Westat

Improving lives through research™

Fine-grained Record Integration and Linkage Tool (FRIL)

DRAFT User Manual

October 8, 2014

Prepared for:
Institute of Education Sciences
National Center for Education Statistics
Washington, D.C.

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

CONTENTS

SECTION			PAGE
1	**Disclosure Analysis Using Probabilistic Linkage Methods**		1
	1.1	Purpose of external linkage for disclosure analysis	1
	1.2	Measure of disclosure risk	3
	1.3	Probabilistic linkage	5
2	**Getting Started: A Primer**		6
	2.1	Identifying variables in microdata file being released for public use	7
	2.2	Review public-use microdata to identify common variable (s)	7
	2.3	Derive variables to make them comparable	8
	2.4	Identify blocking variables (exact matching variables)	8
	2.5	Considerations when preparing matching files and setting parameters	9
3	**Estimating Disclosure Risk Using FRIL**		10
References			29
Appendix A – Explaining the Terminology in the External Matching Procedure			30

FIGURES

Figure 1	Screenshot for metric options	11
Figure 2	Linkage mode screen	15
Figure 3	List of fields in the data source	16
Figure 4	Add to out model options	17
Figure 5	Gear icon in linkage mode screen	18
Figure 6	Join conditions and output columns screen	19
Figure 7	Indicating pairings for comparisons	20
Figure 8	Initiating EM computations	22
Figure 9	Select all-to-all comparisons	23
Figure 11	Assignment of weights	24
Figure 12	SVM join method type selection	26
Figure 13	Beginning the actual linkage process in FRIL	27
Figure 15	FRIL output report	28

1. Disclosure Analysis Using Probabilistic Linkage Methods

It has long been recognized by the Federal government that the release of statistical data for public use may lead to the disclosure of the identity of individual units. Federal and state agencies are struggling with the need and requirement to release study data while, at the same time, maintain the confidentiality of the data provided by individuals and/or institutions. There are several laws to ensure that information provided by individuals is kept confidential. These include the Privacy Act of 1974,[1] the Education Sciences Reform Act of 2002, the USA Patriot Act of 2001, and the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA).[2] Failure to protect the confidentiality of individuals in accordance with these laws can result in a fine and/or a prison term, as well as doing irreparable damage to the study and reputation of the Federal agency.

1.1 Purpose of external linkage for disclosure analysis

The goal for the Federal agency when conducting and disseminating survey data is to maintain the confidentiality of data provided by respondents while limiting data perturbations in a manner that would minimize the impact on data quality. Some Federal agencies provide guidelines and standardized procedures for ensuring data confidentiality and quality. The National Center for Education Statistics

[1] Section 552a protects records maintained on individuals: http://www.justice.gov/opcl/privstat.htm.

[2] Title V of the E-Government Act of 2002, Public Law 107-347 (December 17, 2002). These standards affect all Executive Branch agencies: http://www.eia.doe.gov/oss/CIPSEA.pdf

(NCES) as part of the Institute for Education Sciences (IES) has implemented confidentiality standards for all NCES data releases.[3]

The IES confidentiality standards generally require two separate procedures for public-use microdata files: (1) identify any sensitive variables and records through external matching and mask the variables and records that match using directed (deterministic) data swapping, and (2) introduce an additional measure of uncertainty into the data using random swapping. The software package, *DataSwap*, was reviewed and approved by NCES to satisfy the second requirement for IES data dissemination of both public-use and restricted-use data. The focus of this document is to describe the issues and solutions for satisfying the first procedure (external matching).

When public release of the data is only available through a Data Analysis System (DAS) that has internal Statistical Disclosure Controls (SDC) mechanisms such as the NAEP Data Explorer (NDE), the external matching procedure is no longer required, though the second procedure (random data swap) is still required for fulfillment of IES confidentiality standards. The IES-approved DAS provides the data suppression that ensures confidentiality. However, when public-use microdata are released by the IES, the external matching procedures are normally required prior to these data dissemination. There may be some IES studies that the IES Disclosure Review Board (DRB) exempts from the external matching requirement. In some household surveys the sampling rate is very small, the population is large, and limited geographic information is provided to the public. There may be studies where, based on the data collected, there are no external files available for matching or even if external files are available for matching, there are no data elements collected that can be used to match to the external file(s). In other cases, the data collected for the studies are so specialized and/or limited in size and scope that external matching may not be reasonable. In these cases, data suppression and/or random data swapping may be sufficient for public-use microdata dissemination.

By law, public-use data cannot be used (whether alone, or in conjunction with other available data) to identify any of the respondents, whether they be school, teacher, students (or other respondents like parents). Thus, if one can identify a school (and therefore a teacher or student) based upon matching the data against external files, some of the matching characteristics of the survey data must be perturbed to prevent such identification. The IES DRB has determined that the "Rule of 3" should be used to determine whether a close match is treated as a disclosure risk. In this context, the "Rule of 3" means that the true or actual match cannot be the closest or second closest match (a description of key terms used in

[3] For the detailed standards, see http://nces.ed.gov/statprog/2002/std4_2.asp.

this document is in appendix A). The matching approach, software, and methodology are critical in identifying any records that can be matched (and thus need more masking or perturbation).

One of the problems with the external matching procedures for the IES DRB has been that different IES studies have used different approaches and procedures for generating the external matching routines. Various home-grown Euclidean distance programs have been used to identify matches but there are some problems in terms of accuracy with this methodology. Beginning in 1995, many studies used AutoMatch software (Jaro 1989) that incorporated probabilistic record-linkage matching procedures that could better determine whether the survey record was a likely match with an external file record. The IES DRB prefers standardized software that allows them to better evaluate the validity of the procedures and the results. That is, using approved software that generates standardized reports/results enables the IES DRB to more quickly and confidently determine whether the matching procedures were properly implemented. The IES DRB developed a software package, DRISK, to emulate AutoMatch and provide to all contractors. However, both DRISK and AutoMatch are older programs that cannot run on newer operating systems. Thus recently, a replacement program, Fine-grained Record Integration and Linkage (FRIL), was identified that (1) is universally available in the public domain; (2) is distributed by the Centers for Disease Control (CDC); (3) provides the functionality found in AutoMatch and DRISK; and, (4) is reasonably simple to use and provides a useful documentation through a User's Guide. This document extends the FRIL user manual and provides a school-based example to guide use of the software for IES disclosure analyses. References to AutoMatch are included to help the user transition from AutoMatch to FRIL.

1.2 Measures of disclosure risk

Disclosure analysis balances the risk of re-identification of subjects represented in samples of microdata, contingency tables, or stratified estimates against the loss of information due to attempts to prevent re-identification. Identifying disclosure risk both precedes and follows disclosure-proofing of data and reports released to the public. Prior to discussing measures of re-identification risk through record linkage approaches, other disclosure risk measures are discussed to inform the reader of available approaches if record linkage is not required, or to supplement the record linkage results. Data disclosure risk can be estimated using sample-based and population-based approaches. The sample-based approaches include the exhaustive tabulation approach created by Westat (Li and Krenzke, 2013) and implemented in the *InitialRisk* software at IES, and the Special Unique Detection Algorithm (SUDA) by Elliot, et al. (2002). Two risk measures for re-identification risk include a log-linear modeling approach

by Skinner and Shlomo (2008), and approximations to re-identification risk in Mu-Argus 4.2, which was mainly developed at Statistics Netherlands (see Mu-Argus 4.2 manual[4]).

Domingo-Ferrer and Torra (2001) describe distance-based linkage, probabilistic linkage, and interval disclosure as the prevailing methods for measuring disclosure risk. The authors describe these terms in part as follows:

- Distance-based linkage – distances are computed between the original and masked data sets for pairings of records. The computed distances usually represent the degree of syntactic dissimilarity between two text strings, one from the original and the other from the masked data set. A record is considered 'linked' when the nearest or second-nearest record in the masked data set turns out to be the corresponding original record.

- Probability-based linkage – the matching algorithm described in Jaro (1989) uses a linear sum assignment model to 'pair' records in the two files to be matched. The percentage of correctly paired records is a measure of disclosure risk. The approach requires the user to provide an upper bound of the probability of a false match, and a false non-match.

- Interval disclosure -- where an attacker is completely sure that the original value lies in the interval around the masked value. Reiter (2005) groups 1) interval disclosure into detection of distinct (or nearly so) combinations of variable values, and 2) distance-based and probabilistic linkage into linkage methods; he advocates use of the Duncan-Lambert framework, "indirect probabilistic linkage", to assess the risk of disclosure. The indirect linkage method presupposes that potential intruders have the confidential identifying information that a data provider is withholding from a public release data source.

Risk measures based on a sufficient sample can be used when a statistical matching risk assessment for a certain set of files may not be feasible for all files that may be available to an intruder. Alternatively, record linkage can be used to estimate the probability of a match based on the distributions of identifying values in disclosure-proofed data sets, which gives an indication of which records, variables, and values are at risk. A summary is given in Winkler (1993) and Diniz da Silva, et al. (2010).

[4] Available at http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf (accessed April 14, 2014).

A rigorous probabilistic linkage of public release data to data available to potential intruders amounts to a test of disclosure risk levels. Many articles in peer-reviewed, statistical and information sciences journals have built a strong case for using probabilistic linkage complemented by distance-based similarity measures. It seems unlikely that intruders could find a better method for linking data already available to them to a new data release.

1.3 Probabilistic linkage

Probabilistic linkage methods draw on more recent extensions by Winkler (2002) and Yancey (2004) of Fellegi and Sunter's (1969) theory of record linkage. Linkage of public use data to other publically available data sources begins with a comparison vector. The vector represents, for a product space A X B, the agreement (or not) of attributes (variables) of data sources A and B. In this context, A refers to records representing an entity such as an organization or person and B to public records related to organizations or persons. Pairs of rows, A.r,B.r, partition into unobserved set M of correct links and set U of incorrect links. An intruder seeks to discriminate set M from set U. Data exposure control has the goal of making the risk of that happening quantifiably small.

In probabilistic linkage literature, we call the outcomes of a series of comparisons of variables v in common between data being released (A) and in data already available to intruders (B) an agreement pattern:

$$(A.v_i \approx B.v_j), (A.v_k \approx B.v_l), \ldots,$$

where the subscripts denote specific variables taken in any order from A and B and the ≈ is a comparison operator that yields, in the original sense of a comparison vector, a binary (0,1) agreement pattern that depends on the similarity of the values of the variables. After comparisons of all pairings of records in A and B, a summary of the results reduces the agreement patterns to distinct agreement patterns with frequencies. That is, an "agreement pattern" c, representing outcomes of comparisons of variables (v) in files A and B in a comparison vector, populates a k-tuple of binary values (or more generally in the range {0,1}), e.g., 1,0,1,1,0,0,1… Each distinct agreement pattern has unobserved m (match) and u (unmatch) probabilities:

$$m=Pr(c|\ r \text{ in } M)\ ;\ u=Pr(c|\ r \text{ in } U);$$

The likelihood that a pairing of rows from A and B with agreement pattern c belongs in M = m_c/u_c, so one can rank agreement patterns by $score_c = \log (m_c/u_c)$ and define decision rules to control Type I and II errors.

Since we cannot observe the m and u probabilities directly, we have to approximate or infer them, as follows:

- Approximate: Assuming conditional independence of the elements of an agreement pattern c, for each c_i (where i represents the i^{th} element in the agreement pattern c) estimate m_c' as the expected error in using c_i to predict a correct link, and u_c' as the risk of coincidental agreement of c_i. The sum of log (m_c' / u_c') over the c_i when used to order potential links usually proves robust to small and even moderate errors in estimates;

- Infer: The iterative expectation-maximization (EM) algorithm computes values of the parameters that maximize the likelihood function m/u.

Each pairing of records takes on the likelihood score of its agreement pattern. The linkage algorithm screens out pairs with scores that fall below a level that can be justified as the result of coincidental agreement. Even after screening out pairs with very low scores, a linkage that achieves adequate separation of sets M and U will, for large A and B, have a U-shaped frequency distribution of likelihood scores. Probabilistic linkage minimizes the frequency of pairs with scores between upper and lower score thresholds. Relatively small numbers of "tagged" correct links support further scoring by using logistic regression estimates of parameters, conversions of parameter estimates to estimated probabilities, and receiver operator characteristic (ROC) analysis of linkage results

Initially data linkage software packages such as AutoMatch or FRIL require users to approximate the probabilities that values match, given that they are from the same entity, and it calculates non-match probabilities from data. FRIL also offers different methods for comparisons of different types and content of variables (for example, using a Soundex code rather than an exact match). It and other current data linkage packages contain some features of fuzzy and distance-based deterministic matching, even though they are referred to as probabilistic record linkage software.

2. Getting Started: A Primer

This section provides helpful hints toward linking data slated for public release to external data files and identifying potential disclosure risk. One has to play the role of intruder and look for the best way to compromise the identities of subjects represented in a data release. Some variables or combination of variables may prove to be effective or ineffective in identifying matches, but all potentially identifying variables should be reviewed and tested, and the ones deemed the most reliable should be used in the analysis.

2.1 Identifying variables in microdata file being released for public use

While obvious identifiers such as Social Security Numbers (SSN), driver's license numbers, or names are excluded routinely from data being released for public use, various combinations of other variables may also add substantially to the risk of re-identifying subjects. Even when one excludes records with rare and extreme values, such as annual incomes over $100 million or persons over 95 years of age, the number of subjects in *classes* defined by demographic (e.g., age group, race, gender, county of residence, country of birth) and activity (garden club member, licensed pilot, interest in fishing), for instance, narrows down to very few cases. Access to more than a few of a subject's attributes enables an intruder to use robust linkage methods to work around errors and variations in data, whether they occur naturally or intentionally for the purpose of disclosure control.

Selecting all potentially identifying variables in data being released gives linkage a better chance of controlling excessive disclosure risk. Controlling risk does not guarantee that an intruder will never be able to identify any of the subjects represented in data. Instead it balances disclosure risk with benefits that accrue from releasing useful data to the public. Transforming variables and deriving new variables from multiple variables may (or may not) improve the chances of finding hidden disclosure risks.

2.2 Review public-use microdata to identify common variable(s)

Next, variables in the public microdata file (B) that are in common with the file to be released to the public (A) need to be identified. Common variables are those to use in matching between files A and B. A comparison of variables A.a and B.b would be expected to match two records belonging to the same subject most of the time. As shown in table 1, for each coded variable, a simple two-by-two contingency table of counts of comparison outcome by match type will show the extent each pairing of a variable *a* from A and *b* from B will discriminate true (A.r=B.r) from false matches (A.r≠B.r) of records.

Table 1. Simple matching outcome table.

	Same Subject	Different Subjects
Variable Matches	$n_{i=i,a=b}$	$n_{i=i,a\neq b}$
Variable Does Not Match	$n_{i\neq i,a=b}$	$n_{i\neq i,a\neq b}$

Linkage sensitivity $\equiv n_{i=i,a=b} / (n_{i=i,a=b} + n_{i\neq i,a=b})$

Linkage specificity $\equiv n_{i\neq i,a\neq b} / (n_{i=i,a\neq b} + n_{i\neq i,a\neq b})$

For example, a variable such as age would match almost all of the time for the same subject (rarely it might not match due to recording errors or intentional blurring), while it would not match most of the time for different subjects (unless both happened to be the same age). Meanwhile, a variable like race would match most of the time for the same subject (but may not match for subjects with multiple races, for example), while it would often match for different subjects as well. None of the individual identifying variables has to contribute nearly perfect discrimination. The value of a comparison vector for a match may have some variables that do not match precisely. So long as the overall pattern of the vector strongly suggests a match, probabilistic linkage tolerates uncorrelated errors and variations in identifying variables.

2.3 Derive variables to make them comparable

Data that is derived or estimated, such as the proportion of male to female students or the proportion of each race at the school level, may combine with data from other surveys to re-identify some of the schools participating in the survey. Continuing the example of school data, an intruder could select from Common Core Data (CCD), Private School Survey (PSS) Quality Education Data (QED), and other sources (Census, State websites, any related databases).Transformations, aggregations, and derivations are two edged swords in data linkage. They may blur distinctions that could help discriminate true from false disclosures, or they may reduce distinctions due to incidental variation and error and improve the ability to detect potential disclosures in advance of a data release. Potentially beneficial derivations include

- Summing variables (e.g. Free lunch + reduced lunch);
- Collapsing variables (e.g. urbancentric to urban, suburban and rural breakouts);
- Aggregating variables (e.g. student gender or race to school level).

2.4 Identify blocking variables (exact matching variables)

As the scale of robust linkage increases beyond a few thousands of records, the idea of comparing every possible distinct pair of records in A and B becomes computationally burdensome. Probabilistic linkage methods typically include blocking on one or more of variables such as these:

- Geographic location variables (region, state, etc.);
- Public/private school classification;
- Gender, broad age groups, and other reliable variables with discrete values.

Blocking records during linkage restricts comparisons of record pairs to those within blocks. Imperfect blocking would lead to true matches being assigned to different blocks and eliminated prior to being compared. To avoid this problem, more than one set of blocking variables can be used and the linkage results combined under multiple blocking schemes.

When the scale of linkage allows, with better computing power, or better indexing methods permits, one would do better to compare all possible pairs of A and B records. Blocking merely provides a convenient way to perform robust linkage while working within computing limits.

2.5 Considerations when preparing matching files and setting parameters

A data analyst familiar with the survey data and external data must be involved in the preparation of the data files and setting the parameters for the linkage software. The processing is not "one-size-fits-all" so the software must be adapted to meet the needs of each study. Improper data creation procedures and/or parameter settings could lead to matching different subjects or not matching the same subject. Thus testing is required to optimize the process and maximize the potential for identifying matches.

Some parameters that will be necessary for accurate matching include:

1) **Weights for each matching variable.** Assigning weights to pairs of variables being compared requires knowledge of the reliability and validity of a match. Equal weights can also be used if all variables have comparable or unknown reliability. FRIL has the functionality to generate weights for the variables – sometimes setting some of highly correlated variables to zero. In the test file example in the User's Guide, the FRIL-generated weights are presented. Note that the user should test the various ways to set the weights in order to identify the optimal approach for each study.

2) **Missing value placeholders.** The data analyst must determine how to specify standardized missing values and determine an appropriate matching score (sometimes referred to as a matching weight) to assign to missing values. For example, if age is being compared between two records, and one record has a missing value for age, then it may be more appropriate to assign a score of zero to that variable, rather than considering it a non-match. When comparing variables for which missing values may have some meaning, such as a second level of street address that does not exist in some cases, a more neutral value of 0.5 may be more appropriate.

3) **Tolerance of minor differences in continuous numeric variables.** To enhance the probabilistic linkage results by tolerating minor and often incidental differences in continuous or integer numeric variables, we may accept "fuzzy matches". Counts of persons in a group or ages of individuals, for example, may be accepted as equivalent if lower or upper bounds around their observed values overlap. A plus or minus *delta* (tolerance) value, could define a comparison as a match where an age of 40 in data A falls within the range of 40, plus or minus 15%; that is, between 34 and 46 in data B. This is a common tolerance measure for a continuous integer variable used in NCES data confidentiality studies. While a continuous numeric variable such as the number of students in a school or the time that an event occurred may not serve well as an identifying variable if compared without a fuzz or tolerance delta, it may after applying an appropriate delta value. Generally speaking, any transformation of the values of a variable that reduces noise (incidental variation) and retains some discriminatory power may recast a variable as an identifying variable. Functions that "clean" strings by converting them to lower case and removing punctuation, for example, will often transform too noisy variables into useful identifying variables. When we extend the concepts of linkage sensitivity and specificity from a column variable to a comparison vector, distinct patterns of matches and non-matches have sensitivities and specificities of linkage. The pattern (permutation) of variable matches and non-matches in a comparison vector, weighted to reflect the ratio of the sensitivity and specificity of the variables, adds up to an estimate of the likelihood of correct match, or *score*. Each possible pairing of the records in data A and data B has a score, but relatively few of the pairings have positive scores. Testing a standard, symmetric 15% against other values to maximize matching potential is recommended.

As previously noted, a poorly specified linkage may fail to find any links between microdata and another set of public data, and the user may conclude that means data are disclosure-proofed. Thus finding disclosure-risk cases is proof of validity in the match specifications and data matching design.

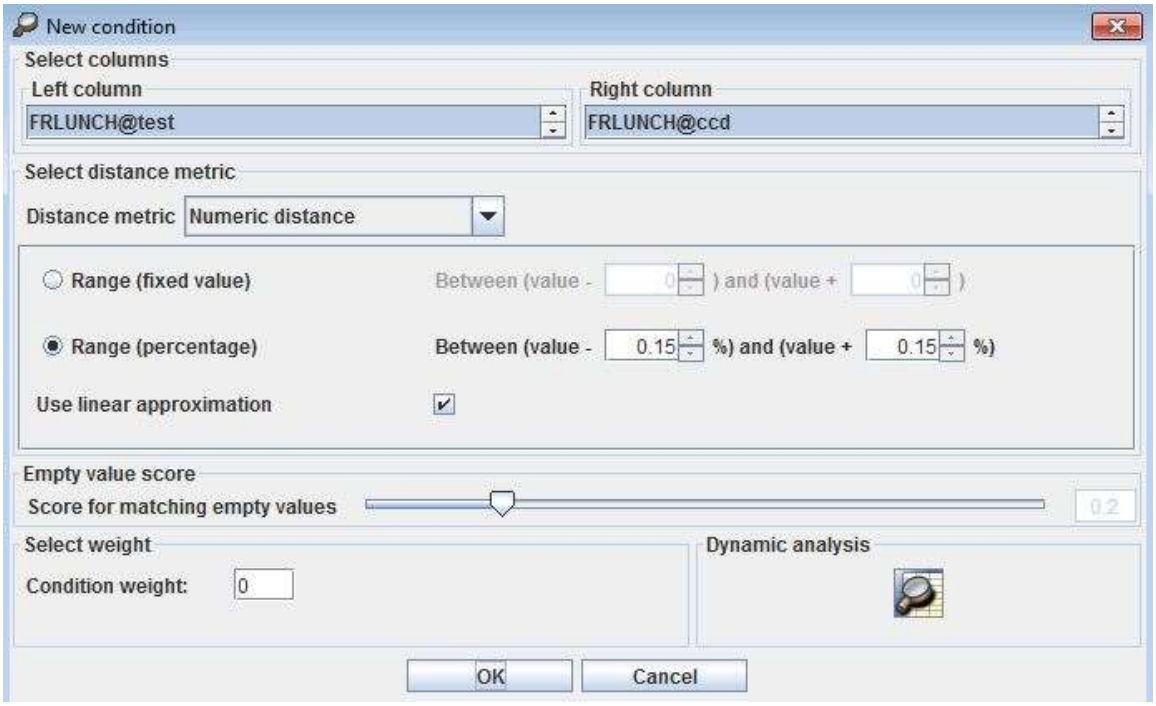
3. Estimating Disclosure Risk Using FRIL

The Web site, http://fril.sourceforge.net/, supports downloads of the Fine-grained Records Integration and Linkage (FRIL). The license.txt file distributed with the FRIL package gives credit to the initial developers of the original code: the Department of Math and Computer Science, Emory University and the Centers for Disease Control and Prevention. It goes on to say that the contents of this file may be used under the terms of either the GNU General Public License Version 2 or later (the "GPL"), or the GNU Lesser General Public License Version 2.1 or later (the "LGPL"), in which case the provisions of the GPL or the LGPL are applicable. The download folders contain technical documentation and tutorials

at http://fril.sourceforge.net/publications.html. The FRIL-Tutorial-3.2.pdf provides an in-depth review of the software. This guide serves as a synopsis of the useful information in conducting and understanding external matching for the purposes of disclosure analysis, though users would benefit from a full reading of the FRIL tutorial. The synopsis has a focus on features of FRIL of special value in disclosure control of releases of statistics and descriptive data to the general public. The more comprehensive FRIL tutorial takes the reader step-by-step through the general architecture of the FRIL system as well as how to bring up the application, define data sources, and specify which variables to add to a comparison vector. The overview of FRIL for use in disclosure control builds on initial steps as presented in the FRIL tutorial.

The FRIL tool offers several features that make it especially appropriate for disclosure-proofing of public use tables, microdata, and other quantitative data releases that carry with them the risk of exposure of confidential or private information. FRIL offers options for linking on numeric as well as character data. The methods for linking numeric data fields support transformations and approximations. In tables that contain counts such as number of students or number of teachers; for example, the Distance Metric options for numeric variables include ranges around a value in levels and percentages as shown in the figure 1 screenshot.

Figure 1. Screenshot for metric options.



Lower and upper ends of ranges do not have to be the same. For example, if the number of students in file 1 is between 5 percent less and 10 percent more than the corresponding variable in file 2, then the full match weight would be given. This non-exact matching rule is sometimes called "fuzzy" matching, and it helps reduce the impact of small variations and errors in the recording of data (technically, improve the sensitivity of linkage).When multiple variables are matched this way, it reduces the chances of a false match (technically, controls specificity of linkage). Fuzzy matches of multiple variables offset the tendency of imprecise matches on a single variable to produce false matches.

To evaluate matches of character variables, FRIL offers special methods for names, postal codes, and similarity of strings. For example, the Jaro-Winkler function computes a value of a distance metric that approaches 0 when strings have neither content nor order in common, and has a value of 1 for identical string values.

Here, the steps performed in the probabilistic matching of data are described below for a mock education study for illustrative purposes. The steps below are necessary for preparing and implementing a Disclosure Analysis Plan (DAP) created for the Disclosure Review Board (DRB) when record linkage is required. For each step, the procedures that were previously designated for AutoMatch are now incorporated into the FRIL process for data matching. For purposes of explaining the steps, data from typical education studies are used as the test case. A typical IES survey data file that requires probabilistic record linkage matching to the CCD would have exact matching variables (blocking), and non-exact matching variables (continuous and derived variables). The assumption made is that if a school from the public released survey data can match the CCD data, then the school, and potentially the students could be identified. The probabilistic functionality of FRIL parallels the key procedures used previously in AutoMatch. Probabilistic matching using sophisticated software such as FRIL is better suited to meet the needs of the DRB than various deterministic and/or Euclidean distance approaches.

Step 1. Prepare source data.

There are two sets of source data: the sampled school databases, and the master school files from which the sampled microdata came. In this example, the TEST DATA third-grade school databases would be the sampled school databases, and the Common Core of Data (CCD) database used to create the sampling frame would be the master school files. If the TEST DATA include private schools, then the Private School Sample (PSS) would be included as well. (Note – other available data (such as QED, Census, etc.) may also be matched against the TEST DATA if the data elements warrant this. For the sake of illustration, this test will be limited to the CCD) We are asking whether a clever intruder using

scientifically valid methods could re-identify a school by linking data in the CCD (B) to the TEST DATA (A).

For AutoMatch users, there would be ASCII file extracts (fixed length) from the TEST DATA school databases: third-grade public schools and CCD ASCII extract of schools containing Grade 3. Variables in the ASCII file are separated by a blank space. For FRIL, there is more flexibility in formats that will be discussed. Each file includes the NCES school identification code (solely used for the verification of true matches) for each record, as well as all variables that are in common with the CCD file. This sometimes requires variables to be derived in order to properly match. For example, Percent Free Lunch in school could be the sum of Free Lunch and Reduced Lunch in the CCD to better match available data from the school questionnaire. Another example is when international educational assessment questionnaires define a school's locale in a way that cannot easily match to what is available on the CCD or PSS. To use this data for confidentiality checks, it is reclassified to rural and non-rural so that the survey and CCD data can be adequately matched.

When using FRIL, there would be a file extract from the external master school files: CCD third-grade schools. Third-grade schools are defined by the respective grade enrollment variable. If the school has a third-grade enrollment of one or more students, then it would be included in the third-grade file. The CCD extracts need to be checked to ensure that they list all TEST DATA from sampled schools. If the CCD incorrectly lists a TEST DATA sampled school as having a zero grade enrollment, then the record is manually added to the appropriate extract file with the enrollment data corrected.

In running FRIL, no additional extract files will need to be created. The TEST DATA third-grade study database can serve as input files, as will the CCD school file. FRIL allows a variety of input file formats, including comma separated value (csv) files, ASCII data files, and Excel files. When defining the data sources, FRIL allows the user to select only those variables that are to be matched. FRIL also has a filtering process for the source data files, so that the user can, for example, select only public or private schools from the TEST DATA school databases using the school sector variable, and selecting third-grade schools from the CCD files using the grade enrollment variables. The user would still need to check that all of the TEST DATA sampled schools have a positive grade enrollment variable within the CCD file. If the CCD listed a TEST DATA sampled school as having a zero grade enrollment, then the record can edited/added during the filtering process with the enrollment data corrected.

FRIL has a graphical user-friendly interface that takes users through an array of options and parameter settings. Users see a range of options and can customize linkage to fit different situations. Given the wide variety of options, it may prove easier to begin with a basic configuration file that carries

forward a typical selection of options and adjust a few to fit a specific situation. FRIL saves configuration files as xml documents. A basic configuration file is included along with the electronic version of this users' guide. The FRIL-Tutorial-3.2.pdf that steps users through examples of preparing source data is included in the Web distribution of FRIL.

Step 2. Calculate match and non-match rates for common variables.

The deterministic swapping procedure involves dividing schools into blocks based on available data. Variables thought to be without error, i.e., those variables that always or almost always agree between the sampled data and the master school files, are used to partition the sampled data files into blocks. In this context the reasoning behind using blocks is that an intruder trying to identify a school would first limit his search based on easily identifiable school characteristics. Variables that have been used as blocking variables may include school sector (public or private), school locale (rural or non-rural), grades in school, and Census region.

The match rate is the percentage of record pairs representing the same subject (true matches) that match on that variable. For example, if 80% of the correctly matched record pairs match on school enrollment, then the match rate for school enrollment is 80%. Variables with match rates of 90 percent or above are generally used as blocking variables. The non-match rate is the percentage of record pairs representing different subjects (false matches) that match on that variable. For example, if 2% of the falsely matched record pairs match on school enrollment, then the non-match rate for school enrollment is 2%.

For FRIL, match rates will be calculated for all common variables between the TEST DATA school data and the CCD, QED, PSS or other matching and/or sampling frame files. The match rates will be used to determine the blocking variables. Because FRIL does not require the calculation of non-match rates for its linking procedures, non-match rates will not be calculated[5].

Step 3. Link the sampled school data with the master school frame data.

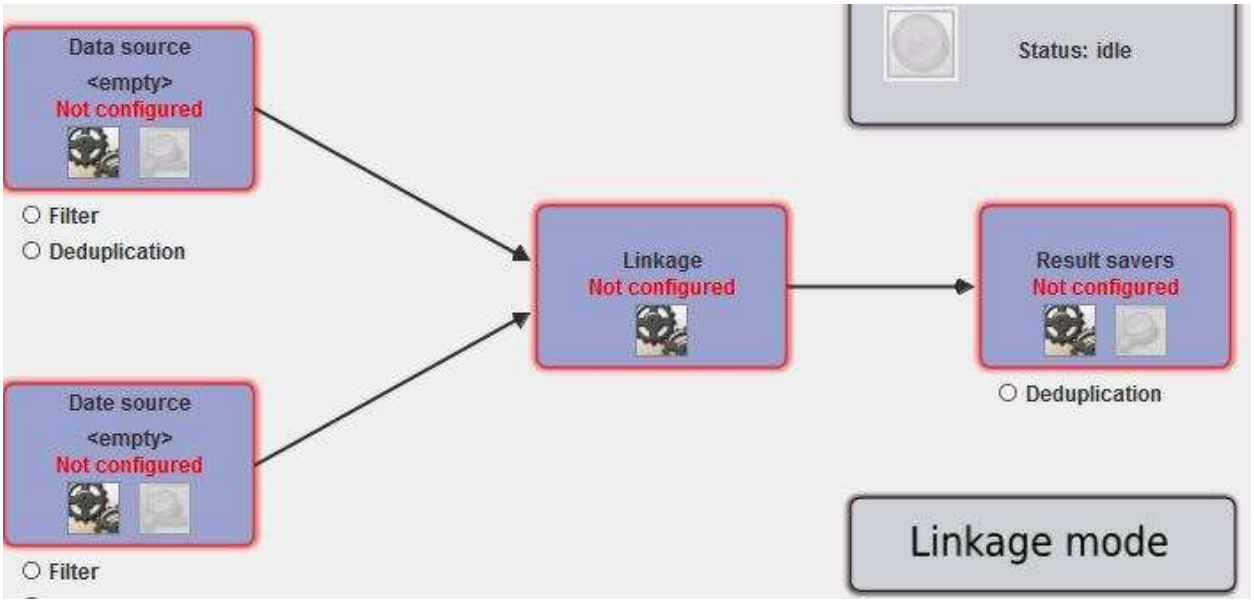
This step involves telling the matching program which pairs of variables should be compared from the sampled school data file to the master school sampling frames. The program must also be told

[5] In AutoMatch, the ASCII file extracts from the TEST DATA school data and the CCD and PSS sampling frames are compared to calculate both match and non-match rates for all common variables. Both the match and non-match rates are used when creating the matching programs in AutoMatch.

how they should be matched; for example, whether an exact match must be made, or whether a partial match can be assigned if the compared values are close but not exact[6].

Using FRIL, the record linkage process starts after the two data sources (i.e. sampled school (A) data and master school frame (B)) have been defined. The FRIL Users' Manual serves as a step-by-step guide to defining data sources and matching up variables from two data sources. If using a standard configuration as a starting point, FRIL will attempt to link to source data file names automatically. To specify or modify file names, click on the "gear" icon in the Data source boxes on the Linkage Mode screen in figure 2.

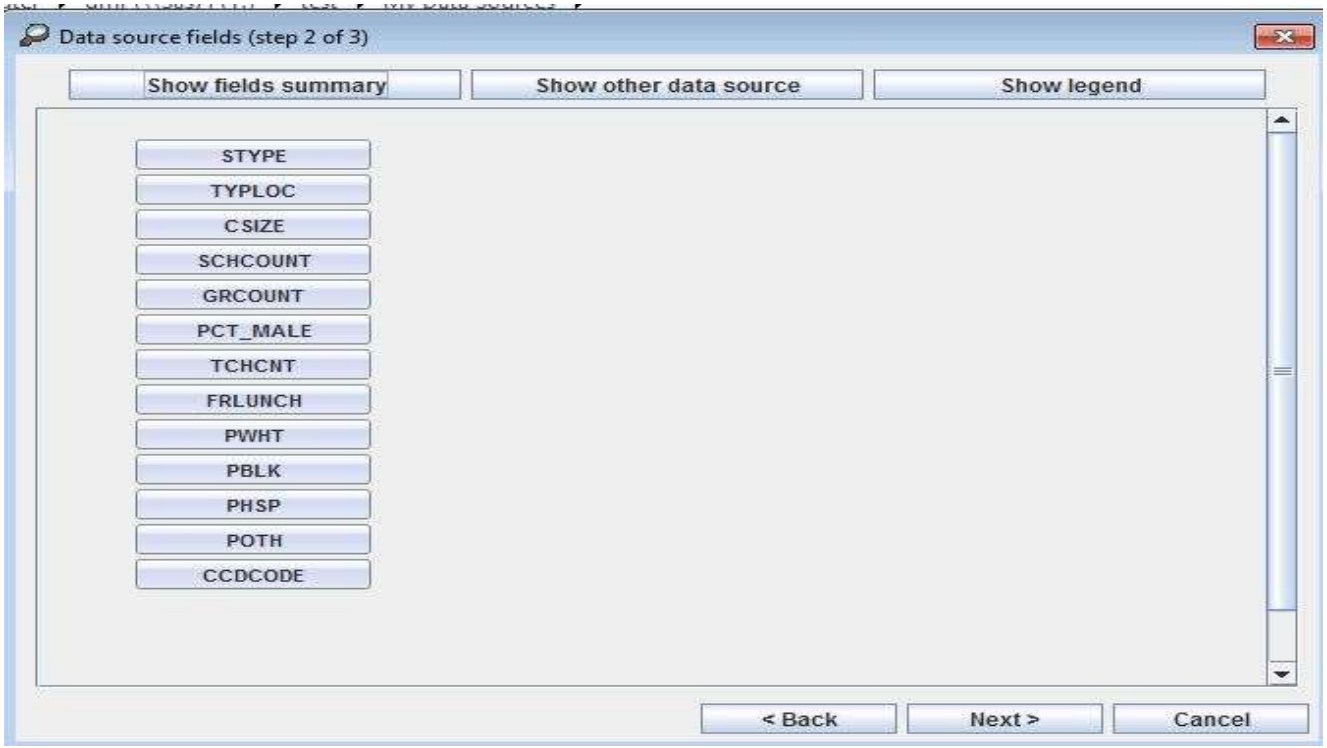
Figure 2. Linkage mode screen.



[6] In AutoMatch, the matching program provides the necessary linking data. Each variable found in the TEST DATA school data ASCII file extract must be matched to a variable found in either the CCD and/or other matching ASCII file extract. A line is written in the program for each pair of variables to be matched. The line indicates the name of the variable in the TEST DATA file, the name of the variable in the CCD or PSS file, the type of comparison to be made, the match rate between the two variables, and the non-match rate between the two variables. For TEST DATA, there are two types of matches used: exact matches, and delta percentage comparisons. For exact matches, a matching weight is assigned if and only if the two values being compared are equal. For delta percentage comparisons, a matching weight is assigned if the difference between the two variables is less than a given percentage. In previous DRB analyses, it is common practice to set the delta to 15 percent but this can be adjusted based on the particular data in the study. The smaller the percentage difference, the higher the matching weight assigned; the larger the percentage difference, the smaller the matching weight assigned. The matching weight that could be assigned to each set of variables being compared is calculated based on the match and non-match rates provided. At the end of the program, the minimum score (threshold) that defines a match between school records is set. As part of a conservative approach to school disclosure, this value was small, as that would guarantee a large number of matches.

As described in the FRIL tutorial, a directory search will set up a different data source for the Linkage process. The [Next >] button at the bottom of the screen will display a list of the fields in the data source, shown in figure 3.

Figure 3. List of fields in the data source.



Right-click on one of the field names to see options:

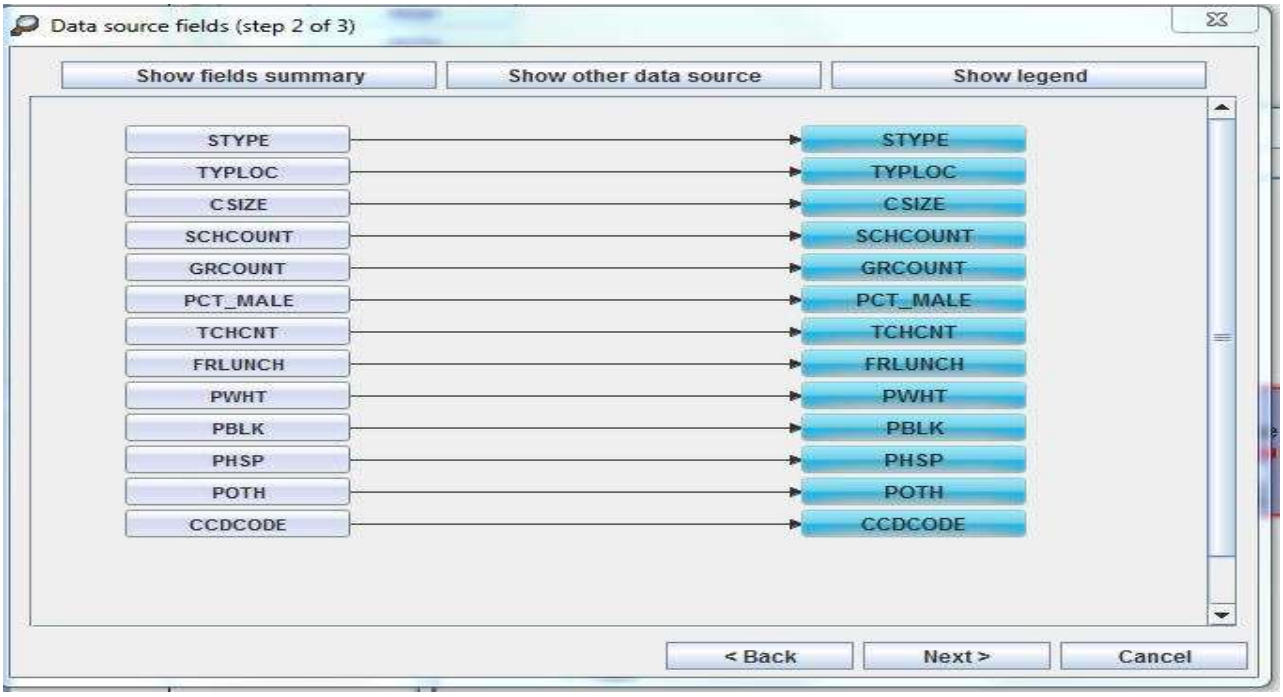
Add to out model

Add all columns to out model

Use converter

FRIL is asking which of the fields to pass along to the next stage. If selecting relatively few fields from among many fields, the "Add to out model" option may suffice. Selecting the "Add all columns to out model" will generally work well since selections of fields to match comes along later, as shown in figure 4.

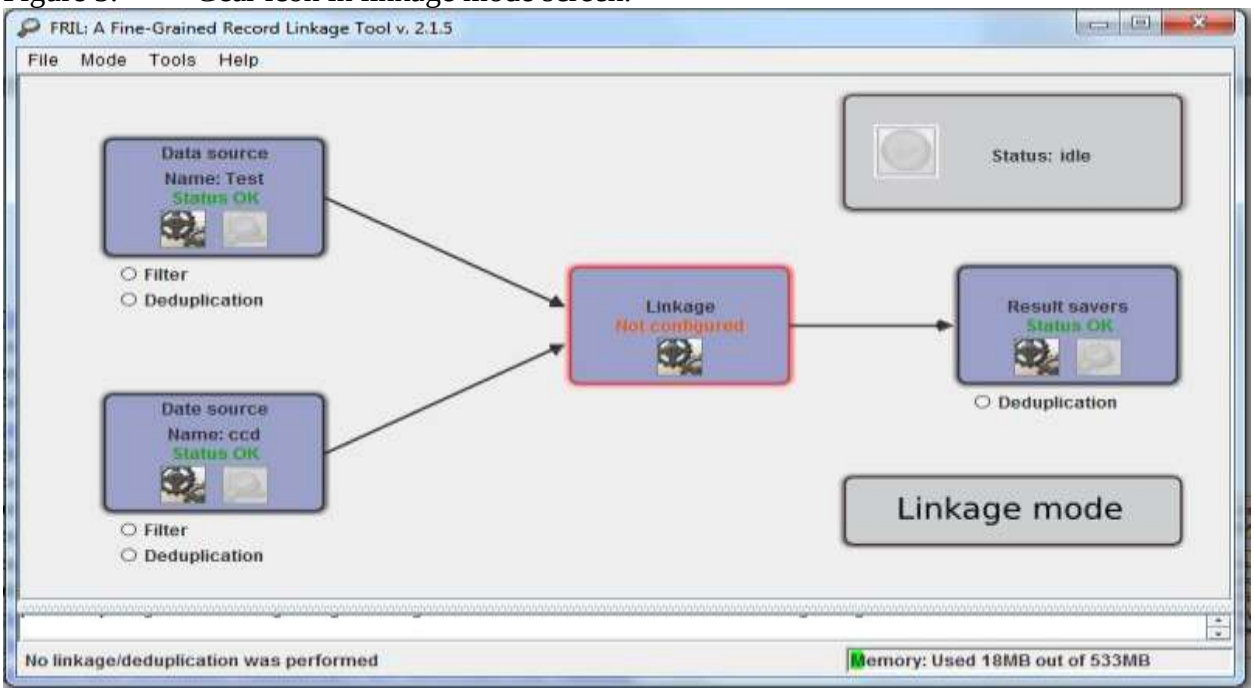
Figure 4. Add to out model options.



Both the sampled school data file and the master school frame need to be linked before the matching of variables can begin.

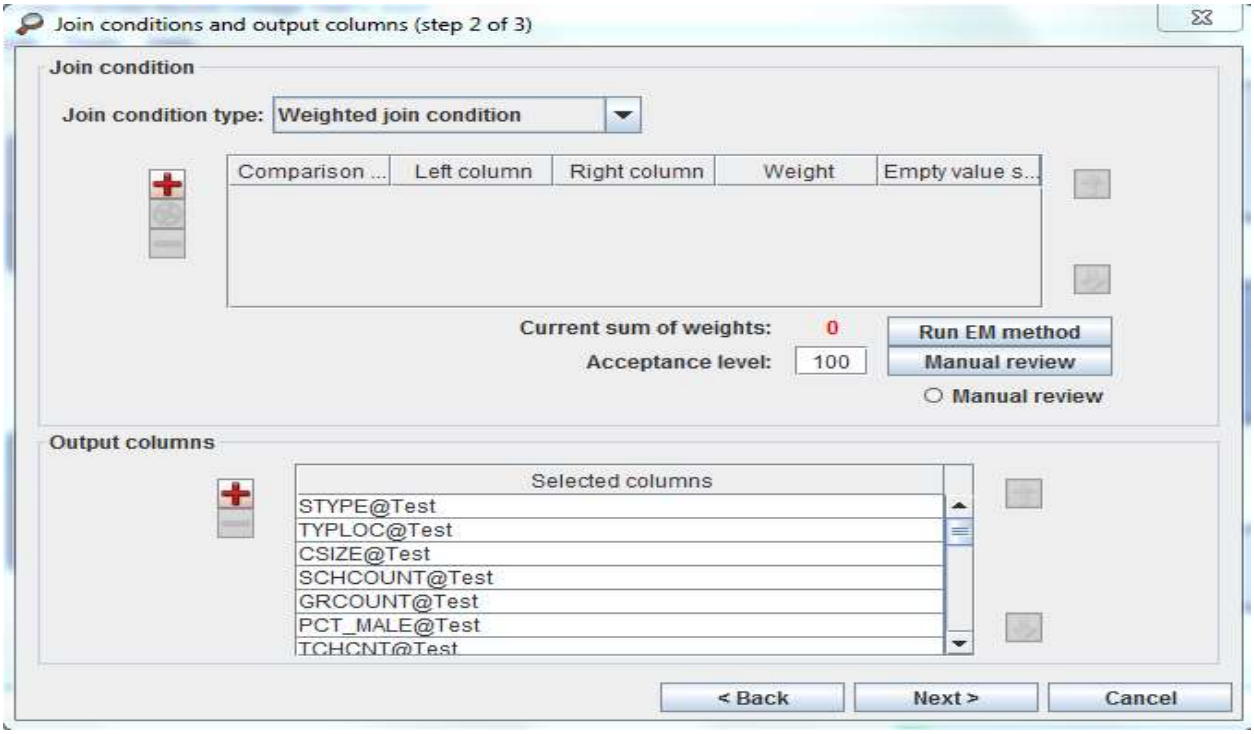
Pairing variables from a data source requiring disclosure control and a data source available to potential intruders usually begins with the Linkage Mode screen. Click on the gear icon in the Linkage box, as shown in figure 5, to specify variable pairs to use in linking.

Figure 5. Gear icon in linkage mode screen.



Use the [Next >] button to skip past the Stratification screen unless planning to stratify the linkage process. The Join conditions and output columns screen comes next, as shown in figure 6.

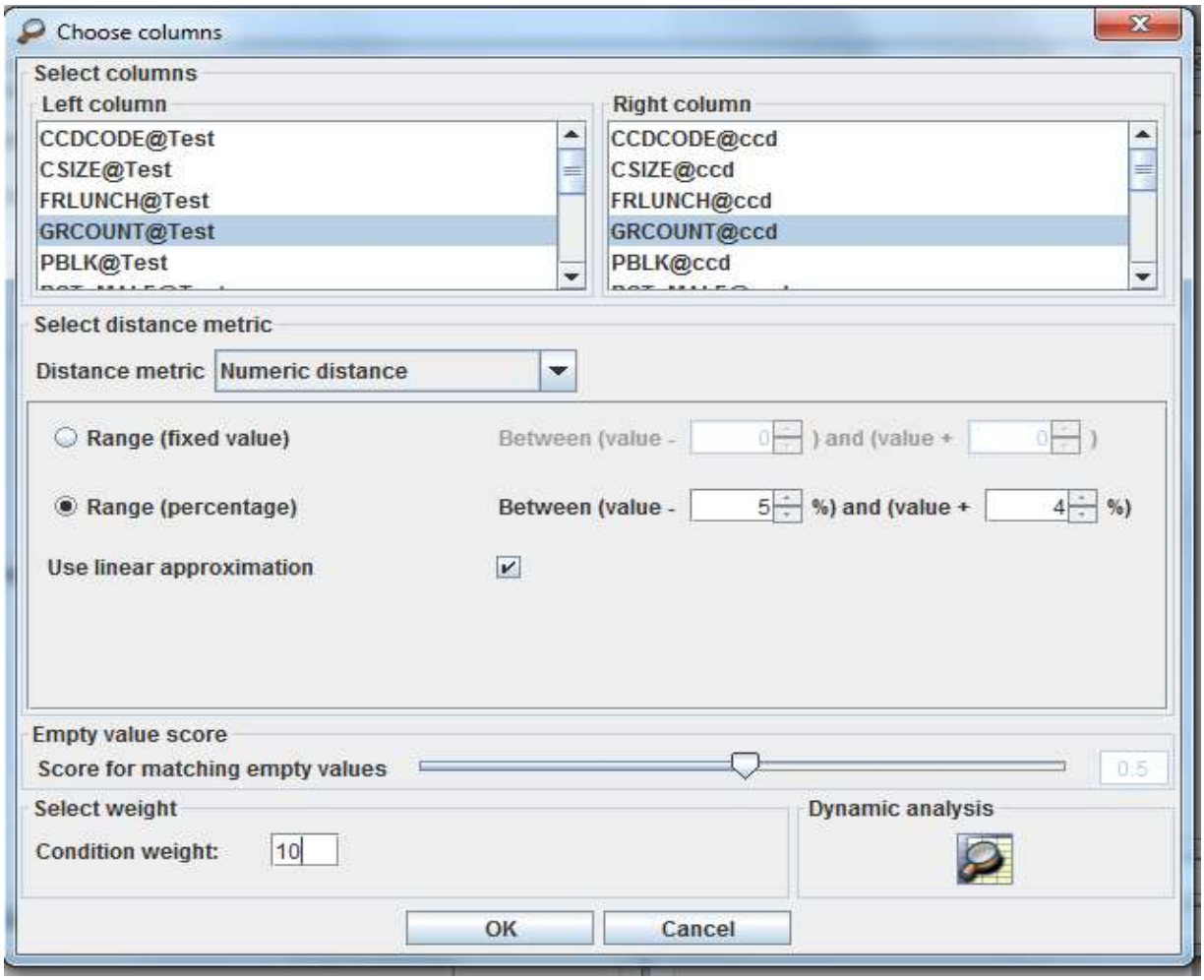
Figure 6. Join conditions and output columns screen.



Use this screen to specify for each pairing of variables the type of comparison, the variables to match from "Left" and "Right" side data sources, weights for the outcomes of comparisons of the variable values, and special values to represent empty (missing) values on the Join condition screen. Do not include pairings of identifiers to be excluded from public release data.

To specify the conditions initially, click on the red plus (+) sign near the top left corner of the screen to bring up the names of variables from the Left and Right side data sources. Highlight a variable from each side to indicate a pairing for comparison, as shown in figure 7.

Figure 7. Indicating pairings for comparisons.



Once one has a pair of variables common to both files, highlighted, specify the type of comparison to be made. The "Equal fields Boolean distance" option is essentially the exact match option, while the "Numeric distance" option is similar to the delta percentage option used in AutoMatch. Other options specify text string comparisons of different degrees of "fuzziness". In both FRIL and AutoMatch, a user can assign weights to each of the matching variables. Not only does this allow the user to determine which variables should be emphasized most in the matching process, but it also allows users to link school identification variables without having them affect the matching scheme (by assigning them a match weight of zero). This latter option simplifies the production of the matching results reports. This technique could backfire in FRIL if a user selects the EM method. FRIL offers an alternative way to include IDs in results and reports. It also allows a minimum match score to be set. It may take several matching runs to establish a minimum match score that allows a large number of matches but also removes obvious non-matches. The expectation is that there will be a large number of false matches but

we are only interested in seeing where a positive (actual) match may occur. Since the purpose of the matching program is to identify a disclosure risk, one need only to look at the top matches and determine whether they are disclosure-risks.

FRIL provides a defined blocking method, but actual matches can still technically occur where blocking values differ[7]. FRIL procedures include the sorted neighborhood method that allows the user to choose the order of importance for the matching variables, and FRIL will use this order to sort the files and then create windows of records which are similar to each other. Records can then only be matched within these windows. For purposes of disclosure analysis, the defined blocking method has normally been the preferred approach for disclosure analyses of education studies. However, the unique characteristics of each education study may provide opportunity to utilize the sorted neighborhood method when blocking variables tend to be less reliable. Unless computation time or cost requires blocking, it would be better to compare all pairs of records, not just those within the same block. That being said, for a national survey, suppose Census region is known to be highly accurate (95% plus). While not using blocking variables should keep all matches within region (since it would be given a high matching weight), one possible approach is to use a blocking approach to ensure that all matches are within the same region.

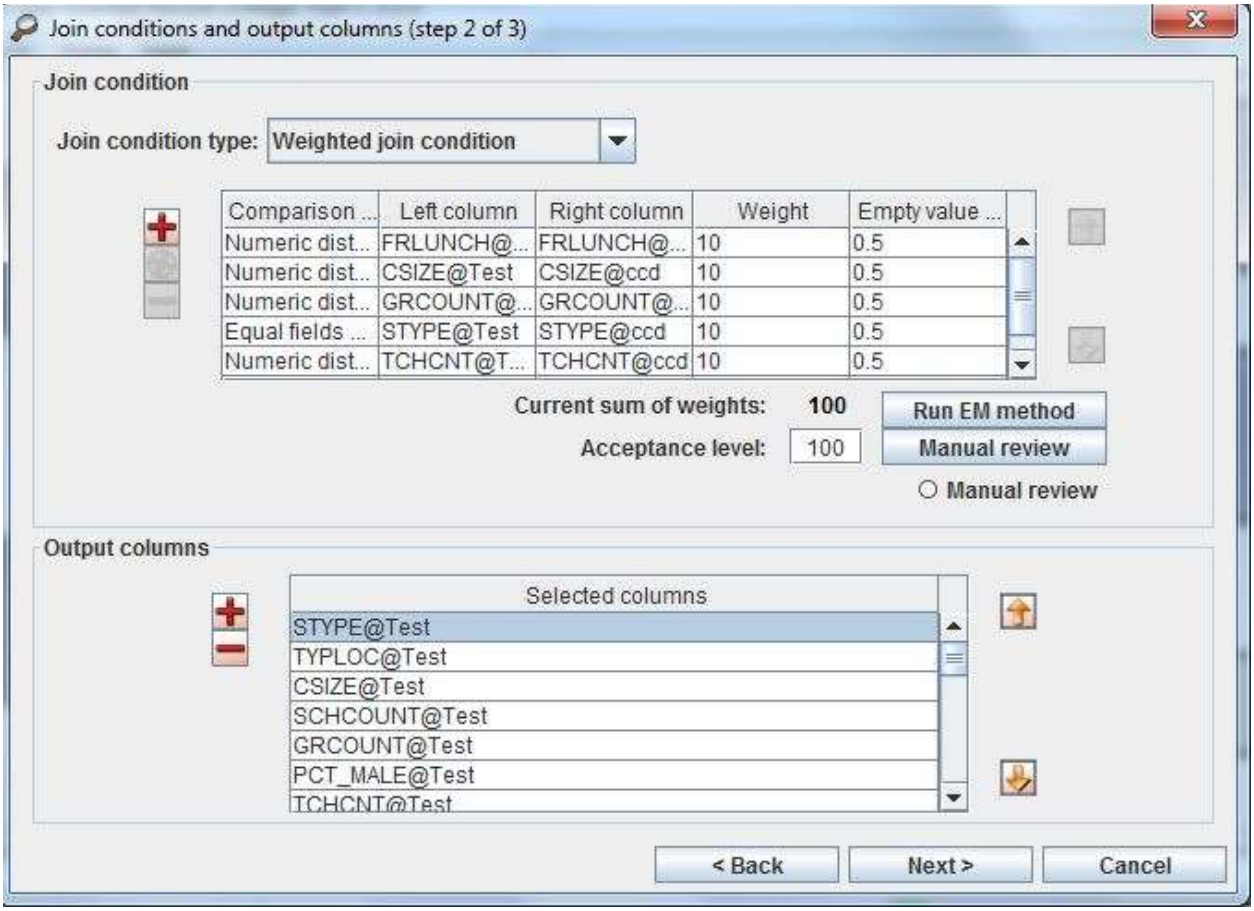
After specifying all pairs of variables to be compared, review the pairings and the weights assigned to each. The weights must add up to 100. Neither the variables paired nor the weights need be exactly correct. The set of weights serves as a starting point. Selecting more pairs of variables common to the two data sources will not hurt and can improve match quality. To change the weights or other specifics for any one pair of variables, highlight the pair and click on the gear icon.

Clicking on the (+) sign under Output columns enables adding or deleting columns from the output of the matching process. The display shows the columns selected for output.

Unless the selection of matching variables and weights seems intuitive and straightforward, the computationally intensive but very powerful and valuable EM method (expectation maximization) comes into play. Starting with the weights assigned to the elements of the comparison vector, the iterative EM method reassigns the values of weights to maximize the likelihood of a correct match. To initiate the EM calculations, click on the [Run EM method] button on the Join condition screen as shown in figure 8.

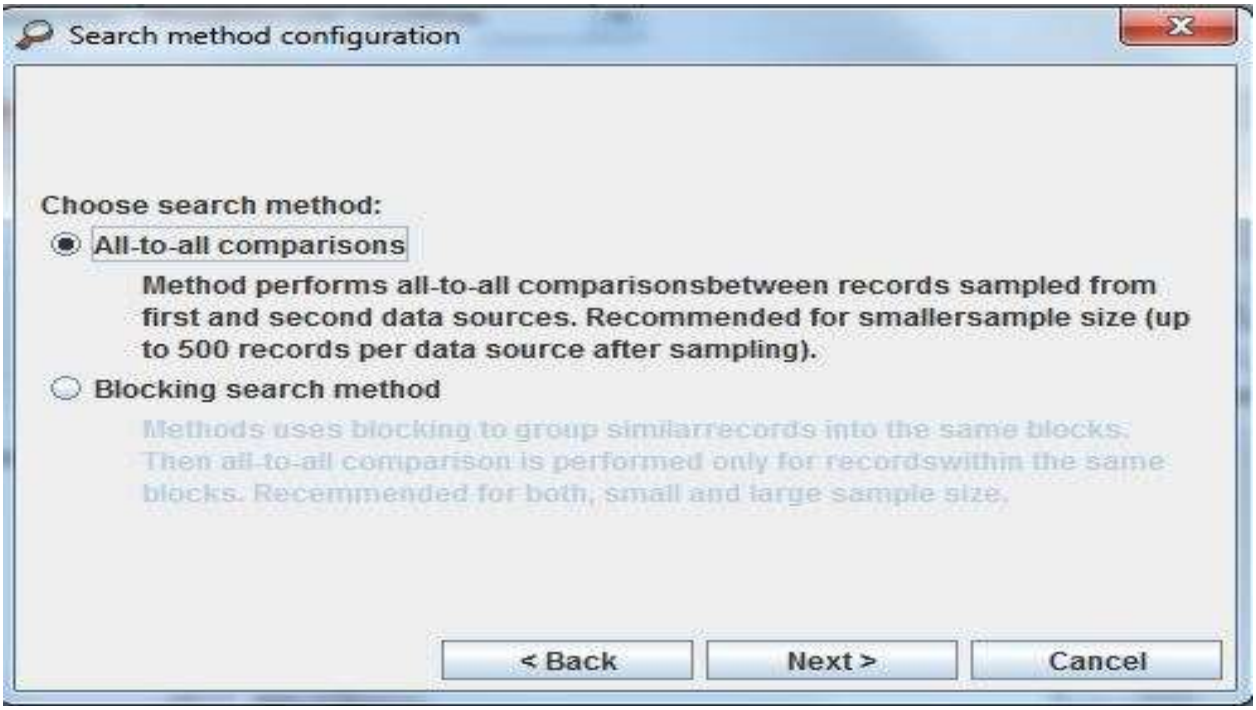
[7] FRIL handles blocking a little differently than AutoMatch. In AutoMatch, the blocking method results in absolute demarcations, where matches are only made within defined blocks.

Figure 8. Initiating EM computations.



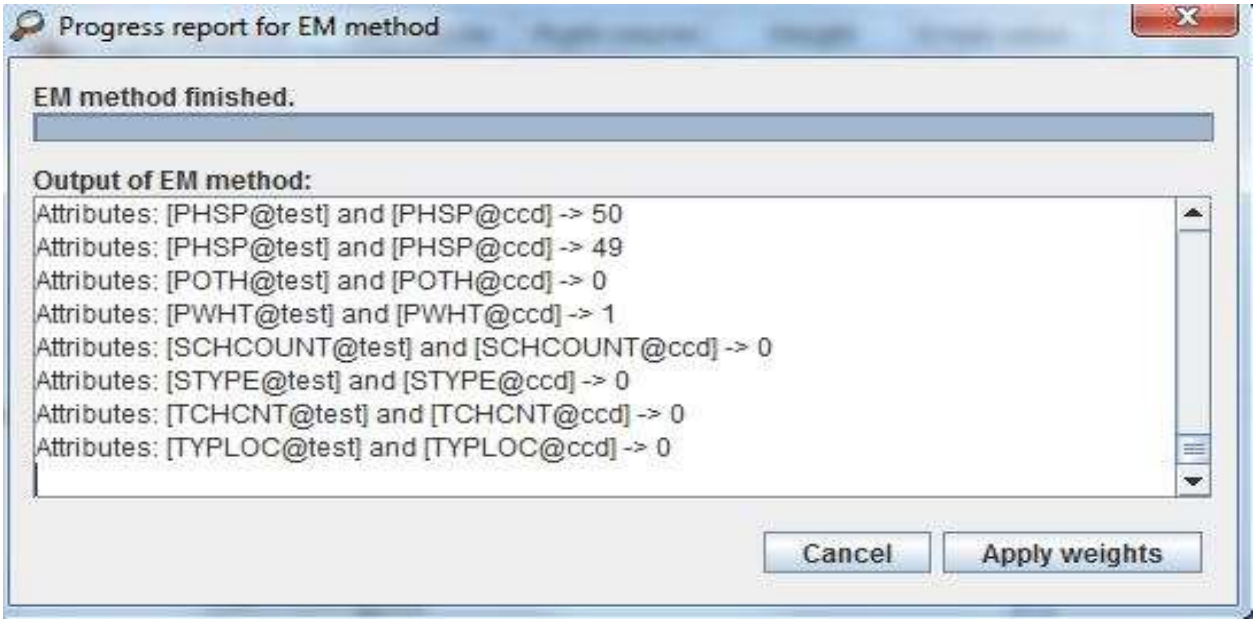
Unless the data sources are very large and ordered randomly, accept the default sampling method in the usual case. The Search method configuration screen pops up next. Select the All-to-all comparisons option as shown in figure 9.

Figure 9. Select all-to-all comparisons.



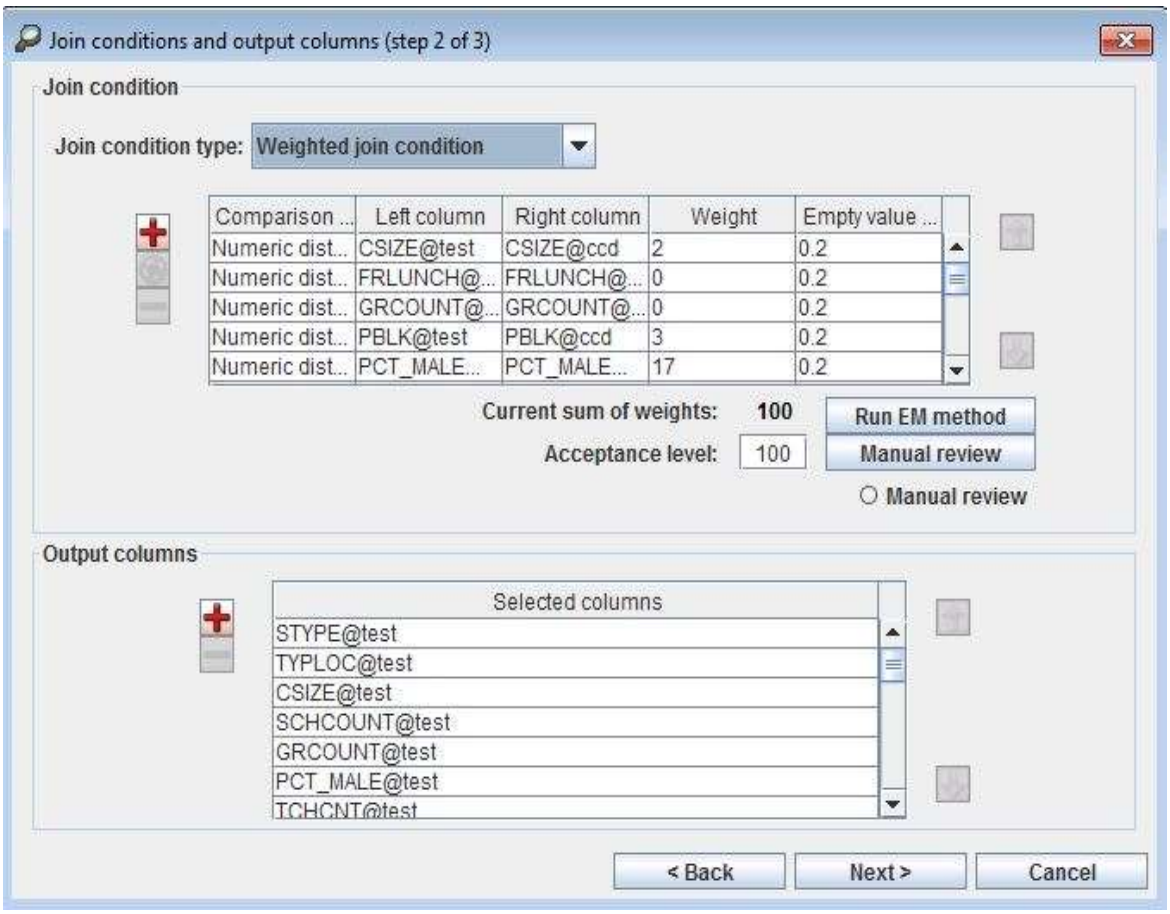
The [Next >] button will start the EM method. The Progress report for the EM method, as shown in figure 10, will log show the current iteration in a series typically of more than a thousand iterations through different combinations of weights as it attempts to maximize the likelihood of good matches.

Figure 10. Progress report for EM method.



The EM method will change the weights assigned to each comparison of matching variables. Note in figure 11 that some weights have increased dramatically, while others have been set to smaller values down to a zero weight. A zero value of a weight in effect suppresses evaluation of an element in the comparison vector. These zero weights are normally only assigned to identification variables that are included to verify the actual or false matches. The EM method assigns zero weights to pairs of variables that do not contribute to the likelihood of a correct match.

Figure 11. Assignment of weights.



Probabilistic data linkage applied to disclosure risk control attempts to identify all potential matches of data available to intruders and public release data. For that reason, this example shows how to implement FRIL methods that optimize the sensitivity of linkage; that is, reduce failures to find potential exposures of public release data and disregard the number of false matches. For the purposes of this example of FRIL linkage, 141 of the 24,677 records in the Test public release data originated in the CCD

and were masked using an initial masking procedure. We are using FRIL to test the effectiveness of the masking procedure.

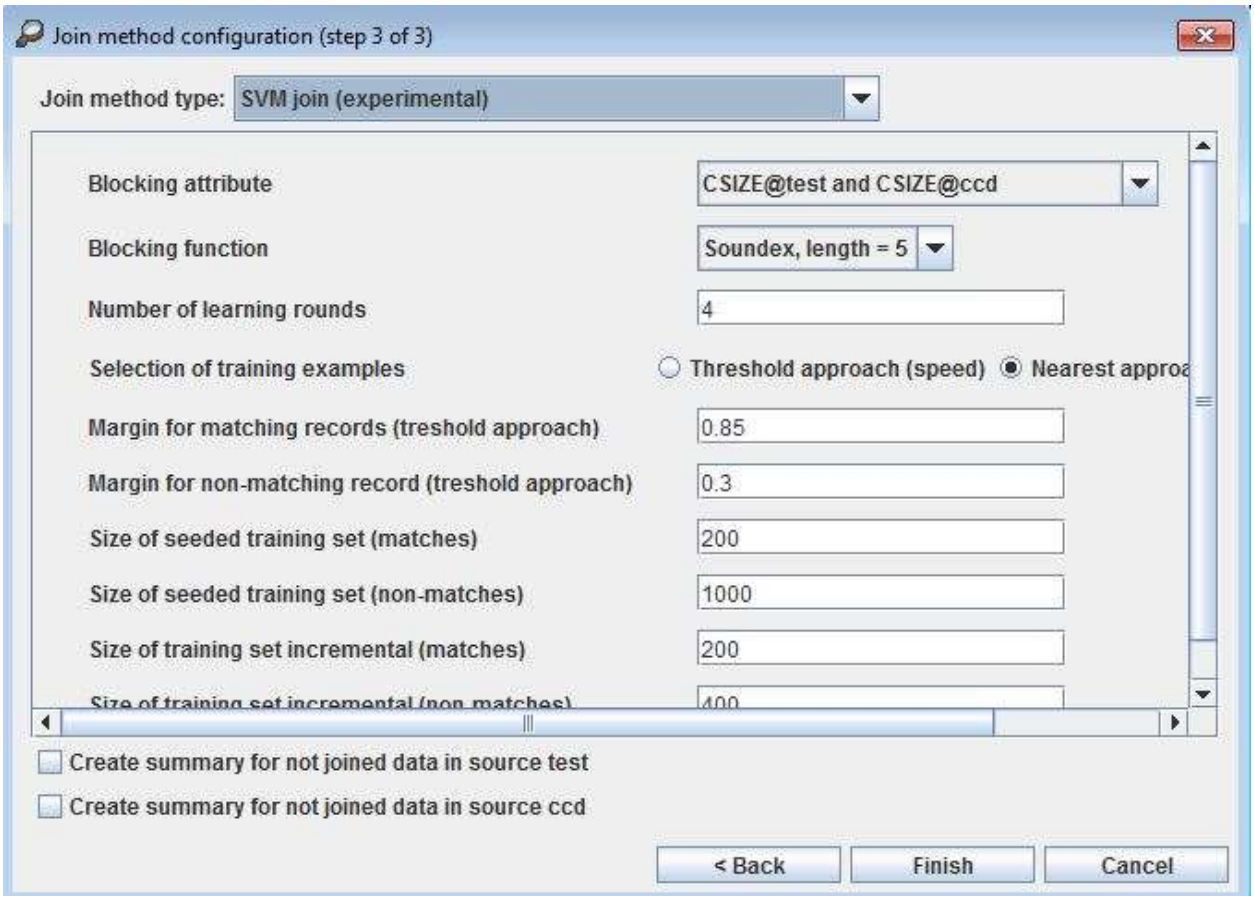
Once specified on the Join conditions screen, the weighted comparison vector sets the stage for linkage of the two data sources. The Join method type screen offers four alternatives:

- Nested loop join (naïve);
- Sorted neighborhood method;
- Blocking search method;
- SVM join (experimental).

The defined Blocking search method would be the usual choice when linking high volume data sources that have one or more reliable blocking variables in common. The Sorted neighborhood method may improve linkage sensitivity by not relying entirely on blocking variables to classify potential matches correctly. The SVM join method applies a mathematical algorithm that separates more likely true links from false ones.

Combining the EM method of assigning weights to elements of a comparison vector and the SVM join method identifies in this extended example more of the actual methods than other methods. As shown in figure 12, select Join method type: SVM join from the drop-down list at the top of the Join method configuration screen. Clicking on the [Finish] button will accept default settings and return to the main FRIL screen.

Figure 12. SVM join method type selection.



Step 4. Match the sampled school data to the master school frame data.

This step performs all the matching of sampled school records to the master school frame records. Matching is done based on the conditions defined during the previous three steps. In this example, we have specified mainly fuzzy comparison of numeric variables; weights determined by the EM method to maximize the likelihood of good matches; and, support vector machine (SVM) matching, to improve discrimination of true from false matches.

As shown in figure 13, when the green Status OK message appears in each of the icons linked by arrows, the arrowhead in the Status icon turns green and Status: idle displays to indicate that FRIL is ready to link the two data sources. Click on the green arrowhead to begin the actual linkage process. FRIL displays the number of records linked as linkage progresses, as presented in figure 14.

Figure 13. Beginning the actual linkage process in FRIL.

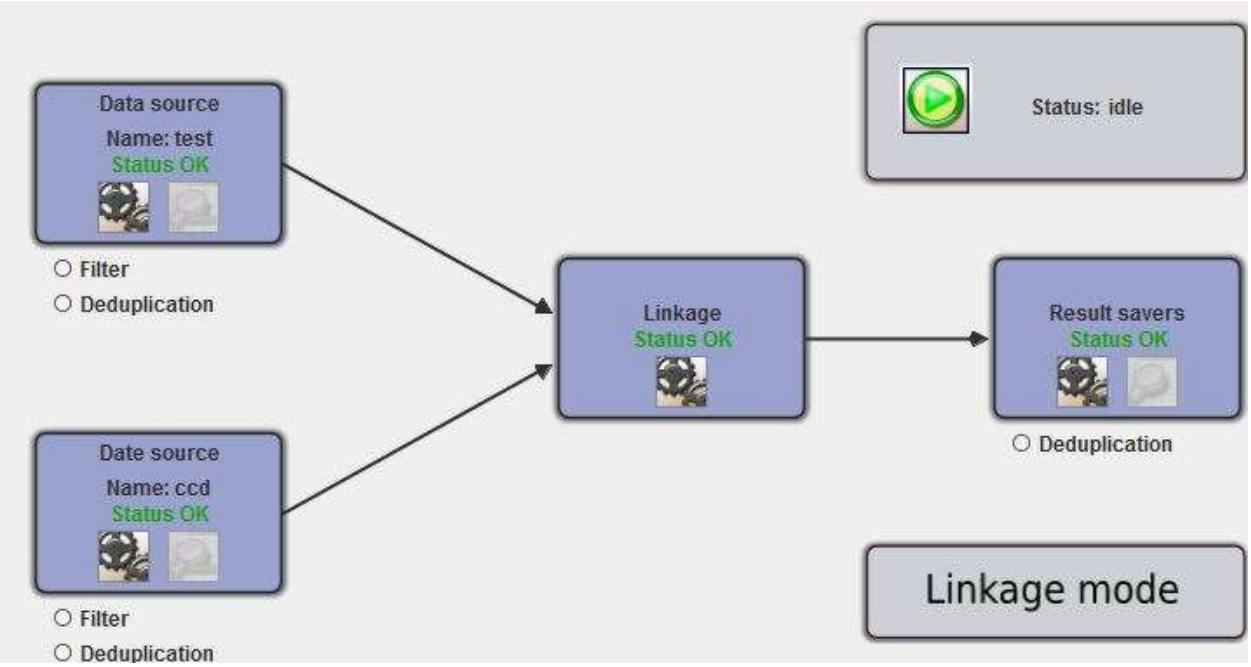
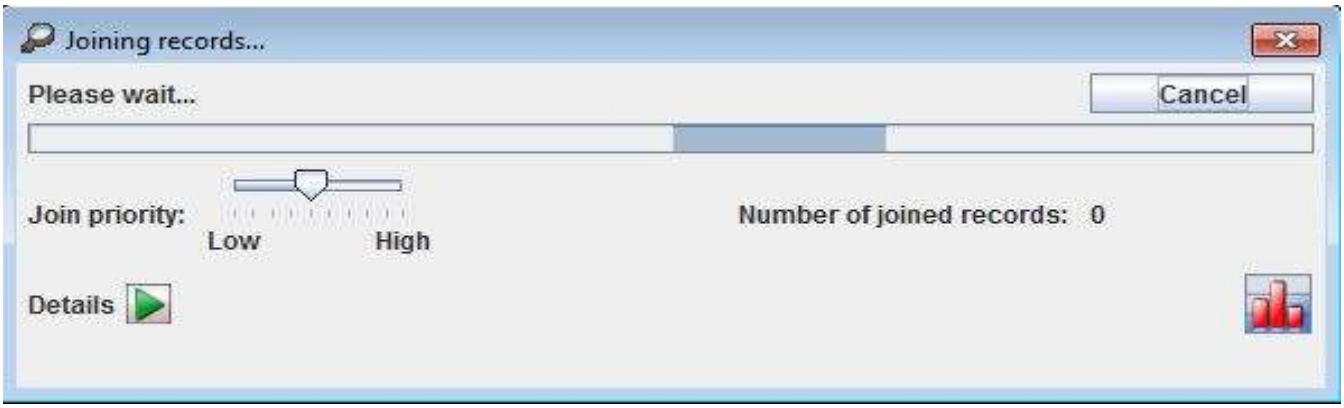


Figure 14. Display of the number of records linked during the matching process.



When complete, the linkage process displays a report presented in figure 15.

Figure 15. FRIL output report.

Parameters also need to be created for each variable[8]. The user can begin the processing by inputting the name of the file in which the matching results will be saved (via the [Result Savers] icon on the main screen).

Step 5. Produce the matching results reports.

Both AutoMatch and FRIL print out listings of matched and non-matched sampled school records. The AutoMatch listings produce ASCII files of the matches, while FRIL can produce CSV and Excel files as well as ASCII files. Because both software packages identify the "best" match and then the closest matches, the manipulation of the results files is needed to produce a clear and readable set of appendices for the final disclosure analysis report. An output manipulation program is developed to basically sort the results by school ID and matching weight and then selects the top 5 or 10 matches for each school.

Overall, from the results of our testing, FRIL and AutoMatch will produce similar results in identifying disclosure risks. FRIL is easier to use than AutoMatch due to its user-friendly graphical interface. FRIL uses a simple Windows-based menu system to define the data sources and establish the linking and matching parameters, while AutoMatch, which under the DOS-based software package, requires the user to create programs to perform the same tasks.

[8] In AutoMatch, the matching process involves running a number of programs. First, the matching program must be compiled. Second, index programs must be created and run so that the software can create the blocks used for matching on both data source files. Third, frequency analyses programs must be created and run so that AutoMatch knows the size of buffers it must create during the matching process. Finally, the matching program can be run, generating output files as specified within that program.

References

Diniz da Silva, A., SanAna Martins Romeo, O., Silva Soares, T. and Layter Xavier, V. (2010). Study of record linkage software for the 2010 Brazilian Census Post Enumeration Survey. *The Survey Statistician*. Book and Software Review, pp. 31-39.

Domingo-Ferrer, J., Torra, V., (2001). A quantitative comparison of disclosure control methods for microdata, confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. Doyle, P.; Lane, J.I.; Theeuwes, J.J.M.; Zayatz, L.V. eds., Elsevier, pp. 111-133.

Elliot, M.J., Manning, A.M., and Ford, R.W. (2002). A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge Based System*, 10(5), 493–509.

Itorralba, M. (2013) Deterministic Matching versus Probabilistic Matching. Blog entry, September 6, 2013.http://blog.infotrellis.com/2013/09/06/deterministic-matching-versus-probabilistic-matching/ (accessed August 30, 2014).

Jaro, M.A. (1989) 'Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida', Journal of the American Statistical Association, 84, pp.414-20.

Li, J., and Krenzke, T. (2013). Comparing approaches that are used to identify high-risk values in microdata. Census Statistical Disclosure Control Research Project 3. Final report. Washington, DC: U.S. Census Bureau.

Reiter, J. P. (2005) Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association*, 100, 1103 - 1113.

Skinner, C.J., and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of American Statistical Association*. 103, no. 483 (2008), 989–1001.

Sweeney, L. (2002). *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.

Winkler, W. (1993) Matching and Record Linkage. U.S. Census Bureau. https://www.census.gov/srd/papers/pdf/rr93-8.pdf (accessed 7/17/2013).

Appendix A

Explaining the Terminology in the External Matching Procedure

Data Confidentiality: Data have been reviewed and perturbed prior to dissemination to ensure that data are adequately, reasonably and statistically demonstrated to be protected against respondent identification.

Deterministic Swapping: In order to ensure that data from Respondents identified as risks from FRIL results are masked, an analyst (manually or automated approach) swaps key variable(s) to nearest partners to maintain overall data integrity and to eliminate the risk of respondent identification.

Blocking Variable: Variables that are, or should be, exact matching variables. They are normally providing specific identifying information. Examples include geographic data such as State or Census Region, School Sector (Public/Private), etc. The probabilistic nature of the matching scores is not utilized – instead, the match is either correct or incorrect.

Matching Variables: The variables that are used for matching are usually ordinal and/or continuous variables. They should be close though not necessarily exact values when matching external with survey data. Examples include school enrollment, percent of school Hispanic, Percent of school male, etc.

Delta or Tolerance (Confidence Interval): The matching scores are adjusted by probabilistic matching based upon distance to the matching values (normally of continuous variables). The delta is applied when calculating the matching scores; if values from the two files fall within the delta, then the matching score is higher than if they are outside the delta.

Matching Weight: FRIL and AutoMatch generate a matching weight (or matching score) for each school that is matched against the base (survey) school. The weights are generated based upon the summing of the scores based on relative proximity of all of the matching variables and their expected level of validity.

Rule of 3: NCES generally required that a school is a disclosure risk if, through matching procedures, the actual (positive) match has the highest or second highest matching weight (score). In certain cases, the criteria can be made more conservative whereby when there are scoring ties, the disclosure risk could be, say, in the top 5 matches.

Deterministic and Probabilistic Data Matching: Data matching can be either deterministic or probabilistic. In deterministic matching, either unique identifiers for each record are compared to determine a match or an exact comparison is used between fields. Unique identifiers can include national IDs, system IDs, and so on. Deterministic matching is generally not completely reliable since in some cases no single field can provide a reliable match between two records. This is where probabilistic, matching comes in. In probabilistic matching, several field values are compared between two records and each field is assigned a weight that indicates how closely the two field values match. The sum of the individual field weights indicates the likelihood of a match between two records. (Oracle website: http://docs.oracle.com/cd/E19182-01/821-0919/ref_sme-deter-probl_c/index.html, accessed August 30, 2014)