An Overview of the Issues Related to the use of Personal Identifiers

Prepared by Mark Armstrong

HSMD, Statistics Canada

July 7, 2000

Table of Contents

List of Appendices

A. Personal Identification Variables

1. Introduction

The purpose of this report is to provide an overview of existing methods and techniques making use of personal identifiers to support record linkage. Record linkage can be loosely defined as a methodology for manipulating and / or transforming personal identifiers from individual data records from one or more operational databases and subsequently attempting to match these personal identifiers to create a composite record about an individual. Record linkage is not intended to uniquely identify individuals for operational purposes; however, it does provide probabilistic matches of varying degrees of reliability for use in statistical reporting. Techniques employed in record linkage may also be of use for investigative purposes to help narrow the field of search against existing databases when some form of personal identification information exists.

The identification of individuals may be done in various ways. If there is a need to follow or track an individual over time or between different databases, the preference is to have a unique personal identifier. This unique identifier is generally in the form of a number such as the Canadian Social Insurance Number (SIN) or the Social Security Number (SSN) used in the United States. In theory, these identifiers are distributed uniquely to an individual and do not change even if an individual elects to change a name, date or place of birth, etc. Numbers are usually distributed in a sequential fashion and simply looking at the number may indicate something about an individual. For example, the first number of the SIN corresponds to the province where the number was issued.

The second type of personal identifier is called a non-unique identifier. This type of identifier can be used to establish the identification of someone based on a number of criteria. These criteria may be based on demographic or social information, physical information or administrative types of data. Demographic or social information would include surnames, given names and initials, date of birth, sex, ethnicity, etc. Physical data would include eye or skin colour, distinguishing features such as a missing finger, fingerprint pattern and other types of biometric attributes. Administrative types of data would include administrative numbers for tax filers, inmate numbers, drivers license numbers, etc. These numbers, once assigned, are likely not to change for a given individual.

The use of unique personal identifiers is not very widespread. In countries where population registers have been established, the use of a unique personal identifier is more feasible. For example, citizens of Denmark possess a unique number that is constantly updated to reflect changes in demographics and location of residence. This unique number is used in a wide variety of administrative areas.

Canadians do not have a unique personal identification number. However, like most countries, there are a number of variables that can help to establish, although not guarantee, unique individual identification. These variables are used in combination to ascertain the identity of a person in both a cross-sectional and longitudinal sense. Many of these variables are in use in a wide variety of administrative data files and databases. Other variables are more administrative specific as is the case in driver licenses, fingerprint numbers or employee numbers.

The ease and accuracy of tracking individuals in whatever sense is very dependent on the existence of unique and non-unique identifiers. The attempt to identify someone using non-unique identifiers involves various processes. It is important to know what these processes are and to understand the consequences of working with different personal identification variables in order to uniquely identify persons.

2. Personal Identifiers in the Canadian Justice System

At the present time, there is no national system of identification that uniquely identifies an individual who is involved in the Canadian justice system. In the United States and in Australia, systems have been developed to track individuals who have entered into the criminal justice system. These systems have adopted a unique number assigned to each arrestee or convicted offender. More detail on the development of these systems is given in Section B.

Most statistical databases used for individuals within the justice system contain common demographic variables. These variables are recorded and updated in all provinces and territories and in federal government departments that have responsibility for justice. Common variables that are reported to Statistics Canada's justice related surveys include surname, first name and initial(s), date of birth, sex, fingerprint system (FPS) number, offense codes and important dates. Not all of these variables are reported in each jurisdiction for each different justice survey. However, the personal identifiers are generally present. The combination of these non-unique identifiers using record linkage techniques can link individuals within a data file or between data files. Table A1 illustrates the current demographic information that is collected by the five annual micro-data surveys conducted by the Canadian Centre for Justice Statistics (CCJS).

Table A1: *Demographic variables collected in the CCJS micro-data surveys*

Survey	Surname	First name	Initial(s)	Name Encryption	Sex	Date of birth
UCR2.0, 2.1	No	No	No	Yes	Yes	Yes
Homicide	Yes	Yes	Yes	No	Yes	Yes
Youth Court	No	No	No	Yes	Yes	Yes
Youth Corrections	No	No	No	Yes	Yes	Yes
Adult Court	No	No	No	Yes	Yes	Yes

Table A1: *Demographic variables collected in the CCJS micro-data surveys*

Survey	Date of offense	Offense code	Juris. Ref. Number	Finger-prints	Other Biometrics
UCR2.0, 2.1	Yes	Yes	Yes	No	No
Homicide	Yes	Yes	No	Yes	No
Youth Court	Yes	Yes	Yes	No	No
Youth Corrections	Yes	Yes	Yes	No	No
Adult Court	Yes	Yes	No	No	No

Not all of the data variables are reported by all respondents for all incoming records. For example, although the FPS number is collected in the Homicide Survey, not all incoming records will give a FPS number. In the Youth Court Survey, one jurisdiction does report to the CCJS the surname, given name and initials of the youth. Other small variances like the above apply to "yes" and "no" indicators in Table A1. Appendix A looks at these variances in more detail. Section 4 and Appendix H present more information relating to the FPS number and its use in Canada.

The consistency of reporting survey data between survey respondents varies. For instance, some respondents attempt to send complete demographic data for all persons and other respondents may be reporting "blank" or "unknown" data. This results in data quality problems that may, or may not, affect the ability to form a unique personal identifier.

Typically, most of the large micro-data surveys at Statistics Canada collect basic demographic information. Common additions to the variables shown in Table A.1 include marital status, education, employment indicators, language, aboriginal status, geographic location (province or lower geographic level). Some of these variables are, to different extents, collected by respondents to the five justice micro-data surveys. However, the use of these kinds of variables to uniquely identify persons over time is typically not very reliable.

Several record linkage studies have been conducted at Statistics Canada using data files from the CCJS. These studies have been documented and one report was prepared that summarizes all of the work done. Each study addressed the use of personal identifiers in terms of linking variables. Issues of data quality were also examined. However, these studies focussed on one or two jurisdictions only. An appreciation of the types of methodological problems that relate to the linkage of justice related data files was realized despite the study limitations. In 1998, the CCJS produced a report that presents record linkage work with justice files in the 1990s. (Canadian Centre for Justice Statistics, 1998)

3. Other Methods of Establishing Unique Personal Identification

In Table A.1, the last column is called "Other Biometrics". The notion of biometrics relates to the measurement, description and classification of physical characteristics. Currently, the survey requirements of the CCJS micro-data surveys do not include any type of biometric data except for the fingerprint system number. The use of biometric data for the identification of people within justice systems in other countries is increasing. This is likely due to the level of accuracy of measurements made, the cost of taking these measurements, and the utility of this type of data within the justice community.

Biometric measurements would include different fingerprint techniques, palm prints, retinal and iris scans, deoxyribonucleic acid (DNA) profiles, voice print recognition patterns, eye and hair colour, mug shot photo, height and weight. Some of these methods will uniquely identify an individual and some will help to identify an individual. Although most biometric data can be more difficult to obtain compared to demographic data, the benefit is that individuals can be identified uniquely at any time during their life. One variable such as the DNA profile is all that is required to uniquely identify an individual regardless of the changes they may experience over time. The reliability of accurately obtained biometric data is greater than demographic data especially in the criminal justice community where persons are known to intentionally deceive justice authorities who are collecting personal identification. The use of alias names, false or altered identification or documents, and other falsifications by offenders makes the use of common personal identifiers more limited.

The report prepared by the Government of Ontario (1998) provides an excellent overview of the biometric methods used in justice systems. The report covers a wide range of methods and elucidates many aspects of usage.

4. Record Linkage of Criminal Data in Canada

There are hundreds of police forces in Canada that maintain data files containing information on criminal activity. The largest of these data depositories is maintained by the Canadian Police Information Centre (CPIC). This Centre is part of the Royal Canadian Mounted police who have responsibility for policing across Canada. About 40% of reported criminal incidents occurring in Canada are directly investigated by the RCMP.

Criminal record files are held in an identification data bank. The data bank is maintained by the RCMP Information and Identification Services personnel on behalf of Canadian police agencies. The originating police agency is responsible for the accuracy of the data it supplies to the data bank.

The key variable for obtaining information from the CPIC data bank is the fingerprint system number. The FPS number is assigned by Information and Identification Services on receipt of fingerprints that are not previously classified or filed. Fingerprints must relate to a criminal code offense only. Fingerprints are obtained in a standard procedure and recorded on a Form C-216.

Queries on the CPIC data bank can be done using only the FPS number. Using this single variable, full criminal records can be obtained. In situations where there is no FPS, queries may be made on the criminal name index (CNI). Using the CNI, searches on persons based on given and surnames, sex, date and place of birth, race, eye colour and physical measurement (metric height and weight) can be done. Querying the CNI using any of the above variables can result in several potential matches. Each potential match is given a weight of likely positive match. The weight has a maximum value that is based on the data entered into the query. The searching software permits searching by the phonetic spelling of the surname and also for spelling variation. Compound names are also searched in more than one way to optimize the number of potential matches. (Royal Canadian Mounted Police, 1999).

Fingerprint numbers do reside on data files other than the CPIC data bank. Many correctional institutions include the FPS number as do some police force data files. Where the FPS number exist on different data files, its use to bring together data from both files is excellent. Appendix H presents several additional issues relating to the FPS number.

B. The Use of Personal Identifiers outside of Canada

The types of personal identifiers of criminals collected in other countries is quite similar to those currently used in Canada. The standard demographic information includes names, date of birth and sex. Age is often used in conjunction with date of birth. The variable "age" requires a specific definition since the value may change at different points within the justice system. For example, the police obtain the age at the time of arrest and the corrections data may capture the age at the time of entry and/or exit from a correctional facility or program.

The United States

In the United States, offender-based transaction statistics (OBTS) have been developing for over ten years. These systems operate on a national as well as state level and depend on criminal justice authorities to provide offender-based data to central repositories during the year. OBTS guidelines outline the types of data to be submitted by each authority. Details submitted about the offender include age, race, sex, ethnic origin, the arresting agency, date of arrest, offense committed, the date and type of police disposition, prosecutor actions, pre-trial actions, court activities such as dates, dispositions, trial type and final plea. The trial outcome and any disposition are also recorded and forwarded to the central repositories.

An important feature of the OBTS data is that records can relate data on arrestees, incidents and charges. Persons arrested are identified through demographic data and fingerprinting. A unique number is then given to the arrested person that can then be linked to any earlier records that may exist. Incident numbers are created to identify each separate incident. This number links incidents where multiple charges are laid or involve many accused in the same incident.

There are several limitations of the OBTS which were the result of development decisions and also due to issues relating to data. Because not all of the states participate in the OBTS, tracking of offenders can not be completely made. Another limitation is that not all offenders are fingerprinted and therefore would not be part of the OBTS. Because youths are not part of the OBTS and only felony offenses are included, the OBTS contains only part of the criminal activity. Another coverage issue is that the OBTS does not include any correctional level data. (Ferrante, 1993).

Master name indexes (MNI) have been created at the state level that contain the names and other identifiers for all persons who have criminal records. These indexes may be used for a variety of reasons including those relating to criminal investigations, the sale of firearms or bail setting. As of 1992, almost all of the state MNIs were automated and virtually complete in the data stored.

The interstate exchange of criminal information can be done using the Interstate Identification Index (III). The Federal Bureau of Investigation (FBI) maintains this identification index of persons arrested for felonies or serious misdemeanors under State and Federal laws. The index includes name, date of birth, race and sex for each individual. Searches are made on the basis of name and other personal identifiers. The III system includes the National Fingerprint File (NNF). Under the NNF procedures, states forward only the first-arrest fingerprints of an individual to the FBI accompanied by other identification such as name and the date of birth. The expansion of the III includes other biometric characteristics such as retinal images and voiceprints. These positive identification methods are preferred over the non-unique variables such as name, sex and date of birth. The US Department of Justice Paper (1997) provides a good summary of other issues related to state and interstate Criminal History Information Systems.

Another method to identify persons within the US criminal justice system is the recently developed Combined DNA Index System (CODIS). As of June, 1998, all 50 states collect DNA samples, primarily from convicted sex offenders. Individual states may extend the collection of DNA samples for other types of criminal activity such as persons convicted of murder, manslaughter, assault and robbery. A DNA profile is unique to an individual. These profiles are entered into the CODIS which then allows both state and local law enforcement crime laboratories to exchange and compare DNA information. These exchanges and comparisons can be done electronically, much like the transfer of fingerprint information. It is unclear how much statistical use is made from the DNA profile data.

The CODIS was developed in response to the DNA Identification Act of 1994. The Forensic DNA Laboratory Improvement Program was then created to improve the capabilities of forensic DNA laboratories to support the investigation and the prosecution of violent crime. (US Department of Justice, 2000)

Australia

Following in the footsteps of the OBTS developed and used in the United States, the Crime Research Centre at the University of Western Australia developed the Integrated Numerical Offender Identification System (INOIS) in the late 1980s. The project's principle aim was to develop a common and unique identifier for offenders so that a longitudinal data base could be established in Western Australia. The tracking of offenders through the criminal justice system over time could then be done.

The INOIS number is based on a docket number that is assigned to an offender after the first arrest. The identification of individuals is validated using fingerprints and fingerprint records. The INOIS number is sequential. It is applied to all offenders and therefore can be used in all of the different criminal justice systems. This includes juvenile justice and the correctional institutions as well as parole. It is the fingerprint identification that ensures the accuracy of the INOIS number as being unique to an individual. (Ferrante, 1993).

The INOIS system operates as follows. Each quarter, or on some other regular basis, cooperating agencies send in offender records that contain name identifiers and other demographic details. These records are systematically matched to police criminal history records. Records are then returned to the criminal justice agencies with an INOIS identifier attached to each individual that was matched. The agencies then give only the INOIS number (name identifiers are not sent) when they supply the unit record data. The Crime Research Centre then adds these records to the longitudinal data base using the INOIS number as the key.

In the situation where exact matching can not be done, the linking system uses a probabilistic approach to determine if records from various sources, which do not have unique common identifiers should be matched. The concepts involved in the matching or linking of records are further examined in Section D.

United Kingdom

The Oxford Record Linkage Study (ORLS) covers 10 million records for 5 million people and spans from 1963 to the present. The dataset is used for the preparation of health services statistics and for epidemiological and health services research. Many of the concepts in the development of the ORLS are also relevant to other domains. In the UK, there has not generally been available a unique personal identifier available to follow individuals over time and in different parts of the health system. Record linkage techniques are therefore used to identify different records relating to the same individual.

Many variables are used in the ORLS to link data from the same person. The primary variable is based on the present surname. Rather than use the surname in the written form, the name is transformed or encrypted. Secondary and subsequent variables include the initial letter of the first given name, second forename or initial, and birth surname. Non-name variables used for identification include date of birth, sex, place of birth, and address information.

Record linkage is done using a probabilistic methodology. The OX-LINK system also detects and removes duplicate information by internal cross matching work. The Oxford name compression algorithm (ONCA) is described in Section 3. (Gill, 1997).

C. Using Names as Unique Identifiers

1. Introduction

The use of names in personal identification is common. The reason for this is because all people have a name. Names may not be consistent in the structure of a surname and given name(s). Some people have only one name and other persons may have many legal names. The structure of a name is related to the ethnicity of people. The use of more than one given name can be common in some parts of the world.

Another reason for the use of names as personal identifiers is that people know them. The name of a person is the most universal personal identifier and likely the most used demographic data that appears on data files.

2. Some Problems with Names

Although the use of names for identification is attractive because of their common existence, they must be used with some caution. This is especially true for use within the justice environment. Although people receive a birth name, this name may change over time. Common and accepted reasons for this are because of marriage, divorce, remarriage, adoption, and personal choice to legally change a name. Other name changes occur because of deceit, use of variations to a name (sex change for example), switching given names or the use of initials rather than given names. Some name "changes" may be the result of stolen identification or fabrication. For these reasons, the use of names as personal identifiers will present some challenges. (Newcombe, 1988).

A problem arises in the justice environment with the accurate reporting of names and the use of alias names. The fact that some persons possess more than one name will hinder the ability to track individuals within a single data file or between them.

Unlike most other demographic variables such as date of birth and sex, names can be difficult to verify. In validating a date of birth, computer systems can be programmed to check that each date component falls within a given range of values and that the relationship between these components meets various conditions. With sex, the data is typically a numeric indicator for which a computer system would accept three values (representing "female", "male", "unknown"). On the other hand, the ethnology of names is such that hundreds of thousands of different names exist.

3. Data Quality of Names

The issues above point to possible data quality problems with using names as personal identifiers. Extending these identified issues into the idea of name data residing on a data file raises the possibility of further difficulties. One problem common to data collection is the accurate asking and recording of a question. Persons arrested by the police may give a nickname or simply give their initials and this is what gets transcribed on the data collection form. Probing by the police officer or verifying the name identification of an individual using documentation in order to determine the real names of people is necessary to obtain good data quality. As an example, the first name of an arrestee could be recorded as "Bill" or "William" on documentation held by the person who says his name is "Billy". If the data collection system allows only one first name to be entered then a decision making process needs to be made to enter the most appropriate first name.

Other errors can and do occur with names. A simple transcription error or a keying mistake can result in problems in establishing unique identification. Transcription errors may occur because the persons recording or gathering the data assumes a particular spelling such as writing "Mark" when the spelling is "Marc".

Decisions on methods to handle characters like hyphens, apostrophes, and accents may also cause problems. The permutations of some names can cause problems, especially in many Asian cultures where multiple names are very common. Some names are very long and can exceed the allowed length of a name field on a data file. In this case, some decision is required. This might entail a simple truncation of the last letters or some other strategy.

Realizing the difficulties of using name as a personal identifier is not unique to justice data. The problems and solutions are well documented in a variety of disciplines. Partial and complete solutions have been developed in various domains and over many decades. Constant improvements and refinements to well-used solutions are still being studied.

There are various methods available to minimize the amount of typographical errors made on getting a name onto a data file. The computer programs are run on the incoming names and output a name file with minimal errors due to spelling or keying mistakes. Three such methods are the Jaro String Comparator, the Winkler method, and the Damerau-Levenstein approach. Appendix I examines the use of these methods. These methods, and others, are not only applied to given names and surnames but they can be applied to street names and those names of businesses and companies. Because the CCJS receives encrypted names, the use of string comparator methods is not directly applicable. However, there is some utility in using string comparators in local data base applications. The result would be better local data and therefore, better quality encrypted data as well.

4. Encryption of Names

In order to take advantage of the name variable for the purpose of personal identification, methods have been developed to minimize the types of problems described above. One of the most common techniques is to use an algorithm for encoding the name. These encryption methods can be used on the surname only or include any given names and initials. A common approach to the encryption of names is to phonetically arrange the name. By doing this, names that sound the same will be grouped together. This will have a direct impact on minimizing many of the problems associated with collecting name information. Name encryption will not solve all of the name-related problems like alias names that are completely different. However, for many of the common problems associated with names, encryption methodologies have been shown to be successful. Encryption methodologies are different than those that simply truncate a name after a fixed number of characters. Algorithms that are based on compression may also be quite different than the methods based on phonetics. (Newcombe, 1988)

There are two different name encryption methods that are used in the CCJS micro-data surveys. They are both principled on phonetic name grouping. These algorithms may also be used on street or company names. A description of each of the two methods presently used is summarised below. Other encryption methods for names are then summarily presented.

4a. The Russell Soundex Code

The first widely used phonologically based system for name encoding was developed by Margaret Odell and Robert Russell and patented in 1918. The system is extensively used today and is simply referred to as the "Soundex" code. Rules for the Soundex code are well known and easy to apply. Manual coding may be done quickly using the standard set of rules. The structure of the Soundex code is four characters consisting of one letter followed by 3 numbers ranging from 0 to 6. For example, using the rules as summarized in Appendix B, the Soundex code for the surname "Hilbert" is H460. The same Soundex code is given to the name "Heilbronn". The names "Rogers" receive a code of R262 and "Rodgers" is R326.

A goal of an encryption methodology is that similar names should be put into the same logical group and that dissimilar names should be allocated to different groups. A problem with the Soundex code is that both of these possibilities exist. In the examples above, the names "Hilbert" and "Heilbronn" have the same Soundex code but are dissimilar in sounds. The names "Rogers" and "Rodgers" are close in sound but fall into different Soundex groups. This type of undesirable outcome is quite common in Soundex coding and is apparent in encryption algorithms that were developed after the Soundex.

As experience with the Soundex code has grown so have the recognized problems and limitations. The Soundex code has been used to code names from as far back as the 1880 US Census. The names are typically "American" and the Soundex code worked fairly well. Unpronounced consonants and compound vowel sounds are not as intrusive in American names compared to British names for example.

The application of the Soundex code to some ethnic groups presents problems as the following example illustrates. The names "Van der meer" and "Van der berg" will each have the same Soundex code (V536). The ability of the Soundex method to discriminate names beginning with "van der", is not very good. With cultural diversity increasing in Canada, the variation of both surnames and given names is steadily growing. The Soundex code in the standard form (surname only, basic rules) is likely loosing the ability to do what is it was developed to do.

There are 6,734 different groups or pockets of names that can be created with the Russell Soundex code (see Appendix B for the calculation). The "range" of Soundex codes is A000 to Z666. Many of the different Soundex codes may be used only on occasion. Other names may be very common. The result is that the Soundex code appears frequently in a large file of encrypted surnames. For example, Soundex codes were produced for persons on the Adult Criminal Court Survey in 1996. The codes "B626", "G255", "L145", and "T651" appeared tens of thousands of times while codes such as "A133" and "A240" were seldom seen. Other Soundex codes had no surnames converted to them.

4b. Henry Code

Because the Soundex rules have problems in bringing together some similar sounding names and bringing together some dissimilar sounding names, another method was developed in Canada at the University of Montreal . The Henry method is designed to better function than Soundex, especially for French names. Compared to the Soundex algorithm, the Henry code has many more rules. The structure of the code as used in the CCJS Youth Court Survey and the Adult Criminal Court Survey (ACCS) is six letters of which the first four are based on the surname and the last two are based on a given name. Other variations have been programmed that use, for example, only a surname. The Henry code is also used to encrypt the names of companies that appear on the ACCS Survey. A summary of the Henry algorithm is given in Appendix C.

It should be noted that the Henry system used for phonetic encryption is not related to the Henry fingerprint system.

4c. Other Encryption Methods Applied to Names

As mentioned, the development of encryption methods for the coding of names continues to be progress. Improvements upon the Soundex method have lead to several other well-known algorithms. The Henry code is one such example that is used in Canada. In 1963, the New York State Identification and Information System or **NYSIIS** was developed. After making many attempts at improving the Soundex coding method, the NYSIIS code was introduced to overcome many of the known difficulties with the Soundex code. Because names are complex, the NYSIIS algorithm looked at improvements developed as a result of the experience since the early 1920s with Soundex.

The NYSIIS code is now used around the world for coding names in many domains including health and justice. The algorithm is more complex than the Russell Soundex although the objectives of the two systems are the same. The NYSIIS code is strictly alphabetical and the code length depends on the length of the names being coded. Ad-hoc rules have been developed for specific applications where ethnic surnames may be of particular interest.

The NYSIIS code is generally preferred over the Soundex code for several reasons. One important reason is that the number of valid NYSIIS codes exceeds the possible variations of the Soundex code. This should allow more differentiation between names and conserve the important name information. The NYSIIS code structure maintains the position of vowels (changes all vowels to “A”) rather than dropping them as in the Soundex method (unless the first letter of the name is a vowel).

The INOIS used in Australia uses the NYSIIS code as part of the matching routine to identify individual offenders (see Section B). The NYSIIS code is also used in large health related data bases maintained at Statistics Canada. Appendix D illustrates the NYSIIS approach.

The **ONCA** (Oxford name compression algorithm) is used in the UK in order to transform the surname of health patients into an encrypted version. The ONCA approach uses an anglicised version of the NYSIIS algorithm and then uses the Soundex algorithm on this to generate the standard four character Soundex name format. As expected, the ONCA method produces "blocks" or names that vary in size from the less common names to the more common ones. Subdivision of these name-based blocks is done using other available information. Researchers who developed the ONCA noted that the standard Soundex approach was not very useful for some common name variants (i.e., Thomson and Thompson gives two different codes) and it does not perform well on short names and names that have a high percentage of vowel letters. This would include many names of Oriental origin. (Gill, 1997).

Other refinements and variations of the original Soundex method have been done over the years. **The IBM Alpha Search Inquiry System** (1970s) builds a phonetic key comprised of 14 digits based on the surname. Using the IBM method, the names "Rogers" and "Rodgers" would both be coded as "04740000000000" while the Soundex system coded these names differently (R262 and R326 respectively). Although the IBM method could be viewed as an improvement over the Soundex, there are still some problems in the grouping and separating of names. One problem with the IBM system is that unless the algorithm is known, there is really no way to know anything about the name. Unlike the Soundex, Henry and NYSIIS methods, IBM codes do not indicate the first letter of the name being coded. For many users, this bit of information is useful. The IBM Alpha Inquiry System method is summarized in Appendix E. (Moore et al., 1977)

Another method for grouping names was developed for Western Airlines in the 1970s. This encryption algorithm was designed to meet a particular need of the airline. Algorithms like those used in the Western Airlines approach consider auxiliary data that resides on data files. Appendix F shows the coding method. The Western Airline approach is somewhat unique but it appeared to solve the problem. An option to developing an entire new method could have been to take existing encryption methods and change or fine-tune them based on characteristics of the population or other collected data. (Moore et al., 1977)

More recently, culturally specific name-matching methods are being developed. Due to the number of compound names, variations of spelling, and use of accented letters, the standard encryption algorithms are generally insufficient. For example, the LAS company in the USA is marketing software and name libraries that use linguistic rules to convert the spellings of names into multiple phonetic transcriptions of possible pronunciations. The software then uses the principles of articulatory phonetics to match names (i.e., the names "Leigh" and "Li"). The company has created software that will classify names based on ethnicity. Specialized processing for Arabic, Hispanic, Mandarin Chinese, and Anglo/European names can be done, and tools/libraries for other cultures are being developed. (Note: Although this software is being reported here, no evaluation was performed and no specific applications of its usage were found in the literature).

An example of a recent development in Soundex systems to improve the coding of specific ethnic names is the **Daitch-Mokotoff Soundex system**. This system was created by Randy Daitch and Gary Mokotoff in the 1990s to help the coding of Slavic and Yiddish surnames. Their system is also said to include refinements on the original Soundex method. The structure of the D-M Soundex is a six digit code with trailing zeros for names that do not have enough coded sounds to have a full six digit number. A detailed coding chart is used to code every letter of a surname. The Daitch-Mokotoff Coding Chart is given in Appendix G.

D. Collecting Personal Identifiers

1. Data Quality Considerations

There are various components that determine what constitutes good quality data. Among these are accuracy, timeliness, relevance, coherence, and interpretability. The desire to have the best quality data typically comes from a balancing of these components. For example, the best possible data may take a long time to prepare which means that the statistics produced may not be as relevant as desired.

Good quality data begins with asking good questions and being a good observer. The recording of demographic data in a police environment must address issues like alias names for example. A court reporter should not guess how a stated name is spelled. In order for the data that appears on a statistical data file to represent the truth, the handling or management of data becomes an important issue. Typically, guidelines are established on what is to be collected and what procedures are to be followed. Documentation that is complete and current is created specifically for this purpose. Situations that arise that are not covered in the guidelines or standards can create problems in ensuring that the data can be used at a later stage.

Whether a complete census or a small sample of records is taken, data quality problems will become important unless methods are taken to reduce the occurrence of errors being made. Accurate data recording and data entry are vital to ensuring that data has the best possible future value. Errors in transcription and key-entry may be minor or potentially cause major problems with the utility of the data. Many statistical datafiles and databases store data that is not the original or raw incoming data. Edit and imputation computer programs are commonly written and applied to incoming data. These programs may check that data values fall within a pre-identified range, that data is logically consistent or that blank data is changed to some value. Edit and imputation procedures will increase the utility of data by decreasing some types of errors in the data. However, the development of these types of programs may also introduce error into the data.

The key to quality data is to ensure that the data is captured at the source in an accurate and consistent way. The use of edit and imputation procedures is minimized and the costs of verifying data are decreased.

Confidence in the quality of data is a result of the processing of data at many intervals. Although the operational environment may not lend itself to collecting statistical data, the utility of the statistical data is solely dependent on persons in these agencies to gather the best possible data. No amount of data analyses can produce good information from poorly collected data. Good information must begin at the source of the data collection and planned data management practices must be in place to ensure that this data reaches the statistical analyst.

2. Confidentiality Considerations

The ideas of unique personal identification and confidentiality of personal information generally conflict. This is especially true when the concept of record linkage is being considered. The collection of survey and administrative data on individuals is typically deemed confidential and available to those who have a specific need to know. Statistical data analyses and reporting do not require personal identifiers to be part of the data files. In an operational environment, this requirement is very likely to be different.

There are many potential statistical uses of justice data based on the integration of data files relating to individuals. Rather than a name, an identification number is all that is really required. The important aspect here is that the same number or character sequence is assigned to the same individual. This applies to studies that are cross-sectional, prospective or longitudinal in nature. The ability to rely on a unique identifier is an especially important consideration over time because individuals can enter into the justice system many times.

In Canada, it is rare to see data presented at the level of the individual. At Statistics Canada and other government agencies in Canada, care is taken to ensure that data on specific individuals, businesses or institutions are not presented or derivable with certainty. Methods such as data suppression, random rounding, cell perturbation, data collapsing, etc. are frequently used so that data remains confidential.

Within the CCJS microdata surveys, incoming name information from the respondents appears encrypted using the Russell Soundex or Henry methods. Each of the two methods creates a coded version of each name that appears in the data file. Variations in the spelling of the reported name may produce a different coded version. These encryption methods only work in one direction. Once a surname is coded, the code can not be uncoded back to the original data. This does not in itself fully guarantee confidentiality of an individual's name. Generally, the larger the data files and the more common the incoming names, the more effective the encryption methods are in establishing confidentiality. On the other hand, the ability to discriminate between individuals for purposes of record linkage or data integration becomes more problematic.

Because of the dual interest in creating and maintaining confidentiality and tracking or linking persons, different record linkage methodologies have been created. These methodologies generally follow sound statistical theories and have been shown to be effective in achieving their purpose. The following section gives an overview of some common record linkage approaches. These methods apply to the linking of person-based records that have a unique or non-unique personal identifier(s). The application of the methods can be made in almost any domain where data at the individual level is collected.

3. Record Linkage

1. Basic Approach

The concept of record linkage is straightforward. A procedure is developed to match or merge data from two or more data files that relate to the same entity. The word "tracking" is also used for record linkage although this word usually relates to the identification of entities within the same data file.

A typical record linkage consists of five steps as follows: (Newcombe et al., 1992)

1. Find exact matches between the two files.
2. Create "pockets" or "blocks" or "groups" of similar records that were not matched,
3. Create pairs of potentially linkable records from the pockets,
4. Weight or assess the probabilities of these created pairs,
5. Calculate threshold values of weights and classify each pair of records.

Exact matching refers to two records that are brought together because they match exactly on all of the record linkage criteria. For data that has a unique identifier, the record linkage procedure will focus on this. All records that match exactly on this one variable are then linked. The Fingerprint System number is an example of where exact matches can occur based on one variable. Any difference in the unique identifier means that the record may or may not get linked. These records then proceed into step 2.

Because data files can be voluminous, data pockets are created. These pockets are commonly based on basic variables such as encrypted names, sex, date of birth or geography reference data. Then, each of the records that does not match exactly gets placed into one of the pockets. This simplifies the linkage process because links are attempted within pockets and not the entire file. This strategy can greatly reduce the number of combinations to try. The quality of these pocket variables needs to be examined and understood prior to their use.

Records within the pockets are then compared to each other using a methodology that could involve a number of variables. For example, if the variable "sex" is a pocket variable, then encrypted name, the date of birth or date of offense may be the linkage variables. Again, the quality of the incoming data for these variables is important in allowing two related records to be linked. Usually, various tests are developed and evaluated before a particular record linkage method is used.

Based on the linkage outcome, weights or probabilities are calculated for each pair of records in the same pocket. The higher the weight, the more likely two records match. Generally, weights are based on probabilities. A weight value of 1.0 would indicate that there is no doubt two records relate to the same entity. Negative weights can also be derived in some linkage methodologies. These weights are useful in indicating that two records definitely do not match each other.

Binit weights can also be used in some record linkage applications. A binit weight expresses the amount of agreement or disagreement between the two variables being matched. Extreme disagreement may result in a negative binit weight. Binit weights are set by the analyst and may result in a scenario where the date of birth has the most influence on matching (10 binits) and sex may contribute only 2 binit scores. Two records with different sex codes could have a binit score of –5 assigned, for example. A record with a high binit score with another record would been considered to likely match. (Gill, 1997)

The final step in record linkage involves the examination of the weights assigned to each linked pair. Threshold values are established that are based on the degree of confidence required in the linkage. The setting of threshold values occurs at two boundary points. One point discriminates between records that are almost certain to match and records that might match. A second boundary lies between those records that might match and those that very likely do not match.

The final classification in step 5 is one of three outcomes. Either two records match, two records do not match, or two records might match. Generally, the records in this last group present the most problems in terms of the time it takes to resolve the record linkage process. Sometimes, the resolution of the "might-match" group of records involves contact with the data provider and referral to the original documents. Changes to data are then required and the record linkage would be run a second time.

2. Record linkage considerations and applications

In the terminology of record linkage, a Type I error is created when a “false-positive” link is made. This means that two records are linked when in reality they relate to different things. A Type II error is created when a “false-negative” link is made. This means that two records that relate to the same thing were not linked. The setting of threshold values will determine the number of Type I and II errors that are made. Insisting on very small Type I and II error levels will mean that there will likely be a large number of records that might match. As mentioned, records in this group could be difficult to accurately place into either the matched group or the unmatched group.

There are a number of papers in the literature that address the issue of record linkage. The seminal paper by Fellegi and Sunter (1969) presents the statistical basis for linking data. The 1992 paper by Newcombe, Fair and Lalonde provides a history of probabilistic record linkage and shows some empirical results using the Fellegi and Sunter approach. Many of the software application programs used today to do record linkage are based upon the approaches given in these papers. If only exact matches are planned for, then the use of the five step methodology is not required. Many different ways to conduct exact matching may be done using common software (SAS, MS Access, etc.).

Large scale record linking using probabilistic (non-exact) matching is done at Statistics Canada using the Generalized Record Linkage System (GRLS). Uses of the GRLS have included the 1996 Census of Population and Housing Reverse Record Check, linking health-based survey and administrative records to the Canadian Mortality Data Base, the Canadian Cancer Data Base and the Canadian Health Data Base, and linking agricultural-based data in the creation of a Central Farm Register. There has also been some experience using the GRLS for linking justice files maintained by the CCJS. The current version of GRLS (Version 4.1) runs in a client-server environment with ORACLE and a C compiler. The software will also run on a PC or workstation which supports the UNIX operating system. The GRLS is particularly suited to applications where there are no unique identifiers on the files being linked. (Fair, 1997). It should be noted that earlier versions of the GRLS were called CANLINK and GIRLS (Generalized Iterative Record Linkage System).

The selection of the variables to be used in the record linkage of individuals is important from the aspect of data quality and in terms of data processing. Because personal identifiers such as surname and given names are not generally available, the encrypted name code is used as a pocket variable. Other common variables are sex, date of birth, date of offense and offense committed. Geographic variables are also potentially valuable for record linkage studies that cover a wide geographic area. In general, the larger the number of variables used in conducting a record linkage, the more problems are encountered. This is due more to issues of data quality and computing time than anything else. Typical record linkage with justice related files include an encrypted name, date of birth and date of offense. Additional variables may be used to confirm linkages. These latter variables are not necessarily required in the actual linkage but may be used to obtain more information about two records, especially records in the "might match" group.

Applications

Record linkage work is being conducted with some success using Health statistics at Statistics Canada. There has been significant work in the past relating to the bringing together of health data using record linkage techniques. Presented here are two examples of previous work and an overview of some work currently in progress in the health domain.

The first example relates to the linking of data from the Canadian Cancer Registry (CCR) and the Canadian Mortality Data Base (CMDB). The CCR is a longitudinal person-oriented database that contains all of the information on cancer patients and their tumours. This database has been in existence since 1992. An important part of the CCR is the Death Clearance Module. This module is a system that was designed to use the death records from the CMDB to confirm the deaths of the CCR patients that occurred during a pre-specified time period. Two types of record linkage were used due to differences in content between provincial death registration files and the data on the CMDB. Direct matching used the year of death, province/territory/country of death and the death registration number. Probabilistic matching used the NYSIIS transformations of the deceased person's surname and the surname of the Father of the deceased. The linkage of the CCR and the CMDB permits the calculation of survival rates for patients diagnosed with cancer, and helps to facilitate epidemiological studies using cause of death. (LaBillois et el, 1997).

In a second example, three data files were merged to meet a specific objective. The objective of the record linkage was to improve the understanding of the mechanisms that influence the utilization of health care services in Manitoba. Using 1996 Census of Population and Housing data, data from the 1986-87 Health and Activity Limitation Survey (HALS) and data from Manitoba Health, an analysis of the association between socio-demographic characteristics and health/health care utilization in the province was done. In this example, a match rate of 74% was achieved for private households. A subsequent study analyzed the overall concordance rate among the matched records to be almost 96%. Identifiers used in linking persons in this study were sex, year of birth, month of birth and postal code. Age of the person may have been considered at a later stage in the matching process. The use of individual names and addresses were not used due to issues of confidentiality and individual privacy. (Houle et el., 1997).

Currently in progress is the Health Information Roadmap Initiative. This is a four-year plan to modernize Canada's health information system. The Initiative seeks to fill gaps that currently exist and to bring together data from a variety of sources that will enable health researchers to access an extensive collection of data.

One of the sources of data is the Person-Oriented Information (POI) database. This database is derived from the Hospital Morbidity System (HMS) files. These files are produce yearly and contain hospital discharge information. The data files are event-oriented and it is therefore up to the user to bring together records that refer to the same person. The POI project takes the HMS data and through a series of linkages, transforms them into person-oriented data. Persons with more than one hospital discharge over time many then have their discharge records brought together. This type of linkage provides valuable data for researchers. Because of the different structure of the provincial and territorial morbidity files, preprocessing (including standardization) is a very important consideration prior to record linkage being done.

Recent advances in desk-top software allow exact record linkage to be done quite easily and probabilistic matching on an elementary basis to be conducted. Many issues on record linkage are described by Newcombe (1988) and many applications (and some theory) are given in the Proceedings of an International Workshop and Exposition on Record Linkage Techniques held in Washington in March, 1997.

These two references will yield many more references into particular aspects of record linkage. Issues such as the choice of variables for matching, including the use of names and name encryption, and the choice of matching rules are well covered. The contributed paper by Scott Meyer (Proceedings) "Using Microsoft Access to Perform Exact Record Linkages" uses justice files from the CCJS. Linkage methodologies and rates are presented for the matching of Adult Criminal Court Survey records to Uniform Crime Reporting Survey 2.0 records for the city of Regina. Although restricted in the scope of study, geographic area and types of offenses, the paper examines many issues concerning record linkage. Variables used in record linkage included the Russell Soundex code of the accused's surname, date of birth, sex, date of offense, and the type of offense committed. Linkage rates varied from 62% to 85% using seven different linkage strategies.

4. Statistical versus Operational Considerations

The requirement to link data files varies in terms of statistical and operational needs. Often, the collection of data required for data analyses may not be available or available only in a different format than what is required. The completeness and accuracy of data may have different levels of importance. For example, data may be collected on an individual's age in an operational setting but the date-of-birth is of more interest for record linkage and data analyses. If date of birth is available for the data provider for their own operational purposes, it may not be in a standard format suitable for statistical use.

A common problem in collecting statistical data is that documentation of how the data was collected and what the data represents is not always available or is out-dated. Frequently, coded values are used as indicators to replace a description. The variable "sex" is often coded to three different values denoting "males", "females" and "unknown sex". Unless the analyst has a description of these codes, then some uncertainly can exist. This problem is extended when a variable can have many more outcomes than three.

Changes in operational procedures and systems may impact upon a statistical application. For example, the collecting of some variables may cease or be modified at some point in time. Statistical users of this data may not be aware of these changes and subsequently encounter problems during data processing or analysis. One particular difficulty with justice data is the inconsistent reporting of the ethnicity of the offender. Although this variable can be very useful in the identification of individuals, the information itself is sensitive to collect and to report. Problems associated with the accuracy of ethnicity data relate to how the question was asked and responded to. The data may be the result of observation and not questioning which can lead to errors.

Incoming data for statistical purposes should be standardized. This can occur at one of two times, at collection or at the processing by the statistical analyst. Most operational administrators have unique data files and data bases to monitor. A police department in one part of the country may indicate the sex of an offender by a numeric code and another department may use an alpha code. A third department may actually have the sex of the offender as “male” or “female” on the data base. These variations in coding/storing data impact on how individuals may be identified over time or even cross-sectionally. An important issue in standardizing data is the cost of doing this. Costs involved in changing existing computer programs for key-entry or editing data may be substantial. Changing the code of a variable on a data file may not be complex but it does take time. Those working in the operational environment will not see the utility of doing this work but there is a large benefit for the statistical analyst.

Non-standardized incoming data could create errors in the data during processing. Dealing with codes that are not known or are in some way problematic, are usually difficult for the analyst to quickly handle. Computer programs developed by the analyst can usually identify problems with incoming data but the editing may not accurately fix the problem. Introduction of errors to basic demographic data during processing could lead to the problems of having some unmatched records or more of the Type I and II errors. For example, a record with no code for the sex variable could mean that any data referring to that record is going to be limited. The creation of data tables with sex as a controlling variable would not include the record above unless a “sex-unknown” category is used. This type of problem lowers the utility of the data and is something that an analyst can not usually resolve with certainty. Unless contact is made with the data supplier, all the analyst can try and do is see if the record can be matched to another record for the same person and use the sex code from that source. This process is time intensive and is usually not common practice in justice surveys at the CCJS. This type of imputation technique can be used if warranted but derived or imputed values are not necessarily correct.

5. Conclusions

The purpose of this report was to present an overview of the issues relating to personal identifiers. The use of these identifiers within the justice context is of prime interest but the report is not written specifically for this domain. It was not the intent of the report to produce a comprehensive review into all of the aspects that may be related to the issues of personal identifiers, encryption methods, record linkage, data confidentiality, etc. Most, if not all, of the main issues approached in the report have been the subject of many years of work by individuals around the world. Some of these efforts have culminated in detailed monographs. The report has been written from a statistical point of view as opposed to an operational environment where many of the issues require additional consideration.

The conclusions presented are written to encapsulate the main issues related to the use of personal identifiers. Many of the cited references are excellent sources for continuing into particular topics. The use of the Internet can also be a good source of information on the history, development and use of the procedures relating to areas covered in this report. Some of the Internet sites visited in the course of writing this report are given after the references. It is suggested that information given on these sites be verified if the information will be used to any extent.

The use of names in personal identification is very common but there are problems. Concerns about confidentiality and quality lead to the use of methods to encrypt names. Although not very useful at the operational level, this is a viable option for helping to uniquely identify individuals on a statistical data file. Using a non-reversible encrypted surname together with other demographic information can permit the linking or tracking of individuals.

The use of biometric measurements in the context of unique personal identification is becoming more widely used. The use of fingerprints is perhaps the most common method but other methods such as DNA profiling and the electronic storage of mug-shot photos is helping to match individuals. Although the use of these types of techniques are obvious in the operation of the justice system, this information can also be used in a statistical sense. Because these newer methods can uniquely identify individuals, the use of these variables in the sense of “data” can allow analysts to match data much more quickly and accurately than is possible using traditional demographic data.

The current complications in the area of statistical matching come when exact matches of individuals can not be made. The importance of collecting and processing the best quality data is important in achieving a high number of exact matches. Statistical theory is used in developing procedures to bring together records that should match exactly but do not. Numerous algorithms exist for increasing the size of matched or linked data files so that more reliable conclusions may be made when doing data analyses. These procedures are well documented with empirical evidence to indicate the successes and difficulties experienced. Generalized or application-specific software is widely available to conduct probabilistic record linkage. The current Generalized Record Linkage System (V3) software at Statistics Canada is the result of over 20 years of research, development and is used on a number of important studies and in the creation of large data bases.

The importance of gathering good quality source data is fundamental to good operational and statistical data. Although the uses of data for these two reasons is likely different, the issue of data quality is the same. The development of integrated justice data bases for statistical purposes is not very much different from those already built in United States within the justice area or the INOIS in Western Australia. These types of experiences make data integration using personal identifiers a real analytical possibility for many applications, including those within the Canadian criminal justice system.

Appendix A

Variances from Table 1: Demographic Variables Collected in the CCJS Micro-data Surveys

The following notes show the variation of reporting demographic variables for each of the five CCJS micro-data surveys from the information shown in Table A.1

Uniform Crime Reporting Survey (2.0 and 2.1) (UCR)

1. The four most serious charges are recorded on the survey and denoted MSO1 for the most serious to MSO4 for the least serious. Other offense charges are not reported to the UCR2.0/2.1 by individual police respondents.

2. Surnames of persons arrested are encrypted using the Russell Soundex code except in Quebec where the Henry code is used.

3. Date of birth and age variables appear on the UCR2.0 and 2.1 surveys. Respondents who report date of birth data will have an age variable derived in the CCJS. In 1998, about 1.9% of the UCR2.0 records were missing date of birth data. For these records, there was an age of the offender given.

Homicide Survey

Different versions of the Homicide Survey have been created from the original survey in 1961. Several changes on demographic questions occurred in the version starting in 1991. The following notes summarize differences in the demographic data collected prior to 1991 and 1991 to the present.

1. Prior to 1991, the surname of the accused was truncated to 10 characters. Two more characters were captured on the given name. The 1991 version to the present captures the entire surname, given name(s) and initials.

2. Prior to 1991, the date of birth of the accused was not available. Rather, the age of the accused at the time of the incident was used. Subsequent versions of the survey received the date of birth.

3. Prior to 1991, the incident data reported only the month and year. Subsequent versions of the survey included the day of the incident.

4. For homicide offenses committed prior to 1974 there was no distinction made on the *Canadian Criminal Code* offenses of murder, manslaughter and infanticide.

5. The variables “marital status”, “employment status” and “aboriginal status” have always been on the Homicide Survey.

6. The variables “Country of residence” and “Occupation” are variables added in 1991.

Youth Court Survey (YCS)

1. Quebec reports the full surname and given name of the accused. However, this name is encrypted on the YCS data base files. The 7 character Henry code is used in Quebec [the Soundex code is used in the other provinces].

2. Date of birth is occasionally missing even after follow-up with the respondent (less than 1% missing).

Youth Custody and Corrections Services Survey (YCCS)

1. The Soundex algorithm is used to encrypt all names. Quebec is currently not reporting to the YCCS Survey.

Adult Criminal Court Survey (ACCS)

1. Surnames and given names of persons are encrypted using the Russell Soundex code. An exception is that the 7 character Henry encryption method is applied in Quebec**.**

2. Names of businesses are also encrypted using the Soundex and Henry methods.

3. In situations where the offender is a business, the sex variable is coded as “3” and the date of birth is coded as “0”.

4. Sex codes are missing in about 15 to 20% of the offender records in Quebec.

Appendix B

Russell Soundex Coding Rules

1. Retain the first letter of the name (usually the surname) and drop all occurrences of the letters A, E, H, I, O, U, W and Y in all other positions [note: the name can be only the surname or include a given name. In this latter case, the given name would be appended to the surname with no spaces].

Assign the following numbers to the remaining letters after the first:

Letters	Codes
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

2. If two or more letters with the same code were adjacent in the original name (before step #1) then omit all but the first.

3. Convert the name to the form of letter, digit, digit and digit. If there are less than three digits then fill with zeros. If there are more than three digits then drop those in positions four and further.

Some preprocessing is typically done on names being encrypted by the Russell Soundex and the other methods that follow. Preprocessing could include the removal of all hyphens, spaces, accented characters, etc. within the name.

Examples of Soundex coding:

Surname	Soundex
Lloyd	L300
Ladd	L300
Harper	H616
Livingston	L152
Rogers	R262
Rodgers	R326
Ho	H000
Jackson	J250

Calculation of the number of different Soundex codes

The first letter can be 1 of 26 alpha values.
The first digit may take a value of : 0, 1, 2, 3, 4, 5, 6
The second digit may take a value of: 0, 1, 2, 3, 4, 5, 6
The third digit make take a value of : 0, 1, 2, 3, 4, 5, 6.

There are some impossible combinations of the Soundex code. For instance, the codes D306 and T061 are not valid. This is because the digits 1 to 6 can not be used after a zero has appeared in the code.

Based on the above the number of different Soundex codes is:

There are 26*7*7*7 or 8,918 different codes without considering the number of impossible codes. Impossible codes can take one of the following forms:

A00X for X = 1, 2, 3, 4, 5, 6 (156 combinations)
A= A to Z

A0X0 for X = 1, 2, 3, 4, 5, 6 (156 combinations)
A = A to Z

A0XY for X = 1, 2, 3, 4, 5, 6 and
Y = 1, 2, 3, 4, 5, 6 (936 combinations)
A = A to Z

AX0Y for X = 1, 2, 3, 4, 5, 6 and
Y = 1, 2, 3, 4, 5, 6 (936 combinations)
A = A to Z

The total number of impossible Soundex codes is 2,184. Subtracting these from 8,918 leaves a total number of 6, 734 different valid Soundex codes.

Appendix C

Henry Coding Rules

The Henry code as used in the CCJS micro-data surveys consists of seven letters. The first five letters relate to the surname and the last two letters relate to the first name. Blanks are created whenever the names are too short to be fully coded.

Henry has three methods for the coding of the last name and the first name;

a) that for the initial letter

b) that for internal letters (involves scanning of each consonant)

c) that for the consonant in the last part.

Henry considers vowels as the letters A, E, I, O, U and Y.

Rules:

A. CONSONANTS

1. The initial letter (or PH or CH) [except H which is considered non-existent]

* B, D, F, J, K, L, M, N, R, T, V are kept

Other initial letters are treated as follows:

Letter	Code
C before A, O, U, L, R	K
C before E, I, Y	S
CH before vowel	C
CH before L, R	K
G before A, O, U, L, R	G
G before E, I, Y	J
GN	N
H	Non-existent
P before H	F
P before another vowel	P
Q before UE, UI, UY, E, I, Y	K
Q before UA, UO, A, O	K
S	S [except in Saint(e), Sainct(e), Sct(e), St(e) and Sain, where it is represented by □X□].
W	V
X	S
Z	S

2. Internal letters (2nd, 3rd and 4th positions of the last name)

(a) Single (consonants between two vowels)

B, D, F, J, K, L, M, N, P, R, T, V are kept

Other cases are handled as follows:

Letter	Code
C before A, O, U	K
C before E, I, Y	S
G before A, O, U	G
G before E, I, Y	J
H	Non-existent
Q before UE, UI, UY, E, I, Y	K
Q before UA, UE, A, O	K
S	S [except in Saint(e), Sainct(e) and Sain, which are represented by "X" after DE].
W	V
X	S
Z	S

b) Group of consonants

Double consonants are considered single and are represented by one letter (NN=N, MM=M and LL=L).

Other cases are handled as follows:

Letter	Code
C before L, R	K
CH before L, R	K
CH before a vowel	C
PH	F

S is silent before all consonants (even H)

H is non-existent in all other cases

Any consonant other than L or R before a consonant other than L or R is silent

L is silent before M and N

ST(E) and SCT(E) after DE are represented by X

3. Final letters

Single (preceded by a vowel):

C, D, H, J, M, N, S, T, V, W, X, Z are silent

B, F, K, L, P are kept

Other cases are handled as follows:

Letter	Code
G	G
Q	K
R preceded by E	silent
R preceded by another vowel	R

b) Group of consonants

Final double consonants are treated as final single consonants

Other cases are handled as follows:

Z = S

S and the end of a name is considered non-existent, whether it precedes or follows one or more consonants

H is always silent, except in CH, which is represented by C, and in PH, which is represented by F

R and P: when penultimate letter is R or P, the final letter is silent

L before F or M is silent, F and M are coded

CQ = K

In all other cases, the final group of consonants is considered silent

B) VOWELS

Only the initial vowel is coded.

2. Vowel is followed by a single consonant or by two or more consonants, the first of which is neither M nor N:

A, E, I, O and U are kept
Y = I

3. Vowel is followed by two or more consonants, the first of which is M or N:

Letter	Code
A	A
E	A
I	E
O	O
U	E
Y	E

4. Vowel is followed by another vowel:

1) Diphthongs: AE = E EI = E OI = O
AY = E EY = E OY = O
AU = O EU = U OU = O

2) Others: See 1

Examples of the Henry Coding (surname):

Surname	Code
Clarke	Klrk
Cyr	Sr
St. Germain	Xjrm
Thivierge	Tvrj
Saint-Denis	Xdn
St. Denis	Xdn
Deschamps	Dc
Auerbach	Orbc
Rodgers	Rj
Rogers	Rj
Szamavitz	Ssmv
Montgomery	Mgmr
Ladd	L
Worthy	Vrt
Harper	Arp

Note: None of the above surnames was long enough to generate the maximum 5 character Henry code.

Appendix D

NYSIIS Coding Rules

I. If the first letters of the name are

MAC	then change these letters to MCC
KN	then change these letters to NN
K	then change this letter to C
PH	then change these letters to FF
PF	then change these letters to FF
SCH	then change these letters to SSS

2. If the last letters of the name are

EE	then change these letters to Y*
IE	then change these letters to Y*
DT or RT or RD or NT= or ND	then change these letters to D* [where * represents a blank space]

3. The first character of the NYSIIS code is the first character of the name.

4. In the following rules, a scan is performed on the characters of the name. This is described in terms of a program loop. A pointer is used to point to the current position under consideration in the name. Step 4 is to set this pointer to point to the second character of the name.

5. For each successive position of the pointer, only one of the following statements can be executed.

   a. If blank go to rule 7.
   if the current position is a vowel (AEIOU)
   and if it is E followed by V then change to AF
   otherwise change current position to A.

   b. If the current position is the letter
   Q then change the letter to G
   Z then change the letter to S
   M then change the letter to N

   c. If the current position is the letter K
   and if the next letter is N then replace the current position by N
   otherwise replace the current position by C

   d. If the current position points to the letter string
   SCH then replace the string with SSS
   PH then replace the string with FF

e. If the current position is the letter H and either the preceding or following letter is not a vowel (AEIOU) then replace the current position with the preceding letter.

f. If the current position is the letter W and the preceding letter is a vowel then replace the current position with the preceding position.

If none of these rules applies, then retain the current position letter value.

6. If the current position letter is equal to the last letter placed in the code then set the pointer to point to the next letter and go to Step 5.

   The next character of the NYSIIS code is the current position letter.

   Increment the pointer to point at the next letter.

   Go to Step 5.

7. If the last character of the NYSIIS code is the letter S then remove it.

   If the last two characters of the NYSIIS code are the letters AY then replace them with the single character Y.

9. If the last character of the NYSIIS code is the letter A then remove this letter.

Examples of NYSIIS coding:

Surname	NYSIIS code
Worthy	Warty
Ogata	Ogat
Montgomery	Mantganary
Costales	Castal
Tu	T

Appendix E

IBM Alpha Inquiry System Personal Name Encoding Algorithm

The coding rules produce a 14 digit phonetic key of the name according to the rules:

1. The first character table recognizes first letters or combinations of letters:

Letter	Value
A	1
E	1
GF	08
GM	03
GN	02
H	2
I	1
J	3
KN	02
O	1
PF	08
PN	02
PS	00
U	1
W (except WR)	04
Y	5

A zero is used in the first digit if the character(s) is not in the above table.

2. The basic table recognizes letters or letter combinations that are phonetically equivalent. All vowels and the letters H, W and Y are ignored.

Code	Letter(s)
0	Z, S, CI, CY, CE, TS, TZ
1	D, T
2	N
3	M
4	R
5	L
6	J, SH, SCH, CH
7	C, G, K, Q, X, DG
8	F, V, PH
9	B, P

3. There exist some exceptions to the basic table in step #2. In certain situations where letters or groups of letters have multiple sounds, a second and third pass through the algorithm is made.

Letters	1st pass	2nd pass	3rd pass
CZ	70	6	0
CH	6	70	0
CK	7	7	6
C	7	7	6
K	7	7	6
DS	0	10	10
DZ	0	10	10
TS	0	10	10
TZ	0	10	10

Examples of IBM Alpha coding:

Surname	Alpha code
Rodgers	04740000000000
Rogers	04740000000000
Kant	02100000000000
Knuth	07210000000000

Appendix F

Western Airlines Match Rating Approach (1977)

The approach developed for use by Western Airlines can be summarized as follows:

1. Deletion of all vowels unless the vowel is the first character of the surname.

2. Elimination of all double consonants by deleting the second contiguous usage of any consonant.

3. Reduce all encoded names to a maximum of six characters. This is done by retaining the first three and the last three encoded characters.

4. The length of each pair of encoded names are examined (one name incoming and one that already exists on a data file [call this name the Personal Numeric Identifier or PNI]). If they differ by more than two, no similarity comparison is done. A minimum acceptable similarity rating is established for each pair of encoded names as follows:

    Sum of lengths is 4 or less; rating of 5
    Sum of lengths is 7 or less; rating of 4
    Sum of lengths is 11 or less; rating of 3
    Sum of lengths is 12, rating of 2

5. Comparison of encoded names is then done. This comparison is done from left to right, character by character. Matching pairs of characters are deleted. This comparison continues until either encoded name has no more remaining characters.

6. All unmatched characters in both encoded names are packed to the right and comparison proceeds from right to left. On completing these comparisons, the number of unmatched characters in the longer name is subtracted from a value of six with the result being the similarity rating for that PNI.

7. Each PNI item that has a similarity rating equal to or greater than the minimum rating (shown in step #4.) is considered to be a good candidate match to the incoming record.

For example: The incoming name to be compared to an existing data file of encoded names is HARPER.

Step 1. HARPER becomes HRPR
Step 2. HRPR stays HRPR
Step 3. HRPR stays HRPR.

If part of the existing data file looked like ...HLDN, HRPR, HRPRD, HRP, HBLTWNS...then

Step 4. Differences in length between HRPR and those above are ...0, 0, 1, 1, 3,... According to the maximum difference rule in Step 4, HRPR would be compared to each of the first names on the existing data file. [the result of the comparison with HBLTWNS would be a value of 3].

Steps 5, 6, and 7 result in the following:

A. Unmatched characters with HLDN are LDN and a similarity rating of 6-3 or 3.
B. There are no unmatched characters with HRPR and a similarity rating of 6-0 or 6.
C. Unmatched character with HRPRD is D and a similarity rating of 6-1 or 5.
D. Unmatched character with HRP is R and a similarity rating of 6-1 or 5.

Because the length of the encoded incoming name was 4 letters, all matches with the data base names having a similarity value of 5 or 6 would be kept for further comparison.

Therefore, the incoming name of HARPER would be compared to the records corresponding to scenarios B, C and D in the above example.

Appendix G

Daitch-Mokotoff Soundex System

1. Names are coded to six digits, each digit representing a sound listed in the coding chart. See Table G.1.

2. When a name lacks enough coded sounds for six digits, use zeros to fill to six digits. GOLDEN which has only four coded sounds [G-L-D-N] is coded as 583600.

3. The letters A, E, I, O, U, J, and Y are always coded at the beginning of a name as in Alpert 087930. In any other situation, they are ignored except when two of them form a pair and the pair comes before a vowel, as in Breuer 791900 but not Freud.

4. The letter H is coded at the beginning of a name as in Haber 579000 or preceding a vowel as in Manheim 665600, otherwise it is not coded.

5. When adjacent sounds can combine to form a larger sound, they are given the code number of the larger sound. Mintz which is not coded MIN-T-Z but MIN-TZ 664000.

6. When adjacent letters have the same code number, they are coded as one sound, as in TOPF, which is not coded TO-P-F 377000 but TO-PF 370000. Exceptions to this rule are the letter combinations MN and NM whose letters are coded separately, as in Kleinman, which is coded 586660 not 586600.

7. When a surname consists or more than one word, it is coded as if one word, such as Ben Aron which is treated as Benaron.

8. Several letter and letter combinations pose the problem that they may sound in one of two ways. The letter and letter combinations CH, CK, C, J, and RS are assigned two possible code numbers.

Table G.1*: The Daitch-Mokotoff Coding Chart*

"NC" means no change to the letter

Letter	Alternate Spelling	Start of name	Before a vowel	Any other situation
AI	AJ, AY	0	1	NC
AU		0	7	NC
A		0	NC	NC
B		7	7	7
CHS		5	54	54
CH	KH (5) + TCH (4)			
CK	K (5) + TSK (45)			
CZ	CS, CSZ, CZS	4	4	4
C	K (5) + TZ (4)			
DRZ	DRS	4	4	4
Letter	Alternate Spelling	Start of name	Before a vowel	Any other situation
DS	DSH, DSZ	4	4	4
DZ	DZH, DZS	4	4	4
D	DT	3	3	3
EI	EJ, EY	0	1	NC
EU		1	1	NC
E		0	NC	NC
FB		7	7	7
F		7	7	7
G		5	5	5
H		5	5	NC
IA	IE, IO, IU	1	NC	NC
I		0	NC	NC
J	Y (1) + DZH (4)			
KS		5	54	54
KH		5	5	5
K		5	5	5
L		8	8	8
MN			66	66
M		6	6	6
NM			66	66
N		6	6	6
OI	OJ, OY	0	1	NC
O		0	NC	NC
P	PF, PH	7	7	7
Q		5	5	5
RZ, RS	RTZ (94) + ZH (4)			
R		9	9	9
SCHTSCH	SCHTSH, SCHTCH	2	4	4
SCH		4	4	4
SHTCH	SHCH, SHTSH	2	4	4

SHT	SCHT, SCHD	2	43	43
SH		4	4	4
STCH	STSCH, SC	2	4	4
STRZ	STRS, STSH	2	4	4
ST		2	43	43
SZSZ	SZCS	2	4	4
SZT	SHD, SZD, SD	2	43	43
SZ		4	4	4
S		4	4	4
TCH	TTCH, TTSCH	4	4	4
TH		3	3	3
TRZ	TRS	4	4	4
TSCH	TSH	4	4	4
TS	TTS, TTSZ, TC	4	4	4
TZ	TTZ, TZS, TSZ	4	4	4
T		3	3	3
Letter	Alternate Spelling	Start of name	Before a vowel	Any other situation
UI	UJ, UY	0	1	NC
U	UE	0	NC	NC
V		7	7	7
W		7	7	7
X		5	54	54
Y		1	NC	NC
ZDZ	ZDZH, ZHDZH	2	4	4
ZD	ZHD	2	43	43
ZH	ZS, ZSCH, ZSH	4	4	4
Z		4	4	4

Examples of the Daitch-Mokotoff Coding:

Surname		Code
Auerbach	A-UE-R-B-A-CH	097500
Lipshitz	L-I-P-SH-I-TZ	874400
Lippszyc	L-I-P-P-SZ-Y-C	874400
Ohrbach	O-H-R-B-A-CH	097500
Shlamowicz	SH-L-A-M-O-W-I-CZ	486740
Szlamavitz	SZ-L-A-M-A-V-I-TZ	486740

(source: http://www.jewishgen.org/infofiles/soundex.txt)

Appendix H

Fingerprint System Number (Canadian)

The FPS number used in Canada consists of a maximum of six numbers followed by one alpha character. An alpha character is assigned sequentially starting at the letter "A".

Fingerprint system numbers are issued uniquely to individuals. The numbers are issued to both youths and adults. A person reentering into the criminal justice system would be identified by the same FPS number. Once a FPS number is entered into the Canadian Police Information Centre data bank it is held until three years after notification of death or the persons reaches the age of 80. Several exceptions exist to the age of 80 criteria including whether the person was serving a life sentence. FPS numbers may also be removed from the data bank depending on various situations including persons found "not guilty" who apply to have their FPS numbers removed.

The use of the FPS as a statistical matching variable is very attractive. The simplicity and the uniqueness of the number makes exact matching possible. The data quality of the number is considered to be very good when the number is captured. Problems can occur when there is an intent to deceive the fingerprint pattern by the person charged. The degree of use , or coverage, of the FPS number is likely the main weakness in linking data files. Typically, incidents that are not indictable do not require fingerprinting of the suspect. Thus, linking an individual between police and courts data may not be possible for many offenses. In recidivism studies this can be an important consideration.

Queries can be made on the Canadian Police Information Centre (CPIC) bank by using just the FPS. If a FPS number is not available for conducting a query, other variables such as name, date and place of birth, sex, race and physical characteristics may be used. Table H.1 summarizes the variables and their attributes for conducting a record linkage to the CPIC data bank. (Royal Canadian Mounted Police, 1999).

Table H.1: *Linking variables used on the CPIC data bank (partial list)*

Variable	Length	Type	Notes
Given name	10	Alpha	Up to 5 given names can be searched at once
Surname	25	Alpha	Phonetic spelling and spelling variations can be managed. Compound names are searched
Sex	1	Alpha	Female, Male, Unknown
Date of birth	8	Num	CC.YY.MM.DD
Place of birth	4	Alpha	Provinces and territories are coded (i.e., Ontario is ONT). All states in the USA are coded. Codes for the UK, Europe and "other" countries are available.
Race	1	Alpha	White, not white, unknown
Age	3	Num	Based on current date of birth
Eye colour	7	Alpha	There are at least 7 types/codes
Height (metric)	3	Num	Based on most recent information
Weight (metric)	3	Num	Based on most recent information

Appendix I

String Comparators

Errors made during the recording and keying of names onto a data file will usually cause problems for record linkage. Depending on the nature of the typographic error, two records may not be matched or matched when they should not be. The use of string comparators minimizes the impact of typographical errors that can change the length or spelling of a name. String comparators are essentially algorithms applied to incoming data and the output is a numerical value that is typically in the range of zero to one. The larger the value, the greater the probability that the two names are the same.

In 1989, Jaro introduced a string comparator methodology to quantify the number of insertions, deletions and transpositions of letters within names. Other empirical studies followed that incorporated changes to Jaro's method. For example, a refinement introduced by Winkler in 1990 gives increased value to agreement on the beginning characters of a string. Other refinements have taken into account the length of names. String comparators are not limited to names of persons but can, and are, used for names of streets, businesses, etc.

Another string comparator method is based on the principle of bigrams. A bigram is simply two consecutive letters with a string. The word "string" has 5 bigrams: "st", "tr", "ri", "in" and "ng". The bigram function gives a value between 0 and 1 with higher values indicating higher agreement between two names.

Details on the methodology and use of string comparators can be found in Porter and Winkler (1997) and Winkler (1995).

Two surname strings, one incoming and one already existing on a data base, can be compared in a number of ways. Table I.1 illustrates the numeric outcomes of five different methods on comparing variations of two surnames. Other methods exist and most methods can be applied to given names, street names and names of businesses.

Table I.1: *Examples of String Comparator Methods*

Incoming name	Data base name	Jaro Method	Winkler	McLaughlin	Lynch	Bigram
Jones	Johnson	0.790	0.832	0.860	0.874	0.000
Massey	Massie	0.889	0.933	0.953	0.953	0.845
Brrookhaven	Brookhaven	0.933	0.947	0.947	0.964	0.975
Abroms	Abrams	0.889	0.922	0.946	0.952	0.906
Hardin	Martinez	0.000	0.000	0.000	0.000	0.000

References

Canadian Centre for Justice Statistics (1998). *Record Linkage in the Canadian Centre for Justice Statistics, 1996-1997.* (Internal Statistics Canada report).

Fair, M. E. (1997). *Record Linkage in an Information Age Society*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 427-441.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association,* 64, 1183-1210.

Ferrante, Anna. (1993). Developing an Offender-Based Tracking System: The Western Australia INOIS Project, *Australian and New Zealand Journal of Criminology*, 26, 232-250.

Gill, L.E. (1997) OX-LINK: *The Oxford Medical Record Linkage System*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 15-33.

Hart, G.E. (1968). ADP- Police Records and Automatic Data Processing Name Indexes, *Police Research Bulletin*, 8, 14-20.

Houle, C., Berthelot, J-M., David, P., Wolfson, M.C. , Mustard, C. and Roos, L. (1997). *Matching Census Database and Manitoba Health Care Files.* Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 305-318.

Gill, L.E. OX-LINK: (1997). *The Oxford Medical Record Linkage System*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 15-33.

Government of Ontario. (1998) Common Positive Identification Technology Assessment and Best Practices. Report prepared by the Ontario Integrated Justice Project.

Labillois, T., Wysocki, M. and Grabowiecki, F. (1997). *A Comparison of Direct Match and Probabilistic Linkage in the Death Clearance of the Canadian Cancer Registry*. Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 203-211.

Moore, G.B., Kuhns, J.L., Trefftzs, J.L. and Montgomery, C.A. (1977). *Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers*. U.S. Department of Commerce (National Bureau of Standards), Washington.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business.* Oxford: Oxford University Press.

Newcombe, H.B., Fair, M.E. and Lalonde, P. (1992). *The Use of Names for Linking Personal Records.* Journal of the American Statistical Association, 87, No. 420. Dec. 1992.

Porter, E. H. and Winkler, W. E. (1997). Approxim*ate String Comparison and its Effect on an Advanced Record Linkage System.* Record Linkage Techniques - Proceedings of an International Workshop and Exposition, March 21-21, 1997, Arlington, VA, USA, 190-199.

Royal Canadian Mounted Police (1999). CPIC Reference Manual (Revision 33). Prepared by the Technical Information Services Section.

US Department of Justice, Bureau of Justice Statistics (2000). *Survey of DNA Crime Laboratories, 1998.* Bureau of Justice Statistics Special Report. Washington.

US Department of Justice, Bureau of Justice Statistics (1997). *Survey of State Criminal History Information Systems, 1977.* Washington.

Winkler, W. E. (1995). *Matching and Record Linkage*. Business Survey Methods. John Wiley & Sons, Inc.

Internet sites visited

Note: Various Internet sites were visited during the research period. Some of the sites visited contain inconsistent information with other sites or with published documents. Variation in the information presented between Internet sites and with original documentation could be of concern. It is always recommended that the source documents be referred to if referred to in an Internet site.

1. "www.las-inc.com/tools.htm". This site relates to the software "MetaMatch", "NameClassifier" and "NameHunter".

2. "home.gnofn.org/~nopl/guides/genguide/soundex.htm". The site relates to the Soundex System.

3. "www.bradandkathy.com/genealogy/overviewofsoundex.htm". This site relates to the Soundex phonetic methodology. There are many other Internet sites relating to genealogy that involve linking or match names.

4. "www.gcis.net/cjhs/aguideto.htm". This site describe the Daitch-Mokotoff Soundex System and shows the algorithm.

5. "www.jewishgen.org/infofiles/soundex.txt". This site highlights the Russell Soundex coding system and the Daitch-Mokotoff Soundex System.