., name and Social Security number) that is made available because the Client has granted permission for its use.

De-duplication involves matching explicitly identified information, such as names; but matching names is horribly problematical. Clients may use nicknames or exchange first and middle names. Misspellings may be common. A well-known de-duplication method used for matching names is Soundex, which matches spellings that may look or sound similar [24]. Using Soundex, the names "James" and "John" are hashed to J52 and J5, respectively, but the names "John," "Jane" and "Jean" are all hashed to the same "J5" value. Therefore, Soundex can frustrate de-duplication.

Of course, consent allows more identifying fields to be shared, so de-duplication problems experienced with name-only matching, for example, may be augmented to exploit multiple fields of information in an attempt to account for recording errors. It should be noted however, that methods that perform such matching reliably are not trivial [25] and should be used with care.

Consent as a UID technology places Clients in the situation of sharing risks and liabilities with Shelters and Planning Offices. The use of explicit UIDs dramatically increases risks for Clients over that of other UID technologies, so standard privacy policy notices discussed earlier in Section 4 are not sufficient; more rigorous versions are needed. It is important to completely and accurately disclose the uses of Dataset and circumstances of sharing. Clients should understand HMIS data uses as well as any secondary data uses of Dataset. (Secondary uses are those situations in which Dataset, in part or whole, is shared beyond the HMIS context.) Clients must be sufficiently informed beforehand of data sharing practices; and conversely, Shelter and Planning Office practices must respect and enforce this originally agreed upon characterization.

Handling situations in which Clients do not grant permission must be considered. Clients cannot be coerced into providing permission, and Clients cannot be denied services for refusing to grant permission. Yet, Clients who do not grant permission deflate the accounting.

Inconsistent permissions may go undetected. A Client may grant permission at one Shelter and not at another, thereby providing an incomplete accounting. These situations should be considered, as well as the ability of a Client to revoke permission previously granted and vice versa.

See Figure 36 and Figure 37 for a gross assessment of using consent as a UID technology. Issues related to utility and the warranty statement appear in Figure 36. Issues related to privacy and the compliance statement appear in Figure 37. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

CONSENT –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information		*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information can be a Social Security number verified to a Social Security card, or a driver's license number. A biometric could also be used.
Client confidence and trustworthiness	░	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Requesting the Client's consent to share captured information tends to build Client confidence because Clients tend to feel in control of their information and believe that the process is transparent. In reality, the consent may place no limits on secondary sharing beyond the HMIS context and intake personnel may learn such. Care should be taken that the accompanying consent form and privacy notices accurately inform Clients of actual data flow, sharing practices, privacy safeguards, and Client options.
Inflated accounting	▒	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Because of the increased Client confidence consent may elicit, Clients may be more willing to provide more sensitive detailed information than with other technologies, but having more information on which to match Client visits does not necessarily lead to more accurate de-duplication. The specifics of how de-duplication is performed matters. For example, name matching can be particularly problematical because of variations in the ways Clients may present their names (e.g., interchanging first and middle names, using nicknames, or different last names), not to mention typographical errors. Using an accurate de-duplication instrument is important.
Deflated accounting	▒	*What are the circumstances under which de-duplication is likely to deflate the accounting?* As was stated above with inflated accounting, having more information on which to match Client visits does not necessarily lead to more accurate de-duplication. The specifics of how de-duplication is performed matters. For example, name matching using crude algorithms like Soundex can inappropriately match names of different Clients together. Using an accurate de-duplication instrument is important. Clients who do not grant consent can deflate accounting, so additional procedures are needed to handle these cases.

…continued on next page …

Handling bad or missing input	░	*What is the effect of bad, incomplete, or missing source information on performance?* While bad or missing information is always possible, more identifying information is typically collected in these environments allowing for a larger number of data elements to be alternatively used for matching in cases where some information is bad or missing. Name matching tends to be problematical, as discussed, but having more fields on which to compare can help.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 36. Gross Warranty assessment of using consent as a UID technology.

CONSENT –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	■	*What vulnerabilities exist for the intimate stalker?* Because consent tends to allow the collection of more sensitive information, anyone with access can be potentially compromised by the stalker to gain access. Further, secondary sharing tends to increase the number of copies of the information appearing beyond the HMIS context, which in turn, increases the number of people having access.
Re-identification: Linking	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because of the increased Client confidence the consent approach may elicit, Clients may be more willing to provide more sensitive detailed information than with other technologies, and the UID itself is explicitly identifying, thereby making linking a serious problem.
Re-identification: Dictionary Attack	■	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* Because demographics and more sensitive information tends to be stored, a dictionary attack per se appears similar to linking the information to a large, population-based database, which can pose serious problems.
Re-identification: Reversal	■	*What is involved in reverse engineering the UID construction method?* The UID is an explicit identifier (e.g., Social Security number), so there is nothing to reverse. The UID itself reveals the sensitive information that would be the object of the reversal.
Exposure	■	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of demographics and sensitive information on Clients can expose Client information to court orders and search by law-enforcement and others. It is more likely to draw requests for research purposes and administrative oversight in its explicitly identified form. Practices and policies for de-identification and secondary use should be considered. A privacy policy informing Clients of potential risks should be considered.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

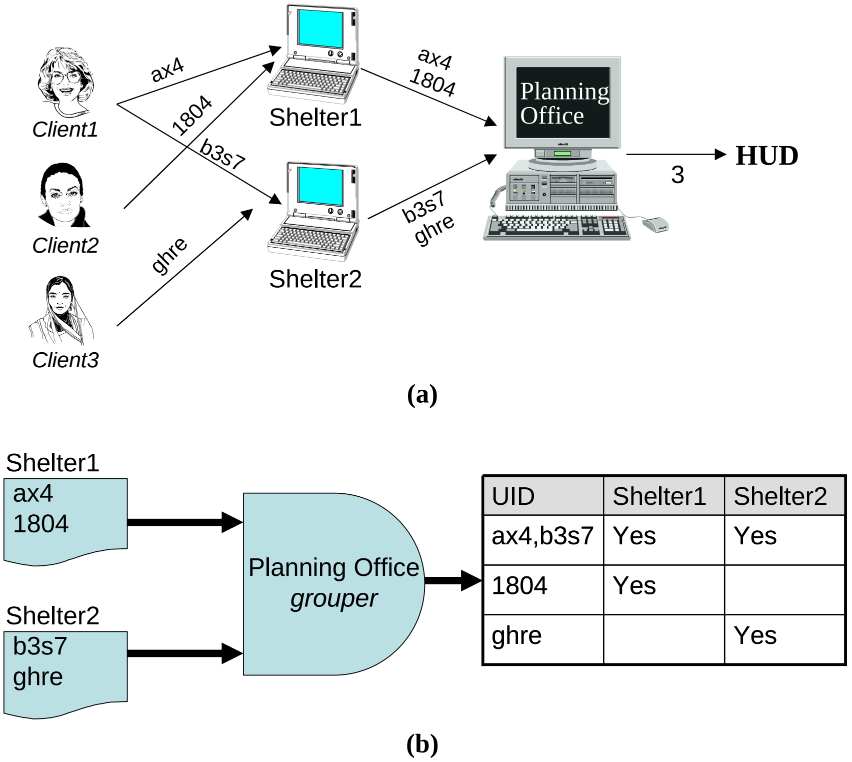
System Trust
Which parties are heavily trusted?

Planning Offices are heavily trusted with the explicitly identified Client data.

Figure 37. Gross Compliance assessment of using consent as a UID technology.

6.7 Inconsistent hashing

Inconsistent hashing works similar to regular hashing (Section 6.2) except each Client gets a different hash number at each Shelter. The Planning Office has a special methods that groups UIDs for the same Clients together ("grouper"). Figure 36 shows different Clients visiting different Shelters. Each Client is assigned a different UID at each Shelter, thereby providing an inability to link information across Shelters without the special grouping method available to the Planning Office. The Planning Office is able to use its grouping method to link UIDs belonging to the same Clients.



(a)

UID	Shelter1	Shelter2
ax4,b3s7	Yes	Yes
1804	Yes	
ghre		Yes

(b)

Figure 38. Depiction of inconsistent hashing used as a UID technology. Above (a) shows Clients assigned different UIDs at Shelters, which are forwarded to the Planning Office. Below (b) shows the Planning Office using a special method to group UIDs belonging to the same Clients.

Inconsistent hashing can be achieved in a variety of ways that primarily differ by the amount of trust given the Planning Office, which holds the grouping method [26].

The most naïve approach, which should be avoided, uses public key encryption. The Planning Office issues a public key unique to each Shelter. UIDs are encrypted with the Shelter keys, making each UID Shelter specific. Because the Planning Office has the matching private key for each Shelter, the Planning Office can reveal the original UID source information, which is then used for direct matching. This approach has the undesirable side effect that the source information (e.g., Social Security number) is revealed to the Planning Office.

A better approach uses strong hashing (Section 6.2) to protect source information from being explicitly revealed, but this approach requires more computation. Each Shelter has a unique strong hash function to generate Client UIDs. The Planning Office holds a copy of each Shelter's hash function. After the Shelters provide their UIDs, the Planning Office hashes the UIDs by every other Shelter's hash function. This takes advantage of the property that the order in which hashes of hash values are performed does not matter. For example, consider Figure 38:

Shelter 1's hash of b3s7 = Shelter 2's hash of ax4

but

Shelter 1's hash of ghre ≠ Shelter 2's hash of ax4 or 1804.

There is concern with this approach. Because the Planning Office has a copy of each Shelter's hash function, a dictionary attack at the Planning Office is possible.

See Figure 39 and Figure 40 for a gross assessment of using consent as a UID technology. Issues related to utility and the warranty statement appear in Figure 39. Issues related to privacy and the compliance statement appear in Figure 40. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

INCONSISTENT HASHING –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information		*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information is a reliably captured biometric.
Client confidence and trustworthiness	▒	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Like hashed UIDs, inconsistently hashed UIDs tend to appear cryptic, which can instill Client and intaker confidence and thereby avoid problems. Further, because UIDs are different across Shelters (and can even be different on multiple visits to the same Shelter), additional Client and intaker confidence can be attained. Problems may emerge based on the sensitivity of requested source information despite the cryptic appearance of the UID itself. Educating Clients and those who perform intake regularly and/or issuing privacy notices may help.
Inflated accounting	■	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Count inflation can occur in cases where a Client provides different source information on different visits. In these cases, different UIDs are generated and therefore will not match to each other even though they are assigned to the same Client. Count inflation can also occur in cases in which a Client provides incomplete or missing information or different source information on different visits, thereby producing different UIDs across Shelters.
Deflated accounting	▒	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation is possible when different Clients provide identical complete and incomplete information. A glaring example occurs for Clients in which all relevant source information is missing. Attention should be paid to how these situations are addressed in UIDs across Shelters. Count inflation is mor likely than deflation.

…continued on next page …

Handling bad or missing input	■	*What is the effect of bad, incomplete, or missing source information on performance?* Typing mistakes and incomplete or missing information can generate different UIDs for a Client than would have been generated with complete and properly entered information. This tends to inflate accounting by generating spurious UIDs for Clients having multiple visits. Incomplete and missing information also tend to inflate accounting. Inflation is more likely than deflation.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 39. Gross Warranty assessment of using inconsistent hashing as a UID technology.

INCONSISTENT HASHING –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	▩	*What vulnerabilities exist for the intimate stalker?* Because each Shelter has a different UID for the same Client, access to Shelter information is limited to a Shelter-by-Shelter basis. Vulnerabilities that are able to be exploited by an intimate stalker are limited to the Planning Office, which controls the grouping method. Vulnerabilities at the Planning Office may be addressed by control and audit of the grouping method and grouped UIDs.
Re-identification: Linking		*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because a different UID is generated at each Shelter a Client visits, and the UIDs are not used outside HMIS data, unauthorized linking is not likely. Practices should limit and account for hash function use.
Re-identification: Dictionary Attack	▦	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* Because a different UID is generated at each Shelter a Client visits, and the UIDs are not used outside HMIS data, a dictionary attack is not likely to be fruitful except at the Planning Office. Colluding Shelters (or access to the Planning Office's grouper) can lead to re-identifications. Vulnerabilities at the Planning Office may be addressed by control and audit of the grouping method and grouped UIDs.
Re-identification: Reversal	░	*What is involved in reverse engineering the UID construction method?* When using strong hash functions, reversal is not usually an issue. But if the Shelters' hash functions are available to unlimited use by the Planning Office, care must be taken to control or limit hash function use to avoid unwanted dictionary attacks (discussed above) or reverse compilations. (A dictionary is more likely than an attempt to reverse compile the function.)
Exposure		*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of inconsistently hashed UIDs used only in the HMIS-context is not likely to expose Clients to additional risks beyond those mentioned above.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

System Trust

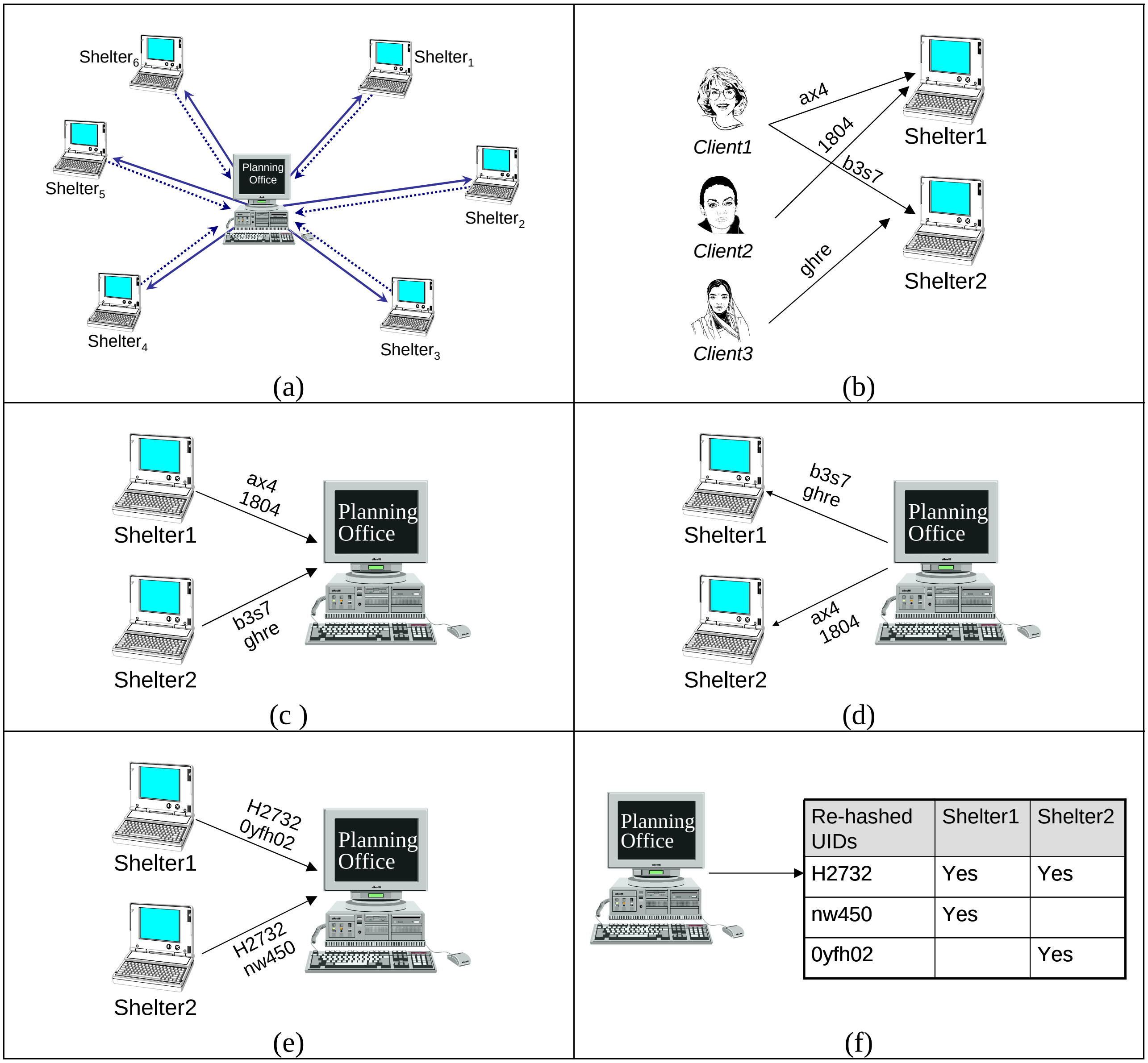
Which parties are heavily trusted?

Planning Offices are heavily trusted to control access and use of the grouping method that links different UIDs to the same Clients.

Figure 40. Gross Compliance assessment of using inconsistent hashing as a UID technology.

6.8 Distributed query

Using distributed query, de-duplication is done on Shelter computers interacting with the Planning Office computer over a network. There are multiple ways this can be achieved. An example analogous to answering AHAR questions (Section 3.6) directly over the network is available at [27]. Another way to use distributed query is described in Figure 41 using an approach that resembles inconsistent hashing (Section 6.7) except the hash functions remain on the Shelter computers.



Re-hashed UIDs	Shelter1	Shelter2
H2732	Yes	Yes
nw450	Yes	
0yfh02		Yes

Figure 41. Distributed query (a) overview showing that Shelter computers communicate directly with the Planning Office computer. A step-by-step example of de-duplication appears in (b) through (f). Clients appear at Shelters in (b). Shelters report inconsistent hashed UIDs to Planning Office in (c). Planning Office requests each Shelter to compute the hash of every other Shelter's UIDs in (d) and Shelters respond in (e). Planning Office then compares results in (f).

In Figure 41 (b), Clients are given unique UIDs at each Shelter using strong hash functions (Section 6.2). Client 1, for example as UID ax4 at Shelter 1 and b3s7 at Shelter 2. UIDs are reported to the Planning Office in (c). The Planning Office then sends the UIDs to all the other Shelters to be re-hashed in (d). This takes advantage of the property that the order in which hashes of hash values are performed does not matter.

Shelter 1's hash of b3s7 = Shelter 2's hash of ax4

but

Shelter 1's hash of ghre ≠ Shelter 2's hash of ax4 or 1804.

In (e), the Shelters provide the re-hashed UIDs back to the Planning Office, which matches them in (f) to show distinct visit patterns.

One concern with this system is the need to have Shelter computers on-line. One never knows when a machine may become unavailable due to repair. One strategy to limit availability problems is to perform the computation monthly, so that interim values can be used to offset any missing information needed for the yearly accounting. In locations where Shelters tend to use commercial or the same service providers to maintain Client data, Shelter information should be reliably available.

See Figure 42 and Figure 43 for a gross assessment of using consent as a UID technology. Issues related to utility and the warranty statement appear in Figure 42. Issues related to privacy and the compliance statement appear in Figure 43. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

DISTRIBUTED QUERY –WARRANTY (UTILITY) STATEMENT

Non-Verifiable source information	■	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Serious de-duplication problems are likely if Clients provide non-verifiable source information inconsistently. On the other hand, source information that is not truthful, but consistently provided, is typically not a problem.
Verifiable source information		*Can problems occur if the UID is based on verifiable source information?* Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not correct, but consistently verified on each visit, no problems are likely. An example of invariant verifiable Client information is a reliably captured biometric.
Client confidence and trustworthiness		*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* The fact that data are minimally shared from locally stored Shelter data tends to build Client and intaker confidence sufficient to avoid problems. Care should still be taken to limit the sensitivity of requested source information regardless. Educating Clients and those who perform intake regularly and/or issuing privacy notices may help.
Inflated accounting	■	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Count inflation can occur in cases where a Client provides different source information on different visits. In these cases, different UIDs are generated and therefore will not match to each other even though they are assigned to the same Client. Count inflation can also occur in cases in which a Client provides incomplete or missing information or different source information on different visits, thereby producing different UIDs across Shelters.
Deflated accounting	▒	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Count deflation is possible when different Clients provide identical complete and incomplete information. A glaring example occurs for Clients in which all relevant source information is missing. Attention should be paid to how these situations are addressed in UIDs across Shelters. Count inflation is more likely than deflation.

…continued on next page …

Handling bad or missing input	■	*What is the effect of bad, incomplete, or missing source information on performance?* Typing mistakes that are go uncorrected, as well as incomplete or missing information, can generate different UIDs for a Client than would have been generated with complete and properly entered information. This tends to inflate accounting by generating spurious UIDs for Clients having multiple visits. Incomplete and missing information also tend to inflate accounting. Inflation is more likely than deflation.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

Figure 42. Gross Warranty assessment of using distributed query as a UID technology.

DISTRIBUTED QUERY –COMPLIANCE (PRIVACY) STATEMENT

Intimate Stalker	░	*What vulnerabilities exist for the intimate stalker?* because information is locally stored at Shelters and UIDs are only generated and used during sharing, a problem is not likely. Access to information is limited to a Shelter-by-Shelter basis.
Re-identification: Linking		*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Because information is kept under Shelter control, unauthorized linking beyond the Shelter itself is highly unlikely. It should be noted that Shelters have always had the ability to link Client data, irregardless of HMIS, because Shelters tend to capture complete, explicitly identified information.
Re-identification: Dictionary Attack		*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* Because information is kept under Shelter control, a dictionary attack is highly unlikely.
Re-identification: Reversal		*What is involved in reverse engineering the UID construction method?* Because strong hashing is used and information is kept under Shelter control, there is no globally available "UID" per se so there is nothing to reverse. If strong hashing is not used, then vulnerabilities may exist (see Section 6.2).
Exposure		*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* The existence of information locally controlled by Shelters is not likely to expose Clients to additional risks than already exists with storage and use of Shelter information.

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
	No problem likely, or not applicable

System Trust
Which parties are heavily trusted?

Shelters are trusted to have computers on-line and available.

Figure 43. Gross Compliance assessment of using distributed query as a UID technology.

6.9 Summary Results

While many other factors must be included to determine which technology is appropriate for the Shelters and Planning Office in a particular region, the gross assessments in the previous section suggested that inconsistent hashing, distributed query and (regular) hashing may be easier to bundle with policies and best practices to get an effective solution. Scan cards, encryption, and biometrics create new kinds of risks to consider. Consent and encoding are technically the simplest to implement but harbor serious dangers to overcome. Biometrics is the only technology that uses source information that does not require Clients to be trusted to provide truthful and consistent source information; all the other technologies tend to require Clients to provide non-verifiable, complete and consistent information (or confirm it) on each visit. Figure 44 contains a quick summary of the results found across the gross assessment of UID technologies. While shadings may identify some problems as being of severe or moderate concern, these problems may be sufficiently addressed with straightforward practices, policies, or technology decisions.

	UTILITY						PRIVACY				
UID TECHNOLOGY	Non-verifiable source	Verifiable source	Client Trust	Inflate Accounting	Deflate Accounting	Bad or missing info	Intimate stalker	Linking	Dictionary attack	Reverse engineer	Expose new issues
Encoding	■	░	▩	░	■	■	■	■	■	■	▩
Hashing	■	□	░	░	■	■	■	■	■	▦	□
Encryption	■	□	▦	░	■	■	■	■	■	▦	▩
Scan Cards/RFID	▩	░	■	■	░	░	▩	▩	░	░	■
Biometrics	□	□	░	▦	▦	░	▦	▩	▦	▦	■
Consent	■	□	░	▦	▦	░	■	■	■	■	■
Inconsistent Hash	■	□	░	■	░	■	▩	□	▦	░	□
Distributed Query	■	□	□	■	░	■	░	□	□	□	□

■	Most severe/difficult problem
▩	Moderate problem
▦	A problem
░	May be a problem
□	No problem likely, or not applicable

Figure 44. Summary of gross assessments of UID technologies, showing utility (warranty) and privacy (compliance) issues.

Of course, details matter. The gross assessments could not provide a complete picture because decisions based on best practices and acceptable policies and particular technology implementations could not reasonably be included in one document. However, the gross assessments that are provided give a framework for reasoning about technical solutions and their issues in generating and matching UIDs. The overall lessons learned appear in Figure 45 and Figure 46.

Non-Verifiable source information	*If a UID is based on non-verifiable source information provided by the Client that is not truthful or is inconsistently used, what happens?* Consistent use of the UID by the Client, irregardless of whether the source information is truthful, is important for avoiding problems. As long as a Client uses the same UID and only that UID, problems can be avoided.
Verifiable source information	*Can problems occur if the UID is based on verifiable source information?* Consistency, not truthfulness, is paramount to avoiding problems. Using invariant Client information that can be consistently verified on each visit is likely to avoid problems. Even if the information is not truthful or correct, but is consistently verified on each visit, no problems are likely. Few sources of invariant verifiable source information are known; however, one such example is a reliably captured biometric.
Client confidence and trustworthiness	*How trustworthy is the UID likely to be perceived by Clients (as well as by those who regularly intake Clients)?* Instilling Client trust in the system can contribute to overall performance because Clients are more likely to provide truthful and consistent information to a system they trust. UIDs that appear to be cryptic (e.g., hashing, encryption, inconsistent hashing) can evoke more confidence than UIDs in which captured information appears transparent (e.g., encoding). Those who conduct the intake of Clients can dramatically influence the perception Clients may have of the system. Intake personnel can encourage Clients to give incorrect information, or even if Clients provide truthful information, intake personnel may record non-truthful information in a belief they are protecting Client privacy. Therefore, educating those who perform intake can be very important to overall performance.
Inflated accounting	*What are the circumstances under which de-duplication is likely to inflate the accounting?* Getting consistent source information can avoid inflated counts and conflicting Client visit information. Also, it is important to test the accuracy of the de-duplication instrument to expose problems and seek better solutions.
Deflated accounting	*What are the circumstances under which de-duplication is likely to deflate the accounting?* Getting consistent source information can avoid deflated counts and conflicting Client visit information. Also, it is important to test the accuracy of the de-duplication instrument to expose problems and seek better solutions.
Handling bad or missing input	*What is the effect of bad, incomplete, or missing source information on performance?* Unintended typing mistakes and missing information are likely to happen in real-world use. While many typing mistakes may be caught by the program in which the information is entered, some allowance has to be made for missing information. Under many real-world scenarios, it may not be possible to accurately answer the information. Therefore, consideration must be given on how to handle these cases.

Figure 45. Summary of Warranty issues found in technology assessments in Section 6.

Re-identification: Linking	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using data linkage?* Linking UIDs and Dataset to other available information requires particular attention to be paid to the demographics on which UIDs may be based. This is particularly important with hashing and encryption if access to the hash or encryption function is not controlled. For example, suppose a voter list is to be linked to a Dataset in which UIDs are hashed or encrypted using Client demographics as source information. The hash or encryption function is used with the records in the voter list to produce a UID for each record; then, the UIDs in Dataset are matched to UIDs in the voter list to re-identify Clients by name. This is a combination dictionary-attack and linking.
Re-identification: Dictionary Attack	*What vulnerabilities exist for re-identification of UIDs (and DataSet) using a dictionary attack?* Dictionary attacks, like linking attacks, can be realized on encoded, hashing, and encryption functions, depending on the source information used and the availability of the source information in other available datasets. Controlling access to the hash or encryption function and key can help. Such control would likely be realized by forcing the function to only run on certain machines for certain named persons. All uses by those people would be logged and the logs routinely checked for inappropriate use. Other security measures can also be implemented.
Re-identification: Reversal	*What is involved in reverse engineering the UID construction method?* Reverse engineering UIDs is not typically the most fruitful kind of attack because cryptographic strong hashing and encryption methods can be used to thwart those attempts, and other approaches tend to require far less technical skill and effort. When considering these kinds of technologies, It is important to use strong methods and not homemade methods whose protection is found in the fact that they are merely unknown or obscure. A highly motivated attacker may be able to defeat these homemade attempts. Additionally, these homemade methods cannot be held to public review (as can the cryptographically strong methods) else they risk being exposed, which further limits the ability to verify the strength of their protection.
Exposure	*What legal or technical risks or liabilities may be introduced based on the existence of the resulting database or UID technology?* Some technologies generate additional kinds of risks by their existence. Scan cards can expose a Client to an intimate attacker. Encryption keys can be back doors to accessing data. The potentially increased collection of data that may be realized from consent makes the data more likely to be requested for secondary uses beyond the HMIS context; and, biometrics, especially fingerprints, can give rise to data sharing with law-enforcement, which is beyond the HMIS context.

…continued on next page …

System Trust
Which parties are heavily trusted?

Individual insiders are heavily trusted when using encoding, hashing or encryption.
System developers are trusted when strong methods are not used (hashing and encryption).
Planning Offices are heavily trusted when using consent or inconsistent hashing.
Shelter computers are heavily trusted when using distributed query.
Clients are heavily trusted when using scan cards.

Figure 46. Summary of Compliance issues found in technology assessments in Section 6.

In summary, this work provides a framework for reasoning about and assessing proposed technical solutions for generating and matching UIDs. Eight categories of technologies (encoding, hashing, encryption, scan cards/RFID, biometrics, consent, inconsistent hash, and distributed query) were examined and a set of recommendations made. While significant differences and trade-offs exist in the use of these technologies, there is no magic solution as much as best practices that must accompany any chosen technology sufficient for it to be shown that there is minimal risk of client re-identification and reasonable correctness in computing an unduplicated accounting when using the technology with accompanying practices.

References

1 U.S. Department of Housing and Urban Development. Homeless Management Information Systems (HMIS); Data and Technical Standards Final Notice. *Federal Register*, Vol. 69, No. 146, July 30, 2004, p. 45888-45934.

2 U.S. Department of Housing and Urban Development. *Homeless Management Information Systems (HMIS) Data and Technical Standards Final Notice; Clarification and Additional Guidance on Special Provisions for Domestic Violence Provider Shelters*. Docket No. FR 4848-N-O3. August 30, 2004.

3 U.S. Department of Housing and Urban Development. Emergency Shelter Grants Allocation History. www.hud.gov/utilities/intercept.cfm?/offices/cpd/homeless/budget/esghistory.pdf as of September 2005.

4 Northeast Ohio Coalition for the Homeless. *Overflowing Shelters: a history and recommended solutions*. April 9, 2005. www.neoch.org/what_to_do_overflowing.htm as of September 2005.

5 U.S. Conference of Mayors, A Status Report on Hunger and Homelessness in America's Cities 2001. www.usmayors.org/uscm/hungersurvey/2001/hungersurvey2001.pdf as of Sept 2005.

6 Markee, P. *Average Daily Census of Homeless Children and Adults Residing in the New York City Municipal Shelter System*. Coalition for the Homeless on behalf of New York City Department of Homeless Services and Human Resources Administration, May, 2002.

7 New York City Independent Budget Office. *Give 'Em Shelter: Various City Agencies Spend Over $900 Million on Homeless Services*. Fiscal Brief, March 2002

8 Conference Report (H.R. Report 107-272) for the Fiscal Year 2002 HUD Appropriations Act (Public Law 107-73).

9 Conference Report (H.R. Report 106-988) for the Fiscal Year 2001 HUD Appropriations Act (Public Law 106-377).

10 Senate Committee Report 107-43 for the Fiscal Year 2002 HUD Appropriations Act (Public Law 107-43).

11 U.S. Bureau of the Census. *1990 Collection and Processing Procedures (Appendix D)* . CD-ROM Technical Documentation Project. University of Michigan. February 1998. www.lib.umich.edu/govdocs/cicdoc/cen90app/append_d.htm as of September 2005.

12 U.S. Bureau of the Census. *1996 National Survey of Homeless Assistance Providers and Clients*. Washington: 1996. www.census.gov/prod/www/nshapc/NSHAPC4.html as of September 2005.

13 M. Burt and L. Aron. *America's Homeless II: Population and Services*. Urban Institute. Washington: 2000. www.urban.org/UploadedPDF/900344_AmericasHomelessII.pdf as of Sept 2005

14 Electronic Privacy Information Center. Comments to HUD on the Matter of HMIS. Sept 2003.

15 The National Network to End Domestic Violence. Comments to HUD on the Matter of HMIS. Sept. 2003

16 National Center for Victims of Crime. Domestic Violence. As of Sept 2005, www.ncvc.org/ncvc/main.aspx?dbName=DocumentViewer&DocumentID=32347

17 U.S. Department of Justice. *Violence by Intimates: analysis of data on crimes by current or former spouses, boyfriends, and girlfriends*. NCJ-167237. March 1998.

18 S. Catania. No safe haven. *Mother Jones*, July/August 2005.

19 National HMIS TA Initiative Documents: AHAR Super Table Shells. As of Sept 2005, www.hmis.info/ta_resources_data.asp?topic_id=11

20 Privacert, Inc. *The Privacert Risk Assessment Server*. Available at www.privacert.com as of Sept 2005. Originally designed and developed by L. Sweeney.

21 Pfleeger, C. *Security in Computing*. Prentice-Hall. Upper Saddle River: 1997

22 Stinson, D. *Cryptography: Theory and Practice*. CRC Press. New York: 1995

23 Ratha, N. and Bolle, R. Automatic Fingerprint Recognition Systems. Springer-Verlag. New York: 2004

24 Russell, R. Soundex. U.S. Patent 1,261,167 April 2, 1918.

25 Record Linkage Techniques -- 1997: Proceedings of an International Workshop and Exposition. National Research Council, 1999.

26 Sweeney, L. *Inconsistent Hashing and the Notion of Single-Use Identifying Numbers*. Carnegie Mellon University, School of Computer Science, Data Privacy Lab White Paper Series LIDAP-WP13. Pittsburgh, PA: 2005.

27 Edo-Eket, S. and Sweeney, L. *Detecting Bio-Terrorist Attacks and Naturally Occurring Outbreaks Over a Distributed Network While Protecting Privacy and Confidentiality: the PrivaSum Protocol*. Carnegie Mellon University, School of Computer Science, Technical Report CMU-ISRI-04-111.

Index