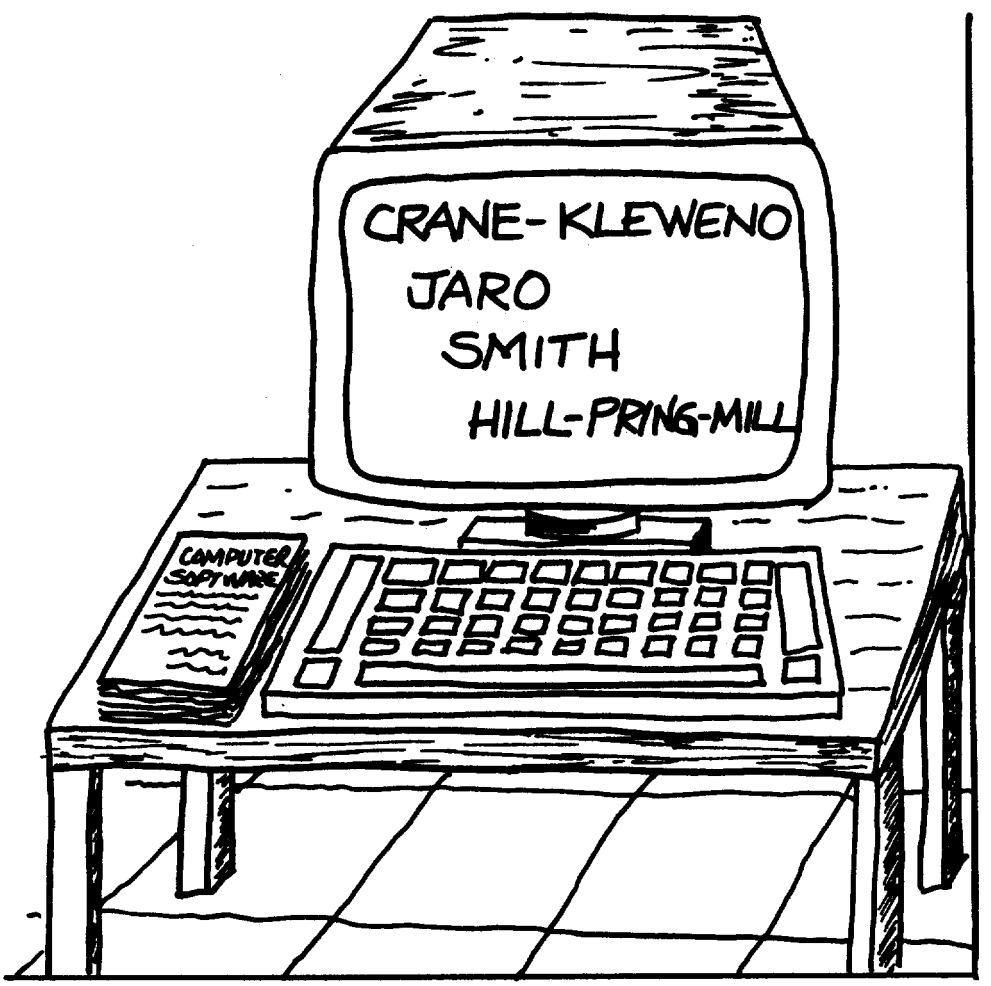
Section VI:
Computer Software



PROJECT LINK-LINK: AN INTERACTIVE DATABASE OF ADMINISTRATIVE RECORD LINKAGE STUDIES

Jane L. Crane, National Center for Education Statistics
Douglas G. Kleweno, U.S. Department of Agriculture

Much information exists on linkage studies using administrative records and, in some cases, survey data. A database called LINK-LINK illustrates the electronic retrieval of linkage study information. This paper is a guide for a prospective user of LINK-LINK. It will briefly describe the database and potential uses of the system, explain how one searches the database for general or specific linkage project information, outline procedures for obtaining copies of the database and address the future direction of the project.

The database is the end-product of a pilot study by the statistical policy committee formed from the Matching Group of the Administrative Records Subcommittee, a standing committee of the Federal Committee on Statistical Methodology. The committee encourages use of the database and solicits comments and suggestions from all users.

A DESCRIPTION OF LINK-LINK

LINK-LINK is an interactive information database devoted to administrative record and survey data linkage studies. The initial database contains 30 studies which were selected for complexity, originality, and diversity of record linkages. Appendix A provides a list of these studies by title.

Information for each study in the database was obtained using a self-administered questionnaire. The questionnaire, designed by the statistical policy committee, was completed for each linkage study. Respondents for the pilot study were contacted by telephone and letter before receiving the questionnaire. After the information was collected, it was edited for clarity and completeness and then it was keyed into the database.

The database is comprised of a series of menu-type prompts to direct the inquirer during the interactive information search. The menu allows the user to choose the search category from a list that appears on the screen. There is considerable flexibility in the database because of a variety of search categories. In addition, the prompts also allow selection of a particular area of user interest.

LINK-LINK was written using a dBASEIII software program. The database, which was developed on an IBM PC/XT personal computer, is on a 5¼" floppy disk.

Equipment requirements for LINK-LINK include: an IBM PC/XT or any other fully compatible personal computer with the MS-DOS or PC-DOS Version 2.0 or greater operating system; a minimum of 256K bytes of memory; two 360K floppy disk drives or one 360K floppy disk drive and a hard disk drive; and a printer with at least an 80 column capacity.

Objectives for Developing LINK-LINK

The primary objectives of the database are as follows:

1) inform and educate data users about record linkage activities;
2) identify and describe major record linkage data files;
3) illustrate procedures to meet confidentiality requirements associated with a particular record file;
4) demonstrate linkage methodology including software limitations, data quality concerns and linkage solutions; and
5) identify a knowledgeable contact person for further linkage information.

Type of Information Available

Each study in the database can be referenced to obtain a broad spectrum of linkage study information including: the linkage purpose; linkage methodology including software used; linkage data files; methods used to meet legal requirements for matching; type of dissemination of the linked data; names of cooperating institutions and their contact person; and titles of supporting linkage publications. A more detailed description of the database contents is given in Tables 2 and 3.

Potential Uses of LINK-LINK

LINK-LINK is a reference source for people seeking information on record linkage studies involving administrative records and/or survey data. The database is a useful tool to:

1) identify new and significant linkage programs using administrative records and survey data, or discover the most recent research activity involving linkage of records;
2) identify the potential uses of linkages involving administrative and/or survey data records;
3) identify the complexity and limitations of data linkages as dictated by public policy;
4) keep abreast of research in administrative record and survey data linkages and avoid redundancy of research efforts; and
5) use as a basis for additional research.

LINK-LINK'S MAIN MENUS

There are two main menus which provide the user with a large selection of information to investigate record linkage studies contained in

LINK-LINK. It is possible to search the database to identify all linkage studies for a certain characteristics such as the linkage purpose or linkage method. It is also possible to search a specific project for detailed study information. The logic of the system flow is from general categories to specific study detail.

Table 1 shows the system's two main menus with the initial user selection categories. Based on the user's interest, the appropriate menu selector value is entered.

Main Menu I is an exploratory menu to give the user a listing of linkage studies by general category. Main Menu II provides detailed data specific to a study in the database. A series of submenus direct the user to the appropriate information of interest within the main menu.

Main Menu I

The user, upon entering the database, keys "do explore" to display the Main Menu I selection categories. As the user responds to additional menu prompts, the search for information narrows until a list of record linkage studies is identified. The format for the list of studies is a five-digit database reference number, a project title, and a brief statement of the study description. The listing is displayed on the computer monitor and is also routed to a printer for hard copy.

Table 2 provides a brief description of the Main Menu I selection categories. For example, to obtain a list of linkage studies used for the construction of a sampling frame, the user keys a "1" in the Main Menu I and a "1" in the submenu. The end point of the Main Menu I is a list of database linkage studies satisfying the conditions as defined by the user in one or more menus.

At the end point of a path search in Main Menu I, the system prompts the user 1) to return for further exploring using major categories in the Main Menu I; 2) to request specific information for one or more studies listed using Main Menu II; or 3) to leave the system entirely with a series of "0" or quit prompts.

Main Menu II

Main Menu II provides the user access to detailed information on a specific linkage study. The user must know the five-digit database reference number which is provided when the listing of studies is printed at the end of Main Menu I. Only one study can be searched at a time. The user can request information on additional studies by entering each reference number as requested. All information displayed on the monitor is again routed to the printer for hard copy.

Table 1: Main Menu Selection Categories in LINK-LINK

Menu	Selector	Category
MAIN MENU I		
	(1)	Identification of Linkage Purpose
	(2)	Restrictions on Access of Files for Linkage Purposes
	(3)	Linkage Methods and Related Issues
	(4)	Data Files Used in Linkages
	(5)	Subjects and Respondents on Files
	(6)	Title and Short Description of Linkage Project
	(7)	Type of Dissemination
	(8)	Documentation of Linkage Studies by Title and Author
MAIN MENU II		
	(1)	Access to Files for Linkage Purposes
	(2)	Linkage Methodology
	(3)	Data File Description
	(4)	Titles/Authors of Written Documentation
	(5)	Contact Person for Study Information

Table 2: Description of Selection Categories for Main Menu I

Selector Category	Description of Contents
1. Identification of Linkage Purpose	Ten linkage purposes are identified. The user selects a category for a list of studies.
2. Restrictions on Access to Files	A submenu with two options are available to the user to identify general study safeguards: 1) studies where access to linkage records is permitted when respondent permission is obtained, and 2) studies where agency policy or legal authority restricts disclosure (general or specific statutes).
3. Linkage Methods	Four options in the submenu permit the user to investigate how database study files were linked: 1) software used for data preparation; 2) software used for matching; 3) data quality problems; and 4) linkage problems. Each submenu prompts the user to select a category of interest.
4. Data File Used in Linkages	Datasets used in all linkage studies contained in the database are listed. Number and title of a study are listed first, followed by the dataset(s).
5. Subjects and Respondents	Four general categories of subject/respondent interest are available.
6. Title and Description	List of linkage studies with database reference number, title, and study description is available.
7. Type of Dissemination	Four dissemination categories in the submenu are available for the user to obtain a list of linkage studies: 1) realeased in aggregate form; 2) public use microdata file; 3) restricted use microdata file; and 4) no dissemination.
8. Documentation of Linkage Studies	List of linkage studies with any published documentation by author, title, and date is available.

The user will generally access Main Menu II after exploring for information in Main Menu I. The user simply enters Main Menu II with the five-digit database reference number for which additional information is requested. Table 3 describes the five selection categories available.

It is possible, if the database reference is known, to skip Main Menu I and go directly to Main Menu II by keying "do lnktomn2." This command will place you at the beginning of Main Menu II where you will be asked to select from the categories identified in Table 3.

THE FUTURE OF LINK-LINK

At this time, the future of LINK-LINK is uncertain. The Matching Group of the Administrative Records Subcommittee is searching for an individual or Agency to assume responsibility for the database. Because the current version of LINK-LINK is a pilot effort still in the development stage, an evaluation of the database design is in order. In addition, the mechanics for updating current linkage studies and adding new studies to the database must be addressed. It is also necessary to support users who request a copy of the database.

Copies of the LINK-LINK database may be obtained by mail. Send two formatted floppy disks for each copy of the database requested and a pre-addressed mailer to return the disks. Specifications for the floppy disks are:

5¼" flexible disk
Double Sided
Double Density
40 tracks

Send correspondence and floppy disks to:

Fritz Scheuren, Ph.D.
Chairperson, Administrative Records
Subcommittee, Federal Committee on
Statistical Methodology
c/o Statistics of Income Division
Internal Revenue Service D.R.S
1111 Constitution Avenue, N.W.
Washington, DC 20224

ACKNOWLEDGMENTS

The success of this project is due to the contributions of members of the Matching Group of the Administrative Records Subcommittee. Group members were: Jane L. Crane, chairperson, National Center for Education Statistics; Mary Bentz and Beth Kilss, Internal Revenue Service; Douglas G. Kleweno, U.S. Department of Agriculture; Wendel Thompson, U.S. Department of Energy, and Tom Reilly, Bureau of the Census. Fritz Scheuren, Internal Revenue Service, provided leadership as the chairperson of the standing subcommittee. Thomas B. Jabine, Consultant, Committee on National Statistics, made many valuable contributions including the identification of the studies for inclusion in the LINK-LINK database.

Table 3: Description of Selection Categories for Main Menu II

Selector Category	Description of Contents
1. Access to Files for Linkage Purpose	Specific information on: parties to the transaction; incentives; how legal requirements were met; how records were obtained; procedures to protect identifiable records during linkages; type of dissemination, if any; and steps taken to prevent disclosure after records have been linked.
2. Linkage Methodology	Specific study information on: software used to prepare data files and to link records; problems in data quality; and problems encountered during the linkage process are listed.
3. Data File Description	Specific linkage study data set names and key variables are listed from each data set.
4. Titles/Authors of Documentation	References of publications by title, author, and date for specific linkage study are provided.
5. Contact Person for Study Information	Specific linkage study resource person including individual's title, employer, address and telephone number are identified.

APPENDIX A: DATABASE STUDIES BY TITLE

Tax Year 1979 Sole Proprietorship Employment and Payroll

Residential Energy Consumption Survey

Developing A Sampling Frame Of Petroleum Sellers

IRS/Census Direct Match Study

Tax Year 1979 Partnership Employment and Payroll

Employer Reporting Unit Match Study (ERUMS)

SRS/ASCS Data Exchange

Intergenerational Wealth Study

Enhancing Data From the SIPP With Economic Data

IRS 1979 Occupational Coding Study

Linked IRS-SSA Data File

Updating of the SSEL

IRS 1982 Estate Collation Study

Deriving Labor Turnover Rates From Admin Records for U.S. and 30 States

Mail List Development for 1982 Census Of Agriculture

High School and Beyond--Third Follow-up Student Financial Aid Record Component

National Health and Nutrition Exam Survey, Epidemiologic Follow-up Study

Census/IRS Link Study

1982 Partnership Employment and Payroll Link Study

1982 Sole Proprietorship Employment and Payroll Link Study

Continuous Wage/Benefit History Project

IRS Mortality Statistics Study

Current Population Survey/ National Death Index Match Study.

Forward Trace Study

Continuous Work History Sample System

Wage and Tax Statement Extract

Information Returns Program Match

IRS/SSA/DOD Match

Special Frame Study

Master Employment List-Unemployment Insurance Records of Texas and Pennsylvania

CURRENT RECORD LINKAGE RESEARCH

Matthew Jaro, U.S. Bureau of the Census

This paper discusses problems involved in the design and implementation of record linkage algorithms for file matching under conditions of uncertainty. Current research activities in this area are summarized, along with a brief survey of some underlying theoretical considerations. This paper stresses techniques that might be used for obtaining confidence in the match decision and algorithm validation. The research being conducted for the 1985 pretest in Tampa, Florida is discussed.

1. SUMMARY OF RESEARCH ACTIVITIES

Record linkage is the process of examining two computer files and locating pairs of records (one from each file) that agree (not necessarily exactly) on some combination of identifiers (or fields). For the Census Bureau this process is typically executed on two files containing individual names, addresses and demographic characteristics. Specifically, record linkage is important for census undercount determination, address list compilation and general census evaluation.

Record linkage research is focused on the development of an algorithm and accompanying manual procedures that will accomplish the above goals in a statistically justifiable manner. To this end the following major activities must be initiated:

A. development of a statistical foundation for the record linkage process;
B. construction of a data base that can be used for calibration, validation and testing of the characteristics of the linkage process;
C. development of methods to obtain information on the discriminating power of the various identifiers and their associated error rates (discriminating power is a measure of an identifier's usefulness in predicting true match pairs); and
D. design and implementation of computer algorithms to perform the actual linking.

The results of this research will be:

A. more accurate undercount determination and coverage analysis;
B. reduction of costly clerical procedures by use of automated methods;
C. a statistically valid process which can replace previous ad hoc techniques; and
D. algorithms that will be useful for overcoverage determination and address list compilation.

2. AREAS OF INVESTIGATION

There are several areas of investigation that must be pursued in order to design and implement a successful matching system. These areas are currently the focus of attention for the Record Linkage Research Staff.

2.1 Blocking and Other Search Restricting Techniques

The set of records that will be examined to find a match for a given record is called a block. Obviously, if an entire file were searched for a match for each record, the probability of finding a true match would be highest, since no records are excluded from consideration. However, the cost of such a process would be prohibitive. As we restrict our search, we exclude records and increase the probability that the "true match" record would be excluded--but the cost of searching decreases.

The ideal blocking identifier would be one which nearly always agrees in "true match" record pairs but nearly always disagrees between pairs which are not valid matches. This ideal blocking identifier must have a large enough number of possible values to insure that the file will be partitioned into many (and therefore smaller) blocks. R. Patrick Kelley of our staff has developed a method for computing an optimal blocking strategy, considering the tradeoffs of computation cost against errors introduced by restricting the search for matches. See [4].

2.2 Weights

The terms "identifier" or "component" represent fields on a computer file (and are used interchangeably). Typical components are street name, street type (e.g., Street, Avenue, etc.) surname, given name, etc. The discriminating power of a component (or identifier) is a measure of how useful that component is in predicting a match. Consider a component such as surname. Common values of surname (such as "Smith") have greater chances of accidental agreement than do rare values (such as "Humperdinck"). Consequently, the frequency of occurrence of a particular value of an identifier is one determinant of the weight or importance of that value as an indicator of matched or unmatched records. Another determinant of the weight is the error rate associated with the value of that component. High error rates diminish the predictive usefulness of an identifier or its values.

Fellegi and Sunter, in [1], presented a general theory of record linkage, including discussions of weight calculations and the development of optimal decision rules. Their basic idea for weighting is summarized below.

The two files (A and B) to be linked consist of a number of components (identifiers) in common. Consider all possible pairs of records. A particular pair is either truly a matched pair (an element in the set M of all matched pairs) or an unmatched pair (an element in the set U of all unmatched pairs).

For all pairs (p) and each component (or component-value state) i let:

m_i = Pr (component agrees | p ε M)
u_i = Pr (component agrees | p ε U).

Weight for the ith component = $\log_2 (m_i/u_i)$.

The above computation would be the same if we were considering specific values of components (such as "Smith" or "Humperdinck") rather than the component as a whole (surname). Similar weights can be computed for disagreements. m_i is computed by examining all matched pairs; u_i is computed by examining all unmatched pairs. For the two files A and B,

$$\{U\} = \{A \times B\} - \{M\} .$$

Since the cartesian product A x B is $O(n^2)$ and M is O(n) (where n is the number of records in the smaller file), then { U } is much greater than { M } and the u_i can be computed by taking the frequency counts of the components in both files.

The calculation of m requires a prelinked set of records M. This fact presents the greatest practical difficulty because of the large sample size necessary, the cost of producing such samples and the inherent error in manual processes.

Fellegi and Sunter, in [1], suggest a method of weight calculation that does not require prelinked pairs. It uses an assumption of the statistical independence of the components and requires the solution of a non-linear system of equations. We plan to investigate the use of this method, which to our knowledge has never been tested.

Another method of weight calculation that we will consider is that of iterative refinement. We propose this method to avoid the construction of costly samples. If there were no errors in a given component, the value "m" for that component would be 1 and the weight for the component could be calculated from the frequency of occurrence of the component value states.

These initial weights can be refined as follows: Whenever a record pair disagrees on a component, that pair would be presented to an operator by the matching program. The operator can then make a decision as to whether the pair is a match or not. This places the pair in either the set M or U and the weights can now be updated (since m is now less than 1 -- because of the detected error -- if this pair is placed in {M}).

The program can obtain information regarding the error rates of each component in this manner, updating the probability as records are processed. The operator supplies the "truth" regarding each record in question (does this pair belong to set { M } or to set { U } ?). This teaches the program to make similar decisions to those of the operator.

The operator can set the level of errors that will control the display of candidate record pairs. In this way, records can be matched automatically despite small errors in components. As confidence is gained, the thresholds for manual intervention can be moved. After all records have been processed, the entire file can be rematched using the new weights and the process can be continued until consecutive iterations produce small differences.

An investigation into this technique is required to determine whether such iterations will converge to a stable set of weights and to determine the amount of bias introduced by such estimation techniques.

A third method of weight calculation that might be explored would involve automatically making the "M" or "U" decisions, instead of relying on human operators. This would be accomplished by considering pairs of records that match on all fields except a specified number. Those pairs could be assigned a match status if the composite weight (Σw_i) for the pair was sufficiently greater than the cut-off threshold. The distance from the cut-off would leave room for weight estimation error without effecting the "M" or "U" decision, and hence, the "M" decision could be made automatically with some degree of confidence. These cases would be used to tabulate the error rate probabilities.

Since the cut-off threshold for a match decision is dependent upon the weights of each field, this threshold would move as weights are revised. The effect of this concomitant variation on the weight estimation must be investigated.

2.3 Composite Weights

If the components are assumed to be statistically independent, then the composite weight is equal to the sum of the individual component weights. Adding the weights is equivalent to multiplying the conditional probabilities. Weights for disagreements can be computed similarly to weights for agreement. Disagreements are generally given negative weights, whereas agreements receive positive weights.

We know that some dependencies exist (such as sex and given name) but the extent to which dependence changes the matching decision rules must be analyzed. For example, "Robert" is principally a male given name, but "Stacy" could be either male or female. Such dependencies could have an effect on the probabilities of agreement given unmatched pairs. If the errors in the fields are dependent, then the probabilities of agreement given matched pairs could change. The disagreement weights would also change proportionally.

We are currently designing simulation experiments to study the effect of covariance on the decision results. It is hoped that a regression analysis will provide information concerning this relationship after a number of runs with differing covariance configurations.

2.4 Error Rates

If a plot were to be made of numbers of observations versus composite weight, a bi-modal distribution would result. Since most pairs are elements of { U } , the disagreement mode is much larger than that for agreement.

For each pair, one of three decisions is made. The pair is said to match if the weight is greater than a threshold μ, or not to match if the weight is less than a second, lower threshold λ . Pairs having weights between these thresholds are classed in the "don't know" category. These pairs must be followed-up using a computer-assisted manual approach.

Once the thresholds are set, bounds on the

probabilities of false matches and false non-matches can be computed by integrating the portions of the distribution tails lying beyond the threshold values. By tabulating weights of candidate pairs, the matcher program could provide information on the error rates associated with the component values. These error rates are useful for verification. The success of this technique will depend upon our ability to fit a curve to the observed tails of each mode in order to perform the integration.

2.5 Component Values

The matcher algorithm will use a table of weights derived from investigations on weight methodologies (see 2.2). One weight would be associated with each predetermined component or identifier value. The algorithm would store the most frequent values of components from tables prepared by other programs and component values not in this list would be given a relatively high weight. Thus, popular names (which have low discriminating power) would receive lower weights than comparatively rare ones, without requiring the construction of exhaustive lexicons. Value tables would only be used if successful results could not be obtained by considering a component to have a single weight.

The weight tables for the program will include expected frequencies of occurrence of component values, error rate information and number of records processed for past data. Information from the current data could be used to update the weight tables as the program gains experience matching.

2.6 Bayesian Adjustment

In addition to keeping records of expected frequencies (based on earlier observed frequencies), the program will also keep observed frequencies of a block for a specific file. If there is much deviation between observed and expected frequencies, temporary modification to the weights can be considered. For example, in a Spanish-speaking area, the name "GONZALEZ" might occur relatively more frequently than it does on the average for the United States.

Missing data values could also result in the reduction of discriminating power of a field within a block.

We have incorporated a Bayesian adjustment technique into our experimental matcher. We have assumed a Beta prior distribution and are investigating parameter estimation techniques for this distribution.

2.7 Distance Metrics

Simple agreement/disagreement patterns of component pairs are not adequate for character strings and numeric data. We are investigating prorating the weight on the basis of degree of agreement.

A number of character-string comparison routines for component values which do not agree completely are available, including the routine designed by Jaro and Corbett, which has been used for 12 years in the UNIMATCH system [3]. Through the use of such a routine, words can be matched despite spelling errors. The UNIMATCH algorithm is an information-theoretic comparator which takes into account phonetic errors, transpositions of characters and random insertion, replacement and deletion of characters. These approaches will be tested in the matcher algorithm.

2.8 Assignment

After blocking, the program uses the various techniques described above to construct a composite weight for each pair in the block. These weights are stored in a cost matrix and the assignments can be made by solving the problem:

$$\text{Maximize} \quad Z = \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij}$$

Subject to

$$\sum_{j=1}^{n} X_{ij} = 1 \qquad i=1,2,\ldots,n$$

$$\sum_{i=1}^{n} X_{ij} = 1 \qquad j=1,2,\ldots,n$$

where C_{ij} is the cost (weight) of matching record i with record j. X is an indicator variable. The matrix is made square by the use of dummy weights.

This problem is the linear sum assignment problem, which is a degenerate transportation problem that can be solved efficiently using only additions and subtractions. Once an optional assignment set is obtained, the Fellegi-Sunter decision procedure is applied to determine whether an assignment represents a match, a clerical review case or a non-match.

3. MATCHER IMPLEMENTATION PLANS

An experimental program has been implemented that incorporates the techniques discussed in this paper so that controlled tests can be conducted without undue difficulty. This program is operating on an IBM Personal Computer.

For production matching it is anticipated that not more than two passes will be required to match nearly all records not requiring professional review. Records failing to match on blocking components in the first pass would have a second chance to match on different blocking components during a second pass. By selecting two high discrimination/low error rate sets for blocking, the probability of intersecting errors is minimized. The high discrimination/low error rate property for a component means there is a high probability that the component can accurately predict a matching record pair. By using two such components, the chance of a successful match is relatively good, since errors on both components would be required to reject a record.

We plan to utilize experience gained by Statistics Canada (the Generalized Iterative Record Linkage System [2]) and others in our investigation into the problems of record linkage. It is our intent to have an operational program for use with the 1985 Census pretest. One of the most important applications will be coverage evaluation for the Decennial Census.

REFERENCES

[1] Fellegi, I.P. and Sunter, A.B., A Theory for Record Linkage, Journal of the American Statistical Association, Vol 64, 1969 pp 1183-1210

[2] Generalized Iterative Record Linkage Systems (GIRLS), Institutional & Agriculture, Survey Methods Division, Statistics Canada, Internal Documentation, Oct. 1978

[3] Jaro, Matthew. UNIMATCH - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty. Spring Joint Computer Conference, 1972, AFIPS -- Conference Proceedings, Vol 40, 1972, pp. 523-530

[4] Kelley, R. Patrick. Blocking Considerations for Record Linkage Under Conditions of Uncertainty. Proceedings of the American Statistical Association, Social Statistics Section, Philadelpia, 1984, pp. 602-605. (Sections 3 & 4 were prepared with the assistance of D. Childers.)

RECORD-KEEPING AND DATA PREPARATION PRACTICES TO FACILITATE RECORD LINKAGE

Martha Smith, Statistics Canada

Lack of adequate personal (or "entity") identifying information and appropriate documentation on what is contained in historical files can be major stumbling blocks in carrying out long-term follow-up studies. Over the past few years, considerable experience has been gained in the use of existing administrative (e.g., industrial employee, mortality, hospital, cancer, marriage, birth) survey and census data files for record linkage studies in Canada [1-3].

The purpose of this paper is to give some practical pointers for agencies and individuals involved in implementing future linkage projects, particularly those where large historical files are being used, and where no unique identity numbers are available. Specific examples will be given here which relate to occupational and environmental health studies, but many of the record linkage problems and their solutions apply also to other areas of statistical research.

Organizationally, the present paper is divided into six main sections. The first section gives the main results and conclusions. The second section outlines the kinds of data files required for occupational and environmental health studies. The third section describes the role that various broad categories of records can play in the linkage process. The fourth section gives examples of the practical problems in the preparation of existing files for linkage, along with the methods and some of the software developed to cope with these problems. The fifth section deals with the probabilistic matching technique and the art of designing an efficient linkage operation. The last section makes recommendations for future record keeping and data preparation practices to facilitate record linkage.

I. MAIN RESULTS AND CONCLUSIONS

A generalized record linkage system has been developed based on the concepts of probability and the use of 'weighted' record comparisons [4-7]. The probabilistic methods developed have several desirable features:

- records can be linked which lack unique numerical identity numbers;
- records are able to link despite discrepancies which may exist between identifying particulars;
- 'weights' can be assigned for agreement, disagreement, and partial agreement; and
- the technique discriminates between rare and common values of a given identifier.

On the basis of fairly extensive experience with computerized record linkage of a probabilistic kind, using the generalized iterative record linkage system (GIRLS), it seems unlikely that the technology and the software will be major limiting factors in the future. The major costs, which can limit the application of the approach, are often likely to be associated with the need to do data entry for additional identifiers in a standard fashion, if these have not already been captured in machine readable form. For historic data files, lack of appropriate documentation and standard data entry rules can cause problems. Some software has been developed to aid in the preprocessing of such files. It is therefore recommended that if the files are to be used for record linkage, sufficient identifying items be captured at the time of the initial data entry. Compromises whereby the amount of identifying information is restricted in order to reduce costs will be reflected in reduced accuracy of the linkages, and of the kinds of uses that can be made of the files.

Certain files may serve in the role of intermediate files that facilitate the linkage of other files.

Procedures to evaluate the quality of the linkage should be planned early. For example, it may be possible to incorporate known alive cases in a mortality search; to carry out independent manual follow-up on a sample of the file and compare with the computer results; or to carry out an alive follow-up to complement the death search.

Improvement of present data sources and the development of new sources would seem to be necessary if further demands for occupational and environmental health statistics are to be met. A checklist of data items to be collected has been described elsewhere [3-4].

Collaboration and co-operation among individuals and agencies are often required to complete studies. Suitable communication networks among investigators must be established, particularly if there is a long geographic distance between the interested groups.

II. KEY ELEMENTS IN A TYPICAL FOLLOW-UP STUDY THE KINDS OF DATA FILES REQUIRED

Certain general principles shape whatever epidemiological studies for long-term health effects are undertaken and influence the nature of the procedures for data gathering and analysis. The data gathering could include examining data systems already available which could facilitate the study. The requirements for identifying information are similar whether one is looking for changes to the exposed individual, or for inherited changes affecting the offspring from such individuals.

The key elements for data collection that should be included in any such study are described in [4]. A typical follow-up study often requires some knowledge of work histories, dose histories, health outcomes and the personal identification of the individuals involved. The software available must be capable of bringing all the various relevant files together at appropriate times.

The kinds of linkages involved may be a series of internal linkages to identify data pertaining to the same individual (e.g., to create individual work histories) as well as two-file linkages (e.g.

to match a work record against a death record). The matching techniques can use individual identification numbers (e.g., Social Insurance Numbers), probabilistic matching techniques, or a combination of the two.

III. THE FUNCTIONS OF BROAD CATEGORIES OF SOURCE RECORDS

The kinds of source records required for studies of delayed health effects may serve one or more of four possible functions in the follow-up process.

First, such records may identify an individual as belonging to an "at risk" population (or to a "control" population with which the other is to be compared). In this case they are referred to as "starting point" records which initiate the follow-up process.

Alternatively they may identify an end effect, such as cancer or death in an individual who is a member of a study population, in which case they are referred to as "endpoint" records. One example of an endpoint file is the Canadian Mortality Data Base consisting of the records of all deaths in that country dating back to 1950. Follow-up thus will consist of using a file of starting-point records to search a file for potential end-point records, and of linking those records from the two files which relate to the same individuals.

The third possible function of a record file is that of an intermediate file which facilitates the searching and the linkage process. For example, where a starting-point record carries the maiden surname of a woman who later married, and the endpoint record contains her married surname, the search of the endpoint file may be more productive and accurate where reference can be made to another file, such as a marriage file or the Social Insurance Number Index which contains both of these names.

The fourth function of record files is as a source of the detailed statistical variables required for the analysis. For example, linkage may be required to bring together individual work histories, dose histories and smoking histories.

In considering the possible uses of various available files, all four functions must be kept in mind.

IV. PREPARING THE INPUT FILES

Prior to linkage of any kind, the records being used need to be brought into the formats that are required for making the necessary comparisons, and into the sequences that are appropriate for the linkages. The quality of the identifiers needed for linkage may also be tested by looking for blank fields and for values of the identifiers that are not permitted (such as day of birth = 32). If data collection and data entry have not been done with record linkage initially in mind, this phase can be quite time consuming and costly.

We have found the Statistical Analysis System (SAS) very helpful at this stage, and as a routine we systematically scrutinize the values of fields in files to be used in linkage. These are compared with any available documentation regarding coded values and their meanings. One can check how many fields have non-missing values, valid values, ranges, codesets, or invalid characters or values. Whereas blank fields can only be filled from other sources, fields which have unacceptable values may sometimes be corrected.

One may wish to create a new field for each record to indicate the "availability" and validity of fields on the same record. For example the value "120112001" could indicate "present and with the valid code range" (1), "present, but with an invalid code" (2), or "absent" (0). A SAS distribution of this word facilitates one's assessment of the likelihood that one will be able to link the files.

It is necessary to obtain copies of the forms of the original source documents, the record layouts and any file documentation, along with detailed information regarding how the administrative system works.

Some problems one may expect to encounter have to do with the quality of the records, and some methods which have been used to deal with the problems are as follows:

(1) Lack of a standard format - particularly for the name and address fields

If name fields have been entered in string format and if a variety of delimiters have been used to separate surnames from forenames, it may be necessary to put the values of the fields into a standard fixed format. It is particularly difficult to separate the components in a name field if blanks have been used as the delimiter. A simple NAMESCAN routine has been developed, which changes all alphabetic characters to "A" and leaves all other characters intact. A SAS distribution can then be made to look at the various patterns on the file.

When standardizing name fields, titles should be put in a separate field e.g., Mrs, Jr, Sr. Two-part surnames can be concatenated (SMITH-JONES to SMITHJONES) and retained along with alternate entries for SMITH and for JONES, special characters may be eliminated (O'CONNOR to OCONNOR) and prefixes concatenated (VAN DYK to VANDYK). A prefix list is shown in Table 1. Geographic and disease codes will usually have changed over time. It may be necessary to recode fields so that all records share a common system of codes, or to use ranges of codes that are comparable.

Table 1. --List of Surname Prefixes

BON	DI	LE	O
D	DO	LES	ST
DA	DU	LI	STE
DE	EL	LU	VAN
DEL	FITZ	LOS	VANDEN
DEN	L	M	VANDER
DER	LA	MAC	VON
DES	LAS	MC	VONDER

(2) Spelling errors

To get around spelling errors in surnames, a phonetic encoding scheme can be used. We currently use the modified New York State Identification and Intelligence System (NYSIIS) surname code [8]. In the 1950-79 Mortality Data Base file, there were about 200,000 unique surnames which mapped into about 40,000 NYSIIS codes. Based on evalua-

tion studies of earlier linkage projects, we are currently considering making modifications to this coding scheme based on some of the phonetics involved with Canadian names (particularly French names).

(3) Incomplete files

Due to the rules regarding cutoff dates for preparation of statistics from certain files, one may find that records are missing due to late registrations. If the files are assigned numbers in an orderly fashion, a sequence and continuity check of the numerically sorted file can be carried out, missing gaps listed, as well as the first and last record numbers of the files. We have done this for the Mortality Data Base file. Where exposure data files have been maintained separately from the Master Identification file, some utilites can be used to match files for "orphan" records i.e. an exposure record with no corresponding record on the master identification file or vice versa.

(4) Missing identifiers

These can be assessed from SAS output of individual fields, as well as using the availability word for a number of variables. It is advisable to split a field into its component parts - for example, for birth date use year, month and day. Sometimes sex code has been found missing from files.A list of all forenames appearing on the Mortality Data Base has been created. This has been used to impute a sex code e.g., 1=male only, 2=female only, 3=either male or female forename. Sex code is required so that appropriate weights can be assigned for forenames in the frequency weighting.

(5) Lack of documentation of old historical files

Here we have found SAS output very helpful, and created documentation regarding the contents of each field.

(6) Possible correlated data items

Certain data fields may be correlated, therefore caution has to be taken when assigning weights to these items e.g., birth place of father, mother, and a child. In certain instances the information relates to identical items (e.g., an address and postal code); in other cases it may reflect multiple wrong guesses (e.g., a birthdate being incorrectly reported).

(7) Duplicate records not properly identified

It is important that for a two-file linkage, all records that are known in advance to relate to the same individual be properly identified. This is to ensure that any groups to which either record of such a pair may belong can be combined by the linkage system. Typical examples are records relating to women who have both a maiden name and a married surname. One is unlikely to want to discard one record and keep the other, because there may be records on the other file that relate to either surname. A field can be added to the record to contain a value of 1,2,3 etc. to indicate whether this is the first, second or third "duplicate" entry for this record. If no duplicate exists, the value of the field can be set to zero. Such duplicate records must all be assigned the same unique number (in the GIRLS system this is referred to as the SEQUENCE number).

If an intermediary file is used, alternative entries can be put in with different versions of the identifying information. These may be either entries from both files separately or in hybrid form (i.e. certain items from one file and other items from the other file).

(8) An internal linkage should have been done first

Any file that is going to be used for a two-file linkage, should first be examined to determine whether an internal linkage is required to bring together all records which refer to the same individual (or entity). If one is uncertain about whether there are duplicates, sometimes a fairly inexpensive first check may be to sort the file by surname, first forename, and birth date and to create a microfiche copy of the file for visual examination. A great deal of work in a two-file linkage can be saved by first unduplicating in this fashion the two files that are to be linked.

(9) Length of data fields

If two fields are to be compared, the lengths of the data fields need to be compatible. For example, as a standard, we encode ten letters of the surname into the NYSIIS code. If the number of letters in one file is less than ten characters, problems can arise when the codes are compared. It is therefore advisable to use a surname field that is ten characters or greater. If special characters were originally used, the data entry of the field should be large enough to allow for the elimination of these special characters in the preprocessing.

(10) Separating out values where the same field was used for more than one purpose

As an example, the same field on some files may be used for maiden as for alias surname. One may wish to try to separate out the two types of surnames that have been entered, so that during the linkage step appropriate rules can be used.

(11) Several unique numbering systems used over time

In certain files, several numbers may have been used over time to refer to the same individual. In administrative sytems, there may be a rather different problem; one often needs to clarify whether such numbers have ever been reassigned to other individuals.

In certain cases, one may wish to chain all the various numbers that were used by the same person over time and use this as a pocket identifier within which a probabilistic match could be made.

V. PROBABILISTIC RECORD LINKAGE TECHNIQUES

The Basic Principle

There are three major difficulties to be overcome in order to achieve efficient record linkage. The personal identifying items are often inadequate to discriminate between the person to whom a record truly refers, and other persons in the population who have similar names. A second difficulty arises because when people report personal identifiers they frequently make mistakes. The third difficulty arises because of the large volume of records involved in record linkage. Some related difficulties include the setting of appropriate threshold values for acceptance and rejection of linkages, deciding how most efficiently to carry out a multi-step operation, deciding on the number of partial agreements to use and the selection of pocket identifiers.

The objective of the Generalized Iterative Record Linkage System was to make it possible for computer procedures to efficiently carry out the data processing involved in the probabilistic

matching of data files, and to do so easily for a wide variety of diverse data requests. The GIRLS system has involved optimizing four major tasks: (1) the search operation, (2) the decision-making step, (3) the grouping of records, and (4) the retrieval of information.

In the searching step, the sequencing information is used as a means of avoiding the many unprofitable pairings that would have to be examined if every record initiating a search were compared with every other record in the file being searched. Generally for searches of the Mortality Data Base, comparison pairs are created only where both the sex and the phonetically coded form of the surname agree.

For other applications, the sequencing may be by one of several systems of numerical identifier or by phonetically coded surname. Regardless of the means by which the record pairs are brought together, the next step will be a detailed comparison of the remaining identifiers. This is necessary even where the numeric identifiers agree, because such identifiers are occasionally used improperly by persons to whom they do not belong, and sometimes even by a relative of the rightful owner who has the same surname.

At the present time, a test is being made to provide a measure of the usefulness of employing personal identifiers from the Social Insurance Number (SIN) index file to supplement those from the work records, for the purposes of carrying out automated death searches. Not only are the names, birth dates and such more likely to be recorded on the SIN record, they are also more likely to be complete, and as well they will frequently include the mother's maiden surname, which carries considerable discriminating power and is quite unlikely to be available from any work record.

In the decision-making step, each of the remaining identifiers is compared in turn, wherever it is represented on both members of the comparison pair of records.

The odds associated with any specified outcome from the comparison of any identifier are:

$$\text{odds} = \frac{\text{freq of specific outcome in linked pairs}}{\text{freq of specific outcome in unlinked pairs}}$$

This applies equally to agreements, disagreements and to any degree of similarity or dissimilarity no matter how it is defined (as long as both definitions are identical above and below the line).

When pairs are sorted in descending order of total weight, a point is reached at which the record pairs should be judged unlinkable or borderline. To calculate where this threshold should be, two further values are required to be weighted for a two-file linkage. These are:
(1) the likelihood that the individual is represented in the file being searched, so that there is a potential for linkage, and (2) the size of the file being searched, since the opportunity for fortuitous agreement increases in proportion to the file size.

The logarithms of both of these values will be negative. When added in with the weight from the identifier comparisons, the resultant sum is known as the "absolute total weight".

$$W^* = W + \log_2 \frac{Na(L)}{Na} + \log_2 \frac{1}{Nb}$$

where,
$W^* = \log_2$ of the absolute odds in favour of a correct linkage;
$W = \log_2$ of the relative odds in favour of a correct linkage = w1 + w2 + w3 .... where these are each logs to the base 2 of the odds ratios for the successive identifier comparison outcomes;
Na and Na(L) are respectively, the total number of records in the file initiating the searches and the number out of these that will be linked with matching records in the file being searched (or a reasonably close estimate of Na(L)/Na may be used initially); and
Nb = the size of the file being searched.

To calculate w1, w2...., for reasons of convenience it is desirable to treat separately the data derived from linked pairs and that which applies to unlinked pairs. If w is the net weight for the particular identifier comparison outcome, $\log_2$ (frequency in linked pairs) is the negative component of this net weight, and $\log_2$ (1/ frequency in unlinked pairs) is the positive component of the net weight.

Because the negative components of weight vary with the quality of the file initiating the searches i.e. with the reliability of the identifiers as recorded on that file, these negative components need to be recalculated for each new linkage before the final weighting is done. The data may be obtained initially from preliminary machine linkage, numerical linkages where available, or from manual linkages. Examples of how the weights are obtained are discussed in reference [9].

The positive components tend to be stable where the files being searched are the same on successive occasions (e.g., the death file) and can usually be calculated from the frequencies of the identifier values in that file.

The Art of Record Linkage

The art of designing an efficient computerized linkage operation depends less upon theory than an intuitive perception of how best to carry out the comparisons and what outcomes from these are most likely to be revealing, so that they ought to be recognized by the computer.

Some of the intuitively obvious refinements that have actually been put to use in Statistics Canada's death searches have to do with:
(1) Recognition of partial agreement outcomes, e.g., of
- surnames (three levels of agreement/disagreement);
- given names (eight levels of agreement/disagreement, including agreement truncation where the initials agree);
- birth year (up to 6 levels of agreement/disagreement);
- birth month (3 levels);
- birth day (4 levels).

(2) Recognition of cross-agreement, e.g., of
- initials (where there is no straight agreement);
- month and day of birth - as for initials.

(3) Recognition of degrees of compatibility/incompatibility e.g., in
- last known alive year versus death date (up to 4 levels);
- marital status (up to 4 levels for each status on a search record).

(4) Comparison of place of work versus place of death.

(5) Calculation of age at the time of the matching death to determine the likelihood of death in a particular year using life-table data.
(6) Use of death file size for that same year as influencing the odds for a fortuitous similarity of the identifying particulars.

A potential refinement may be judged worth retaining as a part of the linkage procedure where it is used often enough in doubtful matches, and makes a large enough difference in the final decision to link or not to link, to justify the possible added complexity in the programming. The GIRLS system makes it possible to gather such data after a preliminary linkage and again after a final production run.

The best tactic when designing a linkage procedure for a specific operation is to gather such empirical data after a preliminary linkage so that the procedure can be revised before the final weighting. The information needed earliest has to do with the frequencies among linked pairs of the different comparison outcomes recognized by the preliminary linkage procedures. The tabulations ("info outcomes") should recognize all the comparison outcomes likely to be useful in the decision process.

We often find that what one learns by looking at some manual linkages first can be very helpful in planning a study. This aids in working out the appropriate methods to use and in preparing cost estimates. One may have to decide whether there is enough identifying information available to do the linkage. To get an overall estimate of this, one can first imagine how strongly unfavourable the odds would be if one did not know whether any of the items agreed or disagreed, and were linking to a file of a given size. Then, as one compares each item, in turn, and assumes they agree, this will demonstrate the possible extent of the increase in likelihood favouring correct linkage. One can use a global overall weight for the items employed in this exercise, and hence get a ballpark impression of whether or not there are enough items available to make it work (see Tables 2 and 3 for an example).

After the linkage status decision has been made, the system can identify groups of records which potentially refer to the same entity and it can indicate where conflicts exist. A conflict exists where groups do not fit your requirement e.g., one record relating to more than one death record. In the GIRLS system there are two ways of resolving these conflicts - automatic resolution by the system based on the 'best' linkage, or by manual resolution. A combination of the two often works best.

The retrieval of information operation of the system is designed to quickly and concisely aid the user in making decisions regarding the future direction of the linkage process. The GIRLS system can produce reports at the detailed level on weight sets, linked pairs, group reports, information about the linked pairs, and it can also produce estimates for updating the weights. One may wish to produce reports based only on links for which a given condition is true (e.g., all links above a given weight) or for which a condition using variables on the source records may be true (e.g., all known dead cases as known earlier on the worker's nominal roll file).

VI. FUTURE DIRECTIONS

There are three main directions for our future endeavours:
(1) The improvement and expansion of existing search and linkage facilities which could include further development and enrichment of our current files (e.g., addition of occupation and industry on the death file). The NYSIIS code routine needs to be evaluated more fully taking into account the kinds of names found in Canada. A dictionary of accredited comparison procedures needs to be developed from past linkage studies that could serve as a guide for future studies. Results from earlier studies need to be carefully evaluated,

Table 2. —Example of a Possible Census-to-Death Linkage — Likelihood of Fortuitously Selecting the Correct Death Record, Using no Identifiers Other than Sex (Assumes enumeration in 1971 at age 42, death in 1979 at age 50, and male sex)

COMPARISON ITEMS	ODDS	CUMULATIVE ODDS	WEIGHT (10 X $\log_2$)	CUMULATIVE WEIGHT (10 X $\log_2$)	NOTES
Random chance of finding death in 1979 male death file, assuming it is there	1/96,532	1/96,532	-166	-166	1
Likelihood of dying in that year, if alive at the beginning of it	1/131	1/12,645,692	- 70	-236	2
Likelihood of being alive at the beginning of 1979 if enumerated in 1971	1/1.04	1/13,151,520	- 1	-237	3

Note: (1) From death file size, for males dying in 1979.
(2) From life tables for likelihood of death in a 12 month period, for a male of age 50.
(3) From life tables, for the likelihood of survival to age 50 among a cohort of males still alive at age 42.

Table 3. --Example of a Possible Census-to-Death Linkage -- Cumulative Effect of Successive Agreements on the Odds in Favour of a Correct Match, when all Identifiers are Present and Agree

IDENTIFIER AGREEING	ODDS	CUMULATIVE ODDS	WEIGHT	CUMULATIVE WEIGHT
			(10 x $\log_2$)	
(Random chance)	-	1/13,151,520	-237	-237
Surname	2,287/1	1/5,745	+112	-125
First initial	14/1	1/410	+ 38	- 87
Second initial	14/1	1/29	+ 38	- 49
Rest of first name	87/1	3/1	+ 64	+ 15
Marital status	26/1	8/1	+ 14	+ 29
Year of birth	56/1	437/1	+ 58	+ 87
Month of birth	12/1	5,242/1	+ 36	+123
Birth prov/country	8.6/1	45.078/1	+ 31	+154
Ethnicity	3.5/1	157,773/1	+ 18	+172
Parental birthplaces	1.2/1	189,328/1	+ 2	+174
Industry, major	6/1	1,135,968/1	+ 41	+215
Occupation, major	11/1	12,495,648/1	+ 31	+246
Residence province	4.4/1	54,980,851/1	+ 21	+267
Residence city	72/1	3,958,621,272/1	+ 62	+329

particularly with respect to the use of intermediate files and the use of alive follow-up procedures as were used in the Ontario miners study [10]. Further refinements are needed in developing a file of non-links to get weight estimates, particularly where the comparisons are fairly complex (e.g., weighting of forenames).

(2) **The development of new and much needed data bases** which would identify, in a more systematic fashion than heretofore, the occupational and environmental circumstances of people, and which could be used as starting point files, to initiate the searches for subsequent health histories. Here data collection rules and forms need to be more clearly developed which could be used by industry. Use of new files such as census of agriculture, farm registers, and census of population files can be exploited. The use of existing files for alive and morbidity follow-up need to be explored.

(3) **The exploration with other agencies of any collaborations** that would be productive for generation of the required statistics, and for setting up the necessary communication network and financial support to implement such recommendations.

ACKNOWLEDGMENTS

The author would like to thank Dr. Newcombe for his contribution to this paper. The opinions expressed in this paper are those of the author and do not necessarily represent the views of Statistics Canada.

REFERENCES

[1] Smith, M.E. and Newcombe, H.B., "Automated Follow-up Facilities in Canada." AJPH Vol. 70, No. 12, pp. 1261-1268, 1980.

[2] Smith, M.E., "Long-term Medical Follow-up in Canada." In: Peto, R., Schneidermen, M. eds. Quantification of Occupational Cancer. Banbury Report 9, Cold Spring Harbor Laboratory, pp. 675-688, 1981.

[3] Smith, M.E., "Development of a National Record Linkage Program in Canada." American Statistical Association - 1982 Proceedings of the Section on Survey Research Methods, pp. 303-308, 1982.

[4] World Health Organization, "Guidelines for the Study of Genetic Effects in Human Popu lations." Environmental Health Criteria 46 (in press).

[5] Howe, G.R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies." Computers and Biomedical Research, Vol. 14, pp. 327-340, 1981.

[6] Smith, M.E. and Silins, J., "Generalized Iterative Record Linkage System." American Statistical Association - 1981 Proceedings of the Social Statistics Section, pp. 128-137, 1981.

[7] Hill, T., "Generalized Iterative Record Linkage System: GIRLS." Research and General Systems, Informatics Services and Development Division, Statistics Canada, Ottawa, 1985.

[8] Lynch, B.T. and Arends, W.L., "Selection of a Surname Coding Scheme for the SRS Record Linkage System." Statistical Reporting Services, U.S. Department of Agriculture, Washington, D.C., 1977.

[9] Newcombe, H.B. and Abbatt, J.D., "Probabilistic Record Linkage in Epidemiology." Red Book Series Report No. 5. Eldorado Resources Limited, Suite 400, 255 Albert Street, Ottawa, Ontario., K1P 6A9, November 1983.

[10] Muller, J., Wheeler, W.C., Gentleman, J.F., Suranyi, G., Kusiak, R., "Study of Mortality of Ontario Miners." International Conference on Occupational Radiation Safety in Mining, Vol. I, pp. 335-343, 1984.

GENERALIZED ITERATIVE RECORD LINKAGE SYSTEM

Ted Hill and Francis Pring-Mill, Statistics Canada

ABSTRACT

The Generalized Iterative Record Linkage System (GIRLS) project was initiated at Statistics Canada in 1978. This paper outlines the concepts behind the system, and summarizes how these have been implemented to provide a powerful tool suitable for a variety of record linkage applications.

1.0 RECORD LINKAGE AND GIRLS

Record linkage is the process of identifying two or more records which refer to the same entity. An entity could be a person, or a business, for example.

In the case where records have unique identifiers (for example, social insurance number), the process of linking is relatively simple as it involves matching on only one field. However in cases where records do not have unique identifiers, information from several fields typically has to be compared to estimate the likelihood that a potential link is a 'true' one. For these cases record linkage is a probabilistic process, and it is for this situation that GIRLS was designed.

GIRLS stands for the "Generalized Iterative Record Linkage System" which has been developed at Statistics Canada, starting in 1978. Since then, the system has been systematically maintained and enhanced on a regular basis.

GIRLS provides a command language in which you can write your own rules for comparing records. Statistically-derived weights are attached to potential links according to the outcomes of these comparisons. Your GIRLS commands are automatically translated into PL/1 (a high-level programming language), compiled, link-edited and executed on the input files to generate an online project database of potential links and the records involved in them. Using other GIRLS commands, you can then query this database to see the results. If these are not satisfactory, you can update the database in various ways, or simply change your comparison rules and try again.

To this end, GIRLS breaks the linkage process into a sequence of distinct phases. Each phase involves deciding on values for system parameters, examining their effect, and adjusting the values as necessary before going on to the next phase. Results from later phases often suggest adjustments to earlier phases. Because phases are distinct, you can easily retrace your steps, run an earlier phase again with new adjustments, run intermediate phases as they are, and quickly catch up to where you were. This is why GIRLS is called an 'iterative' record linkage system.

The principal aims of GIRLS are:

1. To enable you to develop the best comparison rules and statistical weights for the purpose of your linkage project.
2. To provide a convenient framework for this development.
3. To encourage iterative refinement through a sequence of phases and reports.
4. To make the final linkage fast, cheap, and accurate.

Examples of GIRLS applications include:

1. 'Follow-up' studies

   Health Division at Statistics Canada currently runs linkage projects with files provided by employers of individuals exposed to potential health hazards in the course of their work (e.g. uranium miners). These are linked with the Canadian Mortality Database to check that the proportion of matches found is not above normal.

   Such studies can detect risks to health associated with particular occupations, thus pointing the way to causes of disease. They can also aid in testing the long-term effects of curative measures.

2. Building 'case histories'

   Separate records referring to the same person often accumulate in a system. For example, a new record is often made each time an individual is admitted to a hospital. GIRLS can conveniently bring these records together, enabling larger composite records to be made representing 'case histories' for individuals.

2.0 FEATURES OF GIRLS

In the past, record linkage systems have usually been tied to methodologies suited to particular application requirements. GIRLS provides a general solution to developing particular linkage systems.

Its principal features are:

1. A sequence of phases encourages iterative refinement of the linkage process.
2. The full power of database management technology is provided. This includes: automatic maintenance of data integrity across separate files, checkpointing facilities for project recovery, as well as back-up and restore procedures.
3. Both 'one-' and 'two-' file linkages can be performed. (One-file, or internal, linkages can be useful for unduplicating a file or creating composite records.)
4. A variety of samples of records from the input files can be extracted for the purposes of experimenting.
5. A simple but powerful GIRLS command language is provided to write comparison rules, update the project database, and obtain a wide variety of reports at many levels of detail.
6. The commands provided for writing comparison rules can detect full agreement, various levels of partial agreement, disagreement, and missing values. They can also specify cross comparisons of different fields, as well as rules to be executed conditional on the outcomes of previous comparisons.
7. For special purposes you can also write your own PL/1 code and have it included in the Compare program automatically generated from your GIRLS commands.

8. Statistically-derived weights are generated and attached to links to reflect the probability that the records being compared refer to the same entity.

9. Potential links are automatically classified as: rejected, possible, or definite, by comparing link weights against threshold values. You specify these threshold values, and you can easily adjust them. You can also re-classify links manually.

10. Records which refer to the same entity are grouped. Where conflicts exist within groups, these can be resolved either automatically by the system, or manually on a record-by-record basis. (For example, a conflict would exist when records are expected to link to at most one record on the 'other' file, but a group contains some which have linked to several records.)

11. Both batch and online modes are available. Online enables fast iterative adjustment of system parameters by providing quick feedback as to the current state of the project database.

3.0 BASIC OPERATIONS

The phases of the GIRLS system can be grouped into three main operations.

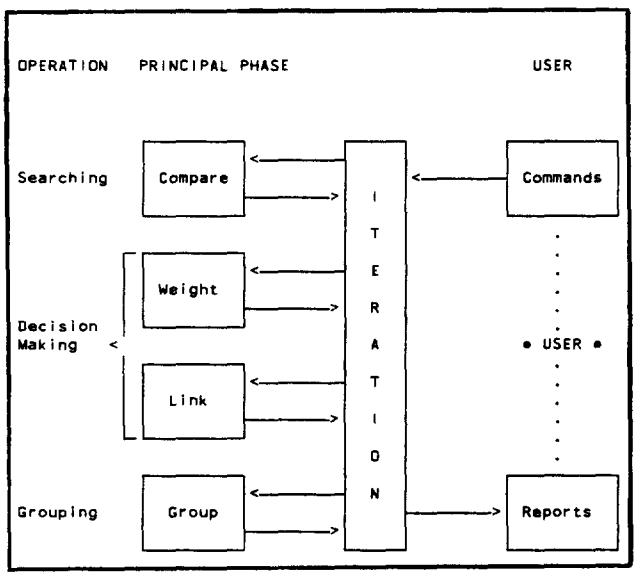
1. Searching.

2. Decision Making.

3. Grouping.

This is shown below:

Figure 1: Basic operations



3.1 Searching

In this operation, pairs of records are compared field by field according to comparison rules you specify. Theoretically, every possible pair of records should be compared. However the number of possible pairs in even a small file is very large. So for practical reasons, records are first blocked into smaller 'pockets' in such a way that it is realistic to look for links only within pockets.

You use GIRLS commands to define your input files, indicate which fields define your pockets, select your sample of records, and specify rules according to which your records are to be compared. Your GIRLS commands are then automatically translated into a PL/1 program, called the Compare program, which is executed on your input files to produce the project database of potential links.

You can write rules to compare fields with values that are: character (e.g. surname), numeric (e.g. birthyear), or coded (e.g. for fields with a small number of discrete values such as birthplace). Your rules can be made conditional on particular outcomes from previous comparisons. You can also specify cross comparisons of different fields (for example, first given name with second given name, in the event that straight comparisons of each field have not already produced an agreement). If your rules do not fit conveniently into the format of the GIRLS command language, you can also write them yourself in PL/1 and have them included in the Compare program.

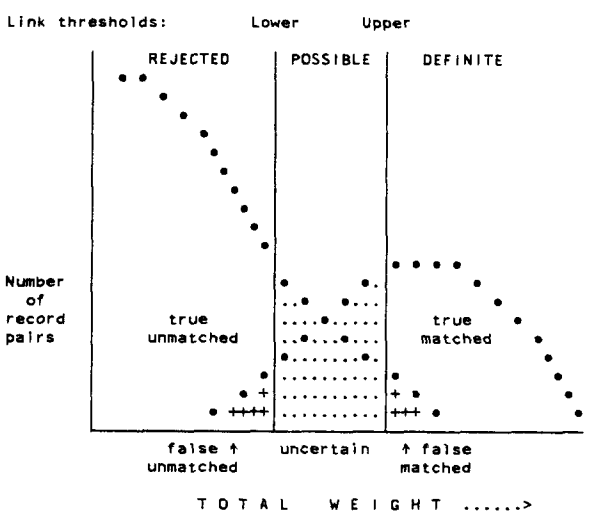
The outcome of having executed a comparison rule can be: agreement, one of various levels of partial agreement, disagreement, or missing. You can specify a 'global' weight to be attached in the event of each one of these possible outcomes.

3.2 Decision Making

In this operation, the potential links generated by the Compare program are evaluated. This involves updating link weights and comparing them against threshold values to decide which to keep and which to reject. Link weights are updated with 'frequency weight sets' which reflect the probability of particular agreements happening by chance. These weights are derived according to formulae developed by Geoff Howe[1], Mike Eagen, and David Binder from methodologies proposed by Howard Newcombe[2], Ivan Fellegi and Alan Sunter[3].

After weight update, the status of links is determined by comparing their total weights against two threshold values. Links with weights above the upper are classified as 'definite', those with weights below the lower threshold are 'rejected', those with weights between the two are 'possible'. This is shown in Figure 2, which is explained as follows:

Figure 2: Link thresholds classify links into three statuses



Let all possible record pairs be divided into two populations: those record pairs which are 'truly matched', and those which are 'truly unmatched'. The goal of the linkage project is then to find the members of the 'truly matched' population. Because it represents all possible record pairs which do not match, the true unmatched population will be far greater than the true matched one. This is shown on the left. The smaller true matched population is shown on the right. The problem is the overlap in the middle, because for these record pairs it is not obvious to which distribution they belong.

The two threshold lines show how GIRLS handles this problem area. Links to the right of the upper threshold are considered 'definite', those to the left of the lower are considered 'rejected', those between the two are considered 'possible'. While permitting flexibility, this approach allows two types of error which any linkage project should aim to minimize.

First is the 'false unmatched' area on the left. These are the record pairs which have been rejected even though they were part of the true matched population. This can happen when information is incomplete or inaccurate on records which 'should' have matched. Second is the 'false matched' area on the right. These are the record pairs which have been accepted even though they were part of the true unmatched population. This can happen when records look very similar even though they refer to different entities, e.g. the different members of the same family. At first glance, these two areas can be minimized simply by setting the thresholds far apart. However this makes for many possible links in between, which will have to be resolved later. By adjusting the thresholds and inspecting various samples of links, however, you can choose the best thresholds for your purposes.

3.3 Grouping

In this operation, the records are grouped according to the status of the links between them. Records may have just one link to another record, or they may have several links to several records. Records joined either by possible or definite links are arranged into 'major' groups - which can be large. Within major groups, records joined by definite links are further arranged into 'minor' groups. A major group may therefore contain several minor groups, and it is the minor groups that contain the best links.

At this stage, 'conflicts' may arise, typically when groups are larger than you want them to be. The system identifies conflicts for you based on your linkage requirement, e.g. one-to-one (i.e. groups are to contain pairs of records only, one from each file). Resolving the conflicts can be done in either of two ways, or both:

1. You can let the system resolve conflicts automatically. This is called 'automatic resolution'. In this case all you specify is your linkage requirement, e.g. one-to-one, many-to-one, or one-to-many.

2. You can resolve the conflicts yourself manually. This is called 'manual resolution'.

You can also use both methods, automatic resolution first followed by an examination of the results and some manual rearrangement where necessary.

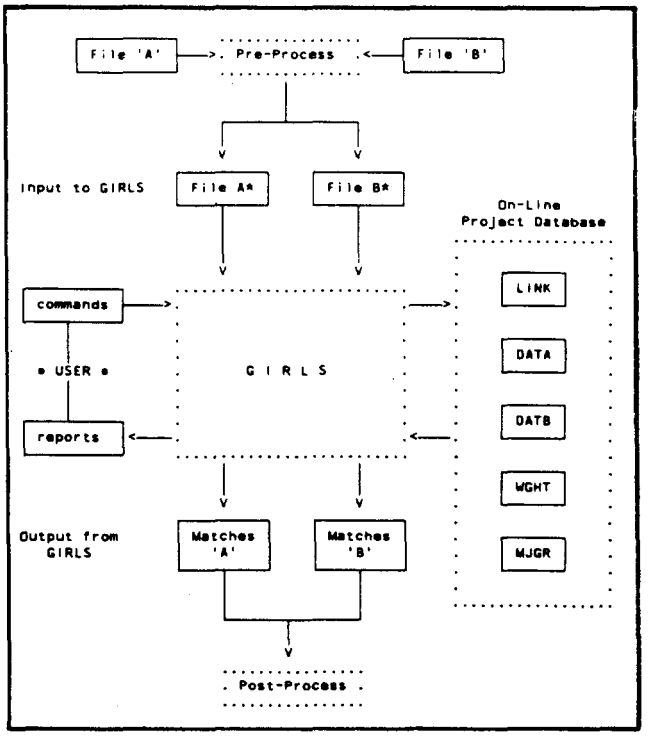
4.0 ENVIRONMENT

4.1 Flowchart

Figure 3 shows a flowchart overview of the system. At the top, two files of records (File 'A' and File 'B') are pre-processed for input to GIRLS. In the middle, records are compared according to your comparison rules, and an online project database is created on the right. This consists of potential links (LINK), the records involved in them (DATA and DATB), together with other files for use later.

On the left, the user is shown interacting with the system via GIRLS commands in the light of the linkage project requirements and feedback from reports as to the current state of the project database. At the bottom, two files of 'matches' are produced. On each output file, each original input record that has been linked is identified by a unique sequence number and has a number identifying the group to which it has been assigned.

Figure 3: Flowchart overview of the system



4.2 Iteration

Iterative refinement of the linkage process can include adjustments to:

1. *COMPARISON RULES*

   From the very many possible links which exist between all possible record pairs, these rules determine which are to be considered the 'potential' links to be written to the project database. These rules can be written, re-written, ordered and re-ordered, so as to produce enough suitable links as efficiently as possible.

2. *WEIGHTS*

   These are attached to links via the comparison rules which applied to the records when the links were formed. It is easy to modify these weights, and thereby select the best ones for your purposes.

3. *THRESHOLD VALUES*

These determine the proportion of definite, possible, and rejected links. The best mixture depends on the aim of a particular linkage project, and is determined by experimenting with the thresholds, and seeing the types of groups which are formed.

For example, for a statistical study it may be satisfactory to find 90% of the links. While for other types of study, it may be necessary not to miss any of them.

4.3 GIRLS Project Files

Making the iterative concept work in practice requires maintaining data integrity across several files when any one of them is being updated. For this reason, an integrated database approach has been taken using the RAPID Database Management System developed at Statistics Canada.[4] The principal RAPID files are:

1. *WEIGHT FILE (WGHT)*

   For each field to be weighted, this contains the values for the field and the frequency weight for each value.

2. *LINK FILE (LINK)*

   For each 'potential' link between a pair of records, this file contains: - the outcomes (agree, disagree) for each comparison rule - the current total weight of the link - the current status of the link (definite, possible, or rejected) - other system control information

3. *DATA FILES (DATA, DATB)*

   These contain the records involved in potential links.

4. *MAJOR GROUP FILE (MJGR)*

   This contains information for each group, enabling reports to be made according to type of group, e.g. "display all groups having more than six records".

4.4 Typical Scenario

A typical (abbreviated) scenario for a GIRLS linkage project might be:

1. Write rules specifying how fields are to be compared.
2. Calculate frequency weight sets (a SAS function is provided to do this job).
3. Use sampling facilities to select a sample of records from the pre-processed input files.
4. Adjust appropriate system parameters, both in batch mode and/or online, until satisfactory results are obtained.
5. Run the full linkage in batch.

Using the system online greatly speeds up the iterative adjustment of linkage parameters. The result can be a linkage process uniquely adapted to the purposes of your linkage project.

Favourable reports from current users include:

- The system is 'comfortable' to use because you remain in control at all stages.
- The command language enables both updates to be made easily, and reports to be obtained to verify intended results.
- Iteration can be continued for as long as it takes for you to be satisfied.

5.0 PHASES

This section briefly outlines the various phases of the GIRLS system. Further details are given in the Strategy Guide and in the User Guide.

5.1 Pre-Process

Purpose: to get files ready for linking

- standardize names and addresses
- validity check
- decide on POCKET
- assign SEQUENCE numbers. (These uniquely identify each record.)
- make duplicate records, when you know records match although they look different. E.g. a record for an individual using her maiden name, and another record for the same individual using her married name.
- recode, e.g. from different codes to common code. (For example, from one hospital coding system to another.)
- encode, e.g. from surname to NYSIIS code
- split files, e.g. by sex, year
- sort files by POCKET

5.2 Weight Creation

Purpose: to create global and frequency weights

- use the provided SAS function to:
  - calculate frequency weights themselves
  - generate GIRLS weight update commands
  - calculate global weights (optional)

"The rarer the value, the higher the FREQ weight."

The frequency weight formula used is:

$$FW_i = 10 \times \log_2 \left(\frac{\text{total number of records}}{\text{No. occurrences of field value } i} \right)$$

where "FWi" is the frequency weight for value "i". For example, the value "SMITH" for the SURNAME field could have a frequency weight of "40".

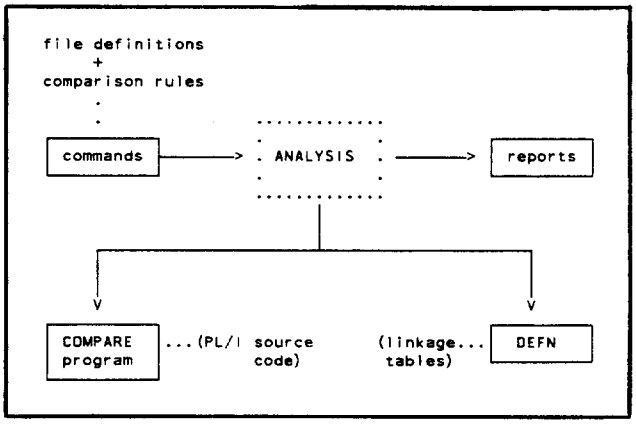
5.3 Analysis

Purpose: to specify comparison rules

- define input files
- choose fields to compare
  - character e.g. surname
  - numeric e.g. birthyear
  - coded e.g. marital status
  - conditional and cross comparisons
  - your own PL/1 code

  choose possible outcomes to weight
  - fully, partially agree
  - disagree
  - missing
- your rules are then translated into a PL/1 program called the 'Compare' program

Figure 4: The Analysis phase



5.4 Compare

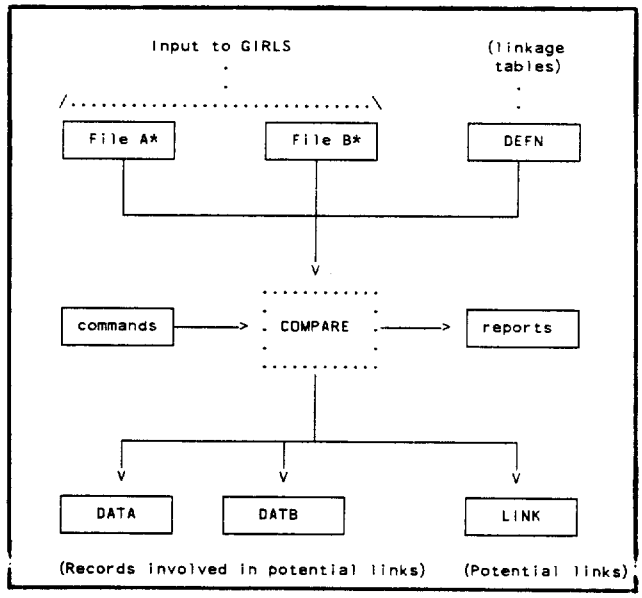
Purpose: to build the linkage database

- set thresholds: upper, lower, and cutoff so as to reject obvious non-matches quickly
- select a sample of pockets with which to experiment
- execute the Compare program

  The comparison rules start assigning global weights to potential links, which are rejected as soon as either current total weight falls below cutoff or if final total weight will be less than the lower threshold.

  The linkage database of potential links and all records involved in them is created.

Figure 5: The Compare phase



5.5 Weight Update

Purpose: to apply and/or modify the weights

- look at link weights 'before'
- apply weights

  You attach frequency weight sets to comparison rules. The system finds all links to which each rule applies and updates the link weights accordingly.
- look at link weights 'after'

5.6 Link

Purpose: to assign statuses to the links

- set a lower and an upper threshold

  The system classifies links by comparing their total weights against these thresholds and assigning a status of definite, possible, or rejected (as explained in Section 3.2).
- inspect results

5.7 Group

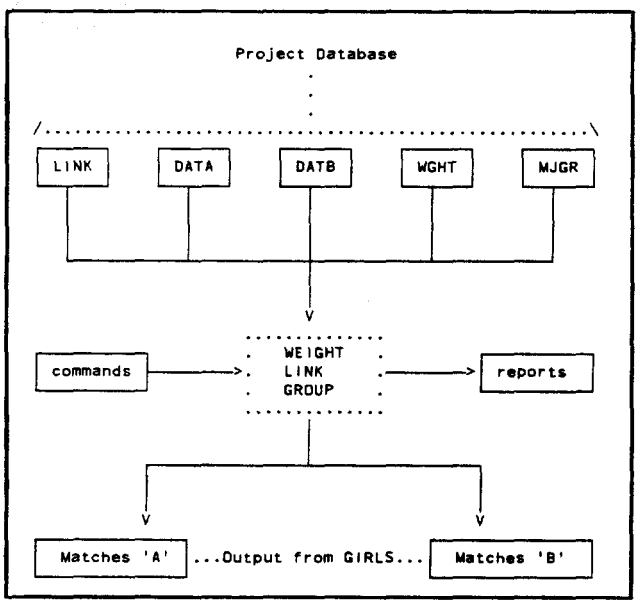
Purpose: to build groups of records

- the system builds 'major' and 'minor' groups of records based on their link status.
  - major groups have both definite and possible links
  - minor groups have definite links only
  - i.e. minor groups contain the best links.
- the system combines groups which share duplicated records. For example, combining a group which contains Mary Smith (maiden) with a group which contains Mary Brown (married).

- resolve group conflicts, either automatically or manually
- output final versions of groups

The Weight, Link, and Group phases are represented below.

Figure 6: The Weight, Link, and Group phases



5.8 Post-Process

Purpose: to use the results of GIRLS

- e.g. for an internal linkage, prepare composite records to represent case histories
- e.g. for a two-file linkage, for each group, generate one record to represent all the members
- create summary files

6.0 EXAMPLE

This is a simple example to show how the GIRLS linkage process works for a two-file linkage.

Part 1 of Figure 7 represents the contents of two files to be linked by GIRLS. File DATA contains 6 records which are to be matched against the 9 records of file DATB. Let the pocket identifier be the SURNAME field (which means that records are compared only if SURNAME agrees on the two records). ROW specifies the row number of the record on the files, and the "..." represents missing data.

Part 2 of Figure 7 shows examples of frequency weights on the WGHT file for the fields SURNAME, MARST and BIRTHYR. (For example, the weight for the surname "Quigley" is "100".) We will be using these weights later to calculate the total weights of links.

Figure 7: Example: two input files and a Weight file

Part 1.-- DATA and DATB file

File	ROW	SURNAME	MARST	BIRTHYR
	1	Barnes	01	1950
D	2	Barnes	..	1950
A	3	Jones	03	....
T	4	Jones	02	1960
A	5	Quigley	03	....
	6	Quigley	02	1960
	1	Barnes	01	1950
	2	Barnes	..	1960
D	3	Barnes	02	1960
A	4	Jones	03	....
T	5	Jones	02	1960
B	6	Jones	..	1960
	7	Jones	02	1960
	8	Quigley	02	1970
	9	Quigley	03	1970

Part 2.-- WGHT file

SURNAME	MARST	BIRTHYR	WEIGHT
Barnes			40
Jones			10
Quigley			100
	01		10
	02		20
	03		30
		1950	10
		1960	20
		1970	30

The table below shows the links we have on the project database LINK file after executing the Compare phase and applying the frequency weights in the WGHT file. The columns in the table are explained below.

Figure 8: Example: the resulting Link file

LINK ROW	DATA ROW	DATB ROW	SURNAME OUTCOME D(-10)	@SURNAME RESULT	MARST D(-20)	BIRTHYR OUTCOME D(-40)	@BIRTHYR RESULT	TOTWGHT	STATUS
1	1	1	A	Barnes	01	A	1950	60	POS
2	2	1	A	Barnes	M	A	1950	50	POS
3	3	4	A	Jones	03	M	....	40	POS
4	4	5	A	Jones	02	A	1960	50	POS
5	4	7	A	Jones	02	A	1960	50	POS
6	5	8	A	Quigley	D	M	.....	80	DEF
7	5	9	A	Quigley	03	M	.....	130	DEF
8	6	8	A	Quigley	02	D	.....	80	DEF
9	6	9	A	Quigley	D	D	.....	40	POS

• THRESH=(40.75) •

Notes:

1. "LINK ROW" identifies the record number of each link. This identifies the link in subsequent reports.
2. "DATA ROW" and "DATB ROW" indicate the File 'A' and File 'B' records that are involved in a link.
3. "SURNAME" and "BIRTHYR" are fields containing the outcomes of comparison. These are "A" (agree), "D" (disagree), "M" (missing on one or both records).
4. For agreement, the "@SURNAME" and "@BIRTHYR" fields contain the result on which the fields agreed.
5. The "MARST" field contains the outcome of the comparison if it is "D" (disagree) or "M" (missing), or the

result on which the fields agreed if the outcome was agreement.

6. For disagreement, the weights are specified under SURNAME, MARST, and BIRTHYR. E.g. for disagreement on BIRTHYR the weight added is "-40".

7. "TOTWGHT" (total weight) is the sum of the relevant agreement and disagreement weights for each link.

8. "STATUS" shows the link status for each link. This is based on the total weight (TOTWGHT) for the link and the current threshold values (THRESH). In this example, the lower threshold is "40", and the upper "75". "POSS" corresponds to 'possible' and "DEF" to 'definite'. (In this example, comparisons resulting in a total weight less than the lower threshold (40) are excluded from further processing.)

For example, for Link 8 we calculate the total weight (TOTWGHT) from the information on the LINK file, and the weights on the WGHT file, as follows:

Figure 9: Example: calculating the weight for Link 8

Comparison	Value	Weight
SURNAME	QUIGLEY	100
MARST	02	20
BIRTHYR	disagree	-40
	TOTWGHT	= 80

The final table below shows the group numbers assigned to the records after grouping. Records with the same group number refer to the same individual. Records having no group number have no matches on the 'other' file. These groups are based on the DATA ROW, DATB ROW, and STATUS values shown on the LINK file.

For example, Group 1 contains three "Barnes" records: A(1), A(2), and B(1), i.e. two File 'A' records have been grouped with one File 'B' record. If our linkage requirement is one-to-one, then this group contains a 'conflict' which will have to be resolved.

Figure 10: Example: group numbers show the linked records

File	ROW	SURNAME	MARST	BIRTHYR	GROUP
	1	Barnes	01	1950	1
	2	Barnes	...	1950	1
DATA	3	Jones	03	....	2
	4	Jones	02	1960	3
	5	Quigley	03	....	4
	6	Quigley	02	1960	4
	1	Barnes	01	1950	1
	2	Barnes	..	1960	...
	3	Barnes	02	1960	...
	4	Jones	03	....	2
DATB	5	Jones	02	1960	3
	6	Jones	..	1960	...
	7	Jones	02	1960	3
	8	Quigley	02	1970	4
	9	Quigley	03	1970	4

7.0 GIRLS TRAINING

As the GIRLS system is relatively complex, we strongly recommend participating in the introductory Seminar, followed by experimenting with an Example Project that has been set up for training purposes.

7.1 GIRLS Seminar

This is a one-day seminar which covers all aspects of the GIRLS system. It is given by the GIRLS system staff on an ad hoc basis. It requires the use of an overhead projector and can be presented at Statistics Canada or elsewhere. This Seminar is a valuable introduction to the system.

7.2 Example Project

This is a miniature GIRLS linkage project with two small files of test data. It consists of a sequence of batch jobs containing examples of the typical use of GIRLS commands. Submitting these jobs one at a time produces a sequence of listings showing the stages by which the records from the two files become linked. You are also encouraged to make a copy of these jobs, change the commands, and then re-submit the jobs to see the effect of your changes. This Example Project is a valuable learning tool.

8.0 HARDWARE AND SOFTWARE REQUIREMENTS

GIRLS requires the following hardware and software:

- IBM 370 compatible hardware with at least two million bytes of storage (real or virtual).
- The OS MVS or MVT operating system.
- The RAPID database management system.
- The IBM PL/1 compiler.
- Direct access storage devices (3330, 3350, 3380 etc.)
- The following are not mandatory but are highly desirable: SAS (Statistical Analysis System) in order to use the Weight Creation function, TSO or ISPF.

NOTES AND REFERENCES

[1] Howe, G.R. and Lindsay, J.(1981). A generalized iterative record linkage system for use in medical follow-up studies. Computers and Biomedical Research, vol 14, 327-340.

[2] Newcombe, H.B. (1967). Record linking: the design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics, vol 19, 335-359.

[3] Fellegi, I.P. and Sunter, A.B. (1969). A theory of record linkage. Journal of the American Statistical Association, vol 64, 1183-1210.

[4] RAPID Database Management System. Informatics Systems Division, Research and General Systems Subdivision, Statistics Canada.