Guidelines for Developing a Hierarchical Record Linkage Process

Prepared by **Statistics Canada**
June 2023

Overview of Record Linkage:

When working with data from different sources which cover a particular population, one often wants to know which records from one file represent the same unit from another file. Being able to use information from various sources would permit the creation of a more complete population while enhancing the analysis of a population, permitting a comparison of similar values between sources, allowing the use administrative data in place of survey data or facilitating the maintenance of a survey frame, etc. Determining which records match between various sources requires a record linkage exercise to be undertaken.

Record linkage is especially important for National Statistical Offices operating under a Statistics Act which grants them access to a rich and wide variety of pre-existing files from other Government offices, organizations or businesses. From a business survey perspective, some files of particular interest include those that record business sales taxes, income taxes based on business revenue, each business' contributions to social security programs and business registration records. These types of files provide good coverage of the current business population while providing different types of useful data.

Concepts related record level linking of two files

Source files refer to the files of data that are available from the various sources. This could include administrative data from other government offices, a pre-existing BR, results of a census of businesses, etc. Each file should contain a unique identifier to record the links as well as data for identifying and verifying links.

Hierarchical matching is a process which starts with very stringent matching criteria which are relaxed with each subsequent matching attempt. For example, a first match could be on the business name, address and telephone number. All pairs of potential links are output to a file for review and the remaining unmatched records pass to the next attempt. Again, the potential matches are output to a separate file for review. The process of relaxing the matching criteria and saving the results to separate files continues until the potential links are mostly false. Saving the results of each pass in separate files implicitly groups the potential links based on the matching criteria which allows the person reviewing the results to accept or reject some groups of matches.

Exact matching produces potential links based on common values from both source files being exactly the same (ex. records with "ABC Inc" on both files would match).

Matching based on measures of similarity produces potential links based on a measure of similarity of the values (ex. "ABC Inc" and "ABD Inc" are 86% similar; 6/7 letters are the same). With this approach, experimentation is require to determine the threshold over which dependable links are generated and do not require manual resolution and the threshold where potential links are mostly false. CARTAC shared information about an Excel based version called Fuzzy matching.

Manual resolution is the process of reviewing potential links to accept the true links and reject the false ones. The results are typically recorded with a flag that is added to the resolution file where 1=True Link and 0=False Link.

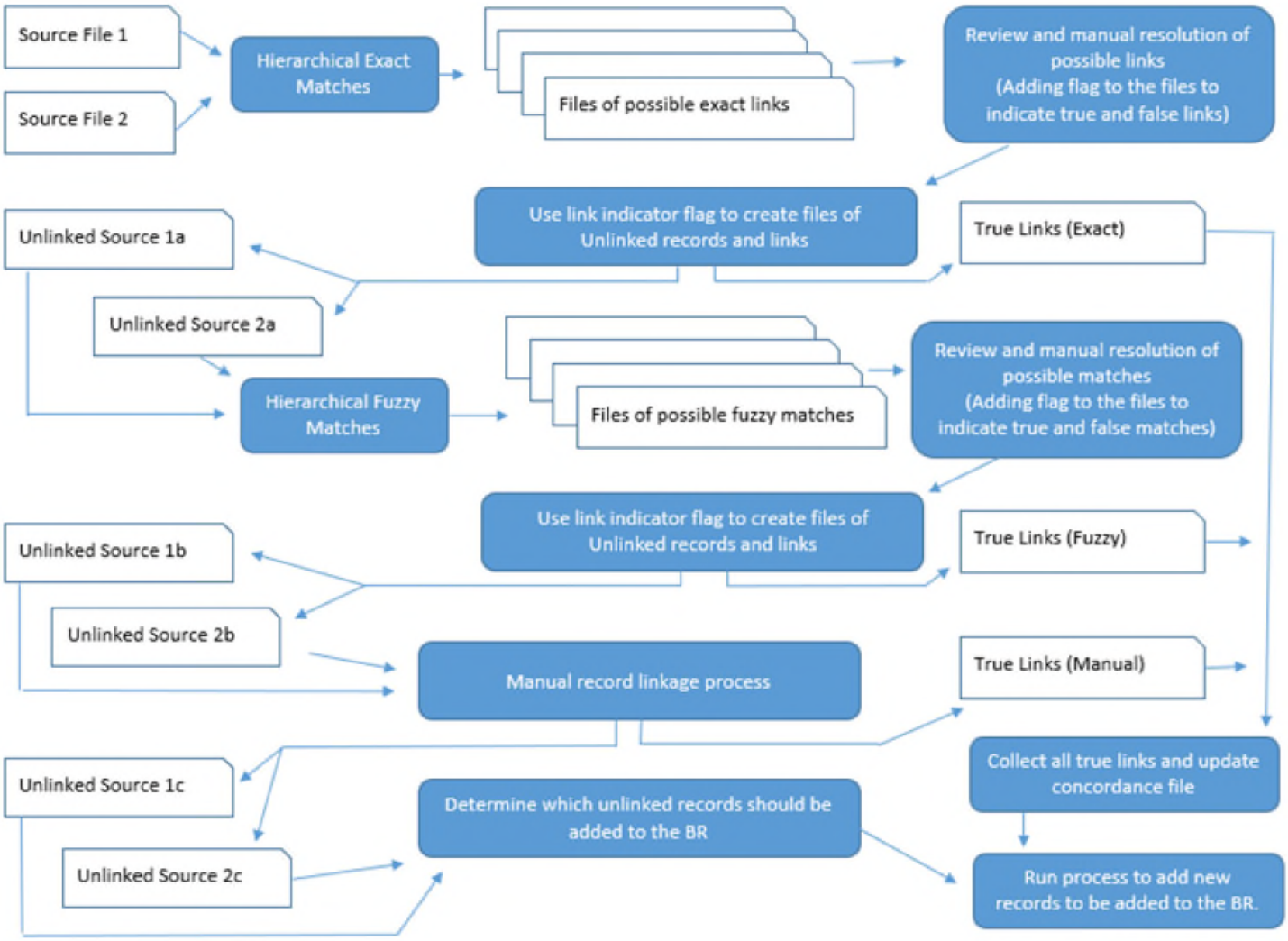
Manual matching is the process of someone reviewing the two source files in an attempt to find links. This is typically the last step in trying to link the last few records that were not linked through the more automated processes. Various sorting and filtering techniques are helpful with such a process.

Concordance file is a file that records the relationships of the unique identifiers from each source file. This is used when the same files will be matched more than once in order to profit from previous linking exercises. When used in conjunction with a BR, each line of the concordance file will contain the BR unique ID in one column and the remaining column will record the identifiers for the same business from the various source files.

Direct Match Key (DMK) is typically a groomed version of a business name where superfluous information is removed in order to increase the chances of finding links. Additional information can be found here:

PRASC_Reference_Ge
nerating a Direct Mat(

General Overview of a Record Linkage Strategy



Preparing files for record linkage:

Prior to conducting a record linkage, it is important to process the files in order to increase the linkage rate. For example:

- At least one groomed version of each **business name** should be generated (The DMK coding; earlier in the document is an option).
- **Addresses** should be reviewed for systematic formatting inconsistencies that can be standardized through an automated process. (For example, Post Office, PO, P.O. can be removed or converted to one common standard such as PO. It may be desirable to move all digits to the start or end of an address, or even move them to their own matching fields. If parish names tend to get imbedded with addresses, move to the correct field. If limited programming abilities exist within the NSO, many rules can be applied through Excel functions).
- All extra characters that do not add value to **Telephone Numbers** should be removed (brackets, spaces, hyphens, etc.) and the recording of telephone extensions should also be removed or standardized.
- A more general 2-digit version of **ISIC** could be generated as a possible matching variable.
- If there is a shortage of matching variables, then revenue and employment groupings can be created.
- All variables that are common to both files should be reviewed to see if they can be used in the matching process (GPS coordinates are expected to become good matching variables).

To improve efficiency of the matching procedure, only the matching variables should be included on the input files. However, all extra information should be added back to the files for the manual resolution process as they can provide insight into which links are true and false. For example, if size measures are available, but not similar enough to match (ex. Number of Employees on one file and Revenue on the other) they will let the person doing the resolution know whether both records are similar or they may provide information for applying a sort order to the possible links. For example, 2-digit ISIC will allow the resolver to process entire industries at the same time and a size measure will allow them to concentrate on the similar-sized units or to put extra effort into the largest and most important units.

Possible hierarchy for a record linkage:

1. When a concordance file exists from a previous linkage exercise, identifying and quickly reviewing the previously linked records should be the first step (this is a special type of exact matching). All records that are part of links that are still true should be removed from further processing.
2. Assuming that Source 1 has the following variables:
    a. Legal name
    b. Operating name
    c. Operating address
    d. Operating telephone number
    e. ISIC

    And Source 2 has the following:
    a. Legal Name
    b. Operating address
    c. Telephone number

    The files can be further enhanced by:
    - Adding DMK versions of the business names and
    - Ensuring that the addresses and telephone numbers have been groomed to increase consistency.

ISIC is available on only 1 file, so it is not of use. If it were available on both files, a 2-digit ISIC could be helpful in creating more combinations of variables (and passes) to the matching strategy.

Note that code values and dollar amounts are not helpful when measuring similarity of values because ISIC 123456 would be quite different from 223456, but would have a high measure of similarity (same with $9,000,000 and $1,000,000)

Based on the information available in the example, the following is a summary of a possible hierarchy of passes that could be executed. It starts with very stringent criteria to first find the most likely matches then moves to more relaxed criteria. If more matching variables are available, then additional passes would be created. After each pass, all possible links are removed from the subsequent matches. However, there needs to be a point where the records involved in failed matches re-enter the process to try to be matched again.

Pass 1:	Legal name	to Legal name
	Operating address	to Operating address
	Telephone number	to Telephone number
Pass 2:	Operating name	to Legal name
	Operating address	to Operating address
	Telephone number	to Telephone number
Pass 3:	DMK Legal name	to DMK Legal name
	Operating address	to Operating address
	Telephone number	to Telephone number
Pass 4:	DMK Operating name	to DMK Legal name
	Operating address	to Operating address
	Telephone number	to Telephone number

Passes 5 to 8, the same as 1 to 4, but without addresses.
Passes 9 to 12, the same as 1 to 4, but without telephone numbers.
Passes 13 and 14, the same as 1 and 2, but with only business names.
Passes 15 and 18, the same as 1 to 4, but without business names.

The hierarchy can be applied to both exact matching as well as matching based on measures of similarity / Fuzzy matching. Through experience, one will gain knowledge of which types of matches produce reliable results and which ones should be removed from the process.

Additionally, it may be desirable to integrate Fuzzy matching with exact matching. For example, conducting Pass 1 with the exact matches, then with the fuzzy match before moving to pass 2. Experimentation should be conducted to determine the most efficient process.

Due to the high amount of computing resources required for the Fuzzy matching, the exact matches are designed to quickly remove the ones that are easy to find, thus allowing the Fuzzy matching to run more quickly. For this reason, once the concordance file becomes more complete and few records remain to be linked, it may become possible to remove some of the exact matching steps and move directly to the Fuzzy matching.

Strategies for manual resolution:

There are various ways that the manual resolution process can be made more efficient and experience will lead to new and updated approaches. The following are just a few suggestions to get a new project started.

Firstly, by saving the results of each hierarchical pass into its own file, the person doing the resolution will be able to accept and reject files of links without reviewing the entire files. That is, if a review of the first 100 records find only true (or false) links, the entire file can be safely accepted (or rejected) without reviewing all of the links. If, after a few linkage exercises, it has been found that certain passes produce only quality links, they can be accepted with only periodic review to ensure that assumptions are holding true. Similarly for the combinations of variables that produce mostly false links, which can be dropped from the processing.

For the files that contain a mix of true and false links, sorting and grouping the potential links allows for similar records to be resolved together. For example, grouping by industry coding (either 2-digit or a more refined level), will allow the links for a certain industry to be resolved as groups. Sorting by a size measures such as the number of employees or revenue will allow a focus on the largest and most important records. Sorting by geographic location will allow for linkages within large office towers to be done at the same time, or for the resolution work to be shared with satellite offices which may have better knowledge of the businesses within their areas.

To make the resolution work more efficient it is important to have strategies that are defined prior to starting the work, and to refine them with local knowledge and experiences.

Strategies for manual linkages:

After the automated linkages, it is important to review the remaining records to see if any other links can be found. If the number of remaining records seems unmanageable for manual procedures, then it would be important to review the entire linkage process to see if improvements can be made at any of the steps. This should include reviewing the process right from the beginning, including the request for administrative data; are there any additional fields that can be requested or perhaps there are ways to work with the other offices to improve the quality of the incoming data. Can the grooming be improved or are there additional derived variables that can be generated? Are there some key combinations of variables that could provide some high quality links?

Additionally, the manual linkage procedures could provide insight on how the automated portion can be improved.

Ideas for conducting a manual linkage can be found in the attached document that was previously written for the PRASC project.

PRASC_Reference_Te
chniques for Manually

Considerations for unlinked records:

At the end of any linkage exercise, there are remaining unlinked records that should be reviewed to determine their fate. In terms of creating a BR population, it is crucial to scrutinize which records should (and should not) be added to the population. Adding records that are already represented on the BR will create an over estimation in any survey that uses the BR as its frame. However, omitting records that are not already represented will have the opposite effect.

Often there are logical reasons why a unit may have not linked to an existing BR but should be added to the population. For example, if there is a date indicating when the business started, it may be possible to determine that this is a new business and it should be added to the BR. Or, if the BR is based primarily on employment data and a tax record (income tax or sales tax) was not linked to the BR, this could be due to the business being active in the country but not having any employees.

On the other hand, sometimes records do not match due to different reporting structures for different data sources. For example, a large hotel could report its rooms and restaurants together for one source and separately for another.

Prior to birthing any large unit onto the BR, it should be well researched and the reasons for it not already being on the BR should be known. For smaller units, it may be sufficient to do some initial research to determine rules for adding new units, as long as the rules are periodically reviewed to ensure that the initial assumptions still hold true.